Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Luxiao Chen Date

Deciphering the Cell Type Specific Activities from High-throughput Omics Data

By

Luxiao Chen

Doctor of Philosophy

Biostatistics

Hao Wu, Ph.D.
Advisor

Karen N. Conneely, Ph.D.
Committee Member

Ying Guo, Ph.D.
Committee Member

Zhaohui Qin, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D, MPH
Dean of the James T. Laney School of Graduate Studies

Date

Deciphering the Cell Type Specific Activities from High-throughput Omics Data

By

Luxiao Chen

M.S.P.H., Emory University, GA, 2018

M.S., Nanjing University, China, 2016

B.S., Nanjing University, China, 2013

Advisor: Hao Wu, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2023

Abstract

Deciphering the Cell Type Specific Activities from High-throughput Omics Data

By Luxiao Chen

There are hundreds of cell types in the human body carrying different functions. Understanding the cell type specific (CTS) activities will greatly enhance our knowledge on the biological and clinical mechanisms. The advancements in bulk and single cell high-throughput omics technologies enable us to study the CTS effects from the genomics perspective.

Bulk high-throughput omics data contain signals from a mixture of cell types. Recent developments of deconvolution methods facilitate CTS inferences from bulk data. Our real data exploration suggests that differential expression or methylation status is often correlated among cell types. Based on this observation, we developed a novel statistical method named CeDAR to incorporate the cell type hierarchy in CTS differential analyses of bulk data. Extensive simulation and real data analyses demonstrate that this approach significantly improves the accuracy and power in detecting CTS differential signals compared with existing methods, especially in low-abundance cell types.

Single cell RNA-seq (scRNA-seq) allows scientists to study gene expression profile of individual cells in one sample. The increasing interest to apply this technique at population level has facilitated appearing of many datasets containing multiple subjects measured by scRNA-seq. In the real scRNA-seq data, we observed that CTS genes may not consistently appear across all subjects, while they are expected to appear consistently. Motivated by this observation, we first designed a statistical model to identify CTS genes that consistently appear in population-level scRNA-seq data. We then designed a strategy to incorporate these consistent CTS genes identified from historical data into analyses like cell-typing. Data analyses demonstrate that the proposed method and strategy can well identify consistent CTS genes and improve downstream analysis performance.

In scRNA-seq data, cells from extremely low-abundance cell types are called rare cell population (RCP), which plays great roles in biological activities. Because its low abundance, traditional clustering methods can hardly identify it. To correctly identify RCPs in scRNA-seq data, methods with different focuses have been developed. This provides great opportunity for RCP studies; meanwhile, it also makes users difficult to choose. Thus, we summarized these methods and benchmarked them with simulated data to provide comprehensive evaluation with different metrics.

Deciphering the Cell Type Specific Activities from High-throughput Omics Data

By

Luxiao Chen

M.S.P.H., Emory University, GA, 2018

M.S., Nanjing University, China, 2016

B.S., Nanjing University, China, 2013

Advisor: Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2023

Acknowledgments

I would like to convey my deepest gratitude to my advisor, Dr. Hao Wu, for consistently providing me with invaluable guidance and unconditional support throughout my time at Emory. Your brilliant ideas, extensive experience, efficient work style, and unwavering passion for both research and life have had a profound impact on me. I wish to express my genuine appreciation to my committee members, Dr. Karen N. Conneely, Dr. Ying Guo and Dr. Zhaohui Qin, whose insightful feedback and suggestions were instrumental in enhancing the quality of my dissertation and provided me with valuable guidance for future research avenues.

Many thanks to Dr. José Binongo for providing me great collaboration opportunities and support in my initial two years at Emory. Additionally, I extend my heartfelt appreciation to Mary Abosi, Angela Guinyard, Melissa Sherrer, Bob Waggoner, and Porchia Arnold for their kindness and help.

Many thanks to my friends for making my life more enjoyable and fun-filled.

Last but not least, I am extremely grateful for my parents. Without their love, understanding, and support, I wouldn't have made it here.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cell type specificity in biological activities

Different cell types have different sizes, shapes, and functions that they play different roles in biological activities. For example, a typical neuron consists of soma, dendrites, and axon, whose structure is highly specialized to function for processing and transmitting cellular signals (Ludwig et al., 2022); differently, the biconcave disk shape of red blood cells in mammals, which facilitates their large reversible elastic deformation during micro-circulation, is necessary to transport oxygen and carbon dioxide (Diez-Silva et al., 2010).

In addition, different cell types also behave differently in response to changes in their living micro-environment that are caused by factors like diseases, drugs or other stimuli from external environments. Moonen et al. (2023) reported that in Alzheimer disease (AD), cell types - astrocyte, microglia and neuron show cell type-specific activation of pyroptosis. Georges and Janmey (2005) reported that given soft substrates surroundings, neurons preferentially branch on it and glia are unable to survive, which could explain why neuron regeneration is limited after injury since some molecules stiffen the injured tissue. Through singles cell RNA sequencing (scRNA-seq) data analysis, Zhao et al. (2021) found that both tumor and non-tumor populations are affected by histone deacetylase inhibitor panobinostat; meanwhile only tumor cells are affected by etoposide.

Thus, well understanding cell type specificity in different biological activities can help scientists uncover complex mechanisms behind their interested phenomenon and provide chances for accurate clinical diagnosis and drug development.

1.2 Cell type specificity analysis with omics data

Through Central Dogma theory, we know genetic information flows between DNA, RNA, and protein that the molecules play important role in determining a cell's fate

(Crick, 1970). So, by studying these molecules we can better understand cell type specificity in various biological activities under different conditions. To achieve this goal, the most straightforward way is to quantify these molecules by omics data. For example, given gene expression or DNA methylation level, we can easily to find cell type specific genes or methylation sites for different cell types and use them to identify/mark corresponding cell type. Furthermore, we can also compare the gene expression or DNA methylation of a cell type under different conditions to study whether and how it responses to the condition change.

"Omics" is a broad concept that it can be used for any scientific field associated with measuring certain biological molecules in a high-throughput way (Micheel et al., 2012). For example, proteomics studies proteins, transcriptomics studies RNA, genomics studies genes, and epigenomics studies methylated DNA or modified histone proteins in chromosomes.

1.2.1 Brief introduction to some types of omics data

It is known that protein is the final product of expressed gene that it executes various biological functions. Instead of direct measuring protein amount to quantify gene expression, measuring mRNA abundance is preferred. It is because mRNA abundance is much easier to measure than protein amount and mRNA abundance is positively correlated with protein amount. To measure the mRNA abundance, the early technique is microarray. An array is a solid surface with a collection of spots on it. A spot contains many copies of same DNA sequences (called "probes") designed to target specific gene. These probes will be hybridized with sequences from their target genes in one sample. The amount of hybridization on each probe represents the amount of mRNA for its target gene, which is measured by fluorescent intensities (San Segundo-Val and Sanz-Lozano, 2016). Given a set of samples, the final result is a matrix with continuous measurement, in which row for probes and column for samples. A log-

normal distribution is often applied on it for downstream statistical analysis (Smyth et al., 2003).

The major limitation of microarray is that the probes are pre-designed that it can only be used to study known genes. This problem can be solved by RNA sequencing (RNA-seq). In one sample, RNA sequences are extracted and converted to cDNA first. These cDNA sequences are then fragmented into short reads and sequenced. Expression level of a gene is quantified by the number of sequenced reads aligned the its corresponding region on genome. Different from microarray, the gene expression level is measured in count. Besides, RNA-seq provides more information than gene expression arrays, such as alternative splicing and gene fusion (Wang et al., 2009).

The gene expression generated by (bulk) RNA-seq is an average count of all cells in one sample. This could be a disadvantage of RNA-seq technique. Because its measured gene expression cannot be used for cell type specificity analysis when cells in one sample have high heterogeneity (a.k.a from different cell types or states). Before the appearance of single cell RNA-seq (scRNA-seq), the only solution is to purify cells from a tissue first, which is expensive and limits novel discovery for new cell types. The scRNA-seq technique can quantify gene expression profile for each cell in analyzed samples, which provides researchers great opportunities for enormous novel biological findings. The major difference of scRNA-seq from RNA-seq is that given a sample, cells will be isolated with methods like Fluorescence-activated cell sorting (FACS) and then submitted for sequencing. So the final result for each sample is no longer a vector but a matrix with row for gene and column for cells.

DNA methylation data is another widely interested omics data. DNA methylation is an epigenetic modification of the DNA sequence by adding a methyl group to the 5-methylcytosine. This process can be affected by environment and it is close related with gene regulation and development. DNA methylation level can also be quantified by arrays (e.g., Illumina Infinium) and sequencing (e.g., bisulfite sequencing)

methods. Arrays appeared much earlier than sequencing, but they are still popular now. The major reasons are that arrays are cheap, simple to analyze, have good reproducibility allowing comparison with previous results (Bibikova et al., 2006), and provide sufficient sensitivity and specificity in most of time (Teh et al., 2016). The advantage of sequencing methods is that they can profile DNA cytosine methylation genome-wide at a single nucleotide resolution. Such high resolution provides the possibility to explore methylation patterns far beyond what arrays can provide (Rauluseviciute et al., 2019). In methylation array data, a common used metric for measuring methylation level at each site is β-value: $\beta = \frac{\text{Max}(M,0)}{\text{Max}(M,0)+\text{Max}(U,0)+100}$, where M is the averaged signal for methylated alleles and U is the averaged signal for unmethylated alleles (Wilhelm-Benartzi et al., 2013). Clearly, the β-value ranges from 0 to 1. In bisulfite sequencing data, at each position, we can have the total number of reads and the methylated number of reads. A beta-binomial distribution is commonly used for downstream statistical analysis on these counts (Feng et al., 2014).

1.2.2 Cell type specificity analysis with bulk omics data

Since the result of bulk omics data is an average measurement of all cells in one sample, the most straightforward way to use bulk omics data study cell type specificity is to make sure the sample only contains cells from one cell type, which needs purification of the sample. However, the purification process is often expensive and time-consuming.

An alternative method is to apply deconvolution analysis on these bulk omics data. The key intuition behind the deconvolution analysis is that bulk sample gene expression is a weighted average of gene expression of different cell types, where the weight is proportion of cell types in the sample. This is shown in Equation 1.1:

$$Y_{gi} = \sum_{k=1}^{K} \theta_{ki} X_{gk} \tag{1.1}$$

In the Equation 1.1, there are K cell types in the i-th bulk sample, Y_{gi} is the g-th gene expression in i-th sample, θ_{ki} is the k-th cell type proportion in i-th sample, and X_{gk} is the g-th gene expression in k-th cell type.

One major type of the deconvolution analysis is cell type proportion estimation. Given bulk samples gene expression/DNA methylation as input, the output is proportion of cell types in given samples. Based on whether a reference gene expression/DNA methylation of different cell types is needed, the cell type proportion estimation can be grouped as reference-based and reference-free. In reference-based methods, the gene expression profile of different cell types is assumed to be known, which can be estimated from external data sets (Tsoucas et al., 2019). In reference-free methods, the cell type composition in samples and the gene expression profile of cell types are needed to be estimated jointly, which is usually based on mathematical framework of non-negative matrix factorization (Teschendorff and Zheng, 2017).

Another type of the deconvolution analysis is cell type specific differential analysis. A simple scenario of this type analysis is given bulk samples from two groups, the result provides information of the differential expressed features between the two groups (e.g., control vs. case) in each cell type. In such analysis, most methods assume cell type composition in samples are known and covariates are incorporated into regression model (Li et al., 2019; Zheng et al., 2018a).

$$E\{Y_{gi}\} = \sum_{k=1}^{K} \theta_{ik} E\{X_{gik}\} = \sum_{k=1}^{K} (\theta_{ik}\mu_{gk} + \theta_{ik} Z_i \beta_{gk}) \tag{1.2}$$

In Equation 1.2, there are K cell types in the i-th bulk sample, Y_{gi} is the g-th gene expression in i-th sample, θ_{ki} is the k-th cell type proportion in i-th sample, and X_{gik} is the g-th gene expression in k-th cell type from i-th sample, μ_{gk} is the mean expression of g-th gene in k-th cell type in control group, β_{gk} is the difference of mean expression of g-th gene between the two groups, Z_i is the indicator variable represents

the sample group information.

1.2.3 Cell type specificity analysis with scRNA-seq data

The scRNA-seq data provides gene expression file for each cell in a sample that it is a natural choice for cell type specificity analysis. Even though it is possible to simultaneously quantifies cell surface protein and transcriptomic data within a single cell readout with CITE-seq (Stoeckius et al., 2017) that can help to identify cell type through the surface proteins, most scRNA-seq data cannot provide cell type information for each cell. So a fundamental step in scRNA-seq analysis is cell type identification, which itself is an interested scientific question (e.g., discover novel cell types) and is also necessary for other downstream analyses. There are two methods to annotate these cells that the first one is clustering based and the second one is reference data based. In the clustering based pipeline, after preprocessing and normalization, cells are clustered with highly variable genes. Each cluster represents a cell type and its identify is confirmed by comparing the clusters' specific highly expressed genes with known marker genes of cell types. In the second reference based method, an external scRNA-seq data with cells well annotated is provided. A cell in target sample will be assigned to a cell type group, in which the cells from the group in reference sample show highest similarity with it (Aran et al., 2019).

1.3 Overview

In this dissertation, we first designed a hierarchical model for cell type specific differential analysis with bulk omics data, which has better performance than existing methods (Chapter 2). We then designed a hierarchical model to identify cell type specific (CTS) genes in population level scRNA-seq data and evaluate these CTS genes' consistency of showing differential expression signal in samples (Chapter 3).

In the last work, we performed a comprehensive benchmark analysis on methods designed for rare cell type identification in scRNA-seq data to help researchers choose appropriate tool for their analysis (Chapter 4).

Chapter 2

Incorporating cell type hierarchy improves cell type-specific differential analyses in bulk omics data

2.1 Introduction

The bulk high-throughput omics experiments are often performed on tissue samples, which are mixtures of different cell types. Traditional bulk data analyses for differential expression (DE) and differential methylation (DM) compare the average signals among different groups. However, it has been reported that certain biological and clinical conditions can alter the DNA methylation or gene expression profile in specific cell types. For example, Grubman et al. (2019) reported that Alzheimer's disease (AD) risk gene APOE shows cell type-specific different expression patterns: it is up-regulated for AD in microglial cells, but down-regulated in both oligodendrocyte progenitor cells and astrocytes. Gu et al. (2021) reported that neuron and glia cells show different DNA methylation pattern within SNCA intron 1 in two synucleinopathies — Parkinson's disease (PD) and dementia with Lewy body (DLB). In PD, decreased DNA methylation within SNCA intron 1 only appears in neuron cells; while in DLB, it only appears in glia cells. These cell type-specific changes are important for understanding biological and clinical mechanisms and potentially provide diagnostic biomarkers and therapeutic targets. Thus, researchers often have great interest in identifying cell type-specific alterations under various conditions.

