Title

Statistical Analysis and Visualization of Single Cell RNA Sequencing Data at Population Scale

Author

Wang, Hao

Publication Date

2024

Peer reviewed|Thesis/dissertation

Statistical Analysis and Visualization of Single Cell RNA Sequencing Data at Population Scale

by

Hao Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Elizabeth Purdom, Chair
Professor Sandrine Dudoit
Professor John Marshall

Summer 2024

Statistical Analysis and Visualization of Single Cell RNA Sequencing Data at Population Scale

Abstract

Statistical Analysis and Visualization of Single Cell RNA Sequencing Data at Population Scale

by

Hao Wang

Doctor of Philosophy in Biostatistics

and the Designated Emphasis in

Computational and Genomic Biology

University of California, Berkeley

Professor Elizabeth Purdom, Chair

The advent of Single-cell transcriptome sequencing (scRNA-Seq) has revolutionized our ability to explore the intricate landscape of cellular diversity within complex biological systems. Initially focused on cataloging cell subtypes and discerning gene expression disparities across cell types, scRNA-Seq has evolved to address broader inquiries, particularly in the realm of human health. While past efforts concentrated on analyzing numerous cells from a few samples, there's now a growing interest in understanding inter-sample heterogeneity and its implications for phenotypic outcomes, notably in cancer and inflammatory diseases. However, existing bioinformatic methodologies inadequately address population-level analyses, with limited consideration for inter-sample variation. The dissertation introduces a novel framework termed GloScope Representation, which is introduced in the first chapter in detail, for representing the entire single-cell profile of a sample. In the second chapter, We applied GloScope across scRNA-Seq datasets spanning diverse study designs, with sample sizes ranging from 12 to over 300. Through illustrative examples, we showcase how GloScope empowers researchers to undertake pivotal bioinformatic tasks at the sample level, with a primary focus on visualization and quality control assessment. In Chapter 3, we demonstrate GloScope's efficacy in evaluating and quantifying batch effects, as well as comparing various batch correction methods' performance in the patient level analysis of scRNASeq data. Furthermore, to assess GloScope 's advantages and effectiveness in detecting different classes of single-cell differences arising from variations in sample phenotypes, we compared GloScope to existing visualization tool and other sample level analysis tool in Chapter 4. We also developed a simulation pipeline for generating single-cell count data. We utilize this simulation framework to conduct quantitative evaluations of GloScope through a series of

simulated experiments.

Contents

List of Figures

List of Tables

Chapter 1

Patient Level Representation of scRNA-Seq

1.1 Introduction to Single Cell RNA Sequencing

RNA-Sequencing (RNA-Seq) is a powerful and versatile technique that appeared in 2008, and is used to analyze the transcriptome of an organism, providing insights into the quantity and sequences of RNA in a sample. This method aims to capture a comprehensive snapshot of all RNA molecules, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNAs. The process begins with the isolation of total RNA from a sample, which can be derived from tissues, cells, or organisms. This RNA is then treated to remove rRNA, leaving primarily mRNA, which is subsequently converted into complementary DNA (cDNA) through reverse transcription. The cDNA is then fragmented into smaller pieces and tagged with sequencing adapters, allowing them to be read by sequencing platforms. Once sequenced, the resulting reads are aligned to a reference genome or transcriptome using bioinformatics tools. This alignment enables the identification of where the RNA reads map within the genome, as shown in Figure 1.1 (Van den Berge et al., 2019). The output of RNA-Seq data for quantitative analysis is usually a gene count matrix where columns represent genes and rows represent different samples. Each cell in the matrix contains the count of sequencing reads corresponding to a specific gene in a particular sample. The matrix provides valuable genetic information for researchers to study gene expression levels, alternative splicing, and the discovery of novel transcripts or gene fusions (Wang et al., 2009; Mortazavi et al., 2008).

High-throughput sequencing technologies, also known as next-generation sequencing (NGS), are advanced methods that enable the rapid and large-scale sequencing of DNA and RNA (Heather and Chain, 2016; Goodwin et al., 2016; van Dijk et al., 2018). These technologies utilize massive parallelization, allowing millions of DNA fragments to be sequenced simultaneously, significantly increasing the speed and volume of data generated. NGS platforms, such as Illumina and Ion Torrent, can produce vast amounts of sequence data in a short

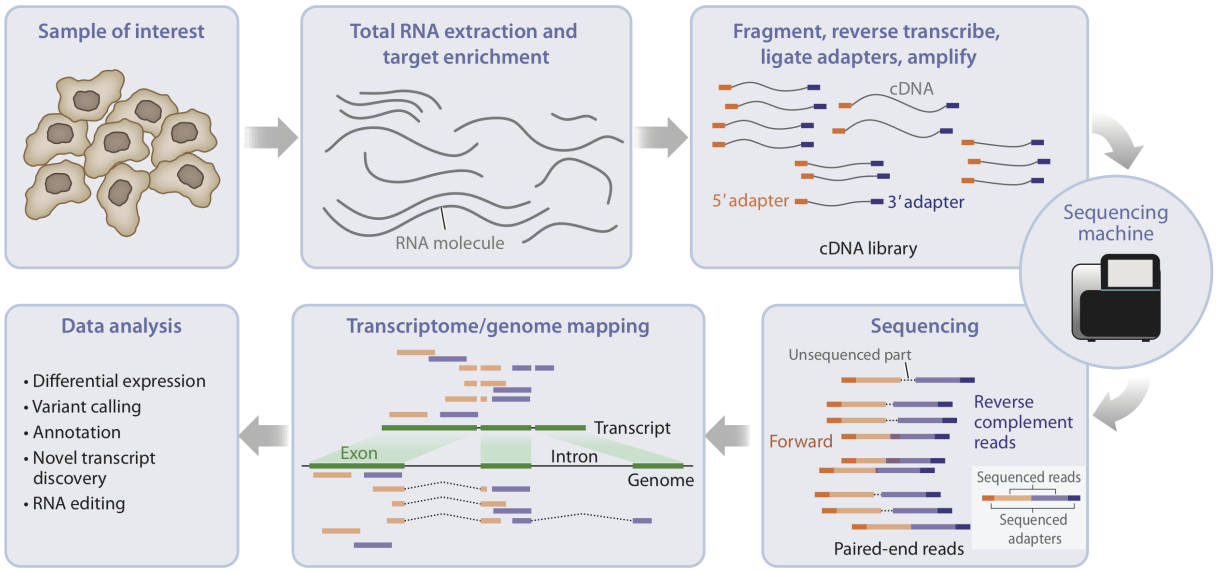


Figure 1.1: **RNA-Sequencing workflow.** Adapted from Van den Berge et al. (2019)

time and at a lower cost per base compared to traditional methods like Sanger sequencing (Goodwin et al., 2016). This capability has revolutionized genomics and molecular biology, making it possible to conduct comprehensive analyses for applications in research, clinical diagnostics, and personalized medicine.

In the past decade, advancements in high-throughput sequencing technologies have revolutionized the field of genomics, enabling researchers to dive deeper into the study of cellular heterogeneity (Reuter et al., 2015; Soon et al., 2013; Weaver et al., 2014). Bulk RNA-seq, which measures the average gene expression levels across a heterogeneous mixture of cells, provided groundbreaking insights into gene expression patterns and regulatory mechanisms. However, it lacked the resolution to detect cellular heterogeneity and subtle differences among individual cells. As technology and computational methods advanced, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to dissect cellular diversity at unprecedented resolution (Soon et al., 2013; Weaver et al., 2014; Luecken and Theis, 2019). Unlike traditional bulk RNA-seq approaches that average gene expression signals across a population of cells, as shown in Figure 1.2, scRNA-seq allows for the characterization of gene expression profiles at the individual cell level (Kulkarni et al., 2019; Li and Wang, 2021).

The fundamental principle of scRNA-seq lies in the isolation and sequencing of RNA molecules from individual cells, thereby capturing the transcriptomic landscape of diverse cell types within a heterogeneous population. This approach not only enables the identification of rare cell populations but also provides insights into cellular states, developmental trajectories, and dynamic responses to various stimuli (Luecken and Theis, 2019; Kulkarni et al., 2019; Li and Wang, 2021). For instance, by mapping the changes in gene expression profiles of individual cells over time, scRNA-Seq allows researchers to reconstruct the lineage

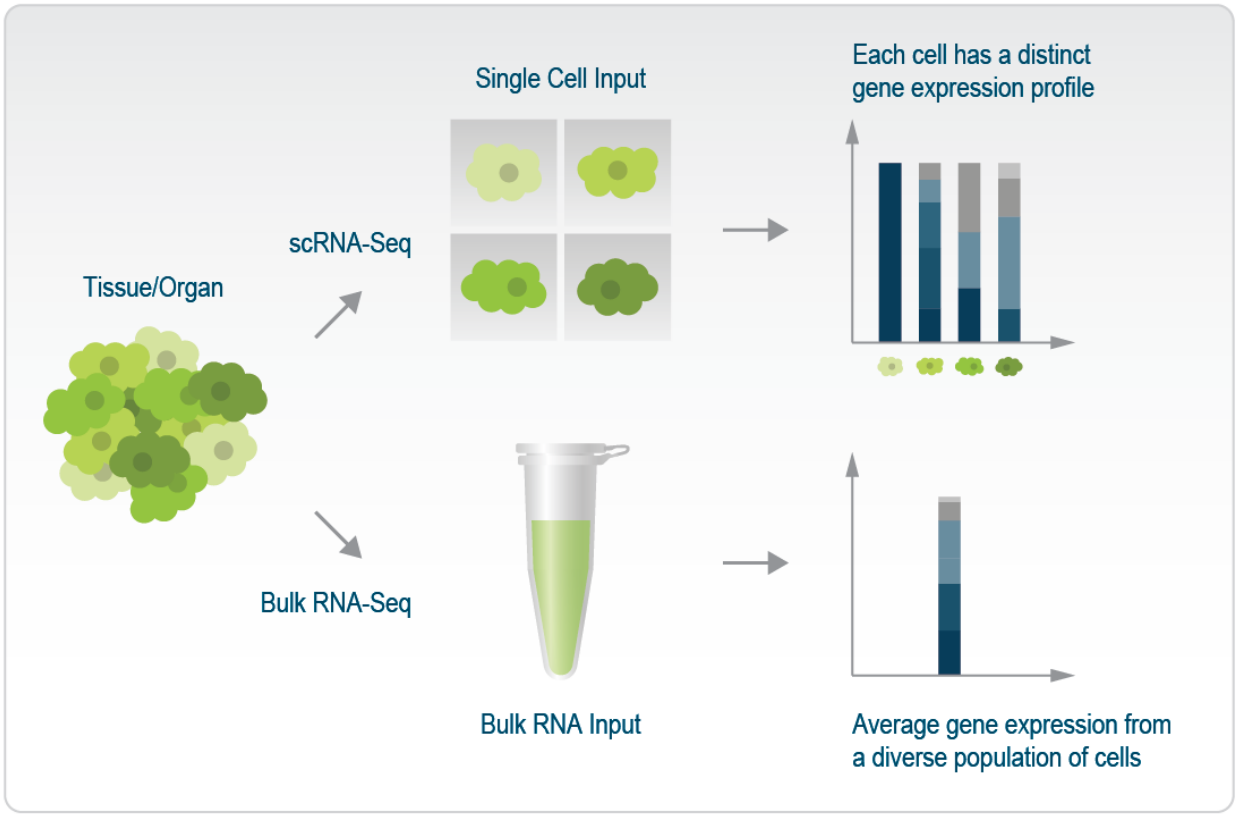


Figure 1.2: **Bulk RNA-Seq vs scRNA-Seq.** Adapted from Lexogen (2024)

and fate decisions of each cell within a developing tissue or organism, offering insights into the mechanisms of cell differentiation, tissue formation, and the impact of genetic and environmental factors on these processes. The application of scRNA-seq spans across diverse fields of biology and medicine, from developmental biology and immunology to oncology and neurobiology. Researchers have leveraged this technology to unravel the cellular heterogeneity underlying complex biological processes, elucidate on disease mechanisms, identify diagnostic biomarkers, and discover novel therapeutic targets (Soon et al., 2013; Paik et al., 2020; van Galen et al., 2019).

1.2 Human scRNA-Seq Study

Early studies of scRNA-seq were predominantly focused on profiling large number of cells from a small number of samples such as mice, which are genetically identical or highly controlled. Such early studies aimed to elucidate shared cell populations and allow researchers to reconstruct developmental pathways and understand the differentiation processes of various cell types within homogeneous cellular populations (Llorens-Bobadilla et al., 2015; Plass

et al., 2018; Van den Berge et al., 2020). For instance, pioneering work by Macosko et al. utilized scRNA-seq to profile thousands of cells from the mouse brain, revealing distinct neuronal subtypes and transcriptional programs underlying neural diversity (Macosko et al., 2015).

Unlike model organisms, human samples exhibit a vast array of genetic backgrounds and environmental exposures, necessitating more sophisticated experimental and analytical approaches to unravel cellular heterogeneity and disease mechanisms (Tanay and Regev, 2017; Stuart and Satija, 2019). As scRNA-Seq technologies have advanced, with improvements in sensitivity, throughput, and computational methods, the focus has expanded to more complex and diverse systems, including human tissues. This progression has led to ambitious projects such as the Human Cell Atlas, which aims to create comprehensive reference maps of all human cells (Regev et al., 2017). Such initiatives leverage cutting-edge scRNA-Seq techniques to catalog the myriad cell types in the human body, uncover their functions, and understand how they contribute to health and disease.

With the motivation from the accessibility and advancement of human scRNA-seq studies, an increasing number of scRNA-Seq investigations target patient populations and emphasize the impact of single-cell variation on human health outcomes. These population-based scRNA-Seq studies typically involve scRNA-Seq data from larger cohorts of individuals who are selected from populations exhibiting various health-related phenotypes. Such clinical relevance of scRNA-seq in human studies demands robust methods for uncovering and deciphering genetic information with diagnostic and therapeutic implications, while adequately accounting for human sample variations.

1.3 Motivation for Population Level Analysis of scRNA-Seq

Although there have been many methodological advances in human scRNA-Seq study, most published methods primarily aim to interpret single-cell level information and do not address population-level analysis adequately. Typically, human scRNA-Seq data is analyzed with individual cells as the primary unit of data such as identifying marker genes for specific cell types, or differentially expressed genes among phenotypes. During the early stage, sample variation is usually omitted and cells from different samples are used together to perform differential expression analysis. However, this raises problems when we deal with multi-sample scRNA-Seq data. For instance, during the exploratory data analysis process of a skin rash disease data (Liu et al., 2022), we discovered population scale heterogeneity issues for patients within the same condition group. Certain genes have higher expression in some of the patients and lower in others, such as the gene `PI3` in Monocyte-Derived Macrophages, shown in Figure 1.3. More importantly, the multi-sample scRNA-Seq data is a nested design for samples and the phenotypic groups where the differentially expressed genes are to be identified. Ignoring the sample effect in such a nested design can lead to an underestimation

of the variance, as it does not account for the hierarchical structure or the correlation between nested observations. This underestimation often results in smaller p-values, which increases the likelihood of incorrectly rejecting the null hypothesis. Some existing tools focusing on detecting differentially expressed genes accounts for population variability by implementing methods such as mixed effects or hurdle models (Crowell et al., 2020; Finak et al., 2015; Tiberi et al., 2020; Zhang et al., 2022).

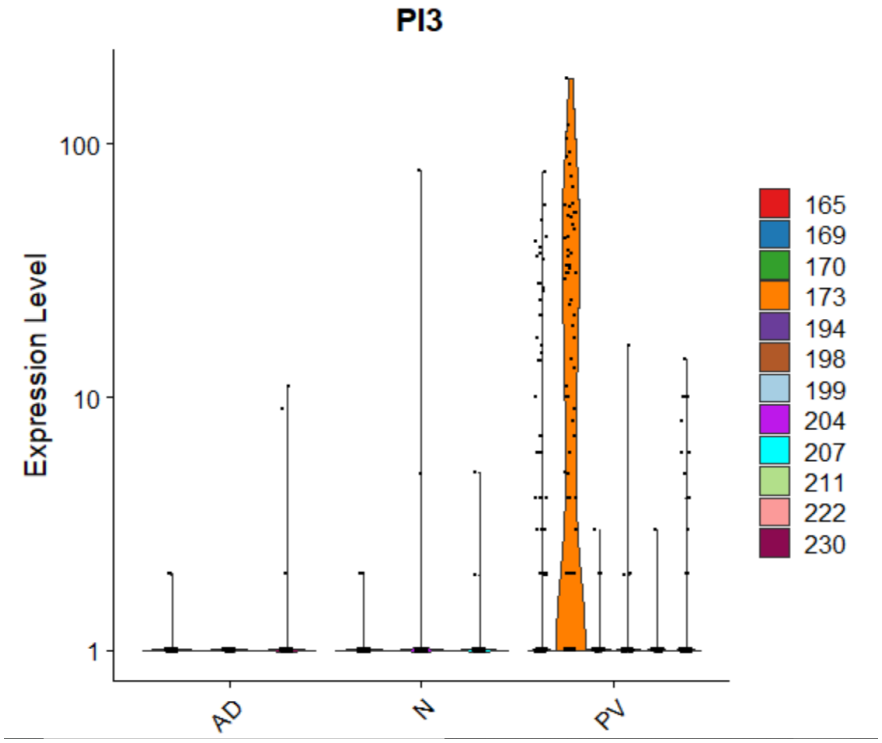


Figure 1.3: **Violin plot of `PI3` gene expression in Monocyte-Derived Macrophages (Liu et al., 2022)** Each column represents a patient, with y-axis showing the gene expression values. AD = Atopic Dermatiti, N = Normal, PV = Psoriasis vulgari. One of the PV patient (`173`) has relatively more abundant `PI3` expression than other patients belong to PV class.

Beyond differentially expressed genes analysis, the available methods for sample-level analysis are mostly limited to comparing the relative proportions of cell subgroups between samples with different phenotypes (Li et al., 2020a), or comparing samples through aggregation methods to summarize sample's single-cell profile into one observation, such as pseudo-bulk. However, such methods diminish the advantage of scRNA-Seq in revealing cellular or genetic heterogeneity, thereby obscuring the detailed diversity of cell populations that single-cell resolution offers. The drawbacks are further discussed in section 4.1.

Because there is such a need to appropriately address sample level variation for multi-sample scRNA-Seq data, we investigate into this issue and propose an analysis protocol:

creating a global representation of samples using their entire single-cell profiles. With such global representation, we could perform essential downstream analysis such as visualization and quality control. We refer to such an approach as the patient-level (or sample-level) analysis.

1.4 Overview of the Population-level Analysis

Common computational approaches assume that each sample is measured using a shared set of features. However, when we first attempt to perform patient-level analysis on scRNA-Seq data, where we consider each patient, instead of cells, as the unit, we note that the format of scRNA-Seq data is not applicable for such approaches. More specifically, for each sample i, we obtain a matrix $X_i \in R^{m_i \times g}$, containing gene expression measurements for that sample across all its cells (here g represents the number of genes and m_i represents the number of cells sequenced from sample i). It is obvious that there is no direct alignment for $m_{i,1}$ cell in sample i to $m_{j,1}$ cell in sample j, or so on. As there is no direct correspondence between the cells in different samples, it is difficult to align data from the samples for use in a statistical model or predictive algorithm.

To address this challenge, we propose an analysis pipeline that uses the entire single-cell profile of a sample instead of focusing on cells as units. More specifically, we propose that there exists such a global profile F_i, which could summarize and represent each patient i, across all cells X_i. Such approach does not require explicitly aligning individual cells across samples, but utilize the properties of observed data to represent each sample in a comparable space. We would note that this pipeline is not designed for gene level study such as differential expression analysis mentioned above. Instead, with such global representation F_i, we would use it as input to perform analysis in population scale such as patients' phenotype prediction, visualization, and quality control tasks. In Chapter 2, we introduce the framework we created for global representation of patient and detailed application to different scRNA-Seq data[1].

[1]The work was accomplished with joint collaboration with William Torous, PhD candidate in Statistics, University of California, Berkeley, and Boying Gong, PhD in Biostatistics, University of California, Berkeley. The work has been accepted by Genome Biology.

Chapter 2

Visualization at population scale with GloScope

2.1 Introduction to GloScope

For generality, we refer "sample" as tissue samples in single-cell studies, instead of "cells", which are collected from each sample. Furthermore, there might be multiple "samples" for one patient where the patient' single-cell profile was measured more than once.

To create the global representation F_i for each sample i as discussed in Chapter 1, we propose to represent each sample as a distribution of cells. More specifically, we consider the gene measurements for each cell to be a random draw from entire population of cells within each sample, and we summarize each sample with a probability distribution describing the statistical behaviors of cells' gene expression within the sample. Such representation allows us to summarize the overall scRNA-profile of a sample into a single mathematical object, while preserving useful information about the variability among single cells. This global representation can be used in a wide variety of downstream tasks, such as exploratory analysis of data at the sample-level or prediction of sample phenotypes. Furthermore, this representation does not require the classification of individual cells into specific cell types (e.g. via clustering) and, therefore, is not affected by any choice made during cell-type identification processes, such as resolution or clustering algorithm.

Probability Distribution

The full population of cells defines a probability distribution we designate as F_i on R^g. F_i is a representation of the sample's entire single-cell profile across all cells and importantly is a mathematical object that can be compared across samples. We do not observe F_i, but we do observe m_i samples from this distribution (the sequenced cells), allowing us to estimate F_i from the data. Thus, we transform each sample from the matrix X_i of observed gene expression measurements to an estimate of the sample's distribution, $\hat{F}_i$. Then we define a measure of divergence d on the space of probability distributions in R^g.

We make the simplifying assumption that the sequenced cells are independent and identically distributed (i.i.d) draws from the sample's full population of cells, F_i. However, even with this assumption, density estimation is complicated in this setting. For scRNA-Seq datasets, g is often in the range of 2,000-30,000 (the number of detectable genes given the sequencing depth). The number of cells per sample, m_i, can vary by experiment, and often ranging from 500 to 10,000 cells per sample. Furthermore, in any given cell, only a small subset of genes is actively transcribed, leading to many genes having zero or very low expression levels across the dataset. Therefore, the data from each cell is high dimensional and sparse, a distributional structure known to be impactful in the analysis of scRNA-Seq data (Pierson and Yau, 2015; Risso et al., 2018; Eraslan et al., 2019; Van den Berge et al., 2018; Jiang et al., 2022).

Defining a Latent Space

Because gene expression data lie in a high dimensional space, with the number of genes g in the thousands, estimating F_i directly from the cells is intractable. Even with several thousand cells per sample, it is infeasible to estimate the density in such a high dimensional space without the assumption of an underlying lower dimensional latent space. Thus, we assume that there exists a lower dimensional representation or latent variable Z_i in $R^{m_i \times d}$ which governs the gene expression of each sample i. Specifically, for each cell c in sample i, there exist a latent random variable Z_{ic} that we predict and we assume that the Z_{ic} is distributed as H_i, a distribution on R^d. Instead of estimating F_i in R^g, we estimate H_i from the m_i cells in the lower-dimensional space R^d and obtain the estimated density $\hat{H}_i$.

Unlike the X_i, which have different, unrelated, dimensions for each sample i, the $\hat{H}_i$ lies in the space of distributions on R^d and can be compared. As probability measures, these representations are now familiar mathematical objects and sample-level analysis can be done in the space of probability measures.

We estimate $\hat{H}_i$ by first estimating a lower dimensional representation of our all our cells. More specifically, let X be the count matrix contains all samples' cell information,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_n \end{pmatrix}, X_i \in R^{m_i \times g} \tag{2.1}$$

By applying dimension reduction techniques on X, we obtain the lower embedding matrix $\hat{Z}$ for all cells:

$$\hat{Z} = \begin{pmatrix} \hat{Z}_1 \\ \hat{Z}_2 \\ ... \\ \hat{Z}_n \end{pmatrix}, \hat{Z}_i \in R^{m_i \times d} \tag{2.2}$$

There, for each sample i, we estimate $\hat{H}_i$ from each $\hat{X}_i$, the m_i observed cells contained in Z_i.

PCA is a common choice for representing the data in lower dimension. But it is not the only option, and our approach can incorporate different dimension reduction methods' outputs. For example, Lopez et al. (2018) propose an alternative method, `scVI`, which uses a variational autoencoder (VAE) that optimizes the encoder and decoder networks to reconstruct the original data in a lower-dimensional latent space $\hat{Z}_i$. In Chapter 4, we elaborate on the comparison of different dimension reduction techniques.

Estimation of Statistical Divergence

Now having estimated each $\hat{Z}_i$ from X, we turn to estimating the density H_i so that we can estimate $d(H_i, H_j)$ to obtain divergence measure on probability measures.

There are many well-known metrics defined on the space of probability measures, such as the Hellinger Distance, Wasserstein Distance, or Bhattacharyya Distance, and downstream analysis can be performed after choosing a metric to quantify pairwise sample differences. For our examples, we implemented the Kullback-Leibler (KL) divergence

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{2.3}$$

to quantify the differences between sample probability distributions P and Q (Kullback and Leibler, 1951). Traditional KL divergence measures the information loss when one probability distribution is used to approximate another, but it is not symmetric: the divergence from distribution P to Q is not the same as from Q to P. To address this, we use symmetrized KL divergence $D_{\mathrm{SKL}}(P, Q) = D_{\mathrm{KL}}(P \parallel Q) + D_{\mathrm{KL}}(Q \parallel P)$. However, though the symmetrized version of KL divergence solve the symmetry issue, it does not satisfy other properties of a true metric (e.g., the triangle inequality). Another popular choice of divergence matrix is Jensen-Shannon (JS) divergence

$$D_{\mathrm{JS}}(P, Q) = \frac{1}{2} D_{\mathrm{KL}}\left(P \parallel M\right) + \frac{1}{2} D_{\mathrm{KL}}\left(Q \parallel M\right). \tag{2.4}$$

JS Divergence is a symmetrized and smoothed version of KL divergence. It is known as an alternative of KL divergence for its adherence to metric properties, including symmetry and the triangle inequality (Lin, 1991). On the other hand, many commonly used distance measures, such as Euclidean distance, are inherently based on the square root of squared differences. Therefore, adapting from JS divergence and square root calculation, we use the square root of the symmetrized KL divergence to make the measure becomes more consistent with these traditional metrics and helps approximate these metric properties better; while not a proper metric, this divergence can be effectively used to create a global representation of probability distributions and has been used in a similar manner in the case of facial recognition (Arandjelovic et al., 2005).

Here we offers two approaches for the estimation of $d(H_i, H_j)$. The first method directly estimates the density H_i for each sample. The second method does not explicitly estimate H_i, but focuses only on $d(H_i, H_j)$

Estimating H_i with Gaussian Mixture models

The first approach involves applying density estimation methods to the $\hat{Z}_{ic}$ to estimate h_i, the density associated with the distribution H_i, and then calculates $d(\hat{H}_i, \hat{H}_j)$ as our estimate of $d(F_i, F_j)$. Here we propose to use a Gaussian Mixture Model (GMM), a widely used probabilistic model for representing normally distributed mixtures within an overall population. In single-cell data, different cell types exhibit distinct genetic profiles. Intuitively, GMMs would make an ideal tool for modeling these variations by representing each potential cell type as a separate Gaussian component.

Single-cell methods utilizing dimensionality reduction, described above, often include a regularizing assumption that the latent variables Z $\sim$ N(0, Σ). This Gaussian regularization in the model and the fact that many single-cell datasets are mixtures of cell type populations, motivates our use of GMMs. Therefore, in our pipeline, GMM is employed to characterize different cell types and states by modeling the gene expression profiles of individual cells as a mixture of multiple Gaussian distributions, with each Gaussian component corresponding to a specific cell subpopulation, defined by its mean and covariance structure, and mixing coefficients representing the prior probabilities of each Gaussian component (or the probability of observing each cell types).

We use the R package mclust (Scrucca et al., 2016) to implement the GMM estimation. As there is no closed form expression for the KL divergence between GMM distributions, we use Monte Carlo integration to approximate the KL divergence between two GMM densities; this is based on R = 10, 000 samples drawn from the estimated GMM distributions, again using the mclust package. Specifically, for R draws of x from H_i, we have

$$KL(\hat{H}_i||\hat{H}_j) \approx \frac{1}{R}\sum_{u=1}^{R} \log \frac{\hat{h}_i(x_u)}{\hat{h}_j(x_u)} \tag{2.5}$$

Estimating $d(H_i, H_j)$ Directly via Non-parametric Nearest Neighbor Approach

We also provide a second approach that estimates $d(H_i, H_j)$ directly using a k-nearest neighbor approach without explicitly estimating the density H_i.