Experiment procedures such as cell sorting or single-cell approaches can directly measure the cell type-specific behaviors. However, the two technologies are laborious and expensive, which limits their large-scale application. While the traditional DE/DM methods for bulk data only compare the average signals, recent development of computational methods makes it possible to perform cell type specific analysis from the bulk data. The cell type-specific analysis on bulk omics data has been an active research field recently. There are several methods developed for signal deconvolution and cell type-specific inference. For example, csSAM (Shen-Orr et al., 2010) adopts a two-step approach: it first estimates pure cell type profiles based on known cell type proportions and then conducts permutation tests to identify cell type-specific

DE (csDE). Both CellDMC (Zheng et al., 2018a) and TOAST (Li et al., 2019) use interaction terms between covariates and cell type proportions in a linear model to test csDE/csDM. This statistical framework has been shown as a generalization of several previous works (Montaño et al., 2013; Westra et al., 2015; Kuhn et al., 2011). TCA (Rahmani et al., 2019) models the cell type-specific methylation levels of each individual and derives a procedure for cell type-specific inference. While CellDMC, TOAST, and TCA mainly focus on continuous methylation or gene expression data measured in microarray, CARseq (Jin et al., 2021) is designed for cell type-specific inference for count data from RNA-sequencing by using a negative binomial (NB) distribution. Different from previous mentioned methods that require known cell type composition as input, HIRE (Luo et al., 2019) jointly perform composition estimation and csDM inference. Even though these methods generally achieve satisfactory performance in detecting differential signals from abundant cell types, their accuracy and power could be low, especially in cell types with small proportions. Using the existing methods, the only way to improve the results for those minor cell types is to increase sample size, which could be infeasible in many settings.

It is known that different cell types in a tissue form a hierarchical structure (Smith and Hodges, 2019; Wu and Wu, 2020). For example, the major groups of lymphocytes include natural killer cells (NK), T cells, and B cells. The T cells can be further divided into many subtypes including CD4+ T cells (CD4) and CD8+ T cells (CD8). Due to the similarity among cell types, it is conceivable that similar cell types could exhibit similar DE or DM patterns, e.g., if a gene is DE in CD4, it is more likely to be also DE in CD8. Correlations of DE/DM states among cell types have been reported in many published works. Mathys et al. (2019) reported that in the late stage of AD, genes up-regulated were common across cell types and primarily involved in global stress response. Tserel et al. (2015) reported that age-related methylation changes (measured by fold change) in CD4 and CD8 have a strong correlation and that all top

sites with the highest methylation differences between younger and older individuals are shared by the two cell types. In a Graves' disease (GD) study, Limbach et al. (2016) reported that a majority of the most significant CpG sites associated with GD had differential methylation in both CD4 and CD8. Conceptually, the similarity of DE/DM status among cell types can be exploited to improve the csDE/csDM results. In this work, we develop a novel and rigorous statistical method to incorporate the cell type hierarchy into the cell type-specific differential analysis in high-throughput bulk omics data. Our proposed method borrows information across cell types through a Bayesian hierarchical model. A key intuition of the proposed method is that the prior probability of one gene being DE in a cell type is impacted by the DE status of this gene in other cell types, for example, if gene A shows strong DE in CD4, its prior probability of being DE in CD8 will be higher due to the similarity between CD4 and CD8. We name the proposed method "Cell type-specific Differential Analysis with tRee" (CeDAR) and implement it in Bioconductor package TOAST (`https://www.bioconductor.org/packages/release/bioc/html/TOAST.html`). We comprehensively evaluate the proposed method with both simulated and real data. The results demonstrate that incorporating the cell type hierarchy in the csDE/csDM framework greatly improves the detection performance, especially in cell types with low proportions.

2.2 Methods

2.2.1 Methods overview

CeDAR incorporates the cell type hierarchy in cell type-specific differential analysis in bulk data. Briefly speaking, for each feature, we define binary random variables to represent its underlying DE/DM states in all cell types, each with a prior probability. Given a realization of the DE/DM states for all cell types, we model the observed

bulk data using a linear model framework similar to TOAST and CellDMC, in which the interaction terms between the cell type proportions and the covariate of interest capture the cell type-specific effects. The unique feature in CeDAR distinguishing it from the existing methods is that the interaction terms are only included for cell types deemed DE/DM. In contrast, TOAST/CellDMC is the full model which implicitly assumes the feature is DE/DM in all cell types, since the interactions are included for all cell types. The marginal likelihood of the observed data can be calculated by summing over all the underlying DE/DM states. Then the posterior probability of a feature being DE/DM in each cell type given observed data can be calculated and used to detect csDE/csDM.

The most important part of the proposed method is the specification of the prior probabilities for the DE/DM status for each cell type. If one only considers the marginal probabilities of DE/DM and assumes independence among cell types, the similarities among cell types cannot be incorporated. In order to take advantage of the correlations among cell types, we make the prior probabilities dependent on the cell type hierarchy. Given a hierarchical tree of cell types, we assign priors for the root and all internal nodes, then compute the priors for the leaf nodes based on the cell type hierarchy. The specification of the prior is graphically illustrated by a toy example in Figure 2.1. Assuming there are three cell types forming a simple tree with one root node, one internal node, and three leaf nodes. All nodes have underlying binary states of being DE/DM (state 1) or not (state 0). Here we define a non-leaf node as DE/DM if any of its direct children's node is DE/DM. Conversely, a child node can be DE/DM only when its direct parent node is DE/DM. The prior probabilities on the non-leaf nodes will implicitly account for the correlations among cell types. For example, even though the marginal probabilities of DE/DM for cell types 2 and 3 are small (0.06, 0.04), their conditional probabilities when the parent node is in state 1 become very high (0.75, 0.5). If a gene shows strong DE in cell

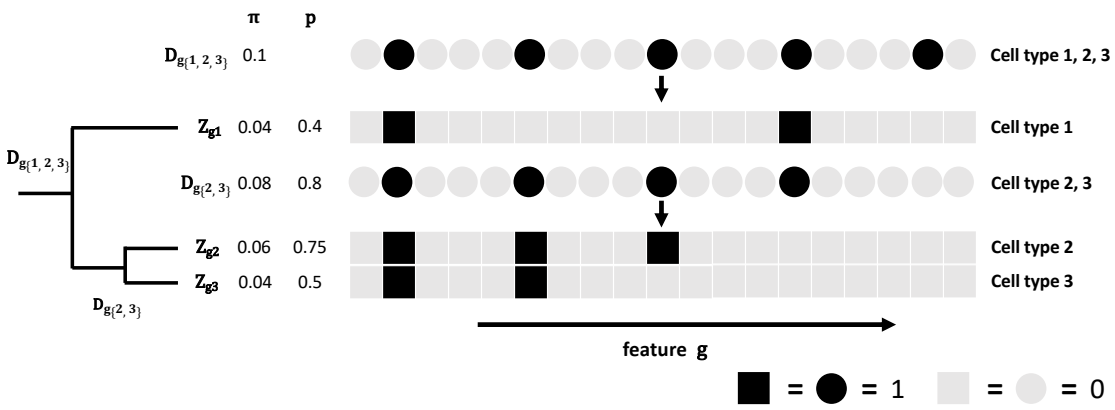

Figure 2.1: Illustration of the specification of the prior probabilities for DE/DM under a cell type hierarchy. The cell type hierarchy is represented by three cell types and a few features (genes or CpG sites). The three cell types form a simple tree (shown in the left). In the array of squares and circles, each column represents a feature. Circles represent root or internal nodes, and the squares represent leaf nodes. Colors represent the differential states of the node (black: 1; gray: 0). The root node $D_{g\{1,2,3\}}$, internal node $D_{g\{2,3\}}$, and leaf nodes Z_{g1}, Z_{g2} and Z_{g3} are binary random variables representing the g-th feature differential states. π represents the marginal probability for a node to be in state 1. p represents the conditional probability of a node to be in state 1 when its parent node is in state 1.

type 2, it will increase the probability for its parent node (an internal node) to be DE, which subsequently increases the prior probability for this gene to be also DE in cell type 3. Thus, the correlation between cell types 2 and 3 is passed through their parent node. On the other hand, the distance between cell types 1 and 3 is larger, so their influences on each other must pass through the root and internal nodes, which is weaker. It is important to mention that the proposed method allows the cell type hierarchy to be any rooted tree, i.e., it does not have to be a bifurcating hierarchical tree.

2.2.2 The CeDAR method

Data model

Suppose the data was generated from a bulk high-throughput experiment, which contains measurement of G features (genes, CpG sites, etc.) in N samples. Let Y_{gi} represent the observed measurement of g-th feature in i-th sample. In each sample, the measurement of each feature is a mixed signal from K different cell types. Let $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{iK})^T$ represent the cell composition of the i-th sample. There are several methods for estimating K and $\boldsymbol{\theta}_i$ in both DNA methylation and gene expression data (Newman et al., 2015; Li and Wu, 2019; Li et al., 2020b). Here we assume both K and $\boldsymbol{\theta}_i$ are known. We assume there are Q confounders to be adjusted in the study. Let $\boldsymbol{C}_i = (C_{i1}, ..., C_{iQ})^T$ represent the confounders of i-th sample. Then $\boldsymbol{C} = (\boldsymbol{C}_1, ..., \boldsymbol{C}_N)_{Q\times N}$ represents the confounders of all samples. Let $\boldsymbol{A} = (\boldsymbol{A}_1, ..., \boldsymbol{A}_N)$ represent the factor to be tested for cell type-specific effects across all N samples. $\boldsymbol{A}_i$ is a scaler if a single continuous or binary factor is involved; if the factor is a categorical variable with multiple levels, it will be coded as a vector of dummy variables.

Now consider the csDE/csDM status for a particular covariate of interest. For the simplicity of notation, we will omit the subscript for covariate. The model described below will be applied to all covariates of interest. Define Z_{gk} as a binary random variable to represent the DE/DM state of the g-th feature in k-th cell type. When $Z_{gk} = 1$, the g-th feature in k-th cell type is DE/DM associated with the factor of interest, and $Z_{gk} = 0$ otherwise. Note that since $\boldsymbol{Z}_g = (Z_{g1}, ..., Z_{gK})$ takes value in discrete space $\{0, 1\}^K$, there are 2^K combinations of DE/DM states for K cell types. Let X_{gik} represent the unknown pure profile of g-th feature in k-th cell type for i-th sample. We assume that given all DE/DM state of g-th feature in k-th cell type, it satisfies $E[X_{gik}|Z_{gk}] = \mu_{gk} + \boldsymbol{C}_i^T\boldsymbol{\beta}_{gk} + Z_{gk}\boldsymbol{A}_i^T\boldsymbol{\delta}_{gk}$. Here μ_{gk} is the baseline profile of g-th feature for k-th cell type; $\boldsymbol{\beta}_{gk} = (\beta_{gk1}, ..., \beta_{gkQ})^T$ are coefficients associated

with confounders, and $\boldsymbol{\delta}_{gk}$ are coefficients associated with the factors of interest. Specifically, for any confounder without cell type-specific effect (C_q), its corresponding coefficients in different cell types are the same ($\beta_{g1q} = \beta_{g2q} = ... = \beta_{gKq}$). It is important to note here that the factors of interest only impact on X_{gik} when $Z_{gk} = 1$ (the cell type is DE/DM). This is a major modeling difference from all other linear model-based cell type-specific methods (TOAST, CellDMC, TCA, etc.), which would always include the impact of $\boldsymbol{A}$. For the observed bulk data, since they are mixtures of cell type-specific signals, the observed measurement Y_{gi} is a weighted average of X_{gik}'s: $E[Y_{gi}; \boldsymbol{\theta}_i] = \sum_k \theta_{ik} E[X_{gik}]$. Thus, given the DE/DM state in K cell types $\boldsymbol{Z}_g$, Y_{gi} satisfies the following linear form:

$$E[Y_{gi}|\boldsymbol{Z}_g] = \sum_{k=1}^{K} \theta_{ik} \left(\mu_{gk} + \boldsymbol{C}_i^T \boldsymbol{\beta}_{gk} + Z_{gk} \boldsymbol{A}_i^T \boldsymbol{\delta}_{gk}\right) \tag{2.1}$$

Since the interactions between mixing proportion and factor of interest are only allowed for cell types showing DE/DM state (e.g., cell type k with $Z_{gk} = 1$), the linear model used in existing methods such as TOAST and CellDMC is a special case in which all cell types are assumed to be DE/DM a priori (the full model).

Given the data model, we can obtain the observed data likelihood and derive the posterior probability for DE/DM calling. Denote $\boldsymbol{Y}_g = (Y_{g1}, ..., Y_{gN})$, the goal of csDE/csDM calling is to compute $P(Z_{gk} = 1|\boldsymbol{Y}_g)$. This posterior probability relies on the prior. In the next subsection, we provide a detailed explanation on how to construct priors based on cell type hierarchy to achieve information sharing.

Prior probabilities for the DE/DM states

As discussed before, a major methodological contribution of this work is the specification of csDE/csDM priors based on the cell type hierarchy. This plays a major role in capturing the similarity among cell types and improving the DE/DM calling

result. For each feature, we define a list of binary random variables for the underlying DE/DM states for all nodes: $\boldsymbol{Z}$ for leaf nodes and $\boldsymbol{D}$ for non-leaf nodes. We assume these binary random variables are independent and identically distributed for all genes. We further assume that the cell type hierarchy is known at this step. The estimation of cell type hierarchy will be discussed in the later section.

The correlation in the hidden DE/DM states among cell types is captured by the joint probability of $\boldsymbol{Z}_g$ and $\boldsymbol{D}_g$. For $g = 1, .., G$, and $k = 1, ..., K$, the DE/DM state of the leaf nodes is represented by binary random variables $Z_{g1}, ..., Z_{gK}$, with $Z_{gk} \sim Bernoulli(\pi_k)$. $Z_{gk} = 1$ means that the g-th feature in k-th cell type is DE/DM, and $Z_{gk} = 0$ otherwise. The states of all non-leaf nodes are also represented by binary random variables. Given a hierarchical tree of the cell types, the state for the n-th node at l-th level ($l = 1, ..., L; n = 1, ..., n_l$) of the tree is denoted by binary random variable $D_{g\Phi_{l,n}}$, where $\Phi_{l,n}$ is a set of cell types represented by corresponding descendant leaf nodes. Specifically, the root node is defined as the first node at level 0, denoted as $D_{g\Phi_{0,1}}$. We assume $D_{g\Phi_{l,n}} \sim Bernoulli(\pi_{\Phi_{l,n}})$. To capture the tree structure, we define that for any non-root node (internal or leaf): if its parent node has state 0, it must have state 0; if the parent node has state 1, its state follows a Bernoulli distribution. Thus, the conditional distribution for the states of the leaf nodes can be expressed as the following, where $D_{g\Phi_{l,n}}$ is the parent node of Z_{gk}:

$$Z_{gk}|D_{g\Phi_{l,n}} \sim Bernoulli(p_k D_{g\Phi_{l,n}}) \tag{2.2}$$

Here, $p_k = \frac{\pi_k}{\pi_{\Phi_{l,n}}}$. Distributions for the non-leaf internal nodes can be expressed in a similar form, that is, the state of a child internal node condition on the state of its parent follows a Bernoulli distribution. Finally, we assume that the sibling nodes are mutually independent if their parent node has state 1.