Denote $r_j(x_i, u)$ as the distance from the u^{th} cell in sample i to its k^{th} nearest neighbor in sample j, the KL divergence can be estimated directly as

$$\widehat{KL}(H_i||H_j) = \frac{d}{m_i}\sum_{u=1}^{m_i} \log \frac{r_j(x_{i,u})}{r_i(x_{i,u})} + \log \frac{m_j}{m_i - 1} \tag{2.6}$$

where d is the dimension of the latent space (Wang et al., 2006; Boltz et al., 2009). We implement this strategy using the `FNN` package to estimate the symmetrized KL divergence between sample i and sample j (Beygelzimer et al., 2024).

2.2 Usage of GloScope

In summary, our proposed representation method consists of representing each sample as a distribution along with a corresponding divergence or distance; we then estimate the distance or divergence between each pair of samples based on their estimated distributions. We call this representation of samples the **GloScope** representation, a global representation of scRNA-Seq data, and we illustrate this pipeline in Fig. 2.1. The detailed implementation of each step of GloScope is available in an accompanying Bioconductor package `GloScope`.

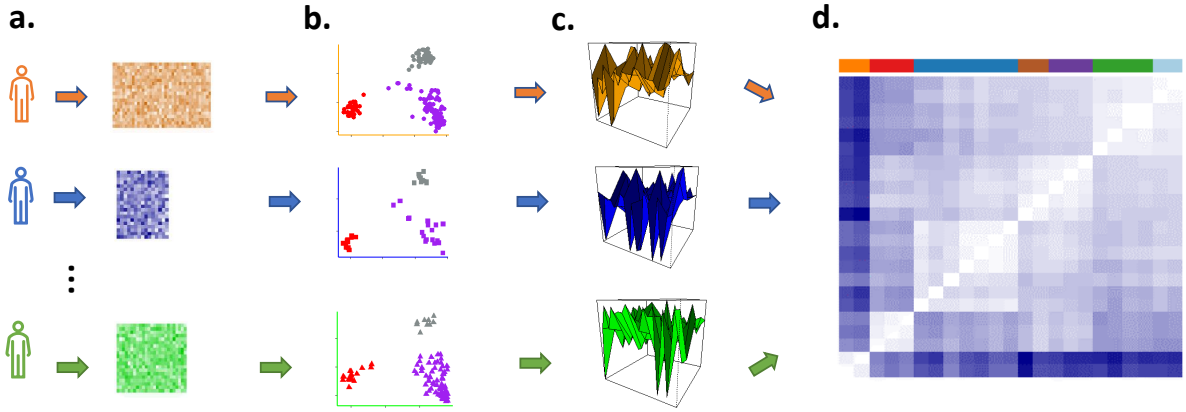


Figure 2.1: **Illustration of the GloScope representing a sample's scRNA-Seq data matrix X_i as a distribution $\hat{F}_i$.** (a) Each sample contributes a $g \times m_i$ matrix of gene expression values. (b) A lower dimensional latent representation is estimated across all cells and samples, resulting in each cell being represented in a lower-dimensional space (c) GloScope estimates the distribution $\hat{F}_i$ for each sample, and then (d) calculates the statistical divergence between each pair of samples, $d(\hat{F}_i, \hat{F}_j)$.

The final output of GloScope is a $n \times n$ matrix D of the pairwise divergences between the samples. The pairwise divergences between GloScope-represented samples can be given as input to canonical divergence analysis methods such as Multidimensional Scaling (MDS) (Cox and Cox, 2001). MDS takes a distance matrix as input and creates a coordinate system, a matrix M in $R^{n \times l}$ to represent the samples in lower dimensions while preserve the pairwise divergences. The primary objective of MDS is to position each object in this low-dimensional space such that the distances between points correspond as closely as possible to the original dissimilarities. Such coordinate system is most widely used across various fields such as

psychology, marketing, and bioinformatics to visually interpret complex data and uncover underlying patterns and relationships, as shown in the remaining chapter. Though we would note here that as D is not a Euclidean distance matrix, there is loss in information, unlike if D was a Euclidean distance matrix. However, such information still provides us with useful inference regarding the data behavior. For example, beyond EDA, the output of MDS can also be used for other important downstream tasks, include clustering of samples and prediction of phenotypes (for example via kernel prediction methods, e.g. Hofmann et al. (2008); Wang et al. (2014)).

We can also use the divergences to numerically quantify the separation of groups of samples using silhouette width or ANOSIM statistics shown in Chapter 3. This allows us to quantify how separated samples are due to a biological condition of interest (e.g. healthy vs diseased samples), or alternatively how separated samples are due to a design artifact (e.g. different processing centers). We will demonstrate that such a representation enables detection of possible batch effects or outliers and exploratory assessment of phenotypic differences between our samples through simulation and hypothesis testing in Chapter 3 and 4.

Furthermore, there are many existing methods for working with scRNA-Seq data, and GloScope is designed to fit into standard pipelines and complement existing quality-control and EDA strategies. Gloscope takes as input low-dimensional latent representations of the individual cells, which can come from PCA, scVI, or from batch-correction methods like Harmony (Korsunsky et al., 2019). As shown in Figure 2.2, GloScope can be performed at different stages of the pre-processing, allowing checks at each stage of whether patient-level artifacts, like processing batches, are inappropriately contributing to differences in the samples.

2.3 Visualization of patient phenotypes using GloScope

In the previous section, we introduced the concept of population-level analysis of scRNA-Seq data and the framework of GloScope Representation. The remainder of this chapter builds on that foundation by exploring the practical application of GloScope across a diverse range of datasets, with study designs varying from as few as 12 samples to more than 300 samples. Through these varied examples, we aim to demonstrate the versatility and robustness of GloScope in handling datasets of different scales and complexities.

We will showcase how GloScope facilitates critical bioinformatic tasks, particularly focusing on the visualization of scRNA-Seq data at the population level. By enabling researchers to visualize complex data in an intuitive and interpretable manner, GloScope helps uncover patterns and insights that might be obscured in cell-level analysis. This capability is crucial for understanding and identifying meaningful biological variations and heterogeneity within large populations.

Moreover, we will discuss specific case studies where GloScope has been applied to differ-

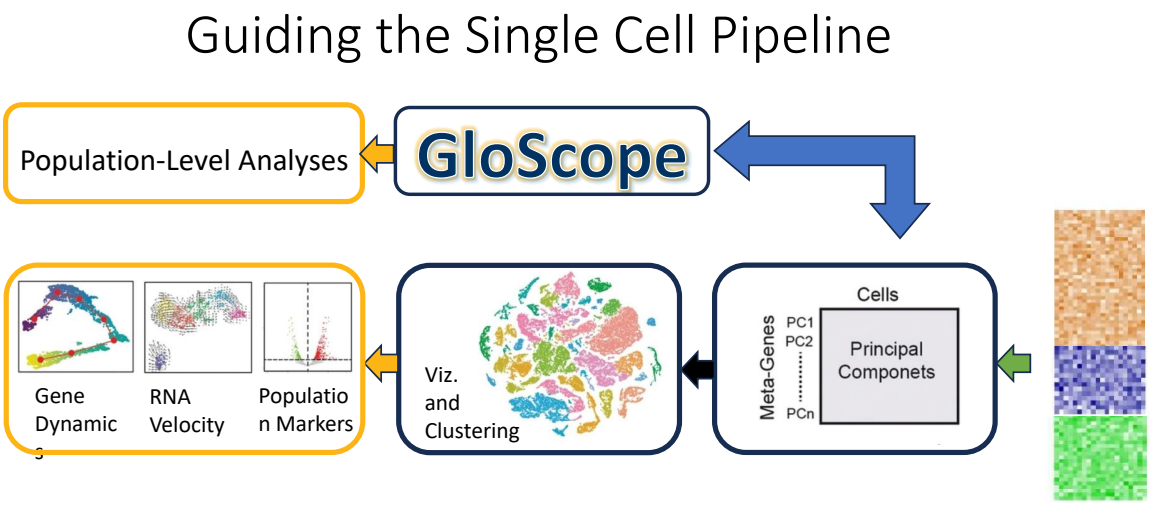


Figure 2.2: **Illustration of the utility of GloScope representing in single-cell data analysis pipeline.**

ent types of scRNA-Seq datasets, highlighting its adaptability to various research contexts and objectives. Whether dealing with small pilot studies or large-scale projects, GloScope provides a powerful tool for researchers aiming to perform comprehensive and insightful analyses of single-cell data. Through these examples, we hope to illustrate not only the technical advantages of GloScope but also its practical benefits in advancing the field of single-cell genomics.

Real World Data Overview

To illustrate the effectiveness of GloScope, we have applied it to several diverse datasets, as summarized in Table 2.1. The datasets vary significantly in size, including small-scale studies with as few as 12 samples and large-scale studies with over 300 samples. The table details each dataset's sample size and key characteristics, providing a clear overview of the diverse applications of GloScope.

Data processing procedures

This section details the steps undertaken to estimate GloScope representations of samples from publicly available scRNA-Seq data. These steps broadly consisted of ensuring the data

Dataset	#Genes	#Cells	#Samples	#Batches
Allenmouse	26,877	1,169,213	59	0
Skin Rash	19,769	92,889	12	0
Covid Lung	29,925	116,313	27	0
Colon Cancer	28,951	359,318	99	0
Covid PBMC	24,929	647,366	143	3
Lupus PBMC	30,933	1,263,676	336	4
Lung Fibrosis	19,680	714,923	144	6
Liver Fibrosis	21,045	326,351	50	6

Table 2.1: Properties of datasets analyzed through GloScope Representation

we used had quality control matching the corresponding paper, estimating the cells' latent embeddings, and applying the GloScope methodology. For most datasets we performed the first two steps with data structures and functions from the R package `Seurat`. For the larger Lupus PBMC and mouse brain datasets, we instead utilized the `SingleCellExperiment` data structure and applied functions from other packages. Code for running these analyses, as well as text files containing data sources and specific processing choices, are available in the following GitHub repository: `https://github.com/epurdom/GloScope_analysis`

Examples of Applying GloScope for Visualizing scNRA-Seq data at Population Scale

In this section we demonstrate the utility of the GloScope representation to visualize and evaluate sample-level phenotypic differences. As an initial illustration, we consider two datasets with replicate samples collected for each phenotype, where the phenotypes have well-known biological differences in cell-type structure. These serve as an initial proof-of-concept of the GloScope representation.

The first dataset is scRNA-Seq data from the mouse cortex (Yao et al., 2021). Here the samples are cells from different regions of the brain with replication in each from three genetically identical mice. This is a dataset where we know the regions have distinct compositions of cell types and gene expressions. When we visualize these samples using the GloScope representation in Figure 2.3A, we see these distinctions clearly. The samples from the two main subdivisions of the cortex, isocortex (CTX) and hippocampal formation (HPF), clearly separate. Furthermore, we see that replicate samples from the same region strongly cluster with each other, while different regions are generally well separated. Within the CTX region, we observe blocks of biologically meaningful brain region groups such as the sensory and visual area: primary somatosensory (SSp), posterior parietal association (PTLp), visual area (VIS), and the Somatomotor areas: primary motor (MOp) and secondary motor (MOs). We also observe clustering of physically adjacent brain regions such as temporal associa-

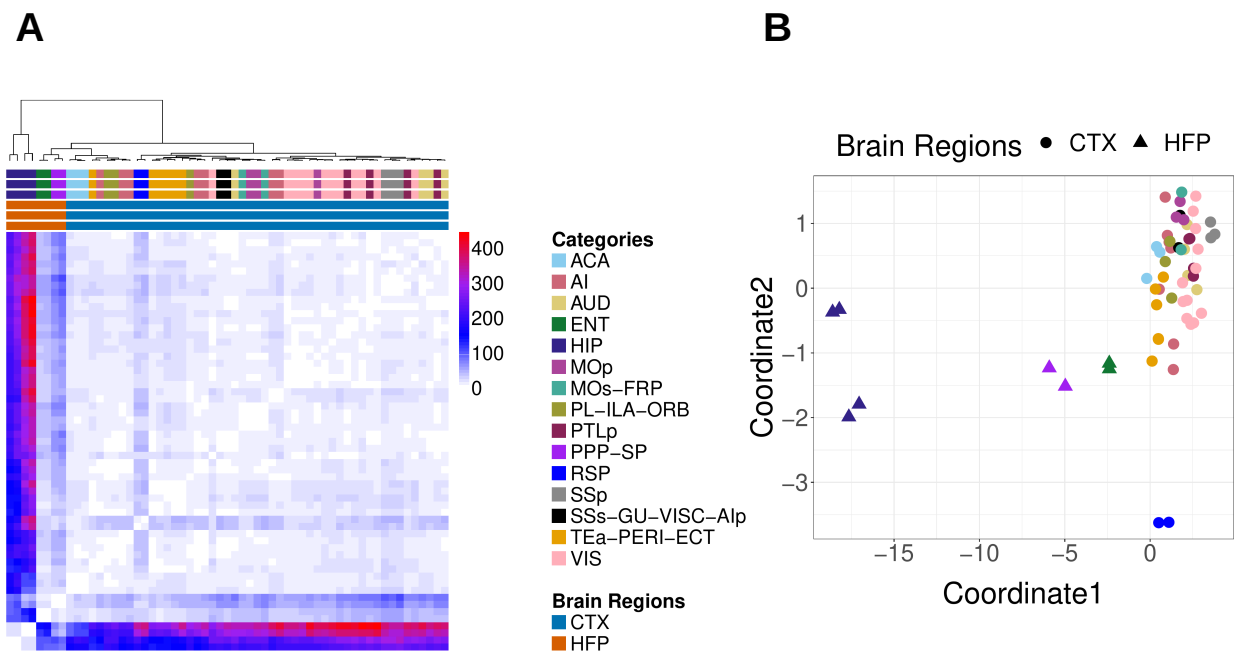


Figure 2.3: **Demonstration of the GloScope representation on 59 mice samples (Yao et al., 2021).** (A) Heatmap representation of the estimate of the divergences between the samples based on the GloScope representation. (B) A two dimensional representation via MDS of the divergences shown in **A**. GloScope used the GMM estimate of the density in the first 10 PCA dimensions. The individual regions represent subregions of two main divisions of the cortex: the isocortex (CTX) and hippocampal formation (HPF). HPF is further divided into hippocampal region (HIP), and the retrohippocampal region (RHP) which is represented by the entorhinal region (ENT) and the remaining RHP, a joint dissection region of postsubiculum (POST)-presubiculum (PRE)-parasubiculum (PAR) region, subiculum (SUB), and prosubiculum (ProS) region (i.e, PPP-SP). The remaining regions are divisions of the CTX.

tion, perirhinal, and ectorhinal areas (TEa-PERI-ECT), agranular insular (AI), prelimbic, infralimbic, orbital area (PL-ILA-ORB) and anterior cingulate (ACA).

Next we consider skin cell samples from a study of twelve patients (Cheng et al., 2018), consisting of nine healthy skin samples from the foreskin, scalp, and trunk alongside three inflamed skin samples collected from truncal psoriatic skin. We expect marked differences between cellular distributions collected at the different locations in the body due to varying proportions of cell types in certain tissues. For instance the authors note different types of main basal keratinocytes and melanocytes dominate in scalp and trunk samples, as compared to foreskin tissues. Our visualization of the GloScope representations of this data in Figure

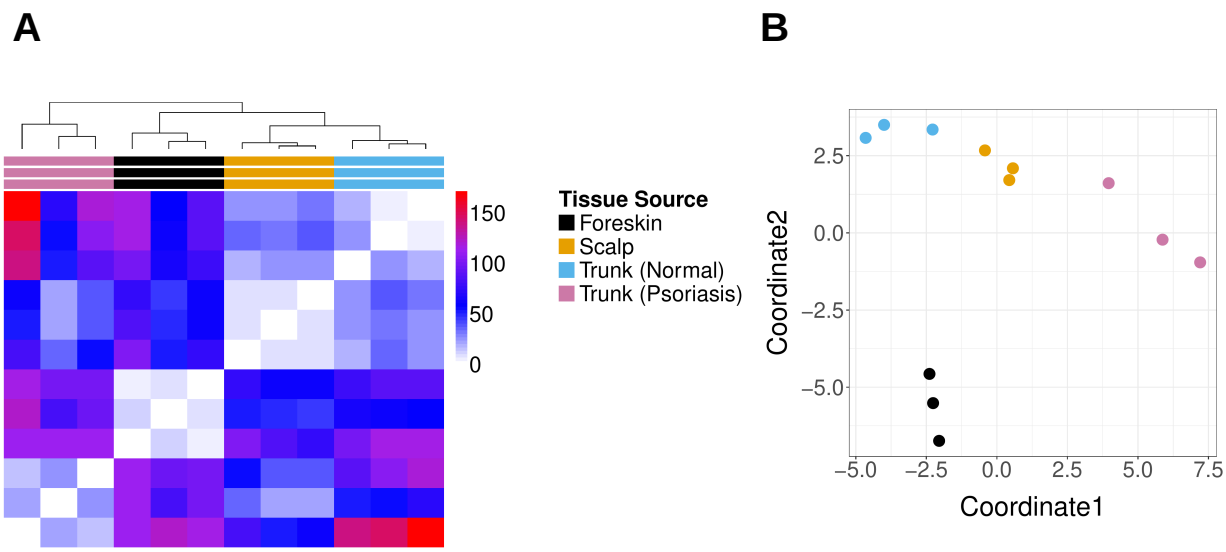


Figure 2.4: **GloScope representation of 12 skin rash patients collected in various locations and conditions in Cheng et al. (2018).**(A) A heatmap visualization of the estimate of the symmetrized KL divergence between the samples' GloScope representation. (B) A two dimensional MDS representation of the divergences. The divergences were calculated using the GMM density estimation based on PCA estimation of the latent space in 10 dimensions.

2.4 shows a clear clustering of skin samples collected from similar locations on the body, and a separation of both the foreskin and psoriasis samples from scalp and trunk samples, echoing the conclusions of the authors who identified a keratinocyte subpopulation which separates these phenotypes from the scalp and trunk control samples (Cheng et al., 2018).

Next we demonstrate the GloScope representation on additional datasets of patient cohorts where the samples are patients with differing disease phenotypes: 1) COVID lung atlas data from Melms et al. (2021), which contains 27 samples, either diagnosed with COVID-19 or healthy control samples, and 2) Colorectal cancer data with 99 samples (after quality control), grouped into three phenotypes: healthy, mismatch repair-proficient (MMRp) tumors, and mismatch repair-deficient (MMRd) tumors (Pelka et al., 2021).

The use of GloScope on these datasets demonstrates its utility for the visualization of both sample and phenotype variability. For the COVID lung samples (Figure 2.5A), we can easily see the separation between COVID-infected and healthy donors, matching the observation of Melms et al. (2021) that lung samples from COVID patients were highly inflamed. For the colorectal cancer data, visualization of the GloScope representation shows healthy samples well separated from the tumor samples (Figure 2.5B). Though the two types

of tumors do not separate in this visualization, an Analysis of Similarities (ANOSIM, Clarke (1993)) test of significance applied to their GloScope divergences between these two groups does find their representations to be significantly different ($p = 0.001$), indicating that the representation is encapsulating systematic differences between the two tumors (see Section 3.2).

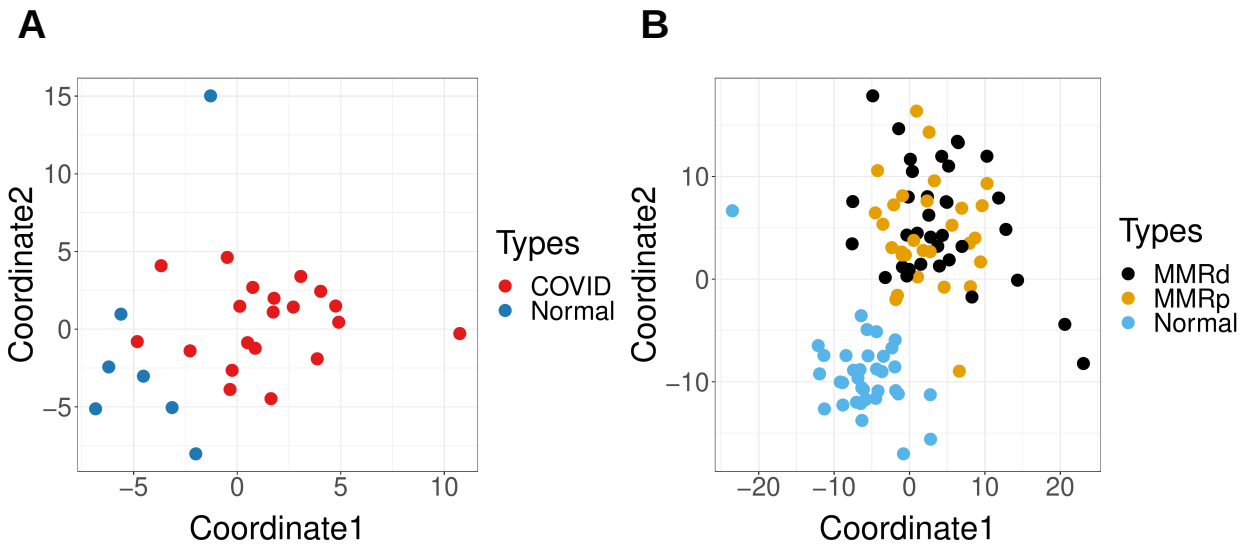


Figure 2.5: **Examples of MDS plot of the dissimilarities calculated from GloScope representation.** (A) 27 samples of COVID lung atlas data that are either healthy samples of COVID patients from Melms et al. (2021); (B) 99 samples colon samples from mismatch repair-proficient (MMRp) tumors, mismatch repair-deficient (MMRd) tumors and healthy samples from Pelka et al. (2021). The dissimilarity matrices were calculated using the GMM density estimate based on PCA estimates of the latent space in 10 dimensions.

2.4 Summary

In this Chapter, we demonstrated the use of GloScope for exploratory analysis, and in particular how the GloScope divergences can be used to create two-dimensional scatter plots of samples, similar to that of PCA plots of bulk mRNA-Seq data.

While we focus on the utility of the GloScope representation to visualize scRNA-Seq data at the sample level, the representation can be used more broadly with other statistical learning tools. For example, we can use the GloScope divergences between samples as input to a prediction algorithm in order to predict a phenotype. With the COVID PBMC data, we apply the SVM algorithm to the GloScope divergences which results in a prediction algorithm

that was able to separate the normal from the COVID samples with a 5-fold cross-validated prediction accuracy of around 0.88. This simple example serves as an illustration of the power of a global representation of the entire scRNA-Seq profile.

Finally, we note that GloScope can easily be incorporated into existing scRNA-Seq pipelines at multiple stages of analysis to assess the progress. Latent-varible representation, via PCA or scVI is a standard initial step in an analysis, while many popular batch correction methods provide low-dimensional representations of corrected data. Even multimodal integrations usually result in a low-dimensional latent space estimation. The output of all of these tasks can be provided to GloScope for evaluation of sample-level similarities, resulting in a flexible tool for exploratory analysis of the results.

In the next chapter, we would demonstrate the ability of the GloScope representation to detect important artifacts in the data, as well as assess batch-correction methodologies. In chapter 4, we would also compare GloScope to the limited available strategies for summarizing the data from a single patient: cell-type composition and pseudobulk.

Chapter 3

Batch Effect and Correction Methods Evaluation via GloScope

3.1 GloScope representation for Quality Control

In this section, we demonstrate the use of GloScope for exploratory data analysis of relatively large sample cohorts and illustrate the utility of having a sample-level representation of the data for exploratory data analysis and batch effect detection.

The first dataset is a study of COVID-19 (Stephenson et al., 2021) consisting of 143 samples of peripheral blood mononuclear cells (PBMC); samples in the study originated from patients that were either identified as infected with COVID-19 with varying levels of severity (COVID), negative for COVID-19 (Healthy), healthy volunteers with LPS stimulus as a substitute of an acute systemic inflammatory response (LPS), or having other disease phenotypes with similar respiratory symptoms as COVID-19 (non-COVID). Figure 3.1A shows these samples after applying MDS to the pairwise divergences calculated from the GloScope representation for the 143 samples of the study.

The visualization shows that both COVID patients and healthy donors are clearly separated from patients with other respiratory conditions (LPS and non-COVID). The other noticeable pattern is that the remaining patients do not show a strong separation between the COVID and Healthy phenotypes, but do appear to separate into at least two groups unrelated to these main phenotypes of interest – an observation that is further strengthened when considering the MDS representation of only the COVID patients and healthy donors (Figure 3.1B). Exploration of the provided sample data from Stephenson et al. (2021) shows that these groups correspond to different sequencing locations, indicating a strong batch effect due to sequencing site, with samples sequenced at the Cambridge site clearly separated from those at the New Castle (Ncl) and Sanger sites. When the individual cells are visualized (Figure 3.2), the distributional differences between these sequencing sites validate these differences, with cells from the Cambridge site lying in quite different spaces from cells of the same cell type from the other sequencing sites. Furthermore, Stephenson et al. (2021)

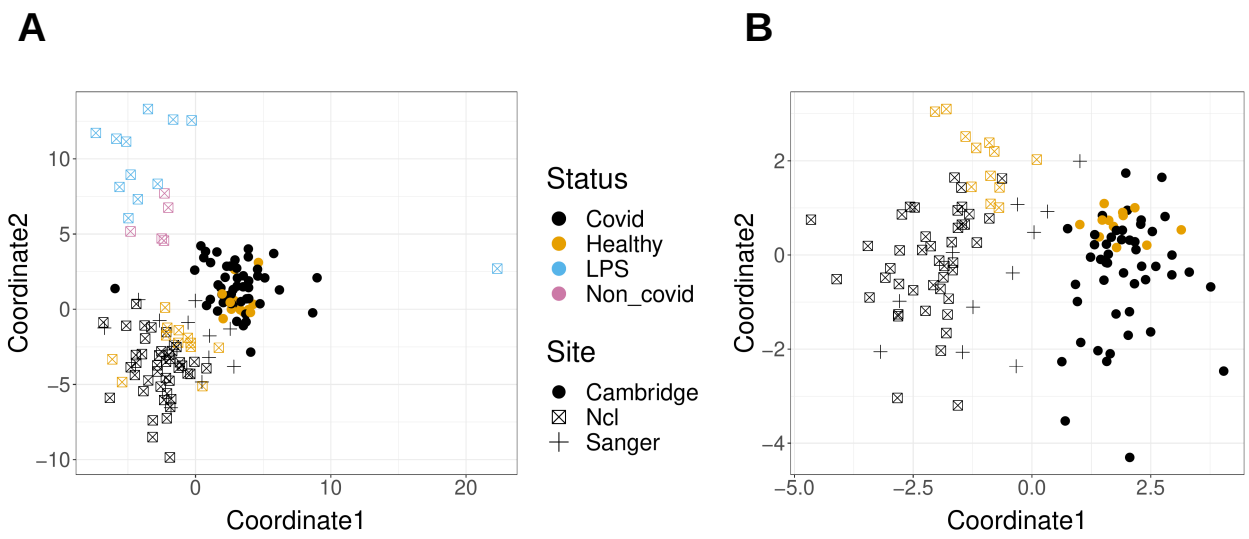


Figure 3.1: **GloScope representation applied to samples sequenced in Stephenson et al. (2021).** Shown are the MDS representation in two dimensions of the KL divergence estimates calculated from the GloScope representation for (A) all 143 samples and (B) the subset of 126 samples that were either healthy or diagnosed with COVID-19 (MDS was rerun on the reduced subset of divergences between these 126 samples). Each point corresponds to a sample and is colored by the sample's phenotype; the plotting symbol of each sample indicates the site at which the sample was sequenced (see legend). Estimated GloScope divergences used the GMM estimate of density and latent variables were estimated with PCA in 10 dimensions.

indicates that samples from these different sites underwent different sequencing steps such as cell isolation and library preparations (and the original analysis in Stephenson et al. (2021) corrected for potential batch effects by applying the batch correction method, Harmony (Korsunsky et al., 2019)).

A similar analysis was applied to a Systemic lupus erythematosus (SLE) dataset, with scRNA-Seq data of the PBMC cells of 261 patients; some patients had multiple samples resulting in total 336 samples (Perez et al., 2022). Again, our GloScope representation clearly shows that there are distinct patterns among different batch sources, in addition to separation of normal samples from the other conditions (Figure 3.3A). After application of Harmony to this data based on the batch, our GloScope representation shows much greater intermingling of the data from different batches (Figure 3.3B). We can quantify the improvement by measuring the separation between samples within a batch compared to those in separate batches using measures such as the ANOSIM R statistic or Silhouette width (see

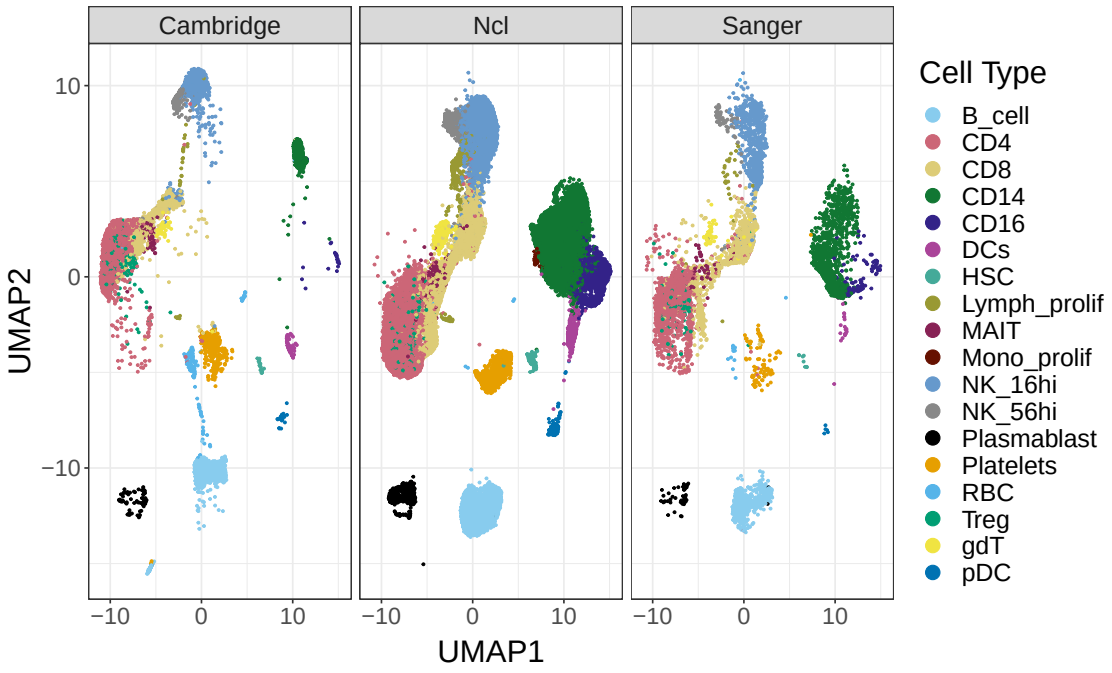


Figure 3.2: **UMAP plot of a subset of 50,000 cells from the original, uncorrected single-cell data from (Stephenson et al., 2021).** The UMAP embedding was calculated based on all cells, and then cells from different sequencing sites were plotted in separate panels. The plot does not indicate the individual samples, but plots all cells from the same sequencing site together regardless of sample or disease status. The cells are color-coded in each panel by the cell-type of the cell, as identified by Stephenson et al. (2021) following batch correction with Harmony. The UMAP was calculated from the first 30 PCA dimensions. This visualization shows the clear differences due to sequencing site in the cells which were identified to be in the same subtype, such as B-cells, CD4 cells and NK_56 high (CD56 bright NK cells).

Section 3.2). We see the improvement due to batch correction, but some loss of separation between biological conditions, which is a common trade-off when correcting for batch effects (Figure 3.3C and 3.3D). This type of exploratory analyses of data is a common task in the analysis of scRNA-Seq data, and the GloScope representation provides a meaningful strategy for evaluating these types of processing choices. We further note that in addition to finding differences amongst the sequencing sites in the Lupus PBMC data, we observe further clustering of samples in Batch 4 (highlighted in Figure 3.3A). These subgroups do not correspond with any patient covariates provided by the authors, but further exploration clearly show strong differences in the gene expression and cell density in certain cell types such as CD4 T cells, Natural Killer cells, and B cells. (Figure 3.4 and 3.5).

Similar concerns are frequently explored when integrating data from different studies.

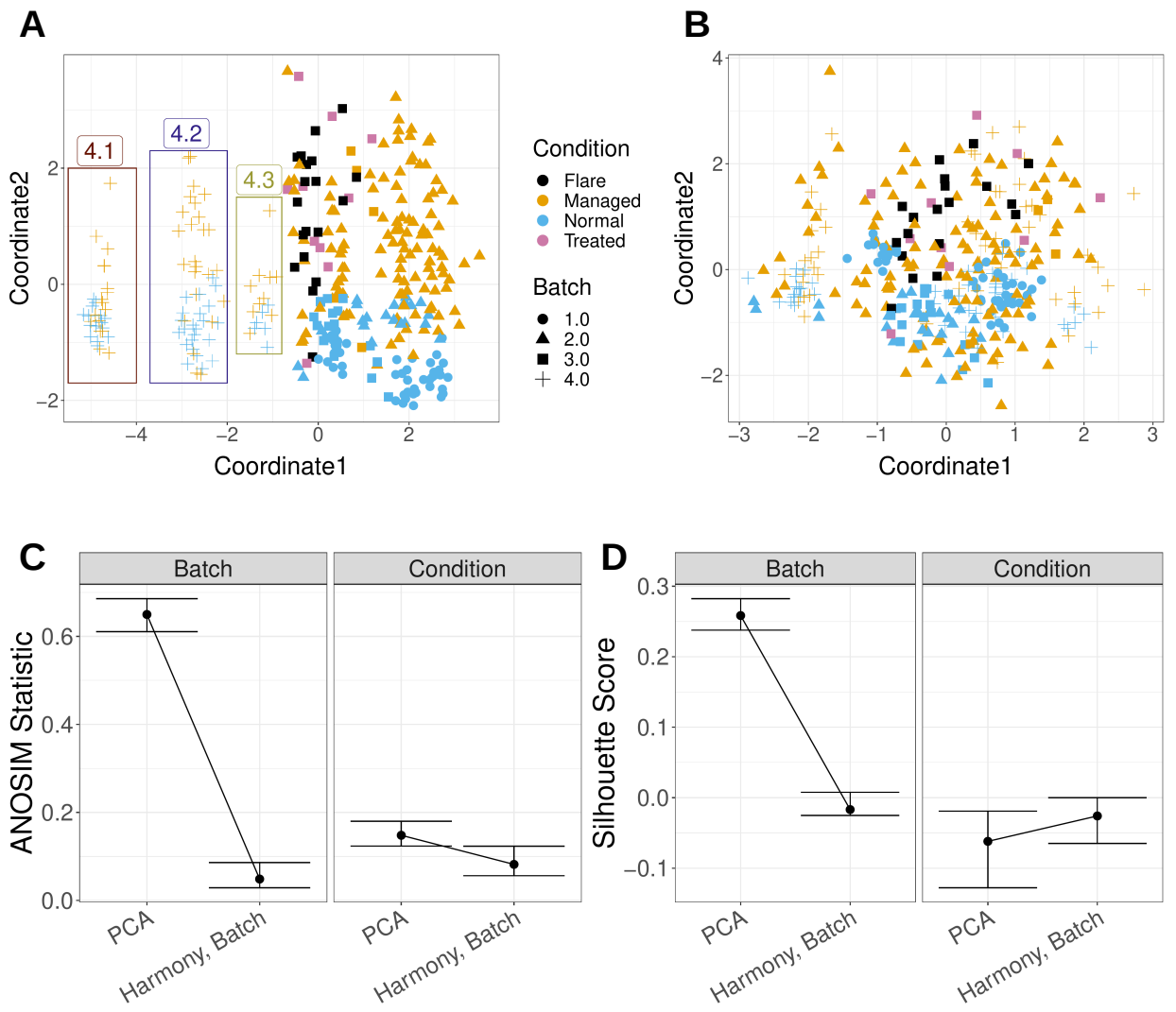


Figure 3.3: **GloScope representation applied to a Systemic lupus erythematosus (SLE) dataset of 336 samples from 261 patients (Perez et al., 2022).** Shown is the MDS of the GloScope representation applied to latent variables defined by (A) the first 10 PCA components of the original data and (B) the latent variables defined by Harmony after normalizing on processing cohort. (C) the ANOSIM statistics changing regarding capturing batch or condition signal, before and after applying batch correction (i.e Harmony) with bootstrap confidence interval. (D) the Silhouette widths changing regarding capturing batch or condition signal, before and after applying batch correction with bootstrap confidence interval.

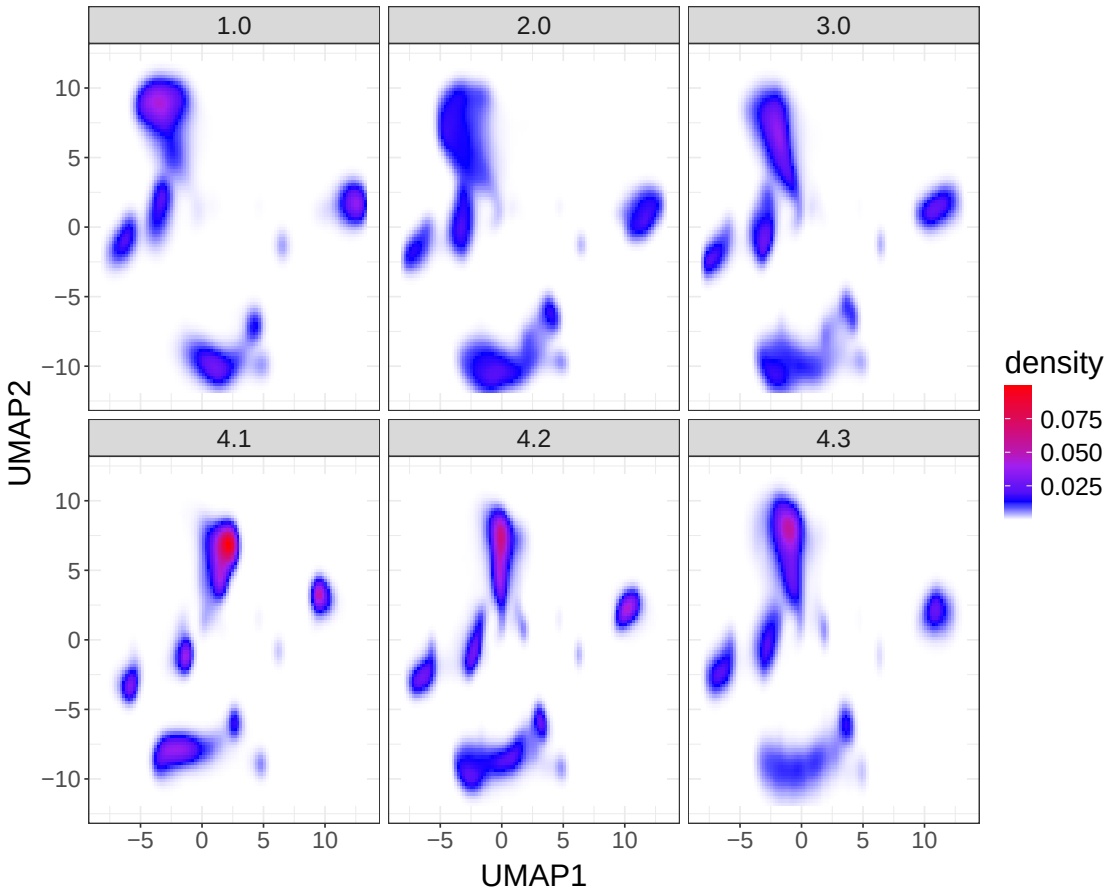


Figure 3.4: **UMAP visualization of the cell density of data from Perez et al. (2022)**, Each panel is the subgroups of batch identified by GloScope Representation.

We applied GloScope on the dataset of Fabre et al. (2023) which integrated six lung fibrosis scRNA-Seq studies, resulting in 144 samples after quality control. Application of GloScope (Figure 3.7A) immediately shows one of the studies (Adams et al., 2020) as quite different from the other five; further investigation shows that the study of Adams et al. (2020) has quite obvious differences in both gene expression and cell type composition than the other five studies. In particular, we observed quite obvious gene expression shifting in myeloid cells and Natural Killer cells in Adams et al. (2020) (See Figure 3.9), and samples collected from Adams have a higher portion of myeloid cells compared to samples from other studies (Figure 3.8). The remaining five studies show relatively smaller differences, but some separation is clearly visible. In addition to large batch effects, we observed a potential outlier (sample `092C_lung`), from the Adams et al. (2020) study detected by the GloScope representation (Figure 3.7A). Further evaluation of that outlier sample shows that `092C_lung` is missing most of the cell types except for B cells and lymphocytes (Figure 3.10).

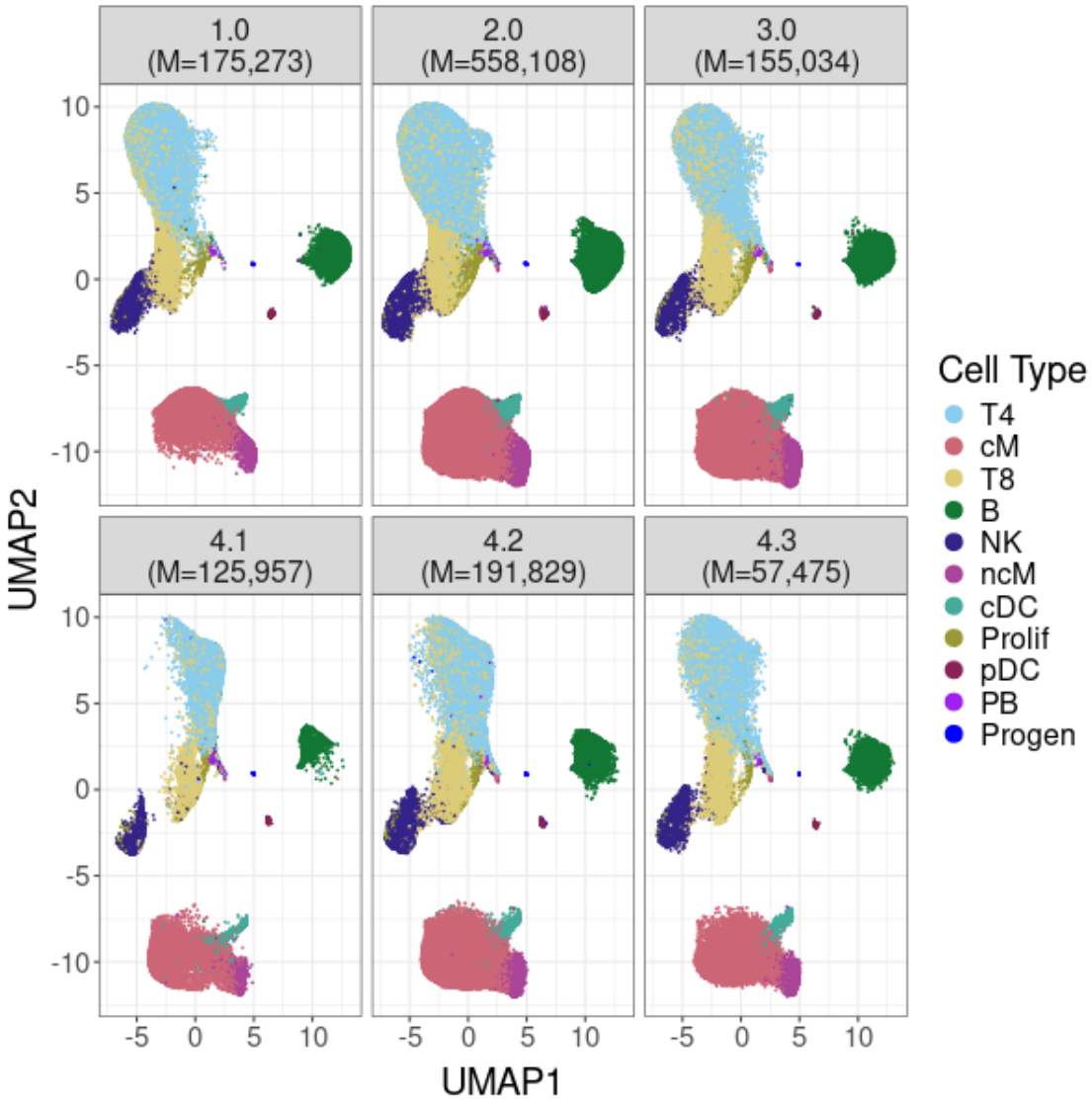


Figure 3.5: **UMAP visualization of the potential subgroups of batch 4 from Perez et al. (2022).** For description of UMAP calculations and color annotations, see Fig. 3.6. The upper panel is the first 3 original processing batches provided by Perez et al. (2022), as shown in Fig. 3.6. The lower panel further separates the fourth batch into the subgroups identified by GloScope.

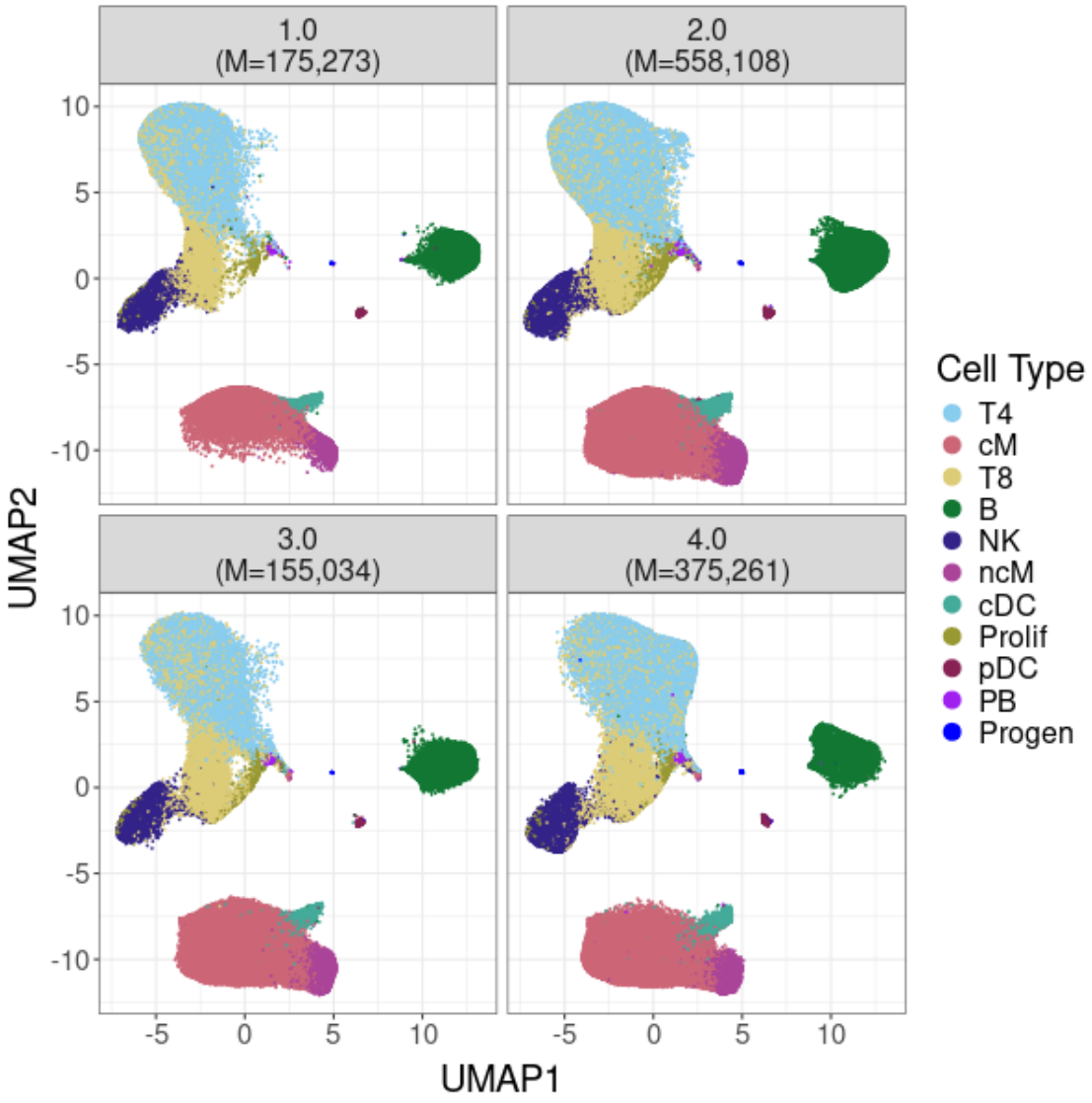


Figure 3.6: **UMAP visualization of the gene expression per batch of data from Perez et al. (2022).** Cells are separated in different panels for batches provided by Perez et al. (2022). The number of cells (M) plotted in each panel are indicated in the panel title. The cells are color-coded in each panel by the cell-type identified by Perez et al. (2022) using canonical marker genes. The first 10 PCs were used to calculate the UMAP representation across all samples.

In contrast, a similar analysis of data from (Fabre et al., 2023) which integrated studies of 50 human liver samples (after quality control) from 6 published scRNA-Seq studies of liver fibrosis shows far less distinction among the studies compared to the lung samples (Figure 3.11). Following application of Harmony for batch correction/integration, the GloScope shows effective integration of the lung studies and a corresponding clearer grouping of biological conditions. (Figure 3.7C,D).

3.2 Quantification of Batch Effects and Evaluation of Batch Correction Methods

Evaluating batch effects involves considering various aspects crucial to ensuring data integrity and accuracy in scientific research. Key considerations include selecting the appropriate batch correction method tailored to the specific experimental setup and data characteristics. Different correction methods, such as Harmony and scVI, offer distinct advantages depending on the nature of the batch effects and the type of data being analyzed (Korsunsky et al., 2019; Lopez et al., 2018). Additionally, choosing the correct batch effect unit, such as individual samples, or batches, is pivotal as it influences the scope and effectiveness of correction strategies. Each unit choice impacts how batch effects are identified, quantified, and ultimately corrected.

On the other hand, although batch correction is a valuable tool in data preprocessing to mitigate technical variations, its application must be approached with caution to preserve biological signal integrity. Blindly correcting for batch effects across datasets runs the risk of oversimplifying data variability and thereby obscuring genuine biological differences. For example, in Section 3.1 where we tried different batch correction methods, we noticed that some biological differences were diminished as well (e.g see Figure 3.3). By assuming that all differences are solely attributable to technical artifacts, batch correction methods may inadvertently attenuate or even eliminate meaningful biological signals that underlie variations of interest. Hence, while prioritizing data alignment for integration, we must also pay attention to the loss of biological signals.

The following sections will provide a comprehensive and detailed introduction to the application of GloScope in both assessing and quantifying batch effects in the sample level.

Numerical metrics for evaluating performance

As further discussed in Section 4.1, various quality-control tools have been developed to evaluate batch effects at the cell level (Korsunsky et al., 2019; Tran et al., 2020). These tools are effective in identifying and quantifying batch effects that may influence individual cell data, ensuring the integrity of single-cell analyses. However, despite the availability of these cell-level tools, there is a significant gap in the field when it comes to evaluating batch effects at the sample level. Currently, there are no dedicated tools designed to assess how batch effects might impact the overall sample, which can be crucial for studies that involve multiple

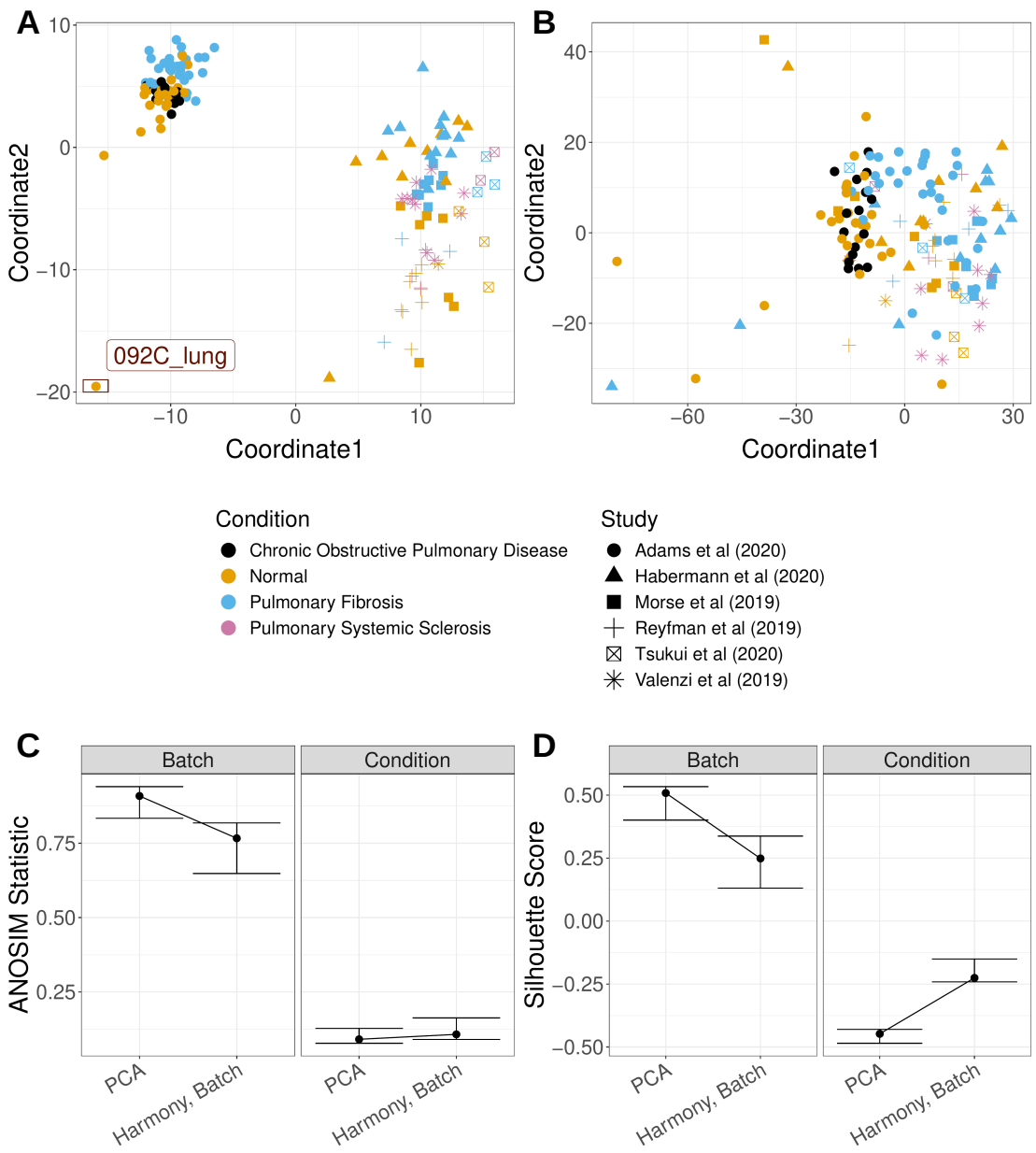


Figure 3.7: GloScope representation applied to samples sequenced in Fabre et al. (2023). Shown are the MDS representation in two dimensions of the KL divergence estimates calculated from the GloScope representation for (A) PCA embedding before batch correction and (B) PCA after applying Harmony batch correction. Each point corresponds to a sample and is colored by the sample's phenotype; the plotting symbol of each sample indicates the studies at which the sample was collected (see legend). Estimated GloScope divergences used the GMM estimate of density and latent variables were estimated with PCA in 10 dimensions. (C) and (D) visualize the ANOSIM R statistics and Silhouette width, quantifying the changes of batch and biological signals before and after batch correction

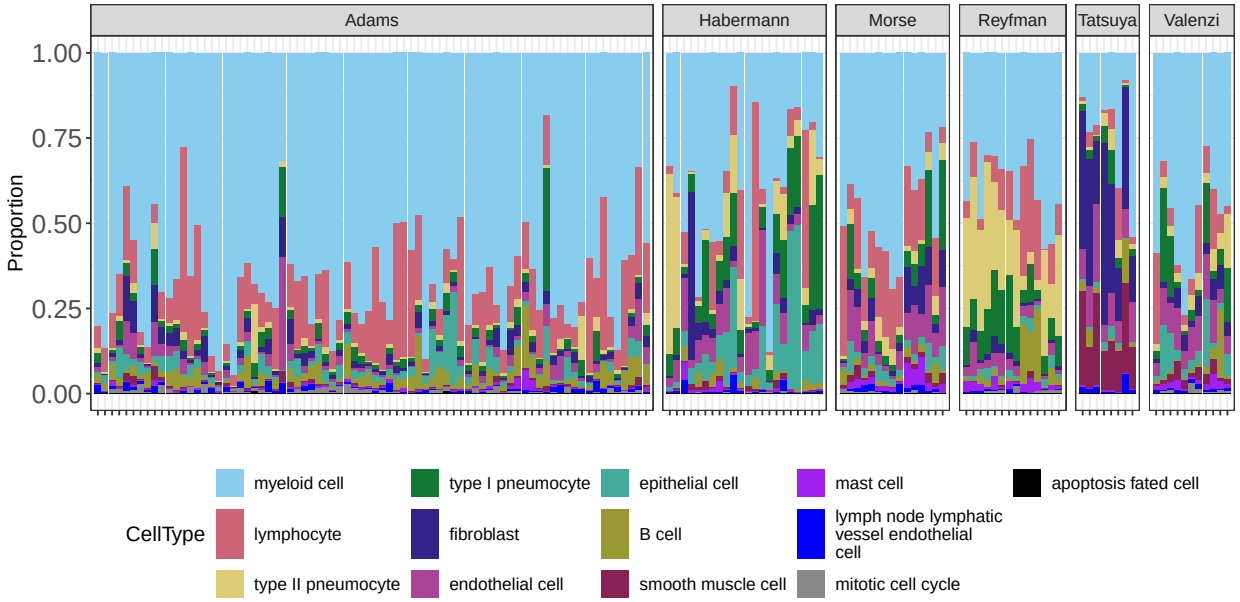


Figure 3.8: **Barplot visualization of celltype proportion per sample of lung study in Fabre et al. (2023).** Each column represent a sample and grouped into different panels by the study where the samples were collected. Bars are color-coded by the cell types identified by Fabre et al. (2023) following batch correction with Harmony. We are able to detect significant cell proportion differences (e.g myeloid cells) between Adams et al. (2020) and other studies.

samples or comparisons across different experimental conditions. Our method GloScope, instead, is capable of addressing this unmet need by utilizing the output divergence matrix as input for various statistical tests, such as ANOSIM.