The specification of the prior probabilities captures the similarity among cell types

according to the cell type hierarchy. Using the structure in Figure 2.1 as an example, there are three leaf nodes with underlying states represented by Z_{g1}, Z_{g2}, Z_{g3}, and two non-leaf nodes represented by $D_{g\{1,2,3\}}, D_{g\{2,3\}}$. The marginal prior probabilities of a randomly picked feature being DE/DM in cell type 2 and 3 are $P(Z_{g2} = 1) = P(Z_{g2} = 1|D_{g\{2,3\}} = 1) \times P(D_{g\{2,3\}} = 1|D_{g\{1,2,3\}} = 1) \times P(D_{g\{1,2,3\}} = 1) = p_2 \times p_{\{2,3\}} \times \pi_{\{1,2,3\}} = 0.06$ and $P(Z_{g3} = 1) = P(Z_{g3} = 1|D_{g\{2,3\}} = 1) \times P(D_{g\{2,3\}} = 1|D_{g\{1,2,3\}} = 1) \times P(D_{g\{1,2,3\}} = 1) = p_3 \times p_{\{2,3\}} \times \pi_{\{1,2,3\}} = 0.04$, respectively. The marginal joint probability of a randomly picked feature being DE/DM in both cell type 2 and cell type 3 is $P(Z_{g2} = Z_{g3} = 1) = p_2 \times p_3 \times p_{\{2,3\}} \times \pi_{\{1,2,3\}} = 0.03$. It is much larger than $P(Z_{g2} = 1) \times P(Z_{g3} = 1) = 0.0024$, which is the probability assuming cell types 2 and 3 are independent. If the root node always has state 1, i.e., $P(D_{g\{1,2,3\}}) = 1$, then cell type 1 will be independent of cell type 2 and 3. Furthermore, if $P(D_{g\{1,2,3\}} = 1) = P(D_{g\{2,3\}} = 1) = 1$, then the three cell types are mutually independent. Importantly, such cell type hierarchy is used merely as a statistical way to capture DE/DM state correlations among cell types. It does not necessarily represent the cell type lineage tree during differentiation or development.

We use $Parent()$ to represent the parent node of a specific node. Then, a prior joint probability of $\boldsymbol{Z}_g = (Z_{g1}, ..., Z_{gK})$ and $\boldsymbol{D}_g = (D_{g\Phi_{0,1}}, ..., D_{g\Phi_{L,n_L}})$ has the following form:

$$
\begin{aligned}
&P(\boldsymbol{Z}_g, \boldsymbol{D}_g) = P(\boldsymbol{Z}_g|\boldsymbol{D}_g) \times P(\boldsymbol{D}_g) \qquad (2.3)\\
&= \left[\prod_{k=1}^{K} P\Big(Z_{gk}|Parent(Z_{gk})\Big)\right] \times \left[\prod_{l=1}^{L}\prod_{n=1}^{n_l} P\Big(D_{g\Phi_{l,n}}|Parent(D_{g\Phi_{l,n}})\Big)\right] \times P(D_{g\Phi_{0,1}})\\
&= \left(\prod_{k=1}^{K}\Big\{[p_k Parent(Z_{gk})]^{Z_{gk}}[1 - p_k Parent(Z_{gk})]^{1-Z_{gk}}\Big\}\right)\\
&\quad \times \left(\prod_{l=1}^{L}\prod_{n=1}^{n_l}\Big\{\Big[p_{\Phi_{l,n}} Parent(D_{g\Phi_{l,n}})\Big]^{D_{g\Phi_{l,n}}}\Big[1 - p_{\Phi_{l,n}} Parent(D_{g\Phi_{l,n}})\Big]^{1-D_{g\Phi_{l,n}}}\Big\}\right)\\
&\quad \times \left[\pi_{\Phi_{0,1}}^{D_{g\Phi_{0,1}}}(1 - \pi_{\Phi_{0,1}})^{1-D_{g\Phi_{0,1}}}\right]
\end{aligned}
$$

Likelihood and posterior probability

Given the data model and the prior probabilities, we are now in position to derive the posterior probability for DE/DM calling. Denote $\boldsymbol{Y}_g = (Y_{g1}, ..., Y_{gN})$, the probability of $\boldsymbol{Y}_g$ given $\boldsymbol{Z}_g$ is:

$$P(\boldsymbol{Y}_g|\boldsymbol{Z}_g) = \prod_{i=1}^{N} P(Y_{gi}|\boldsymbol{Z}_g) \tag{2.4}$$

The joint probability of $\boldsymbol{Y}_g$, $\boldsymbol{Z}_g$, $\boldsymbol{D}_g$ can be derived as the following, noting that $P(\boldsymbol{Y}_g|\boldsymbol{Z}_g, \boldsymbol{D}_g) = P(\boldsymbol{Y}_g|\boldsymbol{Z}_g)$:

$$P(\boldsymbol{Y}_g, \boldsymbol{Z}_g, \boldsymbol{D}_g) = P(\boldsymbol{Y}_g|\boldsymbol{Z}_g) \times P(\boldsymbol{Z}_g, \boldsymbol{D}_g) = \left(\prod_{i=1}^{N} P(Y_{gi}|\boldsymbol{Z}_g)\right) \times P(\boldsymbol{Z}_g, \boldsymbol{D}_g) \tag{2.5}$$

Then, we can have the marginal probability for the observed data $P(\boldsymbol{Y}_g)$ by summing over all combinations of $(\boldsymbol{Z}_g, \boldsymbol{D}_g)$:

$$P(\boldsymbol{Y}_g) = \sum_{(\boldsymbol{Z}_g, \boldsymbol{D}_g)} P(\boldsymbol{Y}_g, \boldsymbol{Z}_g, \boldsymbol{D}_g) \tag{2.6}$$

Similarly, the joint probability of $Z_{gk} = 1$ and $\boldsymbol{Y}_g$ is:

$$P(\boldsymbol{Z}_{gk} = 1, \boldsymbol{Y}_g) = \sum_{(\boldsymbol{Z}_g, \boldsymbol{D}_g)} P(\boldsymbol{Y}_g, \boldsymbol{Z}_g, \boldsymbol{D}_g) \times I(Z_{gk} = 1) \tag{2.7}$$

Based on these, we have the posterior probability of $Z_{gk} = 1$ conditional on $\boldsymbol{Y}_g$ as:

$$P(Z_{gk} = 1|\boldsymbol{Y}_g) = \frac{\sum_{(\boldsymbol{Z}_g, \boldsymbol{D}_g)} P(\boldsymbol{Y}_g, \boldsymbol{Z}_g, \boldsymbol{D}_g) \times I(Z_{gk} = 1)}{\sum_{(\boldsymbol{Z}_g, \boldsymbol{D}_g)} P(\boldsymbol{Y}_g, \boldsymbol{Z}_g, \boldsymbol{D}_g)} \tag{2.8}$$

The joint prior $P(\boldsymbol{Z}_g, \boldsymbol{D}_g)$ derived from Equation (2.3) can be plugged into Equation (2.5) to obtain $P(\boldsymbol{Y}_g, \boldsymbol{Z}_g, \boldsymbol{D}_g)$, and then the posterior probabilities can be calculated for csDE/csDM calling. For all above, we have not made any distribution assumption

on the data. For microarray data, we use normal distributions for the observed data. The same principles apply for other data types with different distribution assumptions.

2.2.3 Parameter estimation

To derive the posterior probability of Z_{gk}, which is shown in Equation (2.8), we need to estimate the cell type hierarchy capturing cells correlation in DE/DM state, the prior probabilities of all nodes in the tree, and the marginal likelihood given different combinations of DE/DM states.

Estimation of the cell type hierarchy

The tree structure describing cell type hierarchy could be estimated by hierarchical clustering of cell types, in which the similarity between cell types is defined based on the Pearson correlation of p-values with the following form:

$$similarity(k, k^{'}) = \frac{1}{2}\left[1 - cor\left(-log10(pval_k), -log10(pval_{k'})\right)\right] \tag{2.9}$$

$\boldsymbol{pval_k}$ are p-values generated by TOAST for testing differential signal in k-th cell type of features satisfying {feature g: for $1 \leq g \leq G$, $\exists k \in \{1, .., K\}$ s.t. $pval_{gk}$ (or fdr_{gk}) $< threshold$}. This step is designed to reduce noise signal from non-DE/non-DM features. The threshold could be arbitrarily defined by users. Users could even define their own rule to select features for estimating the tree structure. Cell types with higher correlations should be more similar.

We want to emphasize that the cell type hierarchy does not have to be a bifurcating tree. In our software implementation, a bifurcating tree will be estimated from the data by default, but users have the option to specify a tree structure according to their prior biological knowledge. In addition, we also have option for using a simplified cell

type hierarchy, in which all cell types are assumed to be independent under the root node. We call this the "single-layer" model, where the correlations among cell types are only captured at the root level.

Estimation of the prior probabilities

Based on the p-values provided by TOAST, the prior probability for an internal node $\boldsymbol{D}_{g\Phi_{l,n}}$ to be DE/DM ($\pi_{\Phi_{l,n}}$) is estimated as the proportion of features deemed significant in any cell type belonging to set $\Phi_{l,n}$ among all G features.

$$\hat{\pi}_{\Phi_{l,n}} = \frac{\sum_{g=1}^{G} I\left(\min_{k\in\Phi_{l,n}}\{pval_{gk}\} < threshold\right)}{G} \tag{2.10}$$

Then the conditional probability of non-root internal node $D_{g\Phi_{l,n}}$ conditional on its parent node $D_{g\Phi_{l',n'}}$ equals to one ($p_{\Phi_{l,n}}$) is simply estimated by plugging in corresponding estimates of marginal probabilities:

$$\hat{p}_{\Phi_{l,n}} = \frac{\hat{\pi}_{\Phi_{l,n}}}{\hat{\pi}_{\Phi_{l',n'}}} \tag{2.11}$$

Prior probabilities of leaf node Z_{gk} can be estimated in a same way, since we can treat it like an internal node, whose set only contains a single cell type k:

$$\hat{\pi}_k = \frac{\sum_{g=1}^{G} I(pval_{gk} < threshold)}{G} \tag{2.12}$$

$$\hat{p}_k = \frac{\hat{\pi}_k}{\hat{\pi}_{\Phi_{l',n'}}} \tag{2.13}$$

Computation of data likelihood

For K cell types, $\boldsymbol{Z}_g$ has 2^K possible combinations. So, totally there are 2^K different linear models to fit. Under each combination of $\boldsymbol{Z}_g$, μ_{gk}, β_{gk}, and δ_{gk} (for $k = 1, ..., K$) are estimated by least square estimators of corresponding linear model in Equation (2.1). By assuming the observed bulk signal follows a normal distribution,

posterior probability of Z_{gk} in Equation (2.8) can be computed by plugging in the least square estimates. In this work, computation of data likelihood is based on normal distribution assumption, which is often used for microarray data. Specifically, for DNA methylation data, we used beta value for analysis. Even though the beta values for all CpG sites follow a bimodal distribution at around 0 and 1, they can be well approximated by normal distributions for one CpG site cross samples (Zheng et al., 2018a; Rahmani et al., 2019). The same framework could be extended to count data by assuming a negative binomial distribution, which would be a future research direction.

Differential signal detection

A feature would be reported showing differential signal in certain cell type if its corresponding posterior probability of DE/DM shown in Equation (2.8) is greater than a user-defined threshold. Higher posterior probability of DE/DM suggests more convincible cell type-specific DE/DM. Besides, the estimated posterior probability of non-DE/non-DM can be viewed as estimated local FDR. The global FDR for a list of features can be derived by averaging their estimated local FDRs.

2.2.4 Simulation

Data simulation

We first estimated cell type-specific mean μ_{gk} and variance σ^2_{gk} for gene $g = 1, ..., G$ ($G = 12,402$) in cell type $k = 1, ..., K$ ($K = 6$) (Neutrophils, Monocytes, CD8, CD4, B cells, and NK cells) from log expression values of microarray gene expression data GSE22886 (Abbas et al., 2005). We defined 10% DE genes between case and control groups in each cell type. Each DE gene has equal probability to be up or down regulated. To maintain the cell type hierarchy, the DE states of genes were generated based on a pre-defined tree structure in Figure 2.3(a). The prior probability of each

node on the tree is $\pi_{\{1,2,3,4,5,6\}} = 0.4$, $p_{\{1,2\}} = 0.3125$, $p_1 = p_2 = 0.8$, $p_{\{3,4,5,6\}} = 0.5$, $p_{\{3,4,5\}} = 0.8$, $p_6 = 0.5$, $p_{\{3,4\}} = 0.78125$, $p_5 = 0.625$, $p_3 = p_4 = 0.8$. For root node, among $G = 12,402$ genes, we used Bernoulli distribution with $\pi_{\{1,2,3,4,5,6\}} = 0.4$ to generate DE state for each feature. Then for one of its child nodes containing cell types 1 and 2, among features with generated potential DE state 1, we used Bernoulli distribution with $p_{\{1,2\}} = 0.3125$ to generate DE state. In this way, we can derive DE state of each cell type (each leaf node) and make sure they share different correlation strengths between cell types. For any non-DE gene g in case and control groups, its expression in k-th cell type of i-th sample, denoted by X_{gik}, follows a log-normal distribution $\log X_{gik} \sim N(\mu_{gk}, \sigma^2_{gk})$. For any DE gene g in k-th cell type of i-th sample in the case group, the pure expression follows a log-normal distribution $\log X_{gik} \sim N(\mu_{gk} + lfc_{gk}, \sigma^2_{gk})$ where lfc is the log2 fold change. For up-regulated genes, the log2 fold change (lfc_{gk}) is randomly drawn from normal distribution $N(1, 0.2^2)$, while for down-regulated genes, it is from $N(-1, 0.2^2)$.

In the simulations setting with six cell types, the mixture proportion of each sample i, $\boldsymbol{\theta}_i$, was generated from a Dirichlet distribution with parameters estimated from the real cell type proportion of six cell types (Neutrophils, Monocytes, CD8, CD4, B cell, and NK cell) (Newman et al., 2019): 27.94, 4.64, 2.47, 4.87, 2.30, 2.21. In the simulation setting for evaluating the impact of different cell type hierarchy, the four cell types selected were Neutrophils, Monocytes, CD8, and CD4, and the corresponding Dirichlet parameter was 27.94, 4.64, 2.47, and 9.38. We assumed there is no cell type proportion difference between the case and control groups.

Finally, we simulated s cases and s controls (s = 50, 100, 200 for different simulations). The simulated measurement for g-th gene of i-th sample, Y_{gi}, is a linear combination of simulated cell type-specific expression $\boldsymbol{X}_{gi} = (X_{gi1}, ..., X_{giK})$ weighted by the mixture proportion $\boldsymbol{\theta}_i$ and added by a random noise ϵ_{gi}: $Y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\theta}^T + \epsilon_{gi}$. We assumed the random noises are mutually independent for each gene and each sample.

To reflect the mean-variance dependence of gene expression, we assumed the variance of the random noise is positively correlated with gene expression: $\epsilon_{gi} \sim N(0, \eta_g^2)$, where $\eta_g = 0.1 \times \max\left(\sum_{i:control} \frac{\boldsymbol{X}_{gi}\boldsymbol{\theta}_i^T}{s}, \sum_{i:case} \frac{\boldsymbol{X}_{gi}\boldsymbol{\theta}_i^T}{s}\right)$.

Cell type proportion estimation

In the second simulation results section, we evaluated robustness of CeDAR to estimated proportions. We estimated the cell type proportion for each sample from the mixture profiles by using reference-based method *lsfit* from the R package *CellMix* (Gaujoux and Seoighe, 2013). The estimated cell type-specific mean from GSE22886, which was used for generating pure cell type expression, was used as a reference profile. Reported marker genes for the six blood cell types (Newman et al., 2015) were used for deconvolution. Proportions of samples in cases and controls were estimated separately.