In order to quantify how well our representation was able to differentiate sample groups or batch effects, and to compare with competing methods, we relied on the following metrics for evaluation:

1. ANOSIM R statistic
2. PERMANOVA effect size ω^2
3. Average Silhouette width using *silhouette* in R package `cluster`.

ANOSIM

The Analysis of Similarities (ANOSIM) test is a non-parametric test based on a metric of dissimilarity to evaluate whether the between group distance is greater than the within group

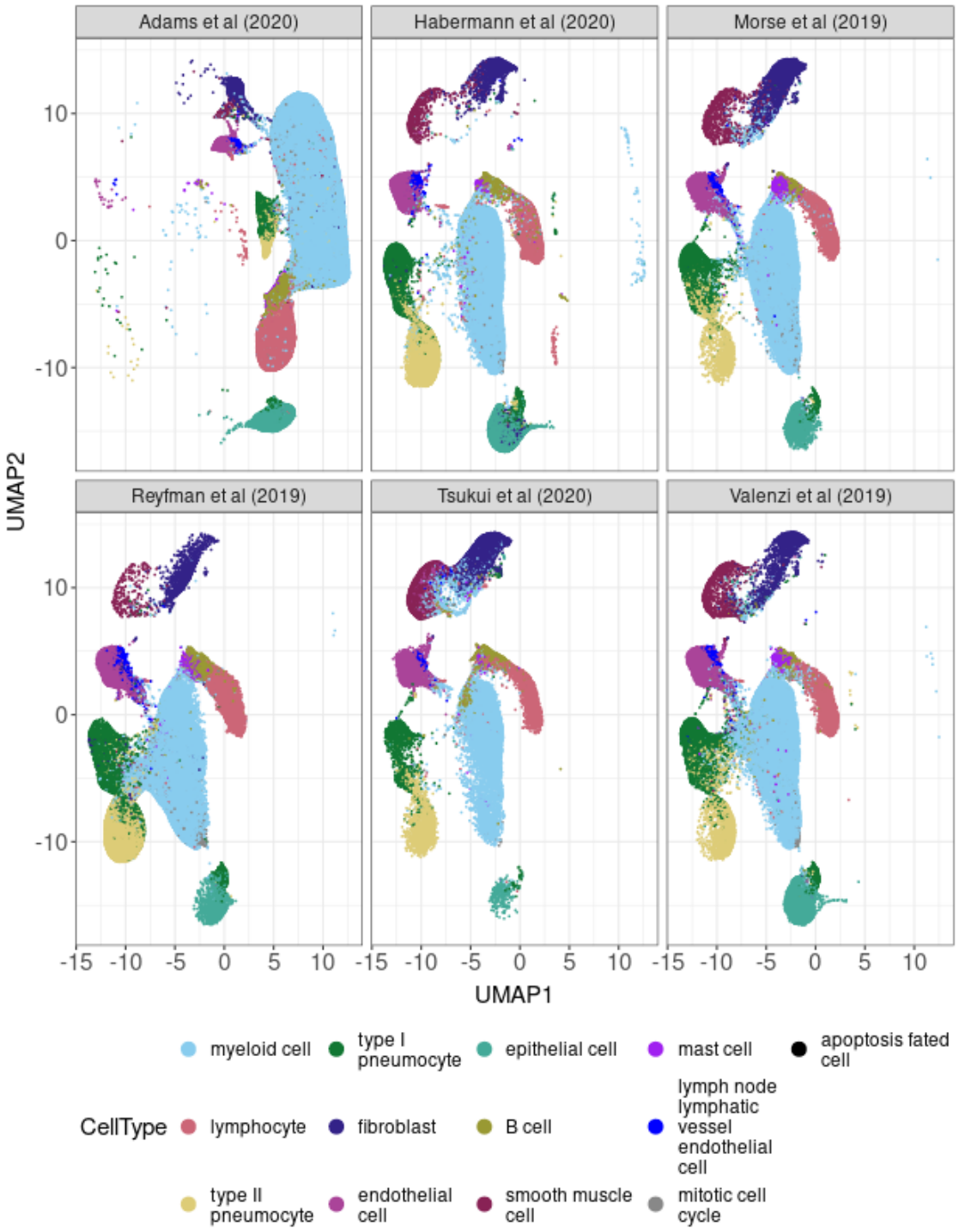


Figure 3.9: **UMAP visualization of individual cells from lung study in Fabre et al. (2023).** Each panel corresponds to cells in the six studies being integrated by Fabre et al. (2023), with Adams et al. (2020) showing widespread differences from the other studies. The cells are color-coded in each panel by the cell-type identified by Fabre et al. (2023) following batch correction with Harmony. The first 10 PCs calculated on all the cells jointly are used for UMAP calculation.

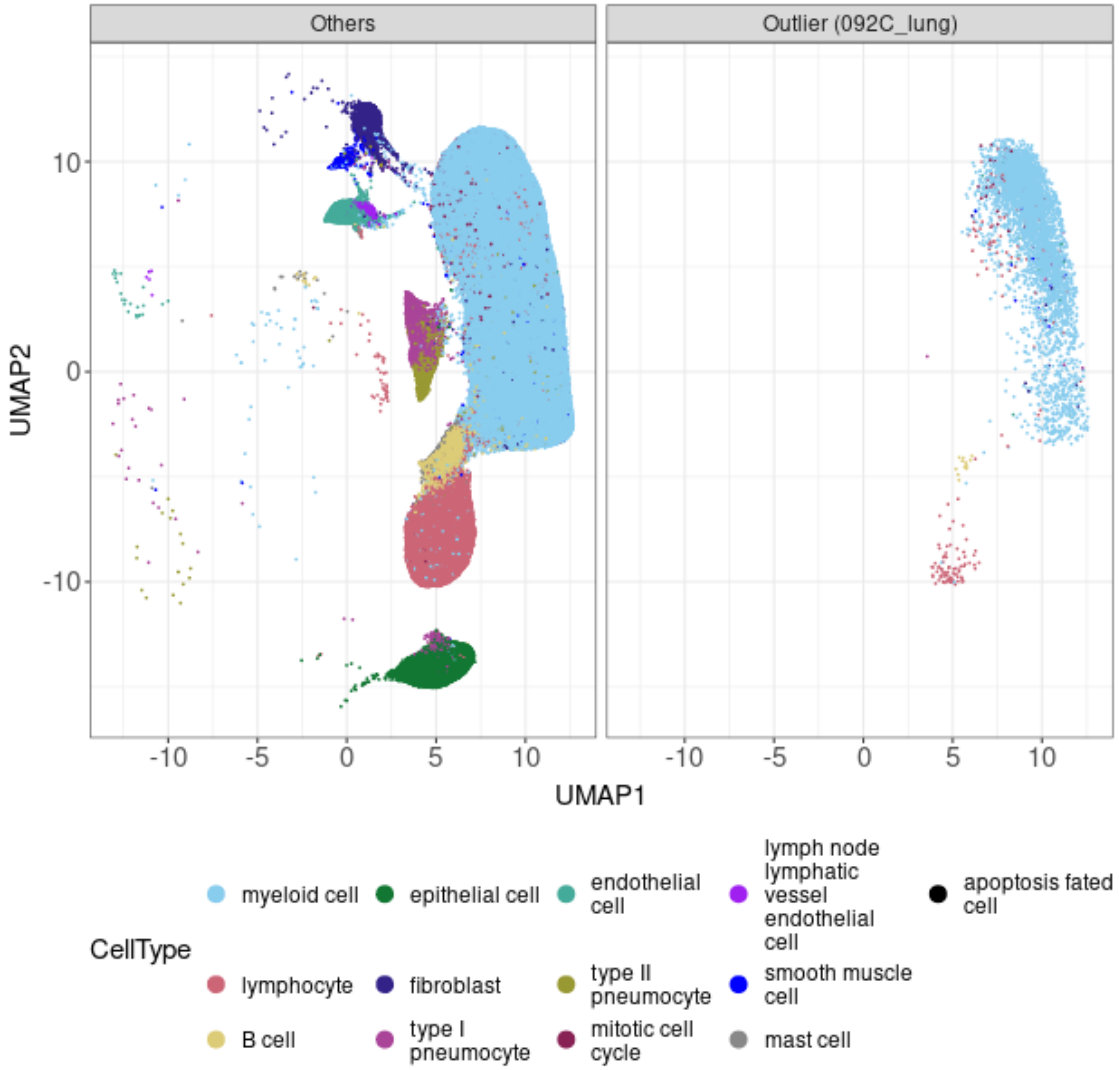
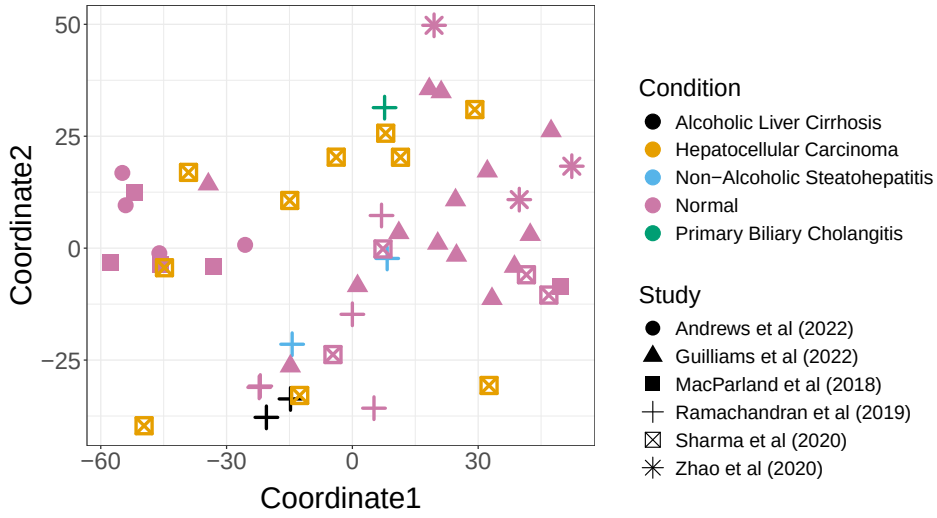
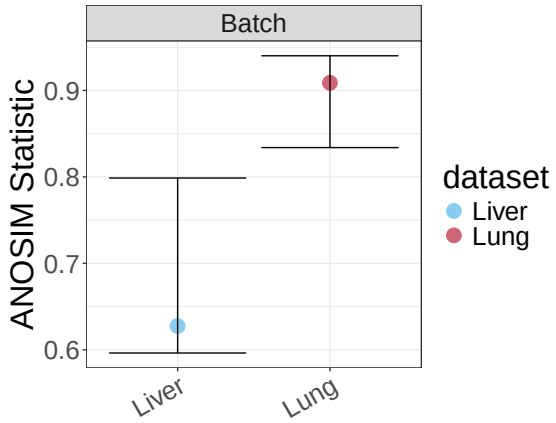


Figure 3.10: **UMAP visualization of individual cells of outlier sample compared to other samples from Adams et al. (2020).** For UMAP calculation and color annotation, see Fig. 3.9. Left panel is the cells from Adams et al. (2020) where the samples are not considered as outliers, and right panel is the cells from the outlier sample (092C_lung). Most of the cell types are missing for the outlier samples.



(A) MDS Plot



(B) ANOSIM Statistics

Figure 3.11: **Visualization of sample from liver study in Fabre et al. (2023).** (A) A MDS plot of the divergences calculated by GloScope, with samples color-coded by their biological condition and with the shape of the point indicating the study of origin. The liver study shows less obvious study effects compared to lung study. (B) Comparison of the ANOSIM Statistic (R) based on GloScope divergences to quantify the separation between samples in different studies for both the liver and lung studies; larger values of R indicate more separation between groups. Individual points show the ANOSIM statistic, with bootstrap confidence intervals indicated by whiskers.

distance. We used the function *anosim* in the R package `vegan` to perform the test (Clarke, 1993). The test statistic is calculated as:

$$R = \frac{r_B - r_W}{N/2(N/2-1)/4} \tag{3.1}$$

where r_B is the mean of rank similarities of pairs of samples from different groups, r_W is the mean of rank similarity of pairs within the same groups, and N is the total number of samples. The test statistics ranges from -1 to 1. Strong positive test statistics means greater between group distances than the within groups; strong negative test statistics means the opposite and may represent wrong group assignments; and test statistics near zero indicate no differences. Finally, p-values are calculated based on a null permutation distribution: the distribution of R recalculated after randomly shuffling the samples' group assignment. The p values are calculated as the proportion of times that the permuted-derived statistics are larger than the original test statistic.

PERMANOVA

A similar metric as ANOSIM test statistics is PERMANOVA test effect size ω^2, which is caculated based on the actual distance values (Kelly et al., 2015). PERMANOVA is a non-parametric method and tests whether the centroid or the spread of samples among the batches are the same. It extends traditional ANOVA by utilizing a distance matrix. PERMANOVA computes the within-group sum of squares SS_W, the average of the squared distances within each group, divided by the number of subjects in each group, and total sum of squares SS_T, from which the between-group sum of squares SS_A is derived as the difference between the total sum of squares SS_T and SS_W. The test statistic, referred to as the pseudo F-ratio, is similar to Fisher's F-ratio. It is calculated as the ratio of the between-group sum of squares to the within-group sum of squares. While the test statistics evaluates the significance of these group differences, we relied on effect size calculation, which measures the strength of the relationship we observed. For PERMANOVA test, the effect size is quantified by

$$R^2 = 1 - \frac{SS_W}{SS_W + SS_A} = \frac{SS_A}{SS_T}, \tag{3.2}$$

though it can be biased due to that it is solely based on the sample sums of squares and does not adjust to accurately estimate the effect size for the general population. Instead, Omega-squared ω^2 offers a less biased measure by incorporating mean-squared error, enhancing accuracy in estimating the effect size for ANOVA-type analyses, which is defined as

$$\omega^2 = \frac{SS_A - (a-1)\frac{SS_W}{N-a}}{SS_T + \frac{SS_W}{N-a}}, \tag{3.3}$$

where a is the number of group or batch.

Silhoutte

The silhouette width is a statistical measure used to assess the quality of clustering in data analysis. It provides a concise yet informative evaluation by quantifying how well each data point fits into its assigned cluster compared to neighboring clusters. Specifically, for each point, the silhouette width is calculated as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.4}$$

where $a(i)$ represents the average distance between the data point i and all other points within the same cluster, known as the intra-cluster distance, and $b(i)$ denotes the minimum average distance from the data point i to all points in any other cluster, referred to as the inter-cluster distance.

A higher silhouette width indicates that the data points are appropriately clustered, with distinct and well-separated clusters. To assess the overall clustering quality of the entire dataset, we obtain the average silhouette width by calculate the mean of the silhouette widths across all data points in the dataset, as shown below

$$\text{Average Silhouette Width} = \frac{1}{n}\sum_{i=1}^{n} s(i). \tag{3.5}$$

Bootstrap

After obtaining the above values from the calculated distance matrix D, we calculated bootstrap confidence intervals for each of the metrics. To do so, we defined the unique combinations of batch and biological condition. For each unique combination, we repeatedly sampled with replacement from samples in that combination; the union of the sampled samples from each combination resulted in a single full bootstrap sample. After obtaining the bootstrap sample for each run, we obtained the bootstrap distance matrix D_{boot} from the original distance matrix D by subsetting to the bootstrap sample ids. Finally, we calculate the two metrics based on D_{boot}. We repeated this for $B = 100$ bootstrap samples. For each of the metrics, we calculated percentile bootstrap confidence intervals by taking the 2.5% and 97.5% quartiles from the empirical distribution of the bootstrap distribution of the metrics.

Batch Correction Methods Summary

Various batch correction methods have been developed to mitigate these unwanted variations. These methods include simple statistical techniques, such as mean or median centering, as well as more advanced approaches like Harmony and Seurat's integration technique, which leverage machine learning and data integration principles to correct batch effects. Each method has its strengths and limitations, and the choice of the most appropriate technique

often depends on the specific characteristics of the dataset and the biological questions being investigated.

In the following sections, we mainly use GloScope and the numeric metrics introduced above on comparing the following popular batch correction methods: Harmony, scVI, fastMNN, and Liger.

Harmony

Harmony is a sophisticated batch correction method designed to handle complex datasets commonly encountered in biological research. Harmony utilizes an iterative algorithm to integrate multiple datasets, effectively mitigating batch effects while preserving the intrinsic biological variation. It operates on the uncorrected latent embedding (e.g PCA), where it harmonizes the data by iteratively adjusting the embedding to ensure that similar cell types from different batches are aligned in the same space. Harmony is particularly adept at handling high-dimensional single-cell RNA sequencing data, making it a powerful tool for researchers aiming to combine datasets from different experiments or sequencing runs without losing critical biological information (Korsunsky et al., 2019).

scVI

scVI (single-cell Variational Inference) is another cutting-edge batch correction method designed specifically for single-cell RNA sequencing data. Utilizing deep learning techniques,scVI employs a variational autoencoder (VAE) framework that captures the underlying biological variation while accounting for technical noise and batch effects. The model uses a negative binomial distribution to model gene expression counts, which helps in dealing with over-dispersion common in single-cell data. By using the VAE techniques, where the encoder maps the observed data to a latent space and the decoder reconstructs the data from this latent space, scVI learns a low-dimensional representation of the gene expression profiles. (Lopez et al., 2018).

fastMNN

fastMNN employs a scalable mutual nearest neighbors (MNN) algorithm that efficiently identifies and corrects batch-specific variations (Zhang et al., 2019). The method first calculates mutual nearest neighbors to identify pairs of cells that are closest to each other across batches. Using those pairs, fastMNN computes a correction vector for each cell, which is the average shift needed to align mutual nearest neighbors. FastMNN operates in a hierarchical manner, iteratively merging batches at a time and correcting them based on the identified MNNs, ensuring that the global structure of the data is preserved.

LIGER

LIGER corrects batch effects in scRNA-Seq data through a mathematical framework involving matrix factorization. It decomposes the gene expression matrix X_i for each batch b into shared factors W capturing common biological signals, batch-specific loadings H_b, and dataset-specific factors V_b representing batch-specific variations. The method minimizes an objective function that balances reconstruction error and sparsity constraints, iteratively optimizing W, H_b, and V_b. By aligning shared components across batches, LIGER harmonizes the data, allowing for accurate cross-batch comparisons and revealing underlying biological insights (Welch et al., 2019).

Tran et al. (2020) has conducted sophisticated comparison of different batch correction methods. However, they mainly focused on cell level comparison and, as discussed in Chapter 4, many potential artifacts stem from variables that differ per sample or patient. Therefore, we rely on GloScope, which addresses this problem by incorporating the sample-level batch comparisons, allowing for a more holistic evaluation and detection of artifacts that impact the data on a larger scale.

Evaluating Batch Unit Choice and Biological Signal

Batch correction methods aim to remove unwanted variation arising from technical sources (e.g processing batchs or sequencing cites) for each cell. Different batch correction methods work by identifying and adjusting for batch effects based on the chosen batch unit per cell (See Section 3.2). When performing batch correction, it's essential to consider the choices between using sample ID or batch ID as the batch unit, as each option represents different underlying factors that can influence the correction results. For instance, cells' sample IDs often contain the information of both the batch group where the sample belongs to and the samples' phenotypes, such as disease vs healthy. Adjusting based on sample IDs would lead to a risk of overcorrecting, which can inadvertently reduce or obscure the biological signal targeted for downstream analyses. Therefore, finding the right balance in batch correction task is challenging and requires careful consideration and assessment of the specific context and the interest of the study.

In the COVID PBMC dataset, the original analysis performed by the author employed Harmony on the initial PCA embedding, utilizing sample ID to correct for batch effects (Stephenson et al., 2021). We redo the analysis by using sequencing site as the correction unit instead. As shown in Figure 3.12, though R statistics' bootstrap confidence interval overlap between the sample ID and Batch ID, PERMANOVA ω^2 value showed that correction based on Batch ID effectively reduces batch effects and also enhances the detection of biological signals between COVID-19 patients and healthy individuals (We also observe that the different values between GMM vs KNN and would discuss such comparison further in Chapter 4, Section 4.2).

Similarly, in the Lung fibrosis study, our findings indicate that correcting for `Study`, rather than sample ID (Figure 3.13), yields more substantial removal of batch effects. This

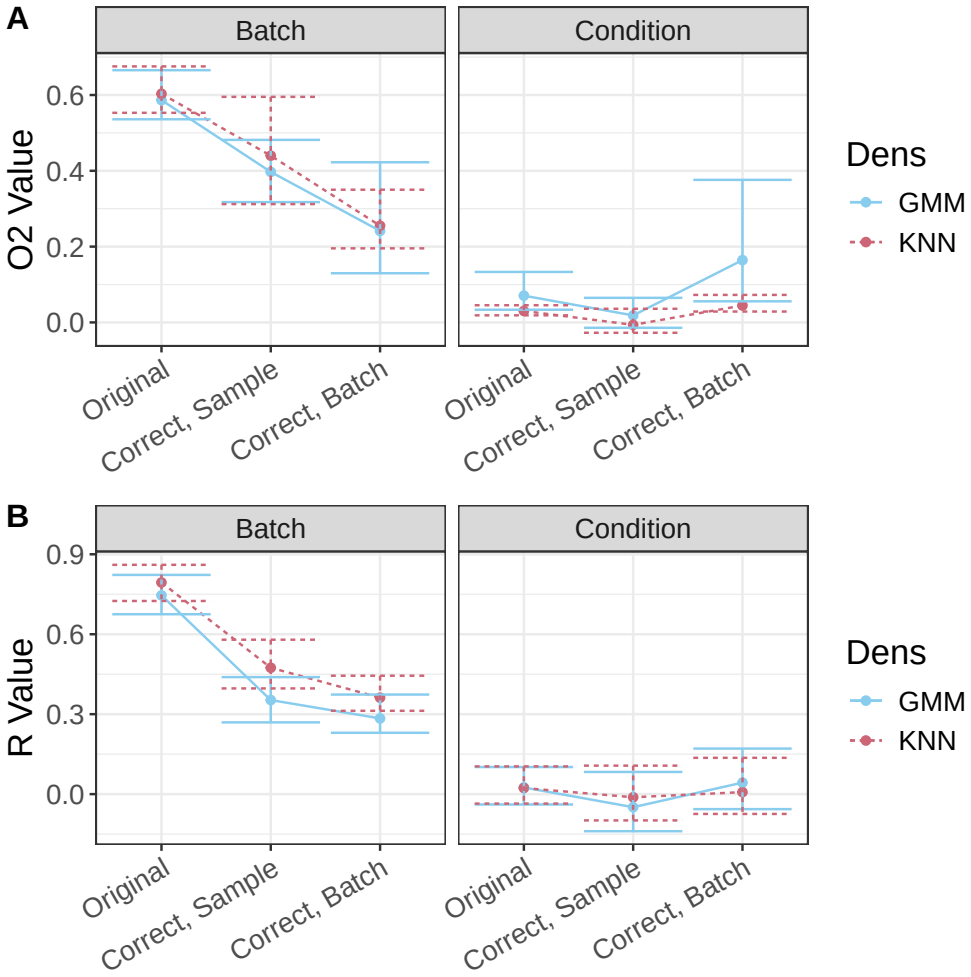


Figure 3.12: **Numeric evaluation of Harmony batch correction applied to COVID PBMC data from Stephenson et al. (2021).** (A) ω^2 values for evaluating batch (Left) and biological signal (Right) among different batch units. (B) R values for evaluating batch (Left) and biological signal (Right) among different batch units.

approach allows us to maintain or even increase the power of detecting biological signals.

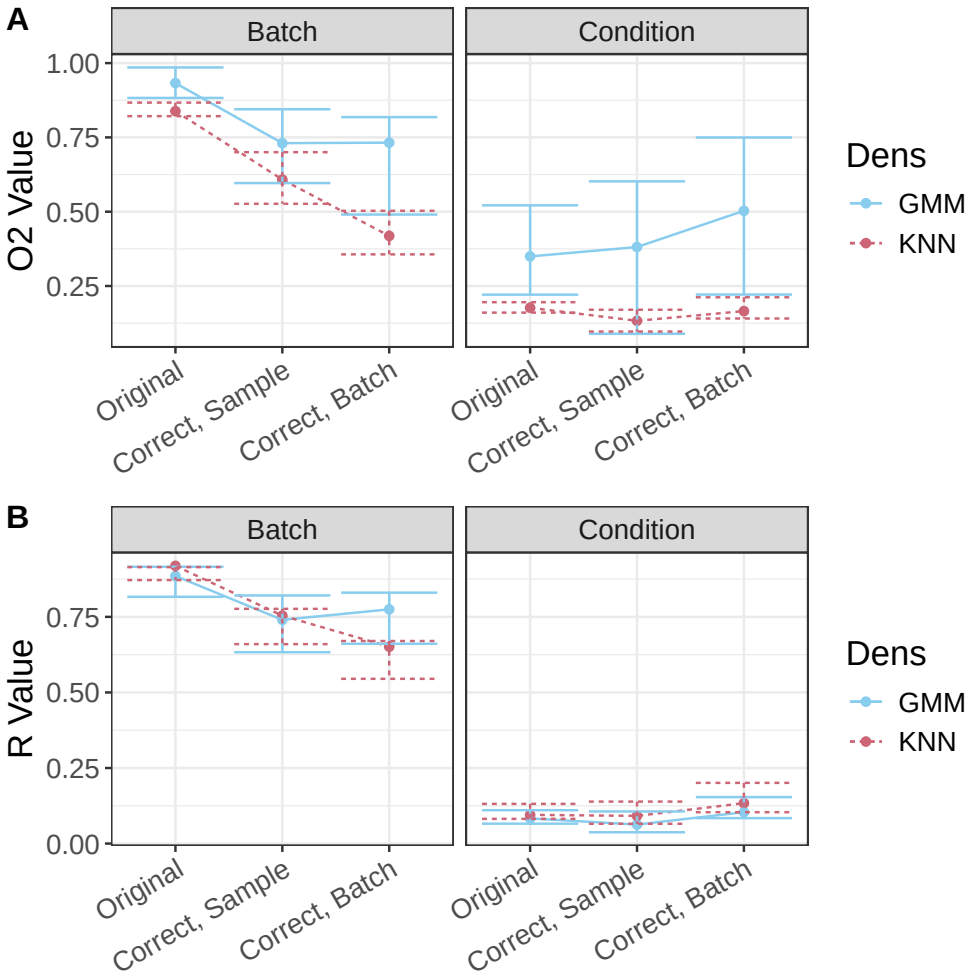


Figure 3.13: **Numeric evaluation of Harmony batch correction applied to Lung fibrosis data from Fabre et al. (2023).** (A) ω^2 values for evaluating batch (Left) and biological signal (Right) among different batch units. (B) R values for evaluating batch (Left) and biological signal (Right) among different batch units.

However, the same conclusion or observation may not be held for all cases. Utilizing batch IDs can present challenges, particularly when there is an uneven distribution of sample phenotypes across batches. This uneven distribution can lead to biased corrections, as the batch correction algorithms may disproportionately adjust certain batches, skewing the results. In such scenarios, the batch correction may fail to adequately address the variability introduced by the batch effects, thereby compromising the accuracy of the data analysis.

In the study of Lupus PBMC, we observed that samples labeled as `Flare` and `Treated` were exclusively present in Processing Batch 3, while Processing Batch 1 contained only

	Flare	Managed	Normal	Treated
1.0	0	0	47	0
2.0	0	120	22	0
3.0	19	4	18	10
4.0	0	52	44	0

Table 3.1: Table of sample distribution among processing batches and conditions in Perez et al. (2022)

`Normal` samples, as detailed in Table 3.1. Given this distribution, we applied the GloScope along with two numeric metrics and we saw in Figure 3.14 that although using Processing Cohort as the batch unit resulted in a slightly more effective reduction of batch effects, it simultaneously diminished the biological signal. Conversely, correcting based on sample ID enhanced the ability to accurately identify sample phenotypes.

Thus, when correcting for batch effects, a careful assessment and quantification of batch effects is required. Researchers should balance the need to remove unwanted batch variation while preserving the true biological variation. Additionally, careful consideration of the distribution of sample phenotypes across batches is essential to avoid introducing new biases or exacerbating existing ones. The above examples demonstrate that GloScope provides a valuable approach for offering researchers insights into better selecting the appropriate correction strategies.

Comparative Analysis of Batch Correction Methods

In Figure 3.15, we noticed that most of the correction methods are consistent on choosing batch unit: using batch id improves the batch effect removal, as well as perserve or improve the biological signals. Among the methods, applying fastMNN on PCA does not yield as satisfying effects as other methods in removing batch effects. While Liger has comparable performance on removing batch effects as Harmony or scVI, it failed in improving the distinguish samples based on biological conditions. Overall, for the particular dataset of Stephenson et al. (2021), we would recommend applying Harmony or scVI and using batch id.

While for the dataset from Perez et al. (2022), one of the methods, fastMNN failed due to excessive computational cost and memory constraints. Hence, we focus on comparing Harmony, Liger, and scVI. Here we noticed that unlike Harmony and scVI, Liger on batch id has better results in removing batch and preserving wanted differences, as shown in Figure 3.16. However, Harmony of scVI on sample ID has comparable perforamnce as Liger on batch ID. Hence, for this particular dataset, we are left with an opening question for what are the best technique and batch ID choice to use. Researchers could choose based on their resources and need, and use our method as the evaluation tool.

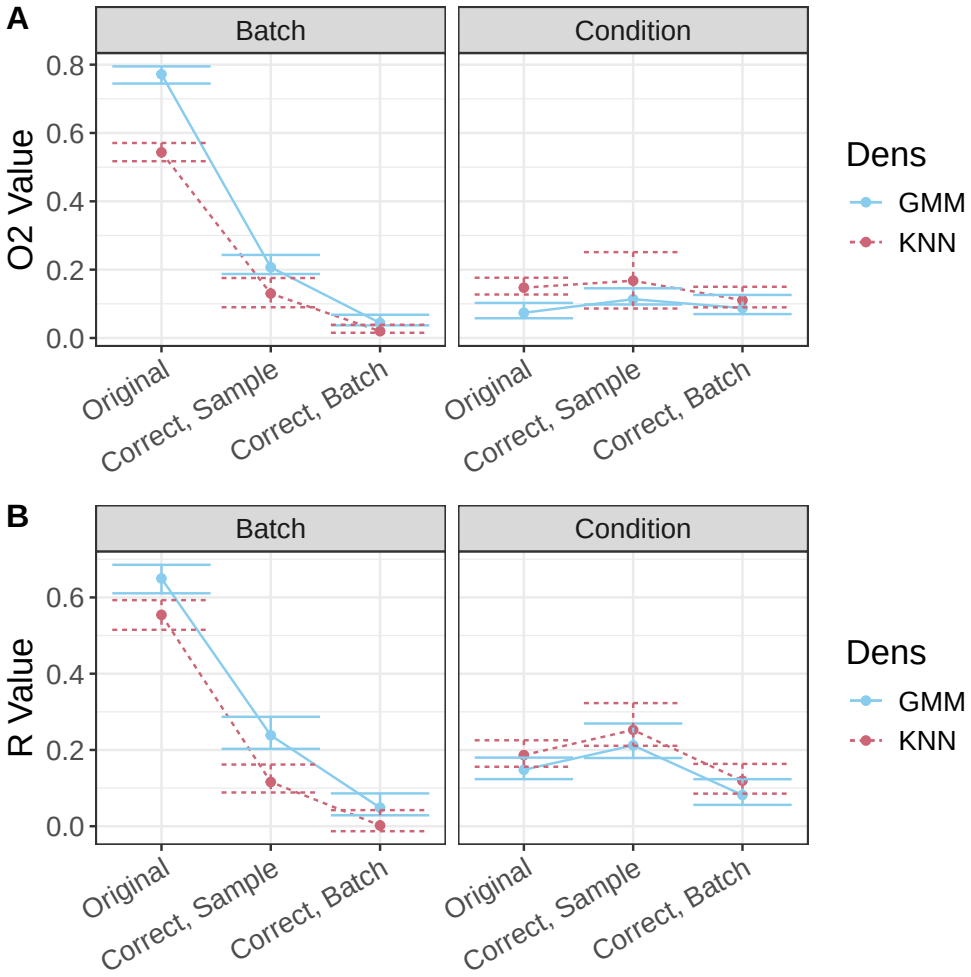


Figure 3.14: **Numeric evaluation of Harmony batch correction applied to Lupus PBMC data from Perez et al. (2022).** (A) ω^2 values for evaluating batch (Left) and biological signal (Right) among different batch units. (B) R values for evaluating batch (Left) and biological signal (Right) among different batch units.

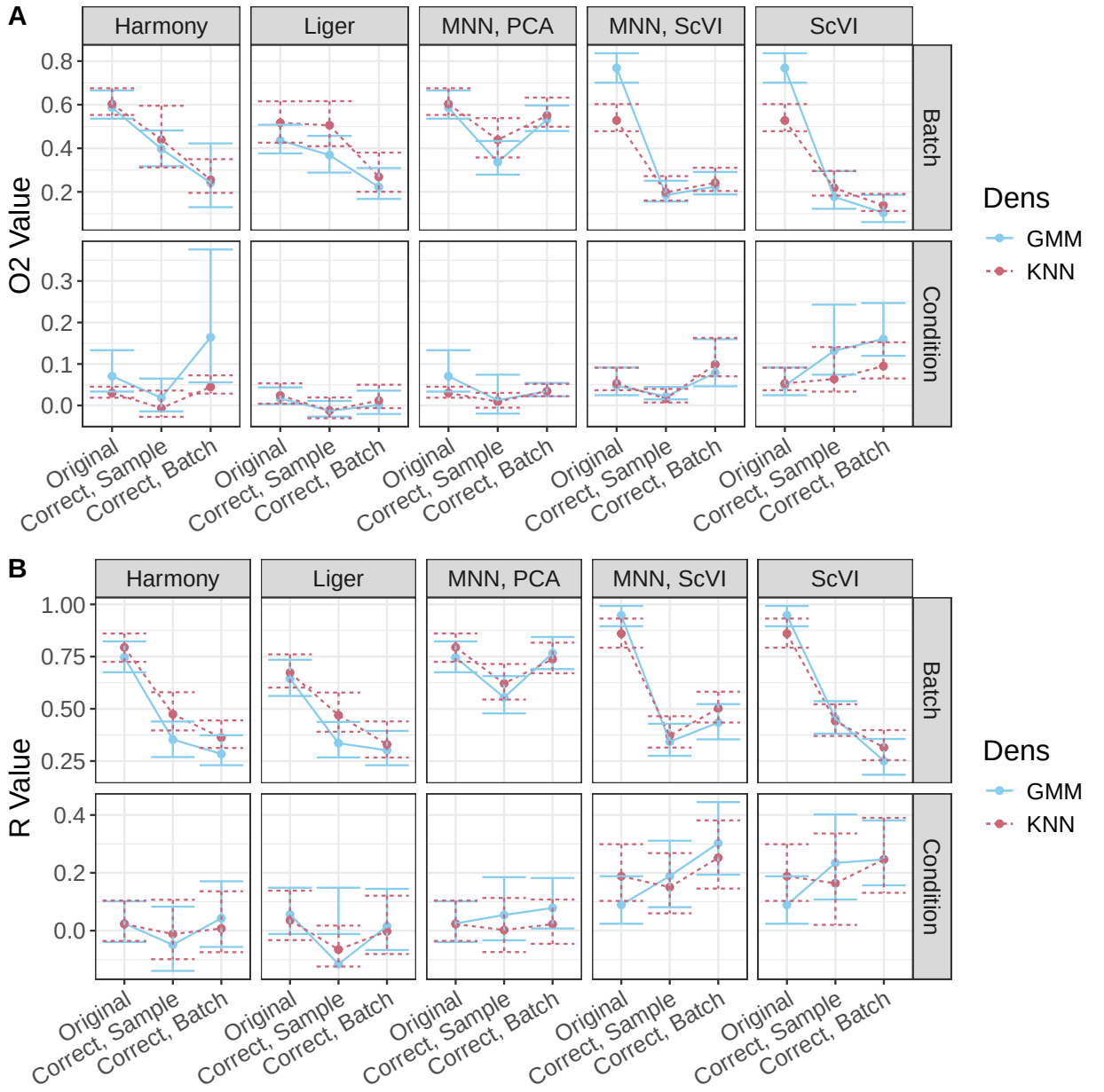


Figure 3.15: **Numeric evaluation of different batch correction techniques applied to COVID PBMC data from Stephenson et al. (2021).** (A) ω^2 values for evaluating batch (Upper) and biological signal (Lower) among different batch units and different batch correction methods. (B) R values for evaluating batch (Upper) and biological signal (Lower) among different batch units and different batch correction methods.

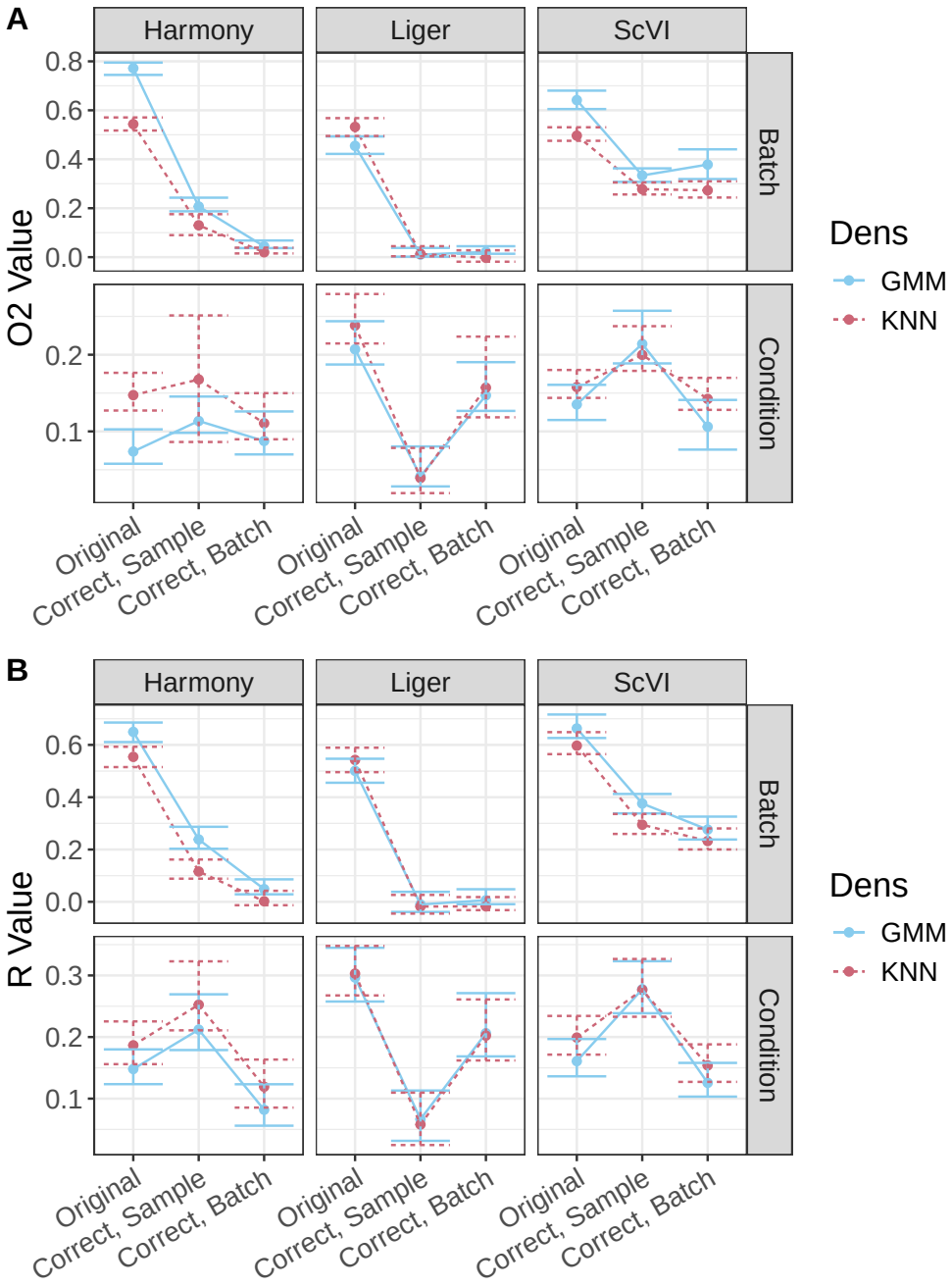


Figure 3.16: **Numeric evaluation of different batch correction techniques applied to Lupus PBMC data from Perez et al. (2022).** (A) ω^2 values for evaluating batch (Upper) and biological signal (Lower) among different batch units and different batch correction methods. (B) R values for evaluating batch (Upper) and biological signal (Lower) among different batch units and different batch correction methods.

3.3 Summary

Batch effects are common concerns with large sets of data, especially in human subject data where the samples are likely to be collected and possibly sequenced at different sites. In this chapter, we demonstrated the ability of the GloScope representation to detect important artifacts in the data. These examples immediately showed the power of our GloScope representation for exploratory data analysis. We also showcase GloScope's strength in performing quantitative evaluation of batch effects and batch correction methods at the sample level. By incorporating GloScope with different numeric metrics, we provide a quantitative framework for evaluating and comparing different batch unit choices and correction methods in the population scale. This approach enables a detailed analysis of how various batch correction techniques and choices might impact the quality of scRNA-seq data, with a focus on sample level integrity, ensuring more accurate downstream analyses and better reproducibility in single-cell studies.

Chapter 4

Evaluation of GloScope with Competing Methods and Simulation

In this chapter, we undertake a comprehensive qualitative evaluation of GloScope. The evaluation is divided into two main sections to provide a thorough understanding of GloScope's effectiveness and reliability. First, we compare GloScope with several established methods in the field using EDA and nuemric metric introduced in Chapter 3. This comparative analysis is crucial for demenstrating GloScope's strength and advantages, and positioning GloScope within the existing landscape of scRNA-Seq data analysis tools, as discussed in Chapter 2. Second, we employ scRNA-Seq simulation to rigorously test GloScope's performance to accurately identify and characterize samples' phenotype heterogeneity in different situations.

4.1 Comparison with Competing Methods

In addition to GloScope, several other methods also tackle the analysis of scRNA-seq data at the sample level. These approaches often differ in their underlying assumptions and computational strategies, providing various ways to handle the complexities inherent in scRNA-seq datasets.

Comparison with other Quality-control tools

Existing tools for EDA and evaluation of potential quality concerns are generally focused on analysis at the level of the individual cell. Numerous metrics exist for evaluating the quality of individual cells and filtering poor cells, such as the the number of detected genes, the number of sequenced reads, or the percentage of mitochondrial DNA (Osorio and Cai, 2020; Ilicic et al., 2016). Yet, many sources of possible artifacts are often due to variables that vary per sample or patient, such as the hospital of collection, the sequencing site, or the laboratory running the experiment. These effects have large-scale effects beyond individual

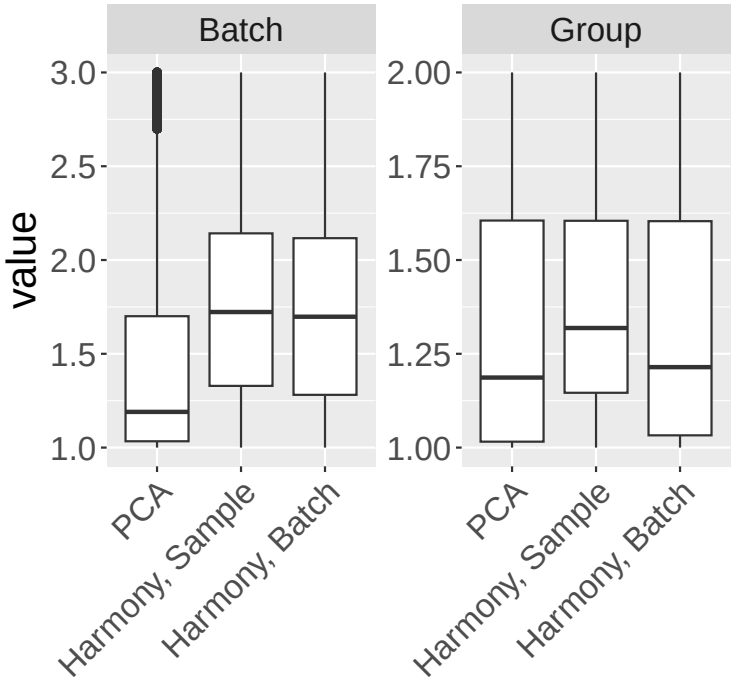


Figure 4.1: **Boxplot of iLISI value for individual cells for data in Stephenson et al. (2021).** The left panel showed the changes of iLISI value of each cell for batch quantification: the closer the values to 1, the more clear batch separation, indicating significatn batch effects; the closer the values to 3 (i.e. the number of batches), the better mixture among cells, indicating better batch correction. We saw that after applying Harmony on sample id and batch id, the iLISI values increased, suggesting the effectiveness of Harmony. The right panel showed the changes of iLISI value of each cells for separation of biological signal (COVID vs Healthy).

cells and are best detected by comparisons of the cells as a group. However, there are limited options for detecting artifacts that vary by sample or individually poor samples.

In particular, analyses at the individual cell-level are less flexible for detecting these sample-level differences. There are metrics at the individual cell-level, such as iLISI (Korsunsky et al., 2019) that can assess the presence of a batch effect *for known batch variables*. These are similar to our use of ANOSIM or Silhouette width to quantify the separation between samples in batches, only these methods are applied to the individual cells. Such methods can highlight similar effects, such as showing an improvement in Harmony corrected data for the Stephenson et al. (2021) data (Figure 4.1), but they are ineffective for discovering effects *de novo*, nor do they provide the ability to compare multiple effects, such as our visualizations of both batch and biological effects in Section 3.1.

A common exploratory visualization strategy for scRNA-Seq data consists of applying tools such as UMAP or tSNE to create a two-dimensional visualization of the individual cells. Individual cells can be color-coded by potential variables or plotted separately per sample for exploration of possible *known* artifacts, as we provided for the Stephenson et al. (2021) data in Figure 3.2. UMAP visualizations can be helpful in retrospect for understanding the nature of the problem, but are not particularly effective in discovering such effects *de novo* given the difficulty in visualizing sample effects for large numbers of cells. The example of the Perez et al. (2022) data is illustrative, where our GloScope representation allowed us to immediately determine unexplained groupings of samples within Batch 4; we were able to follow this discovery with further investigation at the individual cell-level using UMAPs to discover that there were shifts in gene expression and cell density among these subgroups GloScope identified within Batch 4. These differences are not detectable in plotting all cells, and only after identifying the subgroups of patients can a UMAP help in further investigation. Furthermore, differences due to shifts in cell distributions can be tricky to see in UMAP visualizations of individual cells, due to the overplotting of cells. Even after identifying the different subgroups in Batch 4 with GloScope (Figure 3.3), the differences seen clearly in the GloScope representation were subtle to detect using standard UMAP visualization (Figure 3.5, 3.6, and 3.4). This exploratory analysis of the (Perez et al., 2022) data shows the complementary nature of GloScope with other visualization tools. Similarly, outlying individual patients, as we detected in the lung samples of Fabre et al. (2023) (Section 3.1), would require plotting and comparing of UMAPs of each individual sample which is simply not feasible for large cohorts.

There are some limited alternatives to GloScope available for the comparison at the sample-level, and they take different strategies for summarizing the data from a single patient which we next consider: cell-composition and pseudobulk.

Comparison with cell-composition analysis

Grouping patients based on their celltype proportion has been a popular methods for comparing and grouping samples. Reducing each sample to their cell-type composition has been proposed for globally comparing single-cell samples (Orlova et al., 2016; Wagner et al., 2019; Li et al., 2020b; Chen et al., 2020; Joodaki et al., 2023), and there has been some limited work in analysis of data from flow-cytometry using cell-type compositions to globally compare samples which has similarities to using GloScope on the proportions (Orlova et al., 2016; Johnsson et al., 2016; Bruggner et al., 2014; Orlova et al., 2018).

Specifically, if each cell can be classified into one of K subtypes ($K = 1, ..., k$), then we observe for each sample the proportion of cells π_k in each cell-type k. Cells are jointly clustered, and patients summarized and compared by their relative cluster frequencies. A simple version of this strategy is to visualize the proportions per sample in a barplot. Like UMAPs of individual cells, such barplots can be useful tools for greater investigation of differences found by GloScope, but do not scale for easy comparisons of large number of

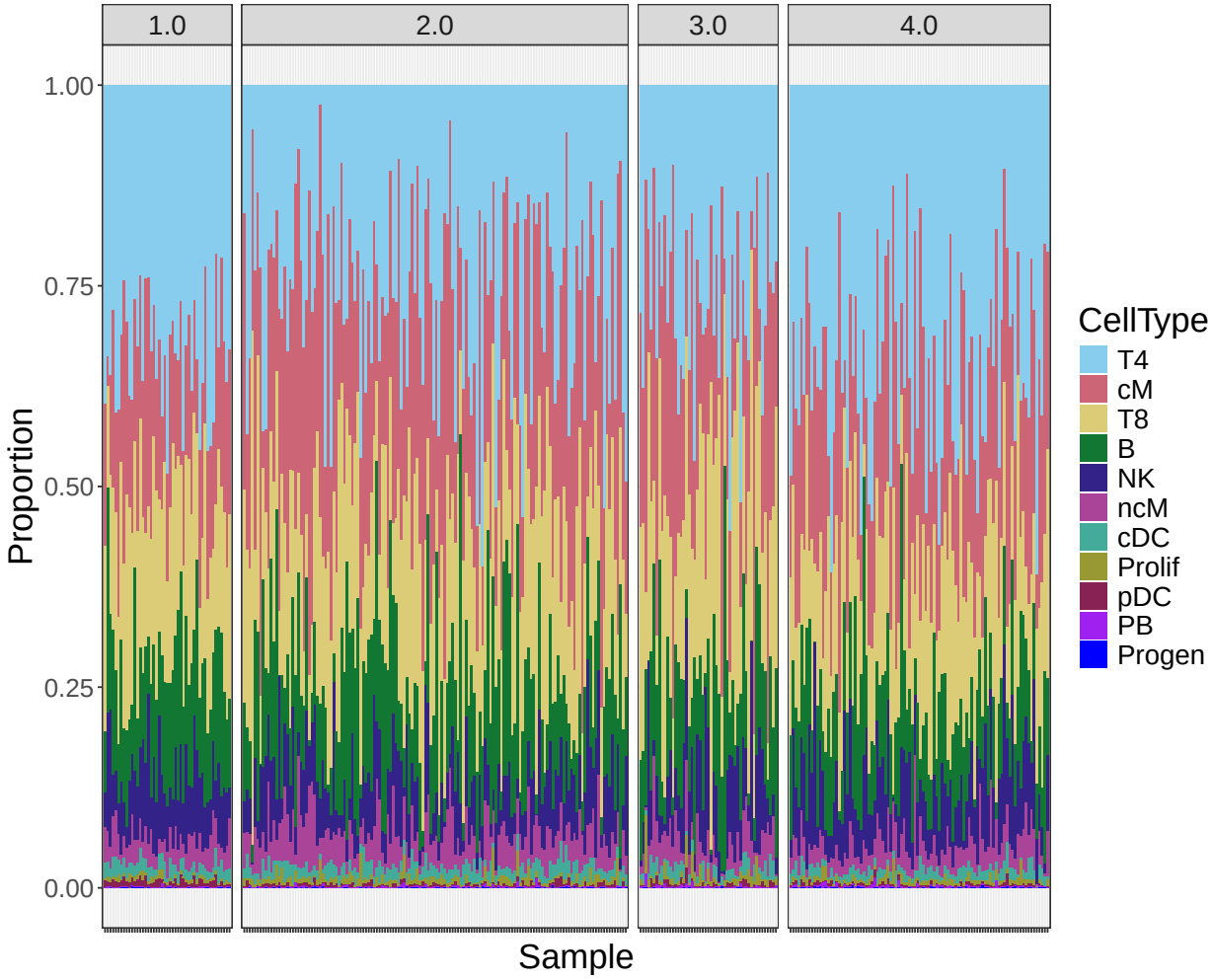


Figure 4.2: **Bar plot visualization of cell type proportion per samples in the original batches of the Lupus PBMC study (Perez et al., 2022).** For plot details and color annotation, see Fig. 4.3. Panels are separated by original batch annotated by the Perez et al. (2022), without further separation of batch 4.0 into subgroups identified by GloScope.

samples and do not aid in discovering possible differences, such as the potential subgroups of batch identified by GloScope (Figure 4.3, 4.2).

The cell-type proportions can also be analyzed more quantitatively– for example the GloScope methodology can also be used for cluster proportions, which we call GloProp, as opposed to our standard implementation which calculates an estimate of the full gene expression density. GloProp take a sample's cluster proportion vector $\pi_i = \pi_{i,1}, ..., \pi_{i,K}$ as input and calculate the symmetrised KL divergences between each sample pair as below

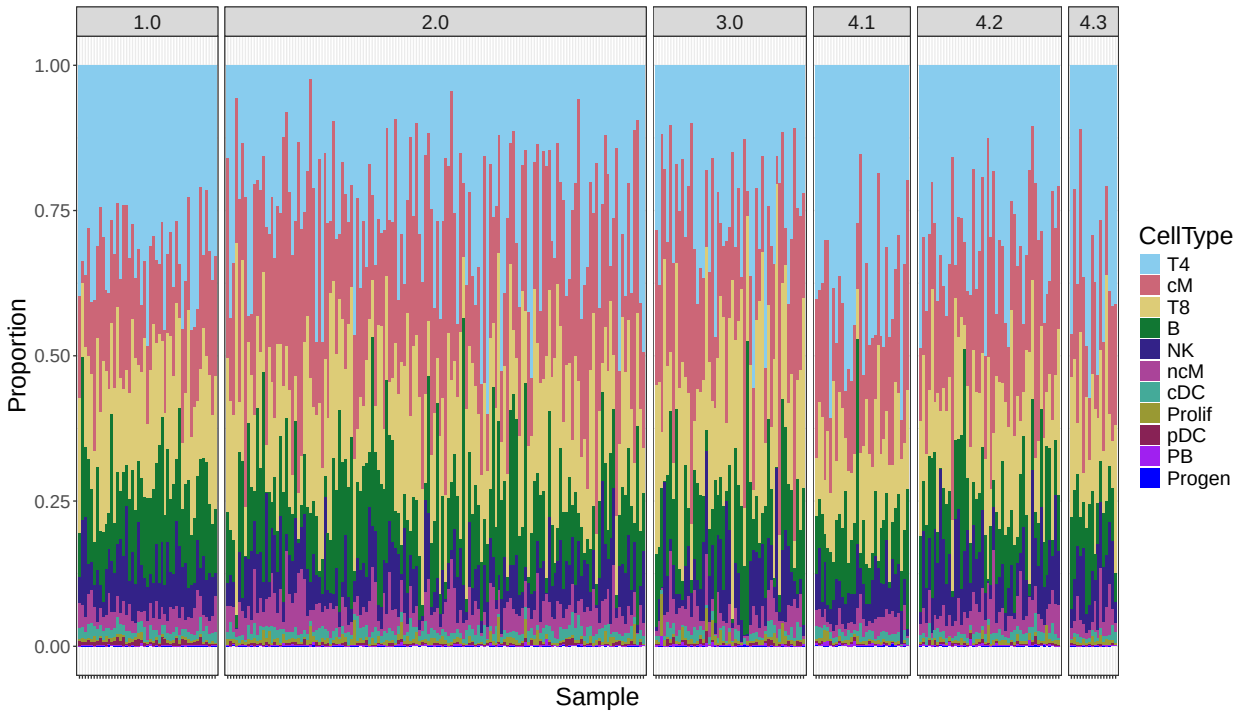


Figure 4.3: **Barplot visualization of cell-type proportion differences in subgroups identified by GloScope for Lupus PBMC study Perez et al. (2022).** Each column/bar represents a sample. The bars are broken into different color-coded segments, with a segment for each cell-type and the size of the segment proportion to the proportion of cells in the data identified with the cell-type. The annotation of individual cells into cell-types are based on the annotation provided by Perez et al. (2022) using canonical marker genes. Samples are separated in different panels based on their processing batches provided in Perez et al. (2022), with the *de novo* subgroups found by GloScope in the fourth processing batch shown separately. For the subgroups of the fourth processing batch, we see samples in batch 4.1 has relatively larger proportion of CD4 T cells than batch 4.2 and 4.3.

$$D_{SKL}(\pi_i, \pi_j) = \sum_{k=1}^{K} \left[\pi_{i,k} \log \left(\frac{\pi_{i,k}}{\pi_{j,k}} \right) + \pi_{j,k} \log \left(\frac{\pi_{j,k}}{\pi_{i,k}} \right) \right] \tag{4.1}$$

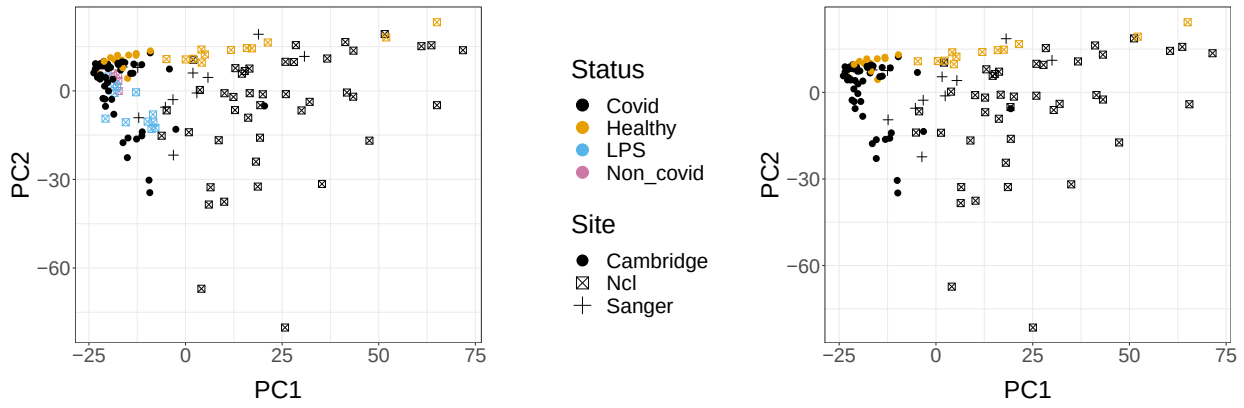
Unlike a full GloScope representation, applying GloScope on the cluster proportion vector requires classifying cells into subtypes before application of the method. Accurate identification of cells into subtypes is often a manual and time-consuming process, which makes this approach less useful for the exploratory data analysis that is often upstream of the subtype identification step. However, GloScope applied to the clusters can be used for more formal hypothesis testing of significant global differences in cell-type composition, as discussed in Section 4.2.