Evaluation of CeDAR method

After deriving the simulated bulk data and corresponding proportion, we compared CeDAR methods with TOAST and TCA. We used ROC to evaluate the accuracy of proposed method and calculated observed FDR at a given cutoff to evaluate type I error control. For the detail of evaluation method used in simulation, please see Section A.1 in Appendix A.

2.2.5 Real data analysis

In this work, we first explored real data to check whether DE state correlation exists between cell types. Then, we compared our designed model CeDAR with previous developed methods by applying them on real data for cell type specific differential analysis.

Cell type correlation calculation from real data

We obtained two data sets from the GEO database. The first data set (GEO accession number GSE166844 (Hannon et al., 2021)) measures DNA methylation profile on Infinium MethylationEPIC microarray for several purified blood cell types, including CD4, CD8, B cells, Monocytes, and Granulocytes, from 30 individuals (18 females vs. 12 males). The second dataset (GSE60424 (Linsley et al., 2014)) provides gene expression from RNA-seq for six immune cell types (CD4, CD8, B cells, NK cells, Monocytes, and Neutrophils) of sclerosis patients before and 24 hours after the first treatment with IFN-beta. In the DNA methylation data (GSE166844), sites with detection p-value greater than or equal to 0.01 in any sample were removed from the processed data set provided on GEO website. We used *minfi* (Aryee et al., 2014; Andrews et al., 2016; Maksimovic et al., 2012; Fortin et al., 2017, 2014; Fortin and Hansen, 2015; Triche Jr et al., 2013; Jaffe et al., 2012) to call DM for male vs. female comparison. CpG sites with q-value less than 0.05 are deemed differentially methylated sites. For the gene expression data, we used *edgeR* (Robinson et al., 2010; McCarthy et al., 2012; Chen et al., 2016) to call DE for before vs. after first IFN-beta treatment. DE genes are defined as genes with false discovery rate (FDR) less than 0.05.

For both data sets, Pearson correlation coefficient depicting cell type correlation in DE/DM state was calculated based on negative log-transformed (base 10) p-values of two cell types and a t-test was applied to test whether the correlation estimate is statistically significant different from zero. Odds ratio of DE/DM in two cell types was calculated based on DMC defined above. Each count of the 2×2 contingency table was added one to avoid infinite OR value. Fisher's exact test was used to test whether the estimated odds ratio is statistically significantly different from one.

Cell type specific differential analysis on real data

We downloaded three DNA methylation data sets (GSE41826 (Guintivano et al., 2013), GSE166844, GSE42861 (Liu et al., 2013)) from GEO database. The methylation level is measured with beta value. R package *minfi* was used to pre-process raw data and call gold standard csDMCs. For data sets with pure cell type samples, we defined gold standard of cell type-specific DM state by setting sites with FDR smaller than 0.01 as true DM, with FDR greater than 0.8 as non-DM. For detecting cell type-specific effects in bulk data, we first used *EpiDISH* (Zheng et al., 2018a; Newman et al., 2015; Teschendorff et al., 2017; Zheng et al., 2018b; Teschendorff and Zheng, 2017; Houseman et al., 2012) to estimate cell type compositions. The DNA methylation reference is mean profile of each cell type for GSE41826 and GSE166844; for GSE42861, which does not have pure cell type samples, DNAm reference consists of 333 immune cell type-specific DMCs (Teschendorff et al., 2017; Zheng et al., 2018b). More details are provided in Appendix A Section A.2, A.3, and A.4.

2.3 Results

2.3.1 Strong correlations of DE/DM states among cell types are observed in real data

We performed real data analyses to explore whether the DE/DM states are correlated among cell types in real data. We obtained two data sets from Gene Expression Omnibus (GEO) database, one DNA methylation (Hannon et al., 2021) and one gene expression (Linsley et al., 2014). Both data sets contain samples of purified cells from individuals under different conditions; thus, the gold standard results are available. We first called DM and DE for each cell type in these two data sets using existing tools. We called DM between males and females in the DNA methylation

data and called DE for sclerosis patients before versus after first IFN-beta treatment. Detailed description for the data and analysis procedures is in the "Methods" Section 2.2.5. Then, we evaluated the pairwise correlation among cell types in terms of their DE/DM status, using both Pearson correlation coefficient (PCC) of log-transformed p-values from the DE/DM tests for all features, and the odds ratio (OR) of being DE/DM from the cell types. The first metric (PCC) evaluates the correlations at the quantitative level that consider the DE/DM strength, while the second metric (OR) evaluates the correlation at the qualitative level since it quantifies the concordance of the binary DE/DM status. Higher PCC and OR indicate stronger correlation among cell types.

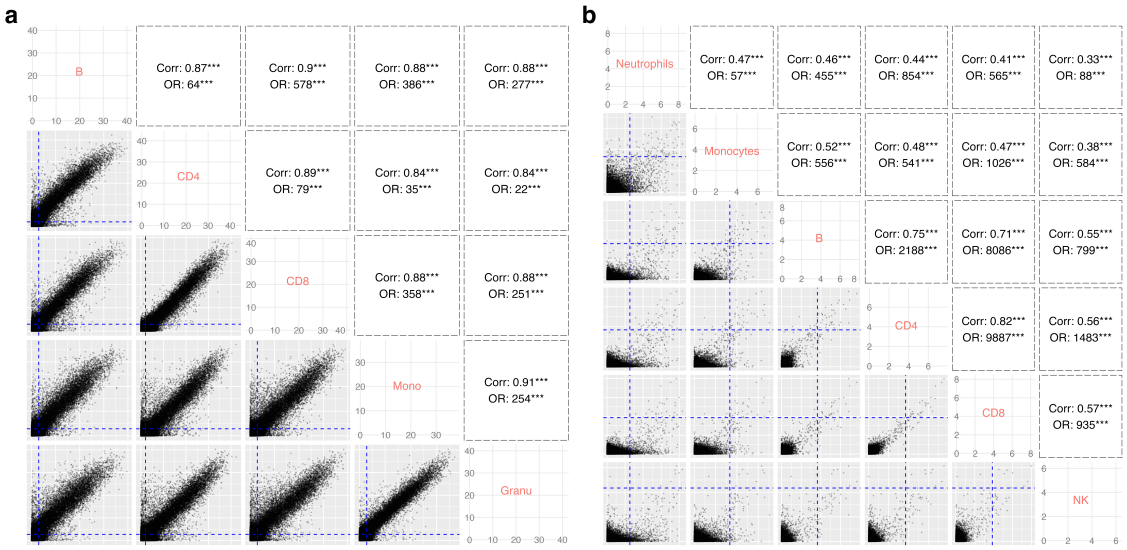

Figure 2.2: Correlations among cell types from cell type-specific differential analysis. (a) Cell type-specific differential methylation analysis and (b) cell type-specific differential expression analysis. DE/DM tests were applied for each feature in each cell type. X-axis and Y-axis represent -log10 transformed p-value from DE/DM tests in corresponding cell types. Each point represents a gene or CpG site. Dashed blue lines represent the thresholds used to define DEG/DMC in each cell type. Pearson correlation coefficients (PCC) of transformed p-values and odds ratio (OR) of differential state are tested for their significance. $***$ represents p-value < 0.01.

The pairwise scatter plots for the comparisons are shown in Figure 2.2. In the DNA methylation data (Figure 2.2(a)), the p-values from all cell types are highly

correlated (all PCCs > 0.83). Besides, the ORs for being DM between any two cell types are all very large. These results indicate very strong correlation among cell types in their methylation difference between males and females. In gene expression data (Figure 2.2(b)), all PCCs are also significantly positive and all ORs are significantly greater than 1. The correlation strength appears to be weaker in the gene expression example than in the methylation data since the molecular differences between sexes (as considered in the methylation data) are likely to be much stronger than the treatment effects (as considered in the gene expression data). Additionally, the gene expression dataset shows different levels of correlation among cell types. For example, B cells, CD4, and CD8 are more correlated with each other compared to others (PCCs > 0.7), suggesting a cell type hierarchy. Similar results are observed by performing the same analyses on three additional real data sets (Section A.5 and Figure A.1 in Appendix A). Overall, these results demonstrate that there are strong correlations among cell types in terms of their DE/DM status.

2.3.2 Simulation results

CeDAR method improves accuracy in cell type-specific differential signal detection

We conducted simulation studies to compare the performance of CeDAR with TOAST, TCA, csSAM, and CellDMC in a two-group comparison. Although TCA was originally designed for bulk methylation data, the method is also applicable to gene expression data (Wang et al., 2021). We incorporated two types of tree structures in the CeDAR test: the first is the simplest tree structure with only one layer (referred to as "CeDAR-S"), where root node is the parent of all leaf nodes. The second is a bifurcating hierarchical tree with multiple layers (referred to as "CeDAR-M"). While CeDAR-M captures a more complex correlation structure among cell types, CeDAR-S avoids the potential negative impacts of the biases in the specified prior

tree structure.

The simulation is constructed based on a dataset (GSE22886 (Abbas et al., 2005)) from whole blood samples with six cell types: Neutrophils, Monocytes, CD4, CD8, NK, and B cells. We simulated gene expression for six cell types based on parameters estimated from the real data to ensure the simulated data has characteristics (pure profiles and cell type composition) matching the real data. Note that we conducted simulation based on gene expression microarray data, but the proposed method can also be applied to DNA methylation microarray data. We made the six cell types have different levels DE state correlation following a hierarchical tree (Figure 2.3(a)). To be specific, we simulated the strongest correlation between cell types 1 and 2 as well as between cell types 3 and 4, both having $\sim$80% DE genes overlapped. Cell types 5 and 6 are made to have slightly weaker correlations with cell type 3 with $\sim$62.5% and $\sim$50% overlapped DE genes, respectively. We simulated the weakest correlation between cell types 1/2 and cell types 3/4/5/6. Between any two of them, only about 12.5% DE genes in one cell type overlap with the other. We used the true proportion to conduct data analyses for the results presented in this subsection and will evaluate the impact of proportion estimation in later sections. The accuracy of detecting csDE genes was measured by ROC curve, the area under the ROC curve (AUC-ROC), area under the precision-recall curve (AUC-PR), and Matthews correlation coefficient (MCC). We also evaluated the type I error controls from different methods by examining their false discovery rates (FDR). All methods were evaluated at different sample sizes (50, 100, 200 per group). The results were summarized from fifty simulations. Detailed simulation procedure is in the "Methods" section 2.2.4.

The simulation result shows that by considering correlation of DE states among the cell types, both CeDAR methods improve the accuracy of csDE genes detection in all six cell types compared to the other methods (Figure 2.3(c) and Appendix A Table A.1). However, the amounts of improvement vary with respect to different

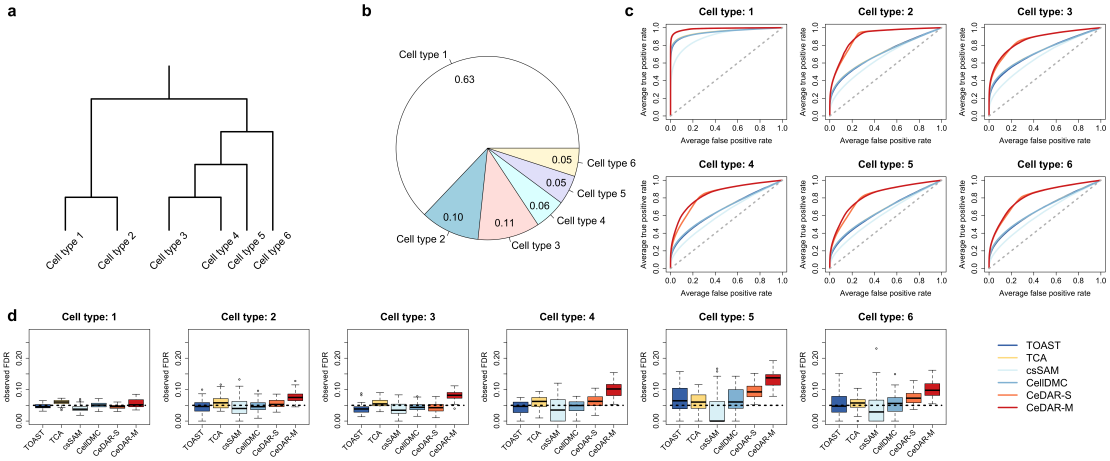

Figure 2.3: Simulation results for comparing different methods in cell type-specific differential expression. The simulation is based on a two-group comparison, with 100 samples in each group. Data were generated as a mixture of six common blood immune cell types (1: Neutrophils, 2: Monocytes, 3: CD4, 4: CD8, 5: B cells, 6: NK cells). (a) Cell type hierarchy used in simulation. (b) Mean proportion of each cell type. (c) ROC curves for csDE detection in six cell types for six methods (TOAST, TCA, csSAM, CellDMC, CeDAR-S, and CeDAR-M). Reported ROC curves are averaged from 50 simulations. (d) Observed FDR for csDE detection from different methods. DE genes are defined with rules: estimated FDR < 0.05 (TOAST, TCA, csSAM, and CellDMC); posterior probability of DE > 0.95 (CeDAR-S, CeDAR-M). Observed FDR from 50 simulations are summarized by box plot

factors, such as cell type proportion and sample size. The improvement in cell types with smaller proportions is greater than in cell types with larger proportions. For example, the improvement in cell type 1 (mean proportion 0.63) is much smaller than the other five cell types (largest mean proportion 0.11). Meanwhile, improvement in cell types with similar proportion could be different. Among the six cell types, cell type 2 and cell type 3 have similar mean proportion (0.10 vs. 0.11), but the accuracy improvement in cell type 2 is greater. A potential explanation is that cell type 2 is clustered with cell type 1 (with large proportion), while cell type 3 is clustered with cell types 4–6 (with smaller proportions). Intuitively, the cell type with small proportion could "borrow" more information from cell types with larger proportion, since larger proportion often leads to more accurate result.

Sample size is another important factor affecting the performance of various methods in detecting csDE genes, especially in cell types with small proportion (Zheng et al., 2018a; Li et al., 2019; Jin et al., 2021). When sample size is small (e.g., 50), both TOAST and TCA have poor performances in cell types of small proportions. However, the improvement of CeDAR methods is more significant compared to scenarios with larger sample size (Appendix A Table A.1). For example, in cell type 2, the AUC-ROC difference between CeDAR-S and TOAST is 0.145 when sample size is 200, while it is 0.235 when the sample size is 50. Additionally, when sample size becomes large (e.g., 200), CeDAR-M has higher AUC-ROC than CeDAR-S in cell types with smaller proportions, such as cell type 2 (AUC-ROC: 0.940 vs. 0.916). This is because larger sample size would lead to more accurate multiple layer tree structure estimation, which helps cell types with smaller proportions to correctly "borrow" information from their closely correlated cell types with larger proportions.

We also investigated the FDR control of the four methods at a given cutoff. While TOAST, TCA, csSAM, and CellDMC use estimated FDR (Benjamini and Hochberg, 1995) 0.05 as cutoff, CeDAR methods use posterior probability of DE 0.95 as cutoff (Leng et al., 2013). In general, all methods have better FDR control for cell types with larger proportions (Figure 2.3(d)). For example, the median of observed FDR in cell type 1 is much closer to 0.05 and the interquartile range (IQR) is much smaller than cell type 6 for all four methods. In cell types with smaller proportion, TOAST, TCA, csSAM, and CellDMC have slightly better performance in controlling type I error than CeDAR. This indicates that the information borrowing across rare cell types tends to mildly inflate the false positives. But overall, all methods do not work well for cell types with small proportions, and the only solution for that is to increase the sample size. Such problem will be alleviated with larger sample size. For example, the observed FDR in cell type 6 from CeDAR-M decreases from 0.247 to 0.065 when sample size increases from 50 to 200 (Appendix A Table A.1).