Concurrently, Joodaki et al. (2023) has proposed a similar metric strategy for comparing cell-type proportions named PILOT, using Wasserstein distance rather than symmetric KL divergence. These approaches require determination of cell-type proportions and can only be run after clustering the individual cells. Such clustering is typically done after EDA and correction of possible batch effects, making it irrelevant for EDA. But in principle clustering could be done earlier in the pipeline for the sole purpose of using PILOT (or GloProp) for EDA (the discovered clusters would not be biologically meaningful until the data has been appropriately pre-processed). We do this clustering on the uncorrected data and compare PILOT and GloProp to GloScope. We see that PILOT performs much worse than GloScope or GloProp in detecting separations between the batches in all of the datasets (Figure 4.6, 4.7).

Comparison with pseudo-bulk analysis

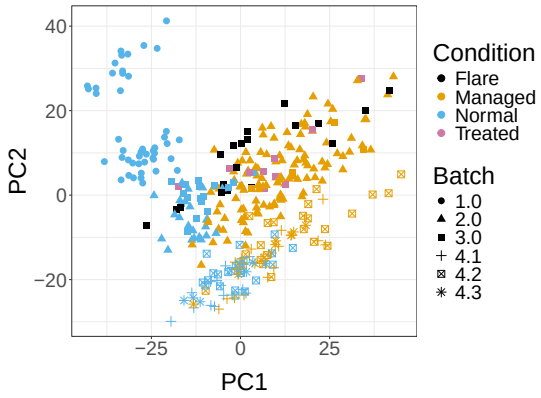
Another potential strategy for sample-level exploratory analysis is using a pseudo-bulk created from the scRNA-Seq data. This is a strategy of aggregating over each sample's cells to obtain a single observation per sample (Crowell et al., 2020); the most common is to simply sum the counts. Then standard methods from bulk mRNA-Seq, such as PCA, can be applied at the sample level. Ramirez Flores et al. (2023) propose a strategy, MOFA, for finding lower-dimensional latent embeddings per sample based on combining pseudo-bulk measures per cell-type, to better reflect cell-type variability.

We create such a PCA visualization of the pseudo-bulk of several of the datasets mentioned above (Figure 4.4, 4.5). For the COVID-19 PMBC samples, for example, the pseudobulk analysis does not clearly separate out the LPS and non-COVID samples, nor is the strong batch effect due to sequencing site as clearly identified. Similarly, for the Lupus PBMC data, the pseudobulk representation does not identify the strong batch effects seen in our GloScope representation. This is borne out by the quantification of the average silhouette width or R statistic (Figure 4.6 and 4.7). On the other hand, these quantification statistics show MOFA to have similar performance in detecting batches as GloScope; however, on closer examination of the visualization of the results of MOFA, we see less clear separation of the effects seen by GloScope. For example, MOFA did not show clear of a separation of all the



(A) Pseudobulk analysis on samples from COVID-19 study

(B) Pseudobulk analysis on COVID and Healthy samples from COVID-19 study

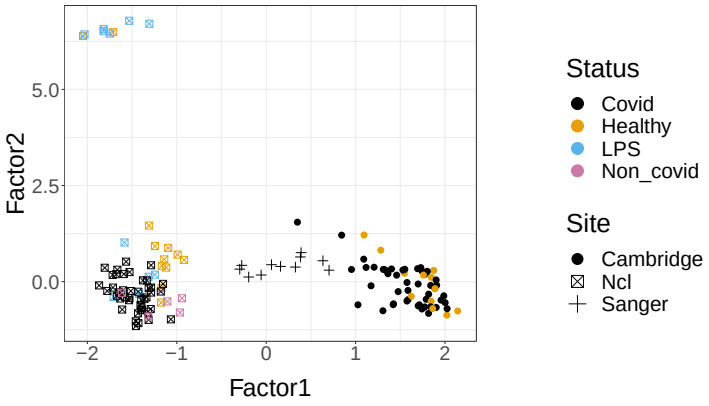


(C) Pseudobulk analysis of samples from lupus PBMC dataset

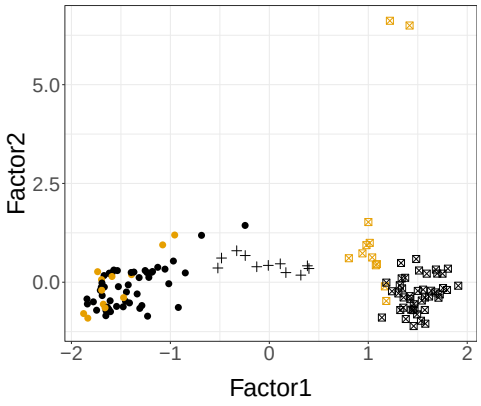
Figure 4.4: **Visualization of the first 2 PC components of the pseudobulk.** (A) samples from COVID PBMC study of Stephenson et al. (2021). (B) Covid and Healthy samples from COVID PBMC study of Stephenson et al. (2021). Removing LPS and non-COVID samples yield similar results as in (A). (C) samples from lupus PBMC study of Perez et al. (2022). Note that the PCA coordinates are equivalent to performing the MDS on the matrix of pair-wise Euclidean distance between the samples.

non-COVID and LPS samples from other samples and the separation of the groupings found *de novo* by GloScope are attenuated and difficult to find (Figure 4.5).

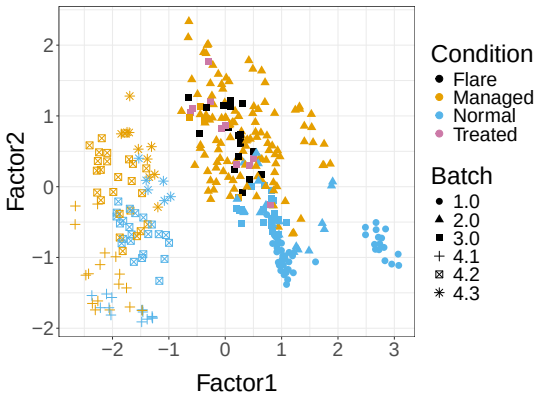
There are other limitations to either of these pseudo-bulk strategies. The pseudo-bulk strategy, including MOFA, is based on summarizing for each gene the expression level of all the cells in a sample, usually the sum of the raw counts. However, in many public



(A) MOFA on samples from COVID-19 study

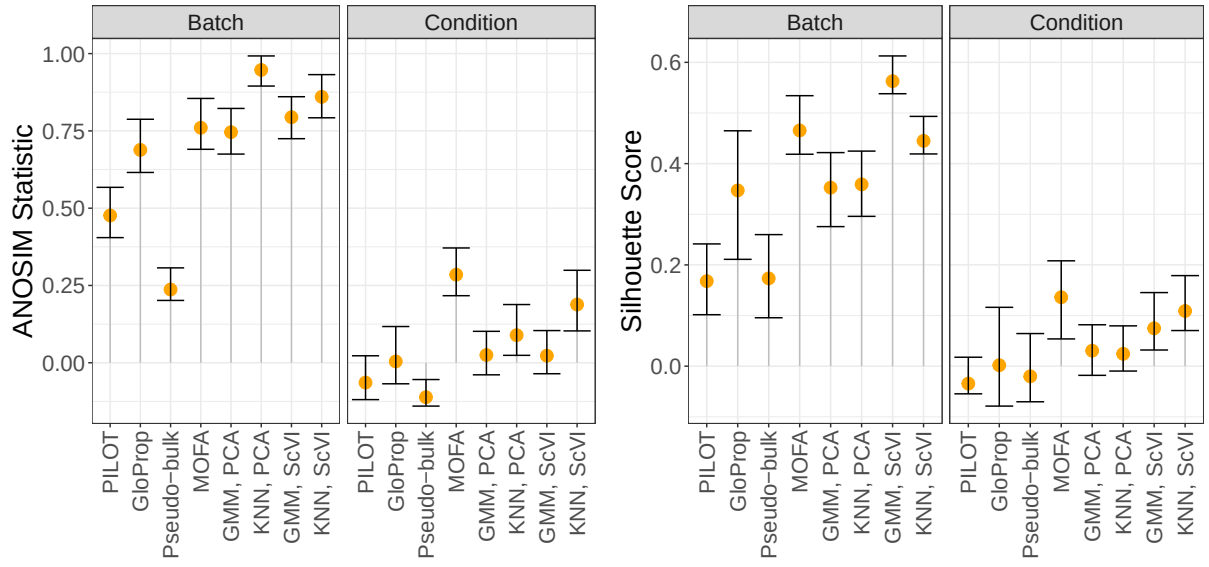


(B) MOFA analysis on COVID and Healthy samples from COVID-19 study



(C) MOFA analysis of samples from lupus PBMC dataset

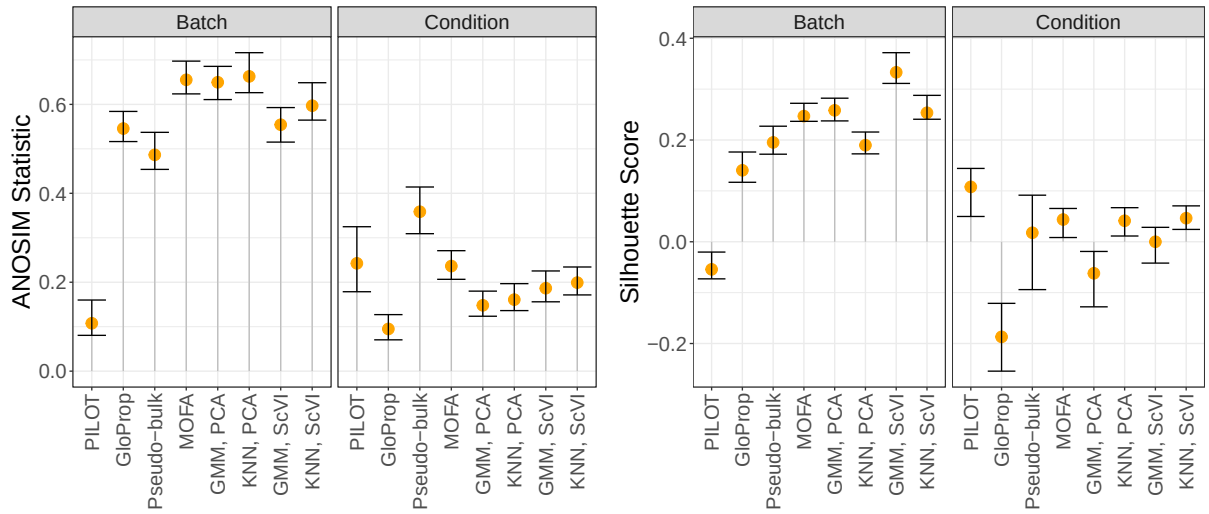
Figure 4.5: **Visualization of the first 2 factors of the MOFA results for data in Stephenson et al. (2021) and Perez et al. (2022).** (A) samples from COVID PBMC study of Stephenson et al. (2021). (B) Covid and Healthy samples from COVID PBMC study of Stephenson et al. (2021). Removing LPS and non COVID samples yield similar results as in (A). (C) samples from Lupus PBMC study of Perez et al. (2022). Each point is a sample, color-coded by their biological condition and with different shapes corresponding to their batch.



(A) ANOSIM (B) Silhouette

Figure 4.6: **Separation of different sample-level methods on COVID PBMC study Stephenson et al. (2021).** The separation of samples in different batches or biological conditions based on the (A) ANOSIM Statistic and (B) Average Silhouette Width. The orange point is the value of the statistic calcualted by the indicated method, along with bootstrap confidence intervals.

datasets provide other normalized versions of the data (e.g. residuals); similarly many batch-correction methods, like Harmony (Korsunsky et al., 2019), provide a batch-corrected latent variable representation. None of these are obvious candidates for either of these pseudo-bulk approaches. Our GloScope representation requires as input only a latent-variable representation per cell and thus is flexible to accommodate all of these types of input. This is important, for example, in evaluating the effect of batch correction methods. With GloScope, we can evaluate the data before and after batch correction with the Harmony algorithm (Chapter 3 Figure 3.3B,C,D), allowing us to confirm that the Harmony algorithm has removed much of the differences between batches. Moreover, the pseudo-bulk methods can often need normalization across samples in addition to normalization that may be done to individual cells so that they do not reflect simply the number of cells, similar to bulk RNA, which adds another layer of complexity since there are many strategies for such a normalization. GloScope summarizes the individual cells as a density, which is a measurement unaffected by the number of cells per sample.



(A) ANOSIM

(B) Silhouette

Figure 4.7: **Separation of different sample-level methods on on Lupus PBMC study Perez et al. (2022).** The separation of samples in different batches or biological conditions based on the (A) ANOSIM Statistic and (B) Average Silhouette Width. Orange point is the value of the statistic calcualted by the indicated method, along with bootstrap confidence intervals.

4.2 Quantitative Evaluation of GloScope via Simulation

scRNA-Seq Simulation

Simulating single-cell RNA sequencing (scRNA-Seq) count data is a crucial step in the development and benchmarking of bioinformatics tools designed to analyze such data. This process involves creating artificial datasets that mimic the complexities and variability inherent in real scRNA-Seq experiments. These simulations allow researchers to evaluate the performance of computational methods under controlled conditions, ensuring they can accurately identify gene expression patterns, cell types, and other biological insights from single-cell data. By generating realistic synthetic scRNA-Seq datasets, researchers can systematically assess the sensitivity, specificity, and robustness of their analytical pipelines, ultimately advancing our understanding of cellular heterogeneity and function.

We proposed to use simulation experiments to quantify GloScope's efficacy at detecting various classes of single-cell differences that might be observed due to differences in samples' phenotype. We simulate sample-level data where different aspects of the single-cell composition of a sample vary depending on their group assignment; for simplicity we consider only

two different phenotypic groups. Count matrices were generated from a pipeline modified from that presented in the R package `muscat` (Crowell et al., 2020).

We focus on two basic biological scenarios that could causes phenotypic-based dissimilarity between scRNA-Seq samples which we would want the GloScope representation to accurately reflect: differential cell-type composition and differential gene expression. By cell-type composition, we refer to the proportion of various cell-types found in a sample; for example an inflammatory disease phenotype might result in a higher proportion of immune cells in the patient than in a healthy sample. Cell-type gene expression differences (DE) refers to differences across samples in the marginal gene expression levels within cells of a certain type. For example the IL2 gene has more expression within the T-cells of inflammation tissue samples when compared to the its expression in T-cells of healthy samples. Both types of differences are biologically plausible and can co-exist. We also note that in practice the distinction between these two can blur: many genes exhibiting sufficiently strong differential expression between phenotypes will result in the creation of a novel cell-type for all practical purposes, thereby corresponding to differential cell-type composition and vice versa.

Motivation from the Muscat Package

To simulate population-level scRNA-Seq data with which we benchmark our methodology, we follow the model introduced by the `muscat` R package.

The `muscat` package is a versatile tool designed for the simulation of single-cell RNA sequencing (scRNA-Seq) data. It provides a comprehensive framework for modeling gene expression patterns at the single-cell level, allowing researchers to generate synthetic datasets that closely resemble real-world scRNA-Seq data. The package offers flexibility in simulating diverse biological scenarios, enabling users to mimic various experimental conditions and cell types accurately. `muscat` incorporates advanced statistical models to capture the complexities of gene expression variability within and between cells, ensuring the generated data are representative of biological processes. Additionally, the package provides functionalities for quality control, visualization, and benchmarking of analysis methods, facilitating rigorous evaluation and optimization of computational tools in the field of single-cell transcriptomics. Overall, `muscat` serves as a valuable resource for researchers seeking to simulate scRNA-Seq data for experimental design, algorithm development, and validation purposes.

We would note that this is a model for simulating count data for each gene, and unlike our GloScope representation does not assume any latent variable representation in generating the data. The `muscat` package assumes a simple two-group setting in which each sample i may come from one of two groups, denoted by the variable $T(i) \in \{1, 2\}$. The m_i cells from sample i come from K different cell-types with the proportion of cells from cell-type k given by $\pi_{i,k}$, where $\sum_k \pi_{i,k} = 1$. Thus the gene expression vector $x \in R^g$ of a cell c from sample i is assumed to follow a negative binomial mixture model :

$$F_{i,c}(x) = \sum_k \pi_k P_{NB}(\mu_{i,c,k}, \phi)(x) \tag{4.2}$$

where $P_{nb}(\mu_{i,c,k}, \phi)$ is a CDF on R^g representing a product distribution of independent negative binomials, i.e. each gene's expression value is independent and follows a negative binomial distribution with mean given by the j the element of the vector $\mu_{i,c,k} \in R^g$ and dispersion parameter $\phi \in R$.

The vector of gene means for cell c in sample i is parameterized in `muscat` as

$$\mu_{i,c} = \lambda_{i,c} e^{\beta_{i,k}} \cdot \theta_{k,j}, \tag{4.3}$$

where $\lambda_i \in R$ is the library size (total number of counts); $\beta_{i,k} \in R^g$ is the relative abundance of g genes in cells belonging to sample i and cell-type k; $\theta_{k,j} \in R^g$ is the fold-change for genes in cluster k if the sample belongs to group $j \in \{1, 2\}$. Notice, as mentioned above, that because of different sequencing depths per cell, each cell within sample i has a different mean $\mu_{i,c,k}$ governed by the sequencing-depth parameter $\lambda_{i,c}$, hence our notation $F_{i,c}$.

Modified Pipeline for Simulating scRNA-Seq Data

We make adjustments to the above model in the `muscat` package to more fully explore sample variability. To explore the effect of library size variation at both the cell and sample level, we introduce the decomposition $\lambda_{i,c} = \bar{\lambda} + \lambda_i + \delta_c$, where $\bar{\lambda}$ is the overall (average) library size, and λ_i and δ_c are variations from that due to sample or cell level differences, constrained so that $\lambda_{i,c} > 0$. We also adjusted the model to allow sample-specific proportions vectors $\pi_{i,k}$, with $\sum_k \pi_{i,k} = 1$. We define proportions per treatment group, $\Pi_j \in R^K$, for treatments $j = 1, 2$, such that $\sum_k \Pi_{j,k} = 1$ and randomly generate probability vectors π_i for sample i from a Dirichlet distribution according to its treatment group, $\pi_i \sim Dirichlet(\Pi_{T(i)} * \alpha)$, with sample level variation parameter α.

Selection of Parameters

The `muscat` package also provides methods for creating these many parameters based on a few input parameters by the user and estimating the other parameters based on reference data provided by the user. We followed their strategy, with the following additions.

We chose the group fold change difference per cell-type, $\theta_{k,j}$ following the schema of `muscat`, which allows for various types and size of changes between the different groups. Briefly, the simulation of $\theta_{k,j}$ is controlled by parameters 1) $\Omega \in R$, which is a user-defined average log2 fold change across all DE genes, 2) $\omega_k \in R^k$, which varies the magnitude of gene expression difference for cluster k, and 3) a proportion vector ρ which is the proportion of genes that follow six different gene expression patterns (see Crowell et al. (2020)); for simplicity, we allowed only the two most typical gene expression patterns, which are EE

(equally expressed) and DE (differentialy expressed) genes for our simulations, resulting in ρ effectively being a single scalar, the proportion of genes that are differentially expressed.

The selection of m_i, the number of cells per sample i, also followed the strategy of `muscat`, where the user provides a value $\bar{m}$, representing the average number of cells per sample across all samples, and the value of each individual m_i for each sample is assigned via a multinomial with equal probability and total number of cells across all samples equal to $n * \bar{m}$.

The parameters ϕ, and initial values of $\lambda_{i,c}$ and $\beta_{i,k}$ were obtained by estimating these parameters from the reference data, following the `muscat` package: after performing quality control, we used the filtered gene matrix and the `edgeR` package to estimate the parameters from the reference data.

Using our modified parameterization described above, $\bar{\lambda}$ was then chosen as the average of the $\lambda_{i,c}$ estimated from the reference samples. Sample-level sequencing depth variability λ_i were simulated as $\lambda_i \sim Unif(-\tau_\lambda, \tau_\lambda)$. Per-cell variability, δ_c, was simulated as $\delta_c \sim Unif(-\tau_\delta, \tau_\delta)$.

Finally, the selection of $\beta_{i,k}$ used in our simulation diverged from `muscat` package strategy. The `muscat` estimates of $\beta_{i,k}$ created overly large differences between the treatment groups and samples (Figure 4.8); furthermore their strategy recycles the same set of parameters $\beta_{i,k}$ if the simulated sample sizes are larger than provided reference sample sizes (i.e. the same value of $\beta_{i,k}$ would be given to multiple simulated samples), resulting in unintended batches of samples. Instead, we estimated $\hat{\beta}_{i,k}$ from the reference data using the `muscat` strategy, and chose a single sample i^* whose initial estimates $\hat{\beta}_{i,k}$ were representative. We then set $\hat{\beta}_k = \hat{\beta}_{i^*,k}$ and created individual $\beta_{i,k}$ with variation per sample by adding noise to $\hat{\beta}_k$, $\beta_{i,k} = \hat{\beta}_k/2 + \xi_{i,k}$, where $\xi_{i,k} \sim N(0, \sigma_\xi)$. σ_ξ controled the degree of sample-level variation.

Figure 4.9 shows the effect of changing different parameters (σ and log-fold change), visualized using UMAP on an illustrative example.

Simulation Settings

In following the above strategy of selecting parameters, we randomly chosen 5 COVID samples from the COVID-19 PBMC dataset, (Stephenson et al., 2021). After estimating ϕ and $\hat{\beta}_k$ as described above from the reference samples, the values were fixed for all simulations. The value $\bar{m}$ was chosen as 5,000, which is similar to the average cell per samples in several datasets (e.g. Stephenson et al. (2021); Melms et al. (2021); Pelka et al. (2021)). The default value for α to control the sample level cluster proportion variability was set to be 100, except where explicitly noted, which keeps the variation in cluster proportions to be relatively small among samples (see Figure 4.10D).

Once these parameters were fixed, the following user-defined parameters were set differently for different simulation settings: n (the number of samples in a single group), the vector group proportions Π_j $(j = 1, 2)$, average library size $\bar{\lambda}$, and the DE parameters Ω, ω, and ρ. With these global parameters chosen for a simulation setting, the remaining sample-specific parameters are generated anew in each simulation:

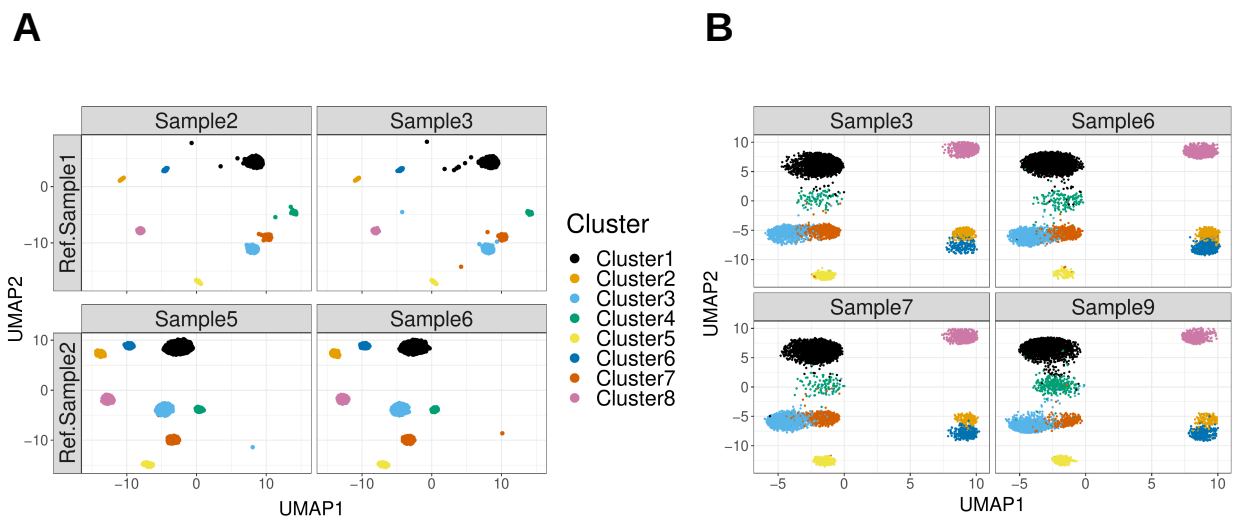


Figure 4.8: **UMAP plot demonstration of original `muscat` simulation pipeline versus modified simulation pipeline. A** shows the umap representation of simulated data from original `muscat` pipeline, where strong sample batch was observed: samples from first row was simulated from the same reference sample and sample from the second row was simulated from the same reference sample. **B** shows that after modifying β_k, some clusters were brought closer to or mixed with each other, and remove the strong sample batch due to the recycled parameters. Such modification allows the simulated data to have more reasonable and similar behavior to the real scRNA-Seq data than the data simulated using `muscat` pipeline.

1. for each cell-type k, n values of $\beta_{i,k}$ as described above based on $\hat{\beta}_k$,
2. for each cell-type k, a single vector $\theta_{k,j} \in R^G$ for the population log-fold-change between groups, based on the parameters Ω, ω, and ρ,
3. for each sample i a single value λ_i and m_i values of δ_c, one for each of the m_i cells from each sample. This results in m_i values of $\lambda_{i,c} = \bar{\lambda} + \lambda_i + \delta_c$ for each sample. (Note that some simulations set λ_i and/or δ_c to 0 for all c and i).

Combining these parameters result in the $\mu_{i,c,k}$ needed for each sample in a single simulation, and then the cell-counts for each sample i are simulated from $F_{i,c}$.

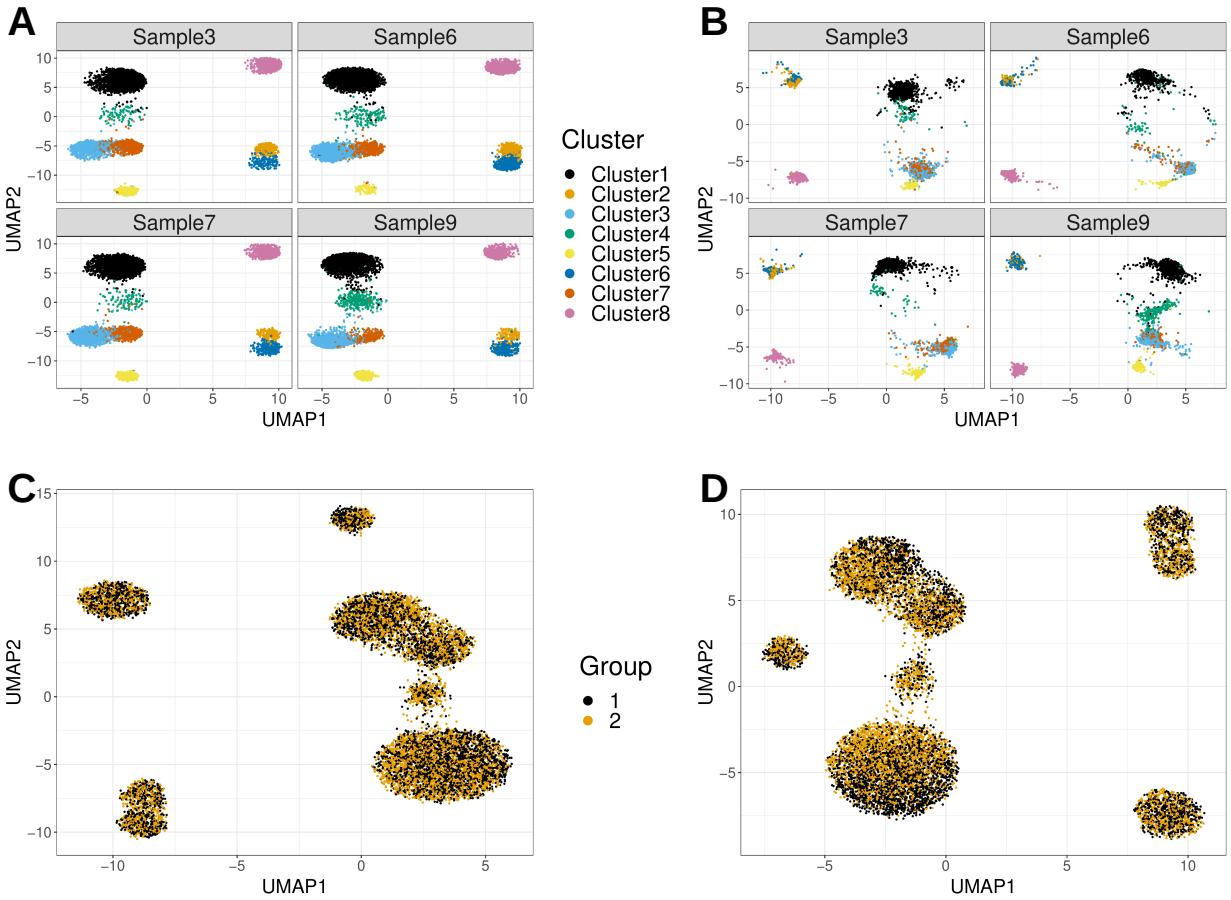


Figure 4.9: **UMAP plot demenstration of different parameter effects, including gene expression changes and sample level variation.** Each plot is drawn from 1 particular simulation realization. **B** shows that increasing σ, the gene expression level variation, leads to more varied expression among samples compared to **A**. **D** shows the increased log-fold change effect compared to **C**.

Results

In our simulations we evaluate how well these two types of differences are detected by GloScope. We create datasets demonstrating either differentially expressed genes or differential cell-type composition. We see that the average differences between samples in different different phenotype groups, as measured by our GloScope representation, appropriately increase in response to both increased differences in global cell composition (Fig. 4.10A) and increased differential gene expression (Fig. 4.10B). This indicates that our representation effectively reflects both types of changes. Similarly, when increased sample variability is added, both in global cell composition and gene expression, our GloScope representation correspondingly shows increased within-group variability (Fig. 4.10C and 4.10D).

We can also use our GloScope representation to compare different choices of the design or analysis of the experiment, based on how well the two phenotypic groups separate in the GloScope representation. To do so, we perform analysis of similarities (ANOSIM), a hypothesis test for differences between groups based on observed pairwise divergences on samples (Clarke, 1993). ANOSIM takes as input divergences between samples and tests whether divergences are significantly larger between samples in different groups compared with those found within groups based on permutation testing (see Section 3.2).

We used the results of ANOSIM to calculate the power in different simulation settings, creating a quantitative metric for evaluating the sensitivity of the GloScope representation in different scenarios. For a choice of input parameters, we repeated the simulation 100 times. For each simulation, we calculated the pairwise distances between all $2n$ samples, then used ANOSIM p-values to determine whether we would reject the null hypothesis. Finally, we calculated the power as the proportion of the 100 simulations' test statistics that have p values smaller than $\alpha = 0.05$. Evaluation of ANOSIM over many simulations gives the power of the test in different settings, resulting in a metric to compare choices in our analysis.

Using these power computations, we also see that changes in the sample variability and sample size are reflected as expected in these power calculations: increasing all of these sources of variability naturally reduces the power (Figure 4.15). These types of simulations, in conjunction with our GloScope representation, can be used to evaluate design choices at the sample-level, such as the number of samples needed to reach a desired power level. Unsurprisingly, differences in cell-composition in large clusters are more easily detected than similar differences in small clusters (Figure 4.16A), and gene expression differences concentrated in small clusters are harder to detect than those found in large clusters (Figure 4.16C).

We can also compare choices in the data analysis pipeline. For example, GloScope relies on a user-provided choice of latent variable representation of the single-cell data. We compare the choice of PCA versus scVI in a wide range of our simulation settings. The most striking difference is in detection of cell-composition differences, where scVI has much less power in detecting differences between the two phenotypic groups than PCA (Figure 4.17). The latent variable representations given by scVI demonstrates much greater variability between

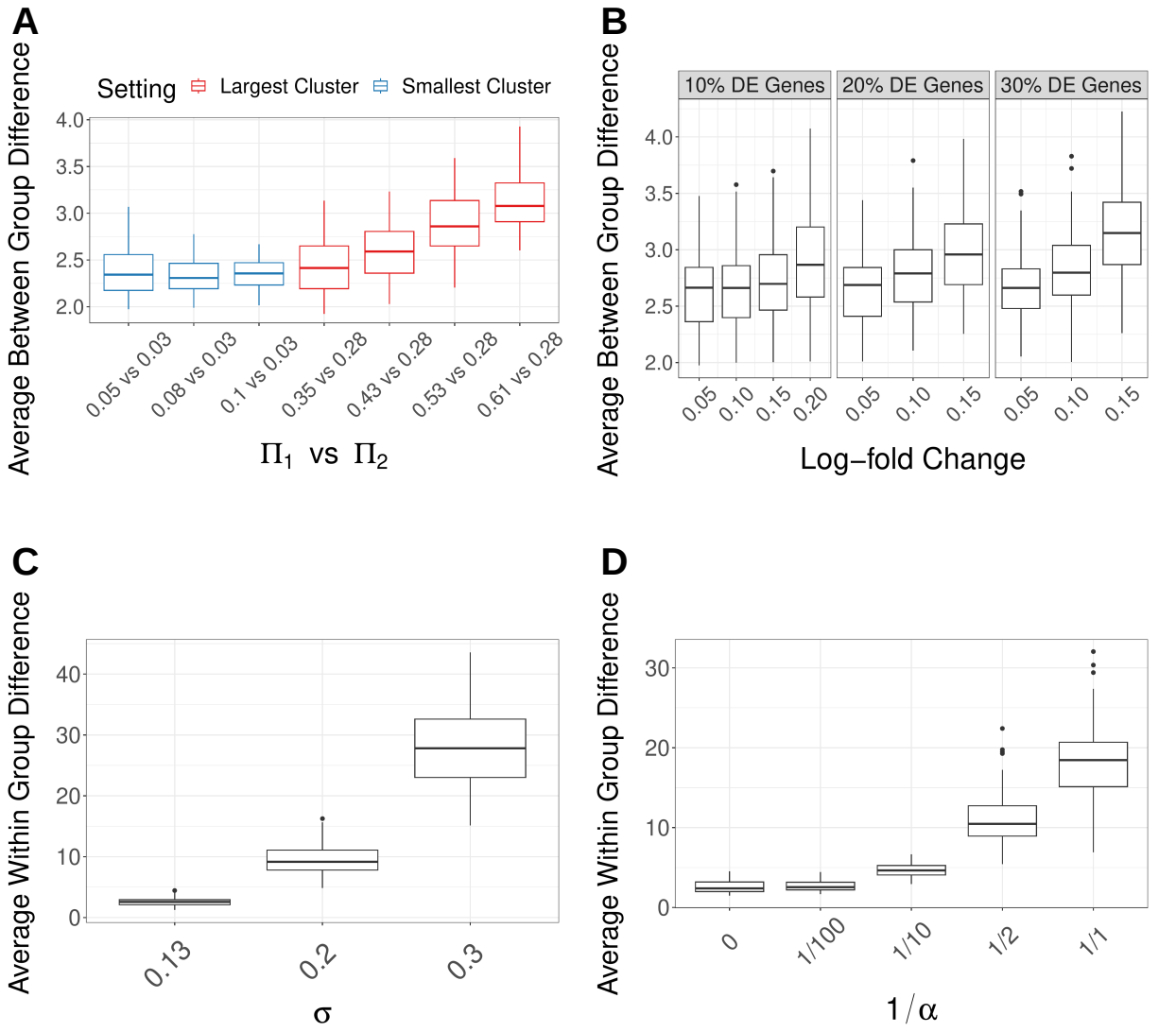
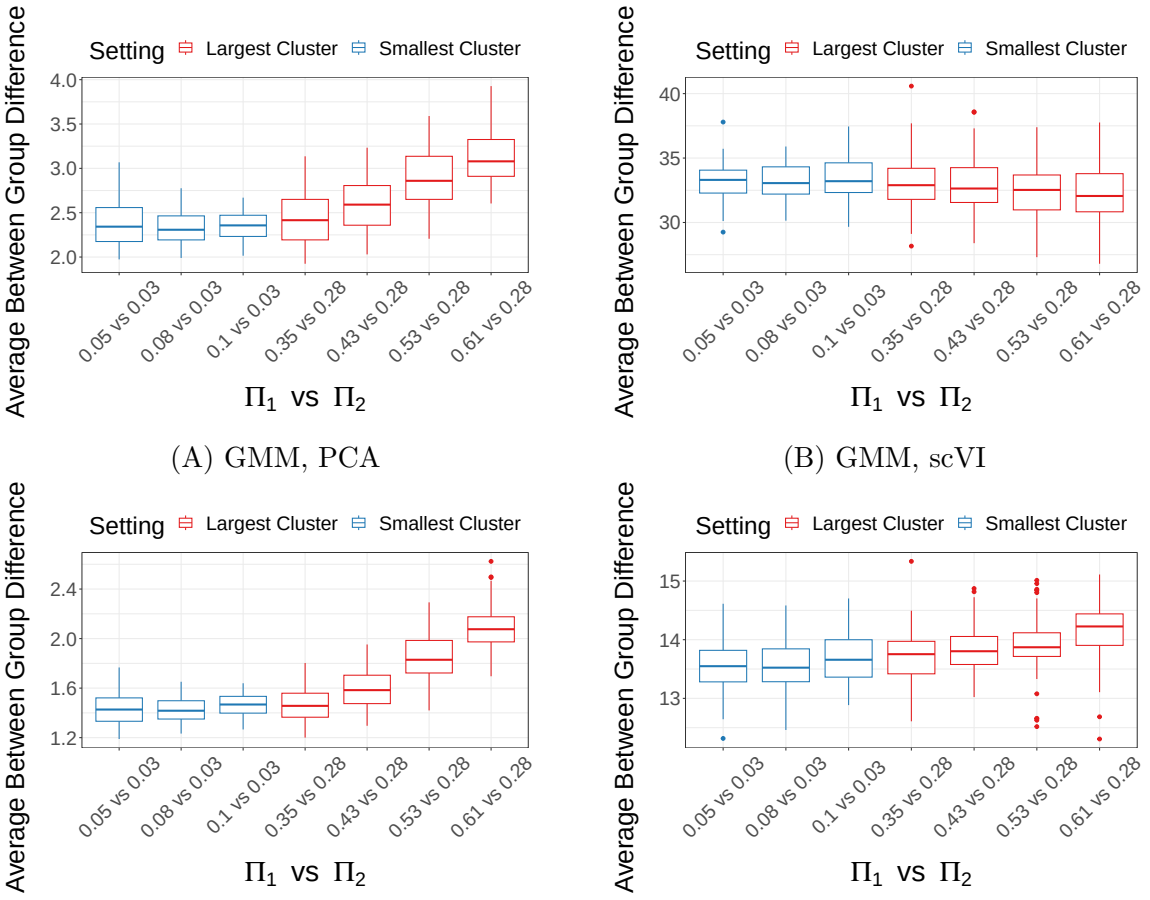


Figure 4.10: **GloScope captures simulated effects** Plots (A) and (B) show how the average GloScope divergence between samples in different phenotype groups increases with (A) increased cell composition differences and (B) increased gene expression differences. The cell composition differences in (A) are color-coded as to whether the major changes were in the two groups' largest cluster or smallest cluster (the actual values of the proportion changes in the largest or smallest group, Π_1 vs Π_2, are labeled in the legends). Plots (C) and (D) shows how the average GloScope divergence between samples in the same phenotype group increases with (C) increased sample variability in gene expression differences and (D) increased cell composition differences. All boxplots show these averages over 100 simulations. The dissimilarity matrices were calculated using the GMM-based GloScope representation based on PCA estimates of the latent space in 10 dimensions. For choices of kNN with scVI or PCA and GMM with scVI, see Figure 4.11-4.14



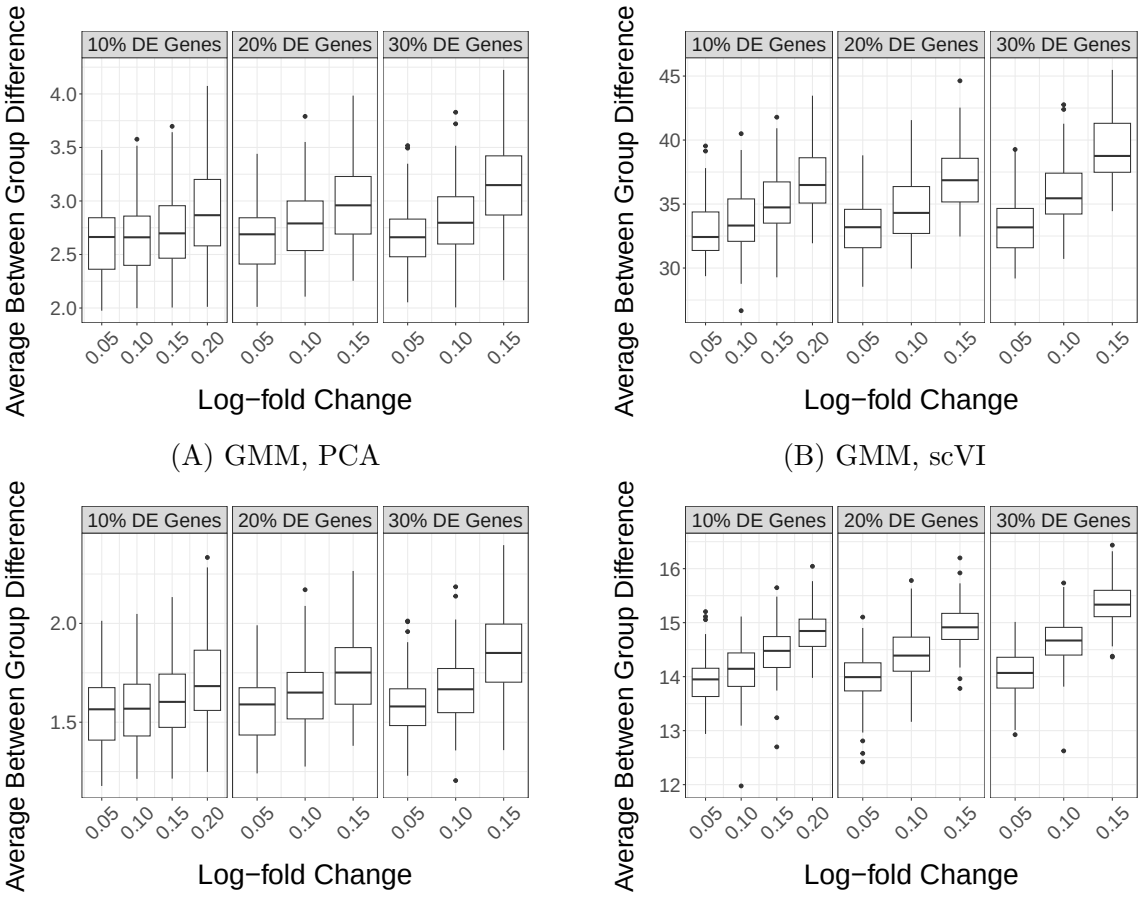
(A) GMM, PCA

(B) GMM, scVI

(C) kNN, PCA

(D) kNN, scVI

Figure 4.11: **Boxplot demonstration of global cell type composition changes detection by GloScope.** The major changes were in the two groups' largest cluster or smallest cluster (the actual values of the proportion changes in the largest or smallest group, Π_1 vs Π_2, are labeled in the legends). Each box is drawn from 100 simulation's average between group distance, calculated using 10 dim embeddings.



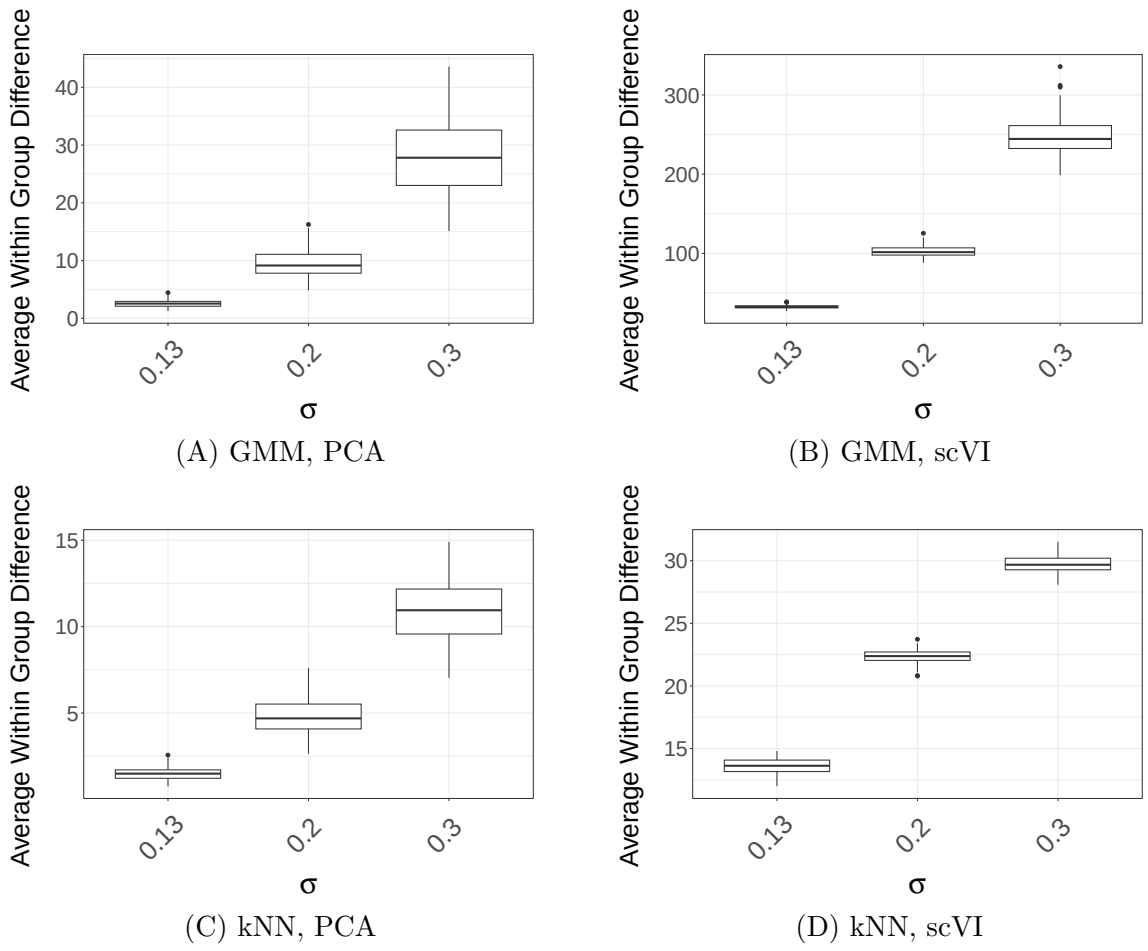
(A) GMM, PCA

(B) GMM, scVI

(C) kNN, PCA

(D) kNN, scVI

Figure 4.12: **Boxplot demonstration of gene expression changes detection by GloScope.** Each box is drawn from 100 simulation's average between group differences, calculated using either GMM or kNN density estimation with either 10 dimensional PCA or scVI 10 embeddings. Upward trend of distance was observed in each combination when log-fold change and percentage of DE genes increase.



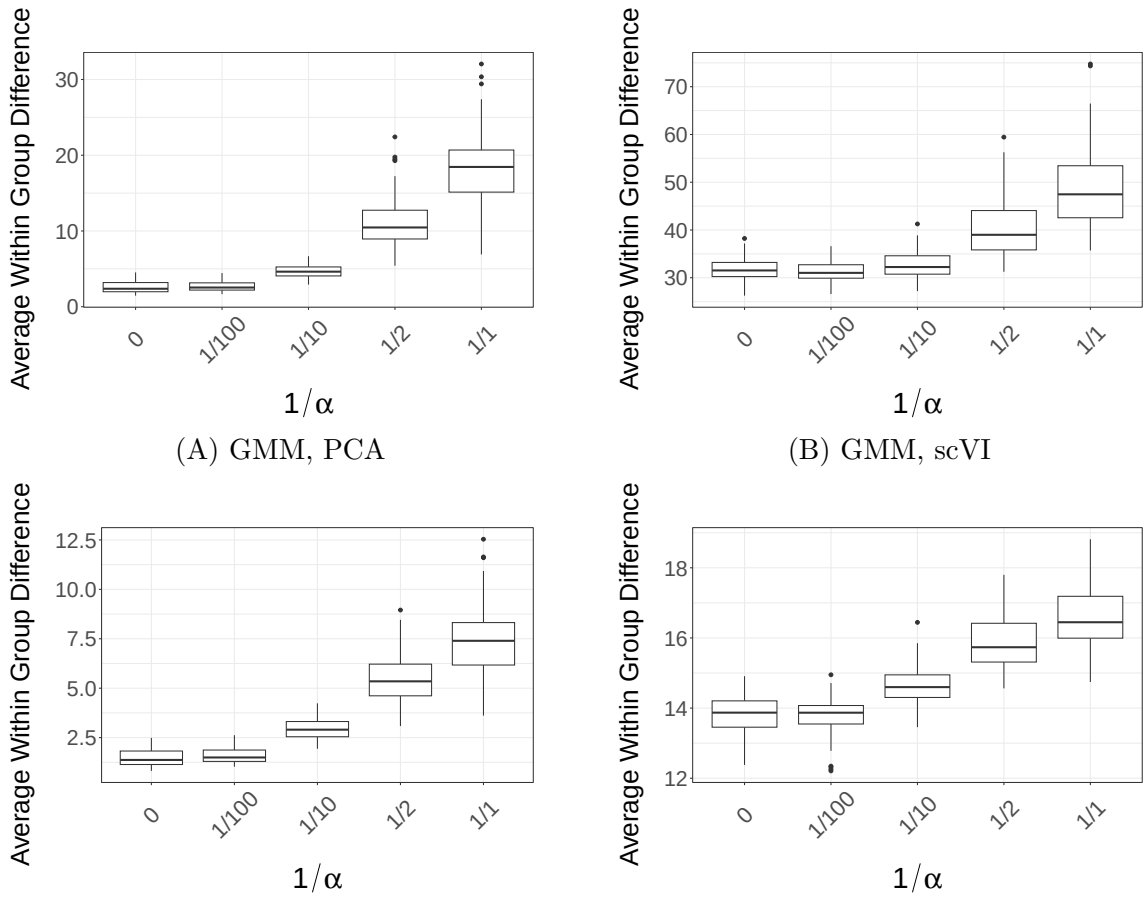
(A) GMM, PCA

(B) GMM, scVI

(C) kNN, PCA

(D) kNN, scVI

Figure 4.13: **Boxplot demonstration of detecting increased sample level variation in the gene expression differences by GloScope.** Each box is drawn from 100 simulations' average divergences among sample within a single phenotype group distance using either GMM or kNN density estimation with either 10 dimensional PCA or scVI 10 embeddings. 10 dimensions. Larger variation of average within group distance could be easily detected in most combination when sample level gene expression variation σ increases.



(A) GMM, PCA

(B) GMM, scVI

(C) kNN, PCA

(D) kNN, scVI

Figure 4.14: **Boxplot demonstration of detecting increased cluster proportion variation α by GloScope.** Each box is drawn from 100 simulations' average divergence among samples within a single phenotype group, calculated using eitherr GMM or kNN density estimation with either 10 dimensional PCA or scVI 10 embeddings. Larger variation in the average within group distances can be easily observed When sample level cluster proportion variation $1/\alpha$ gets larger.

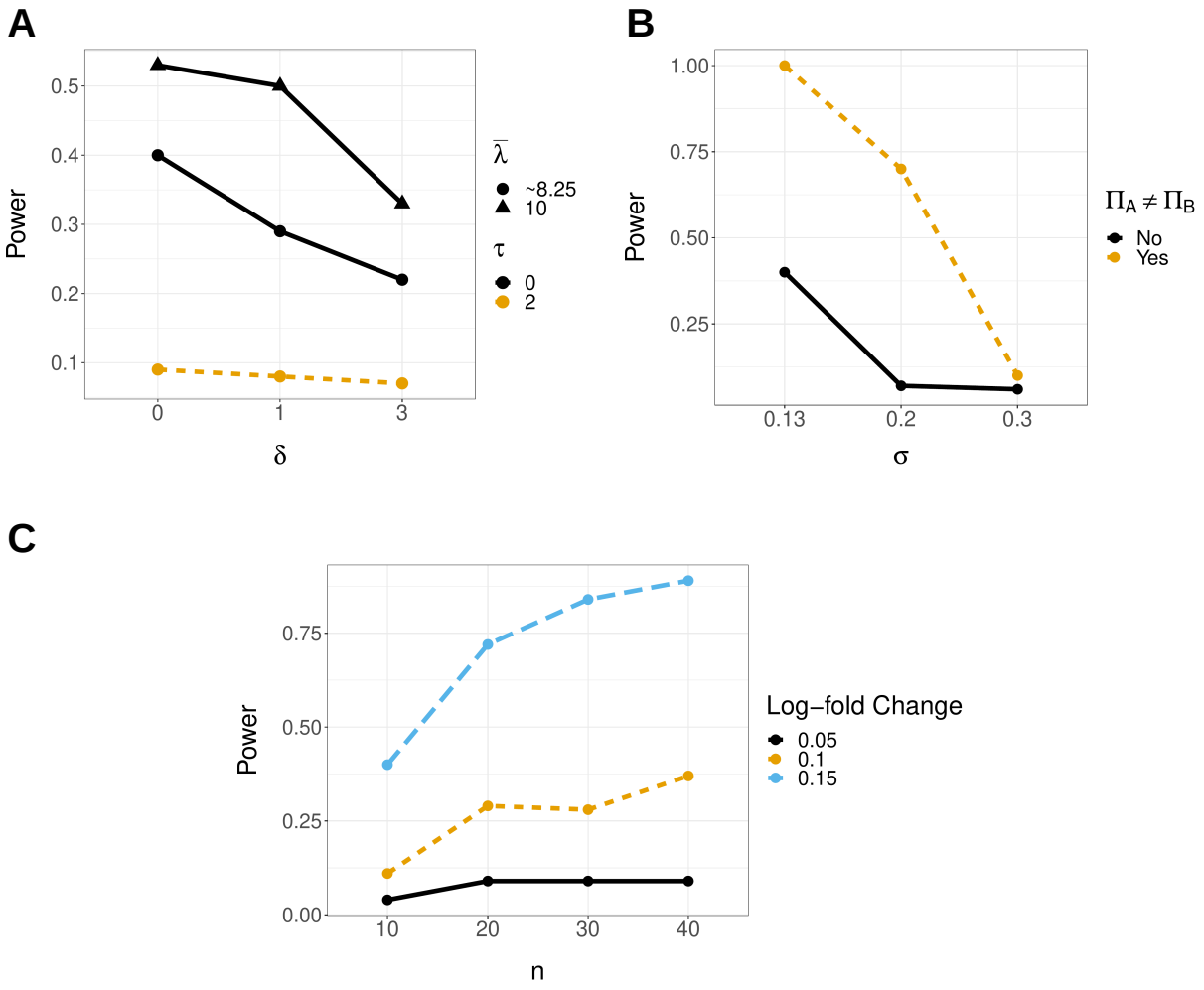


Figure 4.15: **Effect of changing various sources of sample variability on the power to detect group differences.** (A) Power to detect log-fold change differences in the presence of variation in the average library sizes between samples (λ) and individual cells within a sample (τ); (B) Power to detect log-fold change differences in the presence of variation in the baseline expression levels between samples (σ); (A) and (B) have log-fold changes on average of 0.15 in 10% of DE genes. (C) Power to detect log-fold change differences in the presence of variation in the sample size within a single groups (n). Power of ANOSIM calculated based GloScope representation using GMM density estimation and reduced dimensionality representation via PCA with 10 dimensions.