Evaluating the robustness of CeDAR

Robustness to different cell type correlation patterns Due to the complexity of biological system, cell types may show different correlation patterns in their DE/DM status under different conditions. For example, some cell types may not show correlation with each other at all. To evaluate the robustness of CeDAR, we evaluated its performance under different cell type hierarchies. To simplify the simulation but still capture the influences of cell type hierarchy, we simulated data for four cell types (Neutrophils, Monocytes, CD4, and CD8) with different mean proportions (0.6, 0.1, 0.25, 0.05). We evaluated CeDAR methods with six different cell type hierarchies representing various correlation relationships (Figure 2.4(a)-(f)). For hierarchies showing cell type correlation, we evaluated the performance of six methods under two different correlation levels (strong: $\sim$90% DE genes overlapped between two cell types; weak: $\sim$50%). Sample size was set as 200 per group.

The simulation results indicate that when all cell types are independent, CeDAR methods have similar accuracy as TOAST, TCA, and CellDMC, and greater accuracy than csSAM in all four cell types (Figure 2.4(a)). When cell types are strongly correlated, both CeDAR methods have greater improvements over the other methods in cell types with smaller proportions (e.g., cell type 2 in Figure 2.4(b), (d), (e); cell type 4 in Figure 2.4(b), (c), (e)). However, such improvement is not as significant in cell type 1 under all scenarios. This is because cell type 1 has large proportion (mean 0.63) so the data likelihood plays a greater role than prior information; thus, borrowing information from other cell types does not much impact on the result. Additionally, CeDAR-M provides greater performance improvement than CeDAR-S when the cell type hierarchy is more complex than a one-layer tree structure (e.g., cell type 2 in Figure 2.4(d), (e); cell type 4 in Figure 2.4(c), (e)). When correlation is weaker, CeDAR-M has similar performance as CeDAR-S, but the improvement over existing methods (TOAST, TCA, csSAM, and CellDMC) is smaller (Appendix

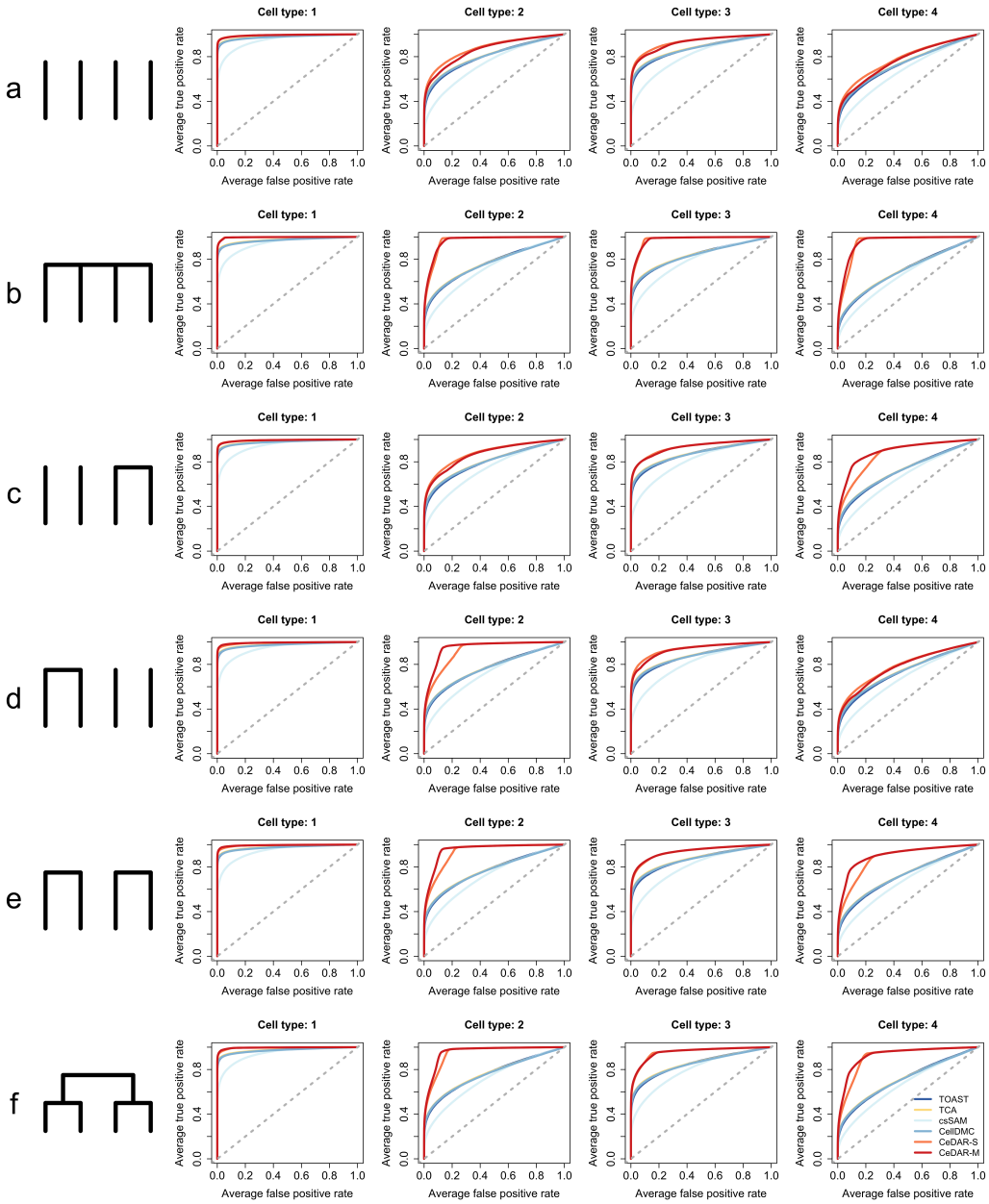

Figure 2.4: ROC curves under different DE patterns (with strong correlation). The simulation is conducted for a two-group comparison with four cell types (1: Neutrophils, 2: Monocytes, 3: CD4, 4: CD8) under six different DE patterns (**a** all cell types are independent; **b** cell types are correlated under the root, but independent conditional on the root (a single layer tree structure); **c** only cell types 3 and 4 are correlated; **d** only cell types 1 and 2 are correlated; **e** cell types 1 and 2 are correlated, and cell types 3 and 4 are correlated, but cell types 1/2 and 3/4 are independent; **f** all cell types are correlated under a multiple-layer tree structure). Methods under comparison include TOAST, TCA, csSAM, CellDMC, CeDAR-S, and CeDAR-M. Reported ROC curves are average over 50 simulations

A Figure A.3, Table A.3). The FDR control result is similar to the simulation result with six cell types in previous section regardless of different cell type hierarchies (Appendix A Figure A.2, Figure A.4, Table A.2, Table A.3).

Robustness to cell type hierarchy estimation In many cases, the cell type hierarchy and/or the prior probabilities of nodes are unknown and need to be estimated from data. We conducted additional simulations to evaluate the impacts of potential estimation biases on CeDAR. We used the same simulation setting as the first simulation result section (six cell types, 100 samples per group) and compared the performance of csDE detection with different combinations of inputs: true tree and true prior probability, true tree and estimated prior probability, estimated tree and estimated prior probability. The result shows that using estimated tree structure and prior probabilities of nodes have very similar accuracy as the other two types of inputs in most cases (Appendix A Figure A.5, Table A.4). The only exception is cell type 2, where the performance is slightly worse by using estimated tree and probability. On the other hand, the observed FDRs from using estimated prior probability as input are closer to the nominal value (0.05) than using true prior portability. Further investigation suggests that the difference in FDR between using true and estimated prior probabilities is associated with data noise. When data noise is large, CeDAR with estimated prior probability has smaller FDR; otherwise, it has larger FDR (Appendix A Table A.5, Table A.6). More details are provided in Appendix A Section A.6.

We further evaluated CeDAR's performance with mis-specified tree structures, which will happen when the tree estimation is inaccurate. We provided mis-specified tree structures to CeDAR and compared the results with CeDAR using the true tree. The results show that CeDAR is robust to mis-specified tree structures and that the major performance decreasing appears in low abundant cell types when they

are mistakenly clustered with other cell types. Detailed procedures and discussions are provided in Appendix A Section A.7 and Figure A.6, Figure A.7, Table A.7. Overall, CeDAR is very robust to potential biases brought by the cell type hierarchy estimation.

Robustness to cell type proportion estimation Although we assumed accurate proportion estimation in previous simulations, the estimation accuracy varies by the data quality and the choice of deconvolution methods. We evaluated the performance of the six methods under the same simulation scenario using estimated proportions from a reference-based (RB) deconvolution method *lsfit* (Abbas et al., 2009) (Appendix A Figure A.8, Table A.8). As expected, using true proportion leads to better results for all methods, especially in low abundant cell types (cell type 3-6). However, these results show that using the estimated proportions, CeDAR methods still have much higher accuracies than the other four methods in all cell types. Another observation is that the observed FDRs from all methods are inflated using estimated proportions. We took a deeper examination of the results and found that the estimated proportions are more variable across individuals compared to the true proportions. Such higher variability makes all methods more sensitive (since proportions are used in the linear model as covariates), but also produces more false positives. The obvious solution to this problem is to have better proportion estimation, or to use a more stringent cutoff in calling csDE/csDM. Overall, these results show that CeDAR still greatly outperforms other methods using estimated cell type proportions.

Computation performance

We benchmarked the computation performance of CeDAR and other methods under the simulation scenario in the first simulation result section (12,402 genes), but

varying the cell type number (4, 6, and 8) and sample size (50, 100, and 200). All simulations were performed on a PC running Linux with 2.80 GHz CPU and 8G RAM. TOAST is the fastest and CellDMC is the second fastest method. For example, they take 0.409 and 24.466 seconds respectively for 6 cell types and 100 samples on average. With default permutation number of 200, csSAM is slower than CeDAR-M with four cell types (sample sizes 50, 100, 200) and six cell types (sample sizes 50, 100), while it is faster with six cell types (sample sizes 200). TCA is the slowest in all scenarios. Overall, even though with K cell types, CeDAR needs to fit 2^K linear regression models, its computation performance is still very good due to efficient implementation. For example, it takes about 36.759 seconds for 6 cell types and 100 samples per group. Computation time for all scenarios is in Appendix A Table A.9.

2.3.3 Real data analysis

Cell type-specific differential methylation in brain

We first evaluated CeDAR on a human brain DNA methylation dataset (GEO accession number GSE41826 (Guintivano et al., 2013)) including both pure (glia and neuron) and bulk samples from 5 males and 5 females. The methylation level is represented as beta values in this study and all following DNA methylation analyses. We applied CeDAR-S, TOAST, TCA, csSAM, and CellDMC on the bulk data to call glia and neuron-specific differentially methylated CpGs (DMCs) comparing male vs. female and used the DMCs identified from the pure cell type as the gold standard to benchmark the results. The gold standard cell type-specific DMCs were detected using *minfi*. To obtain an accurate gold standard and avoid ambiguity in DM calling, we defined sites with FDR < 0.01 as DM and FDR > 0.8 as non-DM. Among all 480,492 CpGs, there were 8475 and 8587 true DM sites identified in glia and neuron respectively. The two cell types share 7622 common true DM sites, indicating a strong correlation between cell types. The true DM and non-DM sites are then used

to evaluate the csDM called from bulk samples. The estimated mixture proportions (by RB deconvolution) and the whole-tissue DNA methylation data were used as inputs for TOAST, TCA, csSAM, CellDMC, and CeDAR-S. Accuracy was measured by true discovery rate (TDR) in top ranked sites. The TDR curves in Figure 2.5 show that CeDAR-S has significantly higher accuracy among the top CpG sites than the other methods in both glia and neuron. For example, in glia, the difference of TDR between CeDAR-S and TOAST among top-ranked 5000 sites is more than 30%.

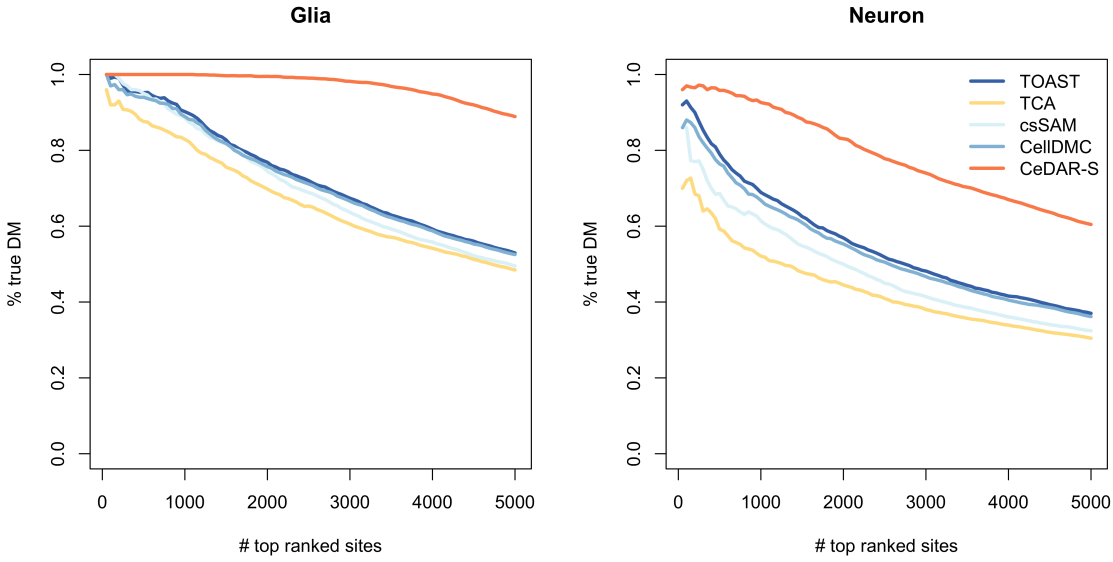

Figure 2.5: Accuracy of detecting csDM in human brain methylation data. The human brain DNA methylation dataset (GEO accession number: GSE41826) contains both bulk samples from postmortem frontal cortex and matched cell type samples of neuron and glia purified by fluorescence-activated cell sorting (FACS). The csDM sites associated with sex were identified between five healthy male and five healthy female samples with TOAST, TCA, csSAM, CellDMC, and CeDAR-S. The results are evaluated by the true discovery rate (TDR) curves, which show the accuracy among different numbers of top-ranked csDM sites from each method.

Cell type-specific differential methylation in whole blood

We further evaluated CeDAR on another set of human blood DNA methylation data (GEO accession number GSE166844 (Hannon et al., 2021)), which contains the profiles of five pure cell types (CD4, CD8, B, Monocytes, and Granulocytes) and the

whole blood samples from 30 individuals (18 females vs. 12 males). We performed cell type-specific differential methylation analyses in the bulk data for male-female comparison. Since there are more cell types in this dataset, we can create a hierarchical tree on the cell types, which allows us to compare CeDAR-M and CeDAR-S. We again defined the gold standard csDM using the pure cell type methylation between males and females by FDR < 0.01; non-DM by FDR > 0.8. There were 27,219 (CD4), 11,155 (CD8), 10,482 (B), 11,325 (Monocytes), and 13,938 (Granulocytes) DM sites identified. The number of overlapped true DM sites among cell types is shown in Appendix A Figure A.9. Again, there are significant overlaps of DMCs in different cell types. The TDR curves for top-ranked csDM sites detected from different methods are shown in Figure 2.6. Both CeDAR-M and CeDAR-S have higher accuracies among the top CpG sites than the other four methods in all five cell types. For Granulocytes (with the largest proportion), all methods have perfect accuracies in top 2000 ranked sites. However, the TDRs of TOAST, TCA, csSAM, and CellDMC in top 5000 sites drop to 90%, while the TDRs of two CeDAR methods are still close to 1, indicating a performance improvement. In cell types with relative smaller proportions (CD8, CD4, Monocytes, and B cells), all methods have worse performance, but CeDAR methods still have much higher TDR than the other methods and the performance improvement is even greater. Additionally, for Monocytes and B cells, CeDAR-M method has higher accuracy than CeDAR-S, since both have small proportions and are clustered together. This suggests that incorporating a detailed tree structure makes information sharing more efficient.

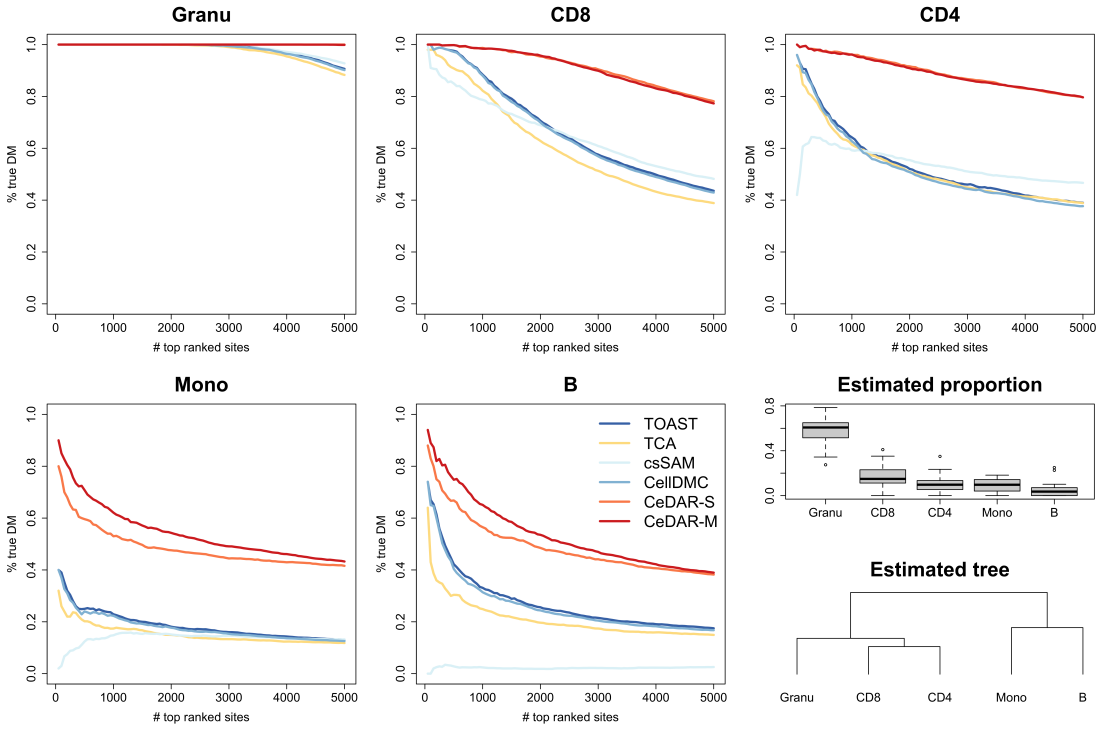

Figure 2.6: Accuracy of detecting csDM in human whole blood methylation data. The human blood DNA methylation dataset (GEO accession number: GSE166844) contains both bulk samples from whole blood and pure cell type samples of granulocytes, CD8, CD4, monocytes, and B cells derived by FACS. The csDM sites associated with sex were identified between eighteen females and twelve males samples using TOAST, TCA, csSAM, CellDMC, CeDAR-S, and CeDAR-M. The results are evaluated by TDR curves. The estimated proportions and estimated tree structure of cell types are shown in the last panel

Cell type-specific differential methylation in rheumatoid arthritis study

Previous two real data sets provide pure cell type data to serve as gold standard. However, the analyses were performed on a rather simple setting: detecting csDM between males and females without other covariates. To fully evaluate CeDAR performance in a more complex experimental design, we analyzed another dataset that provides peripheral blood leukocytes (PBL) DNA methylation from 332 normal individuals and 354 rheumatoid arthritis (RA) patients (GEO accession number GSE42861 (Liu et al., 2013)). After preprocessing, we performed cell type-specific analyses by com-

paring different disease statuses (RA vs. control), treating age as a cell type-specific confounder and smoking status and sex as main-effect confounders. This design contains different types of variables (categorical disease status and continuous age) with potential cell type-specific effects, and other covariates without cell type-specific effects. This analysis showcases the flexibility of CeDAR. All data analysis settings are the same for the six methods except the threshold to call DMC. For TOAST, TCA, csSAM, and CellDMC, sites with FDR < 0.05 were reported as csDMCs; for CeDAR-S and CeDAR-M, sites with posterior probability of DM > 0.95 were reported as csDMCs.

B cell plays an important role in RA (Marston et al., 2010; Wang et al., 2019b; Dörner and Burmester, 2003). From purified B cells, Julia et al. identified ten RA-related DMCs validated in two independent EWAS cohorts (UK and Spain) (Julià et al., 2017). We examined whether the six methods could detect those ten DMCs in B cells from the PBL DNA methylation bulk data. As can be seen from Figure 2.7(a), TCA and csSAM did not report any site out of the ten in B cells; TOAST, CellDMC, and CeDAR-S identified seven of them; and CeDAR-M identified eight sites. CD4 is another cell type reported to be related to RA (van Loosdregt et al., 2016; Chemin et al., 2019). However, there is no experimentally validated DMCs in CD4. To investigate whether the csDMCs detected for CD4 from CeDAR make biological and clinical sense, we performed a series of analyses to evaluate the results. First, Figure 2.7(b) shows a Venn diagram for the overlaps of the reported csDMCs in CD4 by the six methods. We see that CeDAR-M detected much more csDMCs in CD4 that include all csDMCs from CeDAR-S, and a large proportion of csDMCs from other four methods. Furthermore, we performed an enrichment analysis for the csDMCs uniquely identified by CeDAR-M, but not by TOAST, TCA, csSAM, and CellDMC, using *missMethyl* (Phipson et al., 2015). There are six KEGG pathways (Kanehisa and Goto, 2000; Kanehisa et al., 2021; Kanehisa, 2019) significantly

enriched (two with adjusted p-value < 0.1 and four with adjusted p-value < 0.2), which is shown in Figure 2.7(c). The top one, Phospholipase D signaling pathway, has been reported to play a pivotal role in RA. Previous studies showed that abnormal up-regulation of a gene in Phospholipase D signaling pathway, Phospholipase D1 (PLD1), may contribute to the pathogenesis of IL-1β-induced chronic arthritis (Kang et al., 2013). Additionally, genetic and pharmacological inhibition of PLD1 can cause suppression of collagen-induced arthritis symptom, such as induction of the inflammatory response, bone destruction, and osteoclastogenesis (Yoo et al., 2020). The other five pathways are focal adhesion, Wnt signaling pathway, EGFR tyrosine kinase inhibitor resistance, Sphingolipid signaling pathway, and regulation of actin cytoskeleton, which are also reported being related with RA disease (Shelef et al., 2014; Vasilopoulos et al., 2007; Cici et al., 2019; Swanson et al., 2012; Maceyka and Spiegel, 2014). We further investigated whether these six enriched KEGG pathways can be also identified by other competing methods (Table 2.1). We found that among the six pathways, Sphingolipid signaling pathway is uniquely identified by CeDAR. TOAST reports the remaining five other pathways, while TCA, CellDMC, and csSAM report fewer pathways. This result indicates that CeDAR can find unique csDMCs, leading to pathways and biological interpretations related to target phenotype that other methods cannot provide.

Table 2.1: Identification of CeDAR-enriched pathways by TOAST, TCA, CellDMC, and csSAM.

Pathways reported in Figure 7c	CeDAR	TOAST	TCA	CellDMC	csSAM
Phospholipase D signaling pathway	Yes	Yes	Yes	Yes	No
Wnt signaling pathway	Yes	Yes	No	Yes	No
Focal adhesion	Yes	Yes	Yes	Yes	No
EGFR tyrosine kinase inhibitor resistance	Yes	Yes	No	No	No
Sphingolipid signaling pathway	Yes	No	No	No	No
Regulation of actin cytoskeleton	Yes	Yes	No	Yes	No

There are six enriched KEGG pathways (with adjusted p-value < 0.2) based on CeDAR uniquely identified csDMCs. We check whether they can be identified by performing the same enrichment analysis on csDMCs identified by TOAST, TCA, CellDMC, and csSAM. In the table, "Yes" means the pathway is enriched by csDMCs reported by corresponding method, while "No" means it is not.

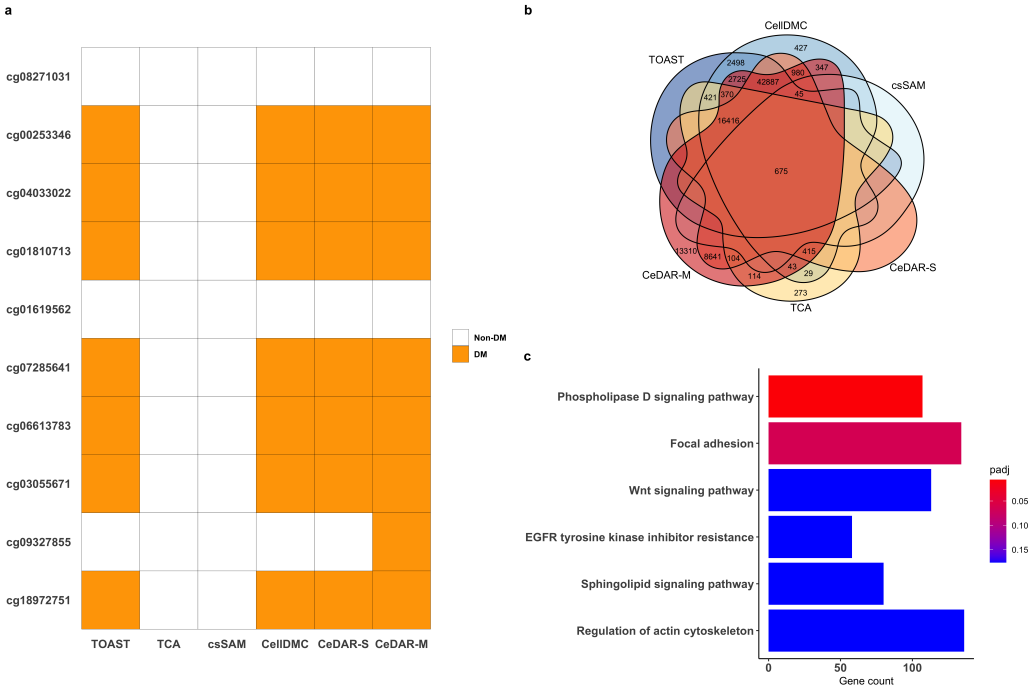

Figure 2.7: Cell type-specific DMC result for PBL DNA methylation data between RA and normal individuals. (a) Examination of six methods in identifying csDMCs of B cells from Liu et al. (2013). The ten csDMCs were identified and validated in two independent cohorts (Julià et al., 2017). (b) Venn diagram showing overlap of reported csDMCs in CD4 cell type from six methods. (c) Top six KEGG pathways enriched by CeDAR-M uniquely identified csDMCs in CD4, but not TCA and TOAST.

Other real data results

In additional to the above results, we analyzed several other real datasets: (1) detecting Down syndrome (DS)-associated csDM sites from frontal cortex gray matter samples (GSE74886 (Mendioroz et al., 2015)); (2) detecting systemic lupus erythematosus (SLE)-associated csDM sites from whole blood samples (GSE118114 (Yeung et al., 2019)); (3) detecting smoking-associated csDM sites from whole blood samples in two independent studies separately (GSE42861 and GSE402079 (Hannum et al., 2013)). All the three results demonstrate that CeDAR methods can achieve much more accurate results than other methods. The details of the analysis procedure and results are provided in Appendix A Section A.8, Figure A.10, Figure A.11, and Figure A.12.

Taken together from the real data analysis results, we conclude that the proposed methods are more accurate and sensitive compared to the existing methods. Particularly, CeDAR-M demonstrates better results compared to CeDAR-S and the results from CeDAR-M can potentially provide more biologically plausible target for future studies.

2.4 Discussion

In this work, we developed a novel statistical model called "CeDAR" that incorporates the cell type hierarchy in the cell type-specific differential analysis. The model is inspired by real data observation that cell types show strong correlation in their DE/DM states. CeDAR is based on a Bayesian hierarchical model incorporating the cell type hierarchy in the construction of prior probabilities for DE/DM. We derived procedures for parameter estimation and used the posterior probabilities for determining features' differential states. Extensive simulation studies and real data analyses demonstrate that CeDAR significantly improves the sensitivity and accuracy

in identifying csDE/csDM compared to existing methods, especially for cell types with low proportions.

We showed that the performance improvement of CeDAR is robust to the specification of cell type hierarchy, for example, when the true structure is not bifurcating or just has a single layer. Even when the cell types are completely independent, CeDAR is not worse than other methods. When the correlation between cell types is strong, CeDAR-M is recommended, since it can capture a complex cell type hierarchy; when the correlation is weak or sample size is small, CeDAR-S is preferred, because it can capture a certain level correlation without the need for the complex tree structure estimation. We also showed that the biases in the cell type hierarchy and cell type proportion estimation may impact the results, but the improvements over other methods are still significant. On the other hand, accurate hierarchy and proportion estimation will lead to better results. With the increasing availability of single-cell genomics data, we envision that such estimation will become more accurate for many biological systems, which will greatly benefit cell type-specific analyses in bulk data.

In this work, we implicitly assumed that the correlations among cell types are consistent for all features. However, in the real world, cell types may show different correlation patterns in DE/DM states among different feature sets corresponding to different biological processes. Thus, a more sophisticated method is to assume cell types have different correlations in different feature sets, which will be our future research direction. Additionally, CeDAR method is currently designed for continuous data, such as gene expression or DNA methylation microarray data. However, the general framework of borrowing information from cell types can be applied to other data types, such as the count data from sequencing. This is another promising future direction for us to explore.

Chapter 3

Investigating the cell type specific genes from population-level single-cell RNA-seq

3.1 Introduction

Single cell RNA sequencing (scRNA-seq) allows the quantification of gene expression levels in individual cells (Kolodziejczyk et al., 2015; Zheng et al., 2017; Macosko et al., 2015). In recent years, the scRNA-seq technologies have been successfully applied to answer a variety of biological questions, for example, to discover new cell types (Li et al., 2020a), estimate cellular tissue composition (Deng et al., 2019), uncover novel biological mechanisms in different biological systems (Jaitin et al., 2016; Fan et al., 2018; Peng et al., 2019), etc. Compared to traditional bulk RNA sequencing (RNA-seq), the major advantage of scRNA-seq is that the single cell expression provides information for understanding the cellular heterogeneity of complex samples. A major source of cellular heterogeneity is the cell types, that is, a complex sample usually consists of many different types of cells which are functionally different. Traditionally, the cell types are defined by their morphological or phenotypical features. For example, an often-used method to define cell type is to use flow cytometry to sort cells according to certain cell surface markers. With the gene expression data, the cell types can be defined by the expression values of some *cell type specific (CTS) genes*, which have distinct gene expression profiles in different cell types.

The CTS genes are defined as the genes with strong differential expression among cell types. These genes are often of great interest because they are closely related to the cellular identity and function, and potentially the pathologies of different diseases (Saul et al., 2022; Velmeshev et al., 2019; Park et al., 2018). They are also very useful in various data analysis tasks including cell type annotation and identification in scRNA-seq (Kim et al., 2021) and bulk data deconvolution (Wang et al., 2019a). For example, in scRNA-seq data analyses, a fundamental step is to identify the cell types for all cells. There are many cell type annotation methods, either unsupervised (Li et al., 2022a; Wang et al., 2017; Miao et al., 2020) or supervised (Aran et al., 2019; De Kanter et al., 2019; Li et al., 2022b; Hu et al., 2020). A majority of these

methods contains a step for feature selection, where only the expression values for the CTS genes are used. The non-marker genes express uniformly in all cell types, thus don't contain information for cell types. Therefore, the feature selection step enhances the signal to noise ratio in the data and will lead to better results.

Studies on CTS genes have a long history. Before the wide application of high-throughput quantification methods such as gene expression microarray, only limited number of CTS genes can be identified with low-throughput techniques such as western blot, northern blot, or RT-qPCR in one study. Researchers have manually curated CTS genes to systematically study cell types under different conditions (Kim et al., 1990). With the advances of high-throughput technology (e.g., microarray or RNA-seq), CTS genes can be identified more efficiently. However, these methods require purified cell types to derive CTS genes. The purification of cell types requires cell sorting methods such as fluorescence-activated cell sorting (FACS), which is expensive and laborious. In addition, accurate separation of cells relies on specific cell surface markers, which is not always available. Compared to the traditional methods, scRNA-seq provides a much easier and efficient way to study the CTS genes. It does not require experimentally purifying cell types but relies on computational procedures. Based on the expression profiles from individual cells, one can first identify cell types for each cell, and then detect CTS genes from differential expression analysis.

Various methods have been applied to identify CTS genes from scRNA-seq data. There are methods based on regular statistical tests for differential expression (DE) analysis, for example, Wilcoxon rank-sum test and Student's t-test that are implemented in Seurat (Hao et al., 2021) or Scanpy (Wolf et al., 2018). There are also more sophisticated methods like Necessary and Sufficient Forest (NS-Forest), which leverages the non-linear attributes of random forest feature selection to identify markers that are highly expressed in one specific cell type only (Aevermann et al., 2021).

Moreover, CTS genes can also be identified by feature selection methods like FEAST (Su et al., 2021) and scTIM (Feng et al., 2020), which are designed to select most representative markers for cell clustering. All these methods ignore one important factor: the between-subjects heterogeneity. Thus, the CTS genes identified with these methods from one subject are not guaranteed to appear in other subjects. In order to consider subject heterogeneity in CTS gene detection, one needs to analyze population-level scRNA-seq data. The results from such analysis are both interesting and important. Biologically, one wants to know the behavior of CTS genes, i.e., whether they would consistently show up in a population, or only appear in a proportion of subject. Computationally, the CTS genes are used in several other tasks such as deconvolution and cell type identification, so their consistency is important. For example in supervised cell type identification, CTS genes are implicitly assumed to appear in both reference and target samples (De Kanter et al., 2019; Li et al., 2022b; Andreatta et al., 2022; Guo and Li, 2021; Zhang et al., 2019b). If this is violated, the result would suffer.

There are some previous works discussed the consistency of CTS genes cross subjects. CellMarker is a manually curated resource that provide CTS genes either from scRNA-seq research or from other experimental research in human and mouse (Zhang et al., 2019a). In the CellMarker database, a CTS gene with more resources reported indicate greater consistency. GeneMarkeR is another database that provide manually curated CTS genes from published results (Paisley and Liu, 2021). It transforms marker gene statistics across publications to a "marker gene score" ranging from 0 to 1. A robust CTS gene should have its marker gene score greater than 0.5 and be specific to at most two cell types. Fischer and Gillis (Fischer and Gillis, 2021) identified replicable CTS genes from Brain Initiative Cell Census Network (BICCN) (Yao et al., 2021b,a) based on two metrics: area under the receiver-operator curve (AUROC) and fold change, and demonstrated that they can improve bulk sample

deconvolution and cell typing performance. Even though these works have provided invaluable information about robust CTS genes, the methods they used are still ad hoc. In addition, the CTS genes provided by these works are limited in specific tissues or species (human/mouse), which cannot satisfy the rapidly increasing demands of scRNA-seq application on various studies.

In this work, we develop a novel statistical method to identify CTS genes and evaluate their consistency from population level scRNA-seq data. We define a CTS gene as the one showing differential expression (DE) between one cell type vs. others. For a gene, we use a hierarchical model to consider both its frequency of being a CTS in a population and the strength of the differential expression (DE). Our model can identify different types of CTS genes, for example, the ones showing strong DE signal in only a small proportion of subjects, or the ones consistently showing weak DE signals across subjects. After detecting these CTS genes, we also design a strategy to utilize their consistency information from historical data for downstream analysis like supervised cell type identification.

The results demonstrate that with our proposed method, CTS genes with different characteristics (e.g., consistency and DE signal strength in subjects) can be identified and the consistent CTS genes information can significantly improve the performance of downstream analysis.

3.2 Methods

3.2.1 Subject-level summary statistics representing cell type specificity of genes

The input data of the model include scRNA-seq expression data from a population, with known cell types for all cells. Suppose there are N subjects from which we want to identify CTS genes. In each subject, there are G genes and K cell types. Let X_{gikc}

be the normalized expression for g-th gene ($g = 1, ..., G$) of i-th subject ($i = 1, ..., N$) in c-th cell ($c = 1, ..., C_{ik}$) of k-th cell type ($k = 1, ..., K$). Here, C_{ik} represents the number of cells for i-th subject in k-th cell type. The normalization is done by computing the read counts per 10,000 reads. We assume the normalized expression X_{gikc} is independent between genes and cells for all subjects. Define $E\{X_{gikc}\} = \mu_{gik}$, and $Var\{X_{gikc}\} = \omega^2_{gik}$ as the mean and variance of the normalized expression value. Then the unbiased estimator for mean expression of g-th gene in k-th cell type of i-th subject is: $\bar{X}_{gik} = \frac{\sum_{c=1}^{C_k} X_{gikc}}{C_{ik}}$. With Central Limit Theorem, when C_{ik} is large enough, $\bar{X}_{gik}$'s approximate distribution is:

$$\bar{X}_{gik} \sim AN(\mu_{gik}, \frac{\omega^2_{gik}}{C_{ik}}) \tag{3.1}$$

for $g = 1, ..., G$; $i = 1, ..., N$; $k = 1, ..., K$.

For the following context, we treat k-th cell type as the "target" cell type for which we want to study its CTS genes. In this work, we focus on CTS genes with expression at a higher level in only one cell type (i.e., one vs. others). Other types of CTS genes (e.g., two vs. others) can also be studied (Appendix B Section B.1).

Let Y_{gik} be the $log2$ fold change (LFC) of the expression for g-th gene in k-th cell type over the average of other cell types in i-th subject. Y_{gik} is computed as shown in Equation 3.2.

$$Y_{gik} = log2(\bar{X}_{gik} + 1) - log2(\frac{\sum_{k' \neq k} \bar{X}_{gik'}}{K - 1} + 1) \tag{3.2}$$

A large value of Y_{gik} indicates that g-th gene is a CTS gene of k-th cell type in i-th subject. Our computation of LFC is different from most existing methods for identifying CTS genes, which test one cell type vs. others. In those methods, when performing test between one cell type and others, the average expressions from the "others" group will be affected by the cell type proportions. Our definition of mean

expression in other cell types in Equation 3.2 excludes the influence of cell type composition, thus will provide more stable results. For the procedures below, we will model Y_{gik} for CTS gene identification. Using Y_{gik} instead of the data from individual cells greatly reduce the computational efficiency without losing much information.

3.2.2 A hierarchical model for CTS genes

We use the hierarchical model shown in Equitation 3.3 to combine the DE information from multiple subjects. We define D_{gk} as a binary random variable representing whether g-th gene is a CTS gene in k-th cell type (1: yes; 0, no). If g-th gene is a CTS gene in k-th cell type ($D_{gk} = 1$), then it has a probability q_{gk} to be DE (higher expression than the average of other cell types) in a randomly picked subject i, which is represented by binary random variable $Z_{gik} = 1$. We further introduce a random variable Δ_{gik} to represent the expected value of the estimated LFC (Y_{gik}), and σ_{gik} is the corresponding standard deviation. If g-th gene is a CTS gene in k-th cell type and shows DE signal in i-th subject ($D_{gk} = Z_{gik} = 1$), then Δ_{gik} should be greater than 0; otherwise, it should have expected value 0 with a small variation. Putting all pieces together, we have following hierarchical model:

$$
\begin{aligned}
Y_{gik}|\Delta_{gik} &\sim N(\Delta_{gik}, \sigma^2_{gik}) \\
\Delta_{gik}|Z_{gik} = 1 &\sim N(m_{gk}, \tau^2_{gk}) \\
\Delta_{gik}|Z_{gik} = 0 &\sim N(0, \tau^2_{gk}) \\
Z_{gik}|D_{gk} = 1 &\sim Bernoulli(q_{gk}) \\
Z_{gik}|D_{gk} = 0 &\sim Bernoulli(0) \\
D_{gk} &\sim Bernoulli(\pi_k)
\end{aligned}
\tag{3.3}
$$

Here, m_{gk} is the population level mean LFC of g-th gene in k-th cell type; τ^2_{gk} is the population level variance of LFC for g-th gene g in k-th cell type. Specifically, we

assume $m_{gk} \geq thres \geq 0$ and $Z_{gi} \perp Z_{gi'}|D_{gk} = 1$. $thres$ is a threshold defined by users, since small LFC is less possible to be a CTS gene and has less interest. In the estimation process, Y_{gik} and σ^2_{gik} are estimated from each individual subject. The detailed procedure is provided in Appendix B Section B.2.

3.2.3 Identification of CTS genes

From the above model, we can obtain several interesting quantities from the model. First, the posterior probability of $D_{gk} = 1$ provides an overall assessment whether a gene is a CTS gene. At the highest level, a gene can be either CTS genes ($D_{gk} = 1$) or non-CTS genes ($D_{gk} = 0$). Next, the conditional probability q_{gk} represents the consistency for a CTS gene to show DE signals cross subjects. The CTS genes are allowed to have different frequencies (q_{gk}) for showing DE in individual subjects and cell types. Finally, m_{gk} represents the conditional population-level DE strength once the gene is deemed CTS gene in population.

If we merely want to identify CTS marker genes, we only need to look at the posterior probability of $D_{gk} = 1$. However, a gene can have large posterior probability of $D_{gk} = 1$ if it has large q_{gk} or m_{gk}, or both. From our model, different types of CTS marker genes can be identified: (1) consistently show strong DE signal in most subjects (large q_{gk} and m_{gk}); (2) consistently show weak DE signals in most subjects (large q_{gk}, small m_{gk}); (3) show strong DE signals in only few subjects (small q_{gk}, large m_{gk}). Usually, the second type of CTS marker genes are difficult to detect from testing on individual subjects one by one, because tests for CTS markers with weak signals have very low statistical power, especially in minor cell types. The third type of markers are difficult to identify by testing on pooled data, because DE signal in partial subjects can be weakened after pooling with other subjects without DE signals. Our proposed method overcome these limitations and can identify all types of marker genes. These different types of CTS marker genes could have distinct

biological meanings and computational utilities. For example, CTS marker genes consistently show strong DE signals in all subjects (have large q_{gk} and m_{gk}) are more preferred for downstream analyses such as cell typing or bulk sample deconvolution, since they can robustly provide clear signal to represent a cell type.

3.2.4 Parameters estimation with EM algorithm

The parameters to be estimated from the proposed model include: m_{gk}, population level mean LFC of g-th gene in k-th cell type; τ^2_{gk}, population level variance of LFC of g-th gene in k-th cell type; q_{gk}, probability of CTS gene g in k-th cell type for a random picked subject; π_k, probability of a randomly picked gene to be a CTS gene for k-th cell type among the subjects. Since there are a number of latent variables in our model (Δ_{gik}, Z_{gik} and D_{gk}), we develop an EM algorithm to estimate these parameters.

Define $\phi(x; m, \tau^2)$ to be the probability density at a point x of a normal distribution with mean m and variance τ^2. We further define following values: $\phi_{ygik} = \phi(Y_{gik}; \Delta_{gik}, \sigma^2_{gik})$, $\phi_{0gik} = \phi(\Delta_{gik}; 0, \tau^2_{gk})$, and $\phi_{1gik} = \phi(\Delta_{gik}; m_{gk}, \tau^2_{gk})$.

Denote $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{q}_k, \boldsymbol{m}_k, \boldsymbol{\tau}^2_k\}$, where $\boldsymbol{q}_k = \{q_{1k}, ..., q_{Gk}\}$, $\boldsymbol{m}_k = \{m_{1k}, ..., m_{Gk}\}$, $\boldsymbol{\tau}^2_k = \{\tau^2_{1k}, ..., \tau^2_{Gk}\}$. We can derive the complete likelihood as following:

$$
\begin{aligned}
L(\boldsymbol{\Theta}) =& \prod_g P(\boldsymbol{Y}_{gk}, \boldsymbol{\Delta}_{gk}, \boldsymbol{Z}_{gk}, D_{gk} | m_{gk}, q_{gk}, \pi_k) \\
=& \prod_g \Bigg\{ \Bigg[\Big\{ \prod_{i=1}^{N} \phi_{ygik} \times \phi_{0gik} \times (1 - Z_{gik}) \Big\} (1 - \pi_k) \Bigg]^{1 - D_{gk}} \\
& \times \Bigg[\Big\{ \prod_{i=1}^{N} \phi_{ygik} \times \big[(1 - q_{gk}) \phi_{0gik} \big]^{1 - Z_{gik}} \times \big[q_{gk} \phi_{1gik} \big]^{Z_{gik}} \Big\} (\pi_k) \Bigg]^{D_{gk}} \Bigg\}
\end{aligned}
\tag{3.4}
$$

Theoretically, the estimation should be done by updating all four parameters jointly. For computation efficiency, we develop the following procedure to approximate the estimate of parameters. The general framework for the modified EM algorithm is as

following:

1. Assume all genes are CTS genes ($D_{gk} = 1$) and then estimate m_{gk}, τ^2_{gk} and q_{gk} with EM algorithm (Z_{gik} is missing data) for each gene $g = 1, ..., G$;

2. Based on estimated m_{gk} and given LFC threshold $thres$ to arbitrarily assign $D_{gk} = 0$ for genes with $m_{gk} \leq thres$;

3. Estimate π_k with EM algorithm, where m_{gk}, τ^2_{gk} and q_{gk} are fixed as estimates derived in step 1; D_{gk} is missing data.

The details of the steps 1 and 3 are shown in Appendix B Section B.3.

3.2.5 CTS gene selection for new subject based on historical data

After obtaining the CTS genes of different cell types from existing public data, we can use them in downstream analyses such as supervised cell type identification (Satija et al., 2015; Kiselev et al., 2017) or bulk data deconvolution (Li et al., 2020b; Li and Wu, 2019; Zheng et al., 2018a). Both tasks include a CTS gene selection step to pick genes containing cell type information. A crucial assumption is that the CTS genes selected from the reference will show cell type specificity in the target data, otherwise the result will suffer. We design the following method to improve CTS gene selection based on the historical data.

Reference based

Suppose we have derived CTS genes from historical datasets and the next step is to use reference sample to perform downstream analysis like cell type annotation. To ensure the selected CTS genes from reference sample (ref-markers) also appear in target sample, we designed a strategy to filter these ref-markers based on CTS genes information derived from historical datasets.

1. Identify CTS genes from reference subject (ref-marker) with any existing methods such as Wilcoxon rank-sum test.

2. For any ref-marker g in k-th cell type, calculate the effect size Y_{gk}^{ref} and sample variance $\sigma_{gk}^{2\ ref}$.

3. Calculate the posterior probability $P(Z_{gk}^{ref} = 1, Z_{gk}^{hist} = 1 | Y_{gk}^{ref}, \sigma_{gk}^{2\ ref}, \boldsymbol{Y}_{gk})$ given estimated m_{gk}, q_{gk}, and τ_{gk}^2, which are derived from historical data with our proposed method in previous section. Z_{gk}^{hist} is the DE state of random picked subject from historical data, which used to represent target subject, since we assume historical data can provide information for the target subject, i.e., if a marker gene appears in historical data, it is more likely to appear in the target data.

4. Call CTS genes based on $P(Z_{gk}^{ref} = 1, Z_{gk}^{hist} = 1 | Y_{gk}^{ref}, \sigma_{gk}^{2\ ref}, \boldsymbol{Y}_{gk}) >$ threshold and sort by effect size. Such called CTS genes are those we believe showing DE signal in both reference and target samples.

Semi-reference based

This is actually reference free, but with historical information, we call it semi-reference based. When there is no reference sample, we select genes with $P(Z_{gk}^{hist} = 1 | \boldsymbol{Y}_{gk}) = P(Z_{gk}^{hist} = 1 | D_{gk} = 1) \times P(D_{gk} = 1 | \boldsymbol{Y}_{gk}) >$ threshold, and sort by effect size. The threshold used here is user-defined and its default value is 0.95 in software.

3.3 Results

In results section, all analyses were performed on PBMC Lupus data (GEO accession: GSE96583) (Kang et al., 2018), which contains twenty-four samples from sixteen individuals. The samples come from two batches: in the first batch, there are eight

control samples from eight individuals with Systemic Lupus Erythematosus (SLE) disease; and in the second batch, there are eight control samples and eight IFN-beta stimulated samples from another eight individuals with SLE disease. In each sample, there are seven cell types: B cells, CD14+ Monocytes, CD4 T cells, CD8 T cells, Dendritic cells, FCGR3A+ Monocytes and NK cells (Megakaryocytes were excluded due to its extremely small composition in samples). We first applied our proposed method to identify CTS genes and evaluated whether publicly available markers reported by GeneMarkeR (Paisley and Liu, 2021) and PanglaoDB (Franzén et al., 2019) consistently appear across samples in PBMC Lupus data. We then applied our proposed model on this data set to identify consistent CTS genes and use them for downstream analyses like cell type identification.

3.3.1 CTS genes do not consistently appear across samples

In each sample, we first performed Wilcoxon rank-sum test for each cell type (one vs. all others). The DE genes (CTS genes) were called by FDR < 0.05. We collected CTS genes of PBMC cell types that are reported by GeneMarkeR (Paisley and Liu, 2021) and PanglaoDB (Franzén et al., 2019) and checked whether these CTS genes can be identified by Wilcoxon rank-sum test.

First of all, we found that only a small proportion of genes were called as DEGs across all samples (Figure 3.1(a)). For example, in CD4 T cells, there are totally 2529 genes are called DEGs in at least one sample among all 6231 genes. However, only 96 genes are called as DE in all 24 samples, while 740 genes are called as DE in only one sample. Same trend can be observed in other cell types (e.g., CD8 T cells, NK cells) with varying number of DEGs called across samples (Appendix B Table B.2). We also found that only part of CST genes reported by GeneMarkeR or PanglaoDB consistently show DE signal in all samples (Figure 3.1(b)). For example, in B cells, CTS genes like CD19, CD79A, CD79B are called DE in all samples; but

other CTS genes like LTB, TMEM156 (Paisley and Liu, 2021) are only called DE in some samples (LTB: 10 out of 24 samples, TMEM156: 17 out of 24 samples). These results imply that CTS genes may not consistently appear in all samples (even under the same experimental condition). Thus, a thorough evaluation of CTS genes consistency across samples is needed for both biological understanding of different cell types and downstream analyses like cell typing or bulk sample deconvolution.

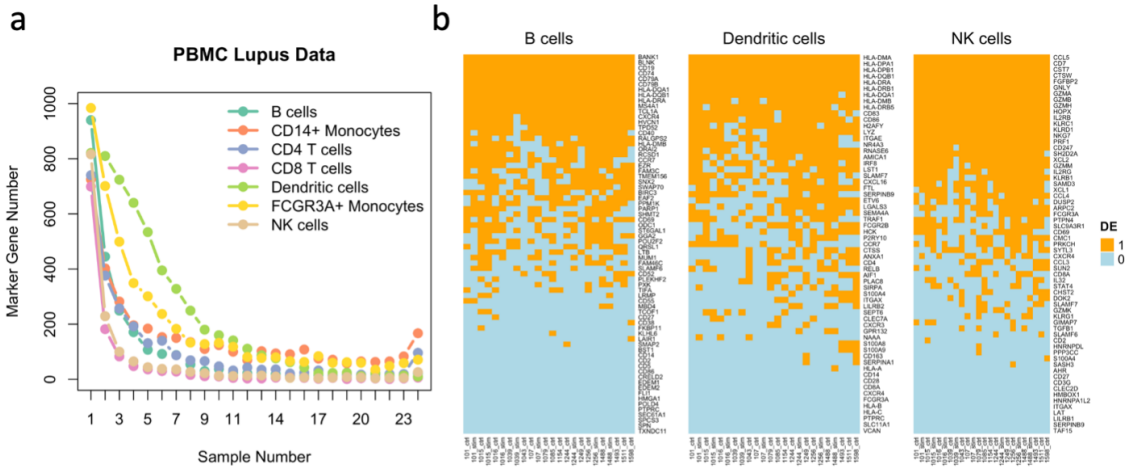

Figure 3.1: CTS genes do not consistently appear in all samples. (a) Numbers of genes called as DE by Wilcoxon rank-sum test in various number of samples for different PBMC cell types (B cells, CD14+ Monocytes, CD4 T cells, CD8 T cells, Dendritic cells, FCGR3A+ Monocytes and NK cells). The y-axis represents the number of genes called DE by Wilcoxon rank-sum test with FDR < 0.05 in different number of samples. The x-axis represents the number of samples (from 1 to 24). Different colors represent different cell types. (b) Heatmap represents DE state of CTS genes reported by GeneMarkeR or PanglaoDB in 24 samples of PBMC Lupus data for three cell types (B cells, Dendritic cells, and NK cells). Genes are sorted by the number of samples showing their DE states. The DE state represents whether the CTS genes can be called ad DEG (one vs. others) by Wilcoxon rank-sum test with FDR < 0.05 in one sample (1: yes; 0: no).

3.3.2 CTS genes with different characteristics

We applied the proposed method on the PBMC Lupus data to call CTS genes for different cell types. We set the threshold for LFC in estimation procedure very loosely as 0 to ensure more CTS genes will be kept. The CTS genes are called by $P(D_{gk} =$

$1|\boldsymbol{Y}_g) > 0.95$ for cell type $k = 1, ..., K$ and gene $g = 1, ..., G$. The genes called as CTS genes for one cell type have different characteristics: probability q_g measuring consistency of DE signal across samples, mean and variance of LFC (m_g and τ_g^2) measuring strength of DE signals across samples (Figure 3.2(a), Appendix B Table B.3, Table B.4).

The proposed method detected different types of CTS genes. First, some of the CTS genes have large LFC ($m_g > 1$) and high consistency ($q_g > 0.9$) in samples, such as CD14, FTL, and TYROBP in CD14+ Monocytes. These three genes are well-known CTS genes for Monocytes. Our method identified all of them and indicated that these three genes have very different LFC variances across samples. Smaller LFC variance represents more stable cell type specific gene expression signal across samples. Thus, with similar mean LFC level, CTS genes with smaller LFC variance is more preferred for analysis like bulk sample deconvolution, in which a fixed gene expression profile is used as reference for all samples. This variance difference (CD14: $\tau_g^2 = 0.442$; FTL: $\tau_g^2 = 0.173$; TYROBP: $\tau_g^2 = 0.003$) can be clearly observed in boxplot of gene expression in CD14+ Monocyte cells across 24 samples (Figure 3.2(b)). Compared to gene CD14 ($\tau_g^2 = 0.442$) and FTL ($\tau_g^2 = 0.173$), gene TYROBP has smaller LFC variance ($\tau_g^2 = 0.003$) and its expression is relatively more consistent across samples in both CD14+ Monocytes and other cell types (Figure 3.2(b)). In contrast, gene CD14 and FTL have greater expression variation in CD14+ Monocytes, but relative consistent expression in other cell types.

In some cell types (e.g., CD14+ Monocytes and Dendritic cells), the proposed method also identified some CTS genes have large LFC ($m_g > 1$), but lower consistency (e.g., $q_g < 0.9$) (Figure 3.2(a)). For example, gene CTSL ($q_g = 0.7$) only shows high expression in CD14+ Monocytes in 16 out of 24 samples, while gene TKT ($q_g = 0.25$) in 8 out of 24 samples (Figure 3.2(b)).

Moreover, the proposed method also identified some CTS genes with small LFC

($m_g < 0.3$) but high consistency ($q_g > 0.9$) (Figure 2a). One example gene is IL6R (estimated frequency is 0.96), which is called DE in only 18 out of 24 samples with Wilcoxon rank-sum test for CD14+ Monocytes. We can observe that IL6R has higher proportion of cells with non-zero counts in CD14+ Monocytes than in other cell types (Figure 3.2(b)). In conclusion, the proposed method can accurately evaluate consistency and differential expression signal strength of identify CTS genes.

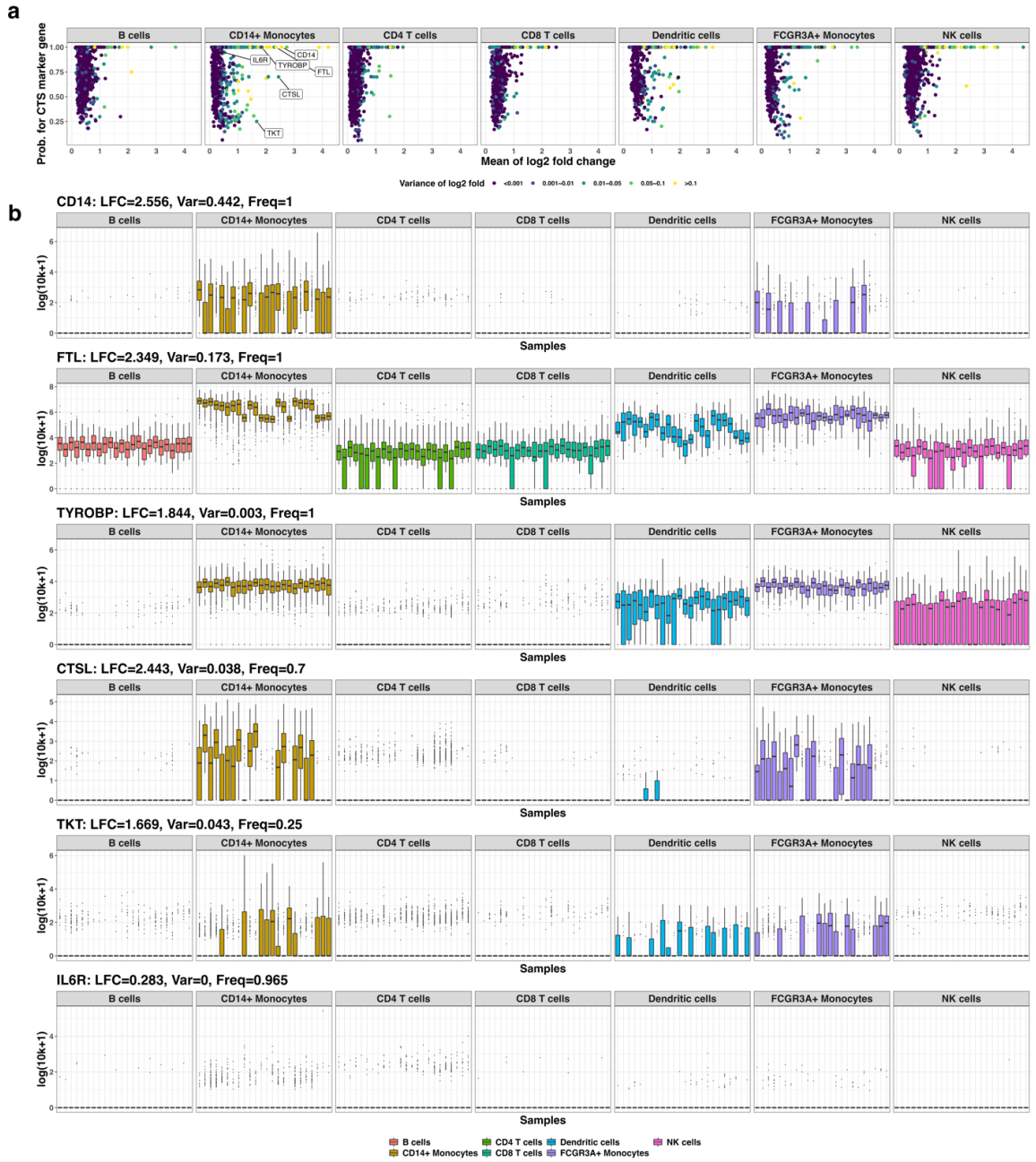

Figure 3.2: Characteristics of CTS genes identified from samples. (a) Scatter plots showing different characteristics of identified CTS genes in PBMC cell types (B cells, CD14+ Monocytes, CD4 T cells, CD8 T cells, Dendritic cells, FCGR3A+ Monocytes and NK cells). The y-axis represents estimated frequency of a CTS gene (q_g) showing DE signal across samples, which measures consistency. The x-axis represents the mean value of log2 fold change (m_g) of CTS genes in analyzed samples. The color of the points represents the variance of log2 fold change (τ_g^2) of CTS genes in analyzed samples (purple: small variance; yellow: large variance). (b) Boxplots of gene expression of all cells in different cell types for 24 samples. Six example CTS genes of CD14+ Monocytes (CD14, FTL, TYROBP, CTSL, TKT, and IL6R) are shown. They have different mean values of log2 fold change (LFC), variances of LFC (Var), and different probabilities to show DE signal (Freq) in samples. The y-axis is the log transformed 10k counts. The x-axis represents samples.

3.3.3 Comparison between Wilcoxon rank-sum test method and the proposed method

Wilcoxon rank-sum test is one of the most common used methods to identify CTS genes in scRNA-seq data (Pullin and McCarthy, 2022). After