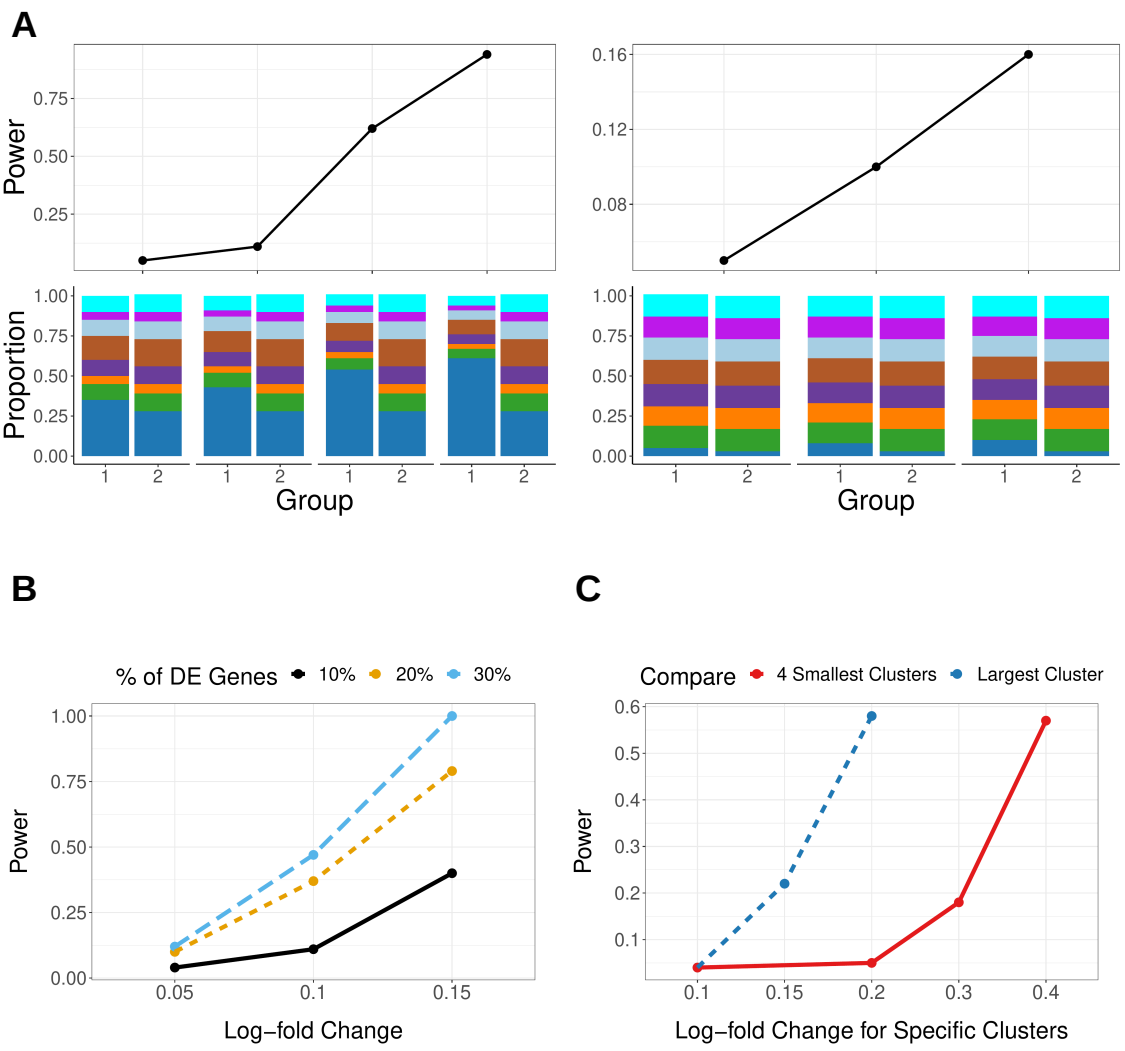


Figure 4.16: **ANOSIM power on simulated data (y-axis) under different conditions** (A) Changes in only the cell-type composition (no DE genes), with major changes in the two groups' largest cluster (left) or smallest cluster (right). The cell-type composition is visualized in the lower panels. (B) Increasing percentage of DE genes (ρ_{DE}) with average log-fold change changing from 0.05, 0.1, and 0.15 (x-axis). (C) Changes of log-fold-changes concentrated in specific cell-types/clusters (ω_k), quantified as relative to the baseline log-fold change θ =0.05; the two lines correspond to whether the log-fold changes were in the largest cluster (representing $\pi_k = 40\%$ proportion of cells) or for the 4 smallest cluster (representing $\pi_k = 30\%$ proportion of cells). Power calculations were done on relatively small groups to show the full range of changes (n=10 samples in each group) with $m = 5,000$ cells per sample; the sample level variability parameter σ is fixed at 0.13, and the sequencing depth $\lambda = 8.25$ (see Methods for details on these parameters). GloScope was calculated based on GMM density estimation with latent space representation via the first 10 dimensions of PCA.

samples than the those of PCA (Figure 4.18), potentially resulting in less power to detect the shared phenotypic differences. On the other hand, scVI representations have more power than their PCA counterparts when the source of differences is due to log-fold changes in genes (Figure 4.19), perhaps due to better accounting for sparse low-count data.

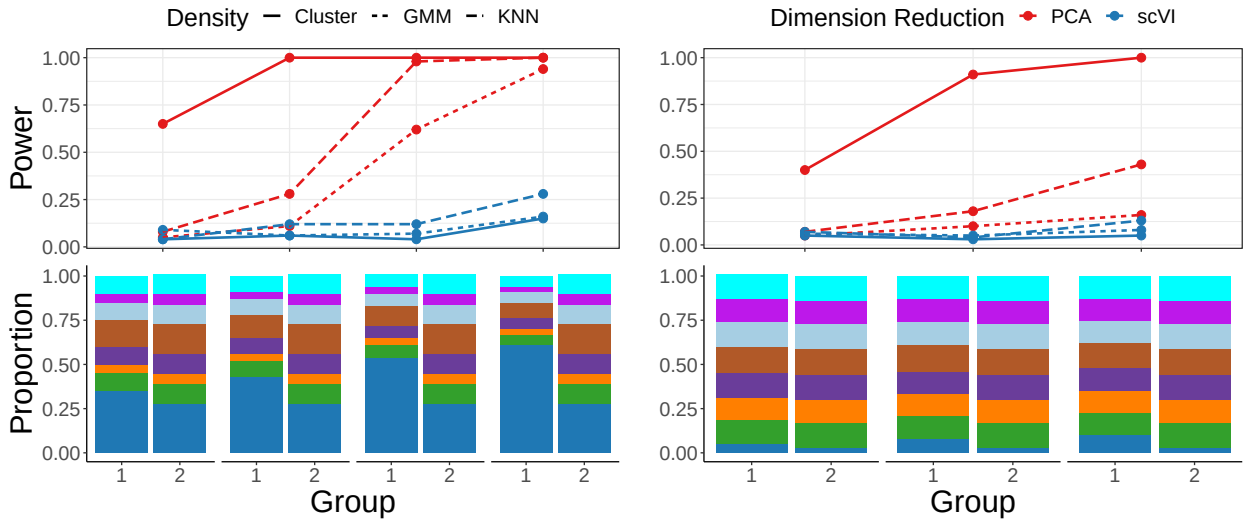


Figure 4.17: **Change in cell-type composition (no DE genes).** Major changes were in the two groups' largest cluster (left) or smallest cluster (right). The cell-type composition is visualized in the lower panels. Each group consists of n=10 samples with m = 5000 cells per sample (the sample level variability parameter σ is fixed at 0.13, and the sequencing depth $\lambda = 8.25$, see Methods for details on these parameters). Power calculated based on cluster proportion vector, GMM or kNN density estimation, and reduced dimensionality representation via PCA or scVI with 10 dimensions.

Finally, we can also consider choices made in implementing GloScope, in particular in the choice of estimation of the density of the latent variables Z in each sample. We consider two popular density estimation strategies as mentioned in Chapter 2: parametric Gaussian mixture models (GMMs) and non-parametric k-nearest neighbors (kNNs). We do not observe large differences in the power of these methods when varying the level of differential expression (Figure 4.19), but kNN is somewhat more powerful in the presence of cell-type composition changes (Figure 4.17).

4.3 Summary

In this chapter, we compared GloScope to the cell-level visualization tool, and limited available strategies for summarizing the data from a single patient: cell-type composition and pseudobulk. We show that alternative methods are not as sensitive in as diverse of settings.

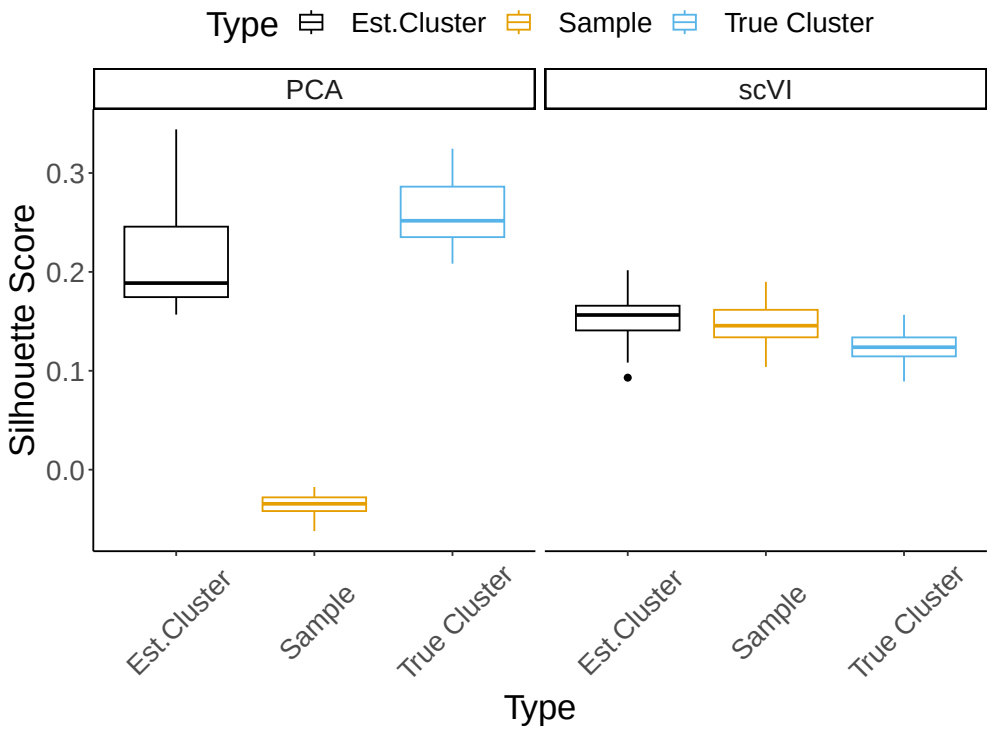


Figure 4.18: **Evaluation of PCA and scVI discrimination of samples and group variability.** Individual cells were simulated from 10 sample with sample-level variability ($\sigma = 0.13$) and reduced to 10 dimensions, either with PCA or scVI. For each simulation, the silhouette score of the reduced dimensionality reduction was calculated at the individual cell-level to assess the similarity of cells within the same sample, compared to the similarity of cells within the same subtype. Larger values indicate larger separation between either samples or subtypes. PCA shows small variation between samples compared to the variation between subtypes, while Each boxplot consists of the silhouette scores for assessing the goodness of clustering different factors for dimension reduction embeddings obtained from either PCA or ScVI. 100 simulations were made to estimate the distance matrics.

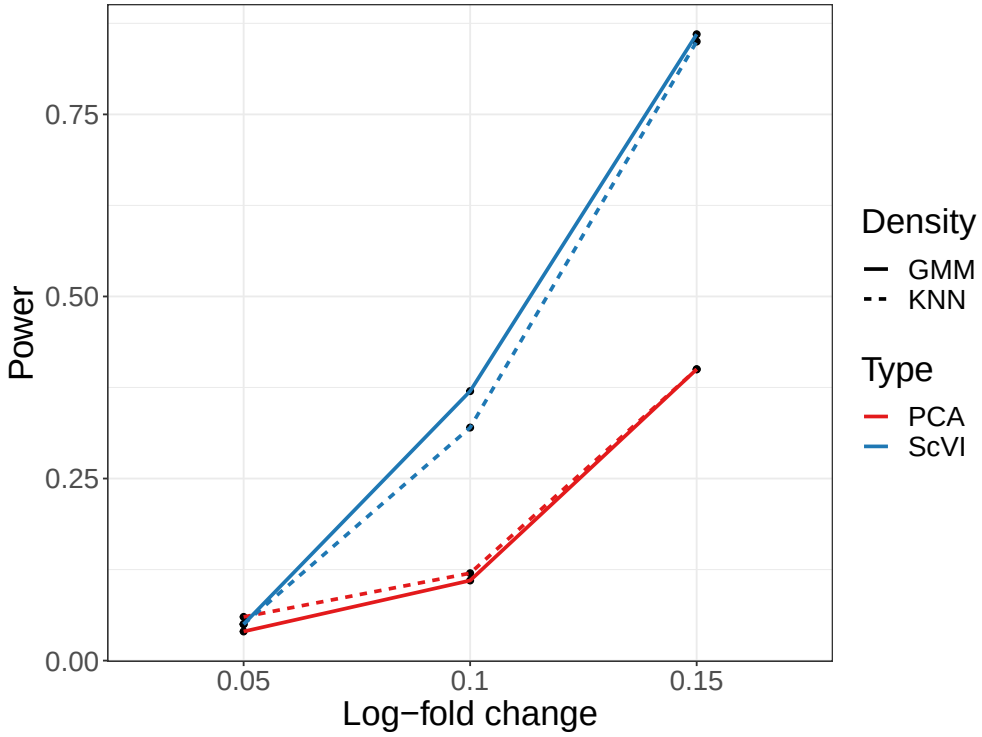


Figure 4.19: **Evaluation of the different choices by calculating the power of detecting gene expressios.** 100 simulations were made to estimate the distance matrics. Power of ANOSIM calculated based GloScope representation using kNN or GMM density estimation and reduced dimensionality representation via scVI or PCA with 10 dimensions. ScVI shows much stronger power of between group difference detection compared to PCA, while there is not much distinction observed when compare GMM vs kNN.

In particular, these competing approaches each focus on one aspect of the sample data (cell-type proportions or gene expression) and are not sensitive to changes found in the other. GloScope uses the entire distribution of the data, thus effectively combining both cell-type proportions and gene expression in a single summary. Furthermore, GloScope is far more flexible for incorporation at different stages of the analysis, whether working with raw counts or normalized data.

We also delves into the application and modification of a scRNA-seq count data simulation pipeline for the quantitative evaluation of GloScope. By integrating specific alterations into the established pipeline by Crowell et al. (2020), we tailored it to better suit the unique requirements of our evaluation process for sample level analysis. Through comprehensive and rigorous simulation experiments, we assessed the accuracy, robustness, and overall performance of GloScope, thereby demonstrating its potential utility of the simulation pipeline in potential hypothesis testing and advanced single-cell data analysis.

Bibliography

Adams, T. S., Schupp, J. C., Poli, S., Ayaub, E. A., Neumark, N., Ahangari, F., Chu, S. G., Raby, B. A., DeIuliis, G., Januszyk, M., Duan, Q., Arnett, H. A., Siddiqui, A., Washko, G. R., Homer, R., Yan, X., Rosas, I. O., and Kaminski, N. (2020). Single-cell rna-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science Advances*, 6(28).

Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., and Darrell, T. (2005). Face recognition with image sets using manifold density divergence. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 581–588 vol. 1. IEEE.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2024). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R package version 1.1.4.

Boltz, S., Debreuve, E., and Barlaud, M. (2009). High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18(6):1266–1283.

Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., and Nolan, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26).

Chen, W. S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., and Krishnaswamy, S. (2020). Uncovering axes of variation among single-cell cancer specimens. *Nature Methods*, 17(3):302–310.

Cheng, J. B., Sedgewick, A. J., Finnegan, A. I., Harirchian, P., Lee, J., Kwon, S., Fassett, M. S., Golovato, J., Gray, M., Ghadially, R., Liao, W., Perez White, B. E., Mauro, T. M., Mully, T., Kim, E. A., Sbitany, H., Neuhaus, I. M., Grekin, R. C., Yu, S. S., Gray, J. W., Purdom, E., Paus, R., Vaske, C. J., Benz, S. C., Song, J. S., and Cho, R. J. (2018). Transcriptional programming of normal and inflamed human epidermis at single-cell resolution. *Cell Reports*, 25(4):871–883.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional scaling*. Chapman and Hall.

Crowell, H. L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1).

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1).

Fabre, T., Barron, A. M., Christensen, S. M., Asano, S., Bound, K., Lech, M. P., Wadsworth, II, M. H., Chen, X., Wang, C., Wang, J., McMahon, J., Schlerman, F., White, A., Kravarik, K. M., Fisher, A. J., Borthwick, L. A., Hart, K. M., Henderson, N. C., Wynn, T. A., and Dower, K. (2023). Identification of a broadly fibrogenic macrophage subset induced by type 3 inflammation. *Science Immunology*, 8(82).

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, 16(1).

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3).

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell rna-seq data. *Genome Biology*, 17(1).

Jiang, R., Sun, T., Song, D., and Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome Biology*, 23(1).

Johnsson, K., Wallin, J., and Fontes, M. (2016). Bayesflow: latent modeling of flow cytometry cell populations. *BMC Bioinformatics*, 17(1).

Joodaki, M., Shaigan, M., Parra, V., Bülow, R. D., Kuppe, C., Hölscher, D. L., Cheng, M., Nagai, J. S., Goedertier, M., Bouteldja, N., Tesar, V., Barratt, J., Roberts, I. S., Coppo, R., Kramann, R., Boor, P., and Costa, I. G. (2023). Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular Systems Biology*, 20(2):57–74.

Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., Bushman, F. D., and Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and permanova. *Bioinformatics*, 31(15):2461–2468.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296.

Kulkarni, A., Anderson, A. G., Merullo, D. P., and Konopka, G. (2019). Beyond bulk: A review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology*, 58:129–136.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lexogen (2024). Demystifying single-cell rna sequencing: A comprehensive guide.

Li, C. M.-C., Shapiro, H., Tsiobikas, C., Selfors, L. M., Chen, H., Rosenbluth, J., Moore, K., Gupta, K. P., Gray, G. K., Oren, Y., and et al. (2020a). Aging-associated alterations in mammary epithelia and stroma revealed by single-cell rna sequencing. *Cell Reports*, 33(13):108566.

Li, C. M.-C., Shapiro, H., Tsiobikas, C., Selfors, L. M., Chen, H., Rosenbluth, J., Moore, K., Gupta, K. P., Gray, G. K., Oren, Y., Steinbaugh, M. J., Guerriero, J. L., Pinello, L., Regev, A., and Brugge, J. S. (2020b). Aging-associated alterations in mammary epithelia and stroma revealed by single-cell rna sequencing. *Cell Reports*, 33(13):108566.

Li, X. and Wang, C.-Y. (2021). From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, 13(1).

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Liu, Y., Wang, H., Taylor, M., Cook, C., Martínez-Berdeja, A., North, J. P., Harirchian, P., Hailer, A. A., Zhao, Z., Ghadially, R., Ricardo-Gonzalez, R. R., Grekin, R. C., Mauro, T. M., Kim, E., Choi, J., Purdom, E., Cho, R. J., and Cheng, J. B. (2022). Classification of human chronic inflammatory skin disease based on single-cell immune profiling. *Science Immunology*, 7(70).

Llorens-Bobadilla, E., Zhao, S., Baser, A., Saiz-Castro, G., Zwadlo, K., and Martin-Villalba, A. (2015). Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell*, 17(3):329–340.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: A tutorial. *Molecular Systems Biology*, 15(6).

Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., and et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Melms, J. C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A. F., Amin, A. D., Schapiro, D., Frangieh, C. J., Luoma, A. M., Filliol, A., Fang, Y., Ravichandran, H., Clausi, M. G., Alba, G. A., Rogava, M., Chen, S. W., Ho, P., Montoro, D. T., Kornberg, A. E., Han, A. S., Bakhoum, M. F., Anandasabapathy, N., Suárez-Farirñas, M., Bakhoum, S. F., Bram, Y., Borczuk, A., Guo, X. V., Lefkowitch, J. H., Marboe, C., Lagana, S. M., Del Portillo, A., Tsai, E. J., Zorn, E., Markowitz, G. S., Schwabe, R. F., Schwartz, R. E., Elemento, O., Saqi, A., Hibshoosh, H., Que, J., and Izar, B. (2021). A molecular single-cell lung atlas of lethal covid-19. *Nature*, 595(7865):114–119.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7):621–628.

Orlova, D. Y., Meehan, S., Parks, D., Moore, W. A., Meehan, C., Zhao, Q., Ghosn, E. E. B., Herzenberg, L. A., and Walther, G. (2018). Qfmatch: multidimensional flow and mass cytometry samples alignment. *Scientific Reports*, 8(1).

Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E. B., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A., and Walther, G. (2016). Earth mover's distance (emd): A true metric for comparing biomarker expression levels in cell populations. *PLOS ONE*, 11(3):e0151859.

Osorio, D. and Cai, J. J. (2020). Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell rna-sequencing data quality control. *Bioinformatics*, 37(7):963–967.

Paik, D. T., Cho, S., Tian, L., Chang, H. Y., and Wu, J. C. (2020). Single-cell rna sequencing in cardiovascular development, disease and medicine. *Nature Reviews Cardiology*, 17(8):457–473.

Pelka, K., Hofree, M., Chen, J. H., Sarkizova, S., Pirl, J. D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W. H., Xu, K. H., Chao, S. X., Zollinger, D. R., Lieb, D. J., Reeves, J. W., Fuhrman, C. A., Hoang, M. L., Delorey, T., Nguyen, L. T., Waldman, J., Klapholz, M., Wakiro, I., Cohen, O., Albers, J., Smillie, C. S., Cuoco, M. S., Wu, J., Su, M.-j., Yeung, J., Vijaykumar, B., Magnuson, A. M., Asinovski, N., Moll, T., Goder-Reiser, M. N., Applebaum, A. S., Brais, L. K., DelloStritto, L. K., Denning, S. L., Phillips, S. T., Hill, E. K., Meehan, J. K., Frederick, D. T., Sharova, T., Kanodia, A., Todres, E. Z., Jané-Valbuena, J., Biton, M., Izar, B., Lambden, C. D., Clancy, T. E., Bleday, R., Melnitchouk,

N., Irani, J., Kunitake, H., Berger, D. L., Srivastava, A., Hornick, J. L., Ogino, S., Rotem, A., Vigneau, S., Johnson, B. E., Corcoran, R. B., Sharpe, A. H., Kuchroo, V. K., Ng, K., Giannakis, M., Nieman, L. T., Boland, G. M., Aguirre, A. J., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A., and Hacohen, N. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18):4734–4752.e20.

Perez, R. K., Gordon, M. G., Subramaniam, M., Kim, M. C., Hartoularos, G. C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., Thompson, M., Rappoport, N., Dahl, A., Lanata, C. M., Matloubian, M., Maliskova, L., Kwek, S. S., Li, T., Slyper, M., Waldman, J., Dionne, D., Rozenblatt-Rosen, O., Fong, L., Dall'Era, M., Balliu, B., Regev, A., Yazdany, J., Criswell, L. A., Zaitlen, N., and Ye, C. J. (2022). Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589).

Pierson, E. and Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1).

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391).

Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B., and Saez-Rodriguez, J. (2023). Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., and et al. (2017). The human cell atlas. *eLife*, 6.

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9(1).

Scrucca, L., Fop, M., Murphy, T., B., and Raftery, Adrian, E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289.

Soon, W. W., Hariharan, M., and Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, 9(1).

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J. S., Wilson, N. K., Mende, N., Jardine, L., Gardner,

L. C. S., Goh, I., Horsfall, D., McGrath, J., Webb, S., Mather, M. W., Lindeboom, R. G. H., Dann, E., Huang, N., Polanski, K., Prigmore, E., Gothe, F., Scott, J., Payne, R. P., Baker, K. F., Hanrath, A. T., Schim van der Loeff, I. C. D., Barr, A. S., Sanchez-Gonzalez, A., Bergamaschi, L., Mescia, F., Barnes, J. L., Kilich, E., de Wilton, A., Saigal, A., Saleh, A., Janes, S. M., Smith, C. M., Gopee, N., Wilson, C., Coupland, P., Coxhead, J. M., Kiselev, V. Y., van Dongen, S., Bacardit, J., King, H. W., Rostron, A. J., Simpson, A. J., Hambleton, S., Laurenti, E., Lyons, P. A., Meyer, K. B., Nikolić, M. Z., Duncan, C. J. A., Smith, K. G. C., Teichmann, S. A., Clatworthy, M. R., Marioni, J. C., Göttgens, B., and Haniffa, M. (2021). Single-cell multi-omics analysis of the immune response in covid-19. *Nature Medicine*, 27(5):904–916.

Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272.

Tanay, A. and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338.

Tiberi, S., Crowell, H. L., Samartsidis, P., Weber, L. M., and Robinson, M. D. (2020). distinct: A novel approach to differential distribution analyses.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21(1).

Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., Patro, R., and Robinson, M. D. (2019). Rna sequencing data: Hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science*, 2(1):139–173.

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., Robinson, M. D., Dudoit, S., and Clement, L. (2018). Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome Biology*, 19(1).

Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1).

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681.

van Galen, P., Hovestadt, V., Wadsworth II, M. H., Hughes, T. K., Griffin, G. K., Battaglia, S., Verga, J. A., Stephansky, J., Pastika, T. J., Lombardi Story, J., and et al. (2019). Single-cell rna-seq reveals aml hierarchies relevant to disease progression and immunity. *Cell*, 176(6).

Wagner, J., Rapsomaniki, M. A., Chevrier, S., Anzeneder, T., Langwieder, C., Dykgers, A., Rees, M., Ramaswamy, A., Muenst, S., Soysal, S. D., Jacobs, A., Windhager, J., Silina, K., van den Broek, M., Dedes, K. J., Rodríguez Martínez, M., Weber, W. P., and Bodenmiller, B. (2019). A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, 177(5):1330–1345.e18.

Wang, Q., Kulkarni, S., and Verdu, S. (2006). A nearest-neighbor approach to estimating divergence between continuous random vectors. In *2006 IEEE International Symposium on Information Theory*, pages 242–246. IEEE.

Wang, X., Xing, E. P., and Schaid, D. J. (2014). Kernel methods for large-scale genomic data analysis. *Briefings in Bioinformatics*, 16(2):183–192.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Weaver, W. M., Tseng, P., Kunze, A., Masaeli, M., Chung, A. J., Dudani, J. S., Kittur, H., Kulkarni, R. P., and Di Carlo, D. (2014). Advances in high-throughput single-cell microtechnologies. *Current Opinion in Biotechnology*, 25:114–123.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17.

Yao, Z., van Velthoven, C. T., Nguyen, T. N., Goldy, J., Sedeno-Cortes, A. E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., Ding, S.-L., Fong, O., Garren, E., Glandon, A., Gouwens, N. W., Gray, J., Graybuck, L. T., Hawrylycz, M. J., Hirschstein, D., Kroll, M., Lathia, K., Lee, C., Levi, B., McMillen, D., Mok, S., Pham, T., Ren, Q., Rimorin, C., Shapovalova, N., Sulc, J., Sunkin, S. M., Tieu, M., Torkelson, A., Tung, H., Ward, K., Dee, N., Smith, K. A., Tasic, B., and Zeng, H. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.e26.

Zhang, F., Wu, Y., and Tian, W. (2019). A novel approach to remove the batch effect of single-cell data. *Cell Discovery*, 5(1).

Zhang, M., Liu, S., Miao, Z., Han, F., Gottardo, R., and Sun, W. (2022). Ideas: Individual level differential expression analysis for single-cell rna-seq data. *Genome Biology*, 23(1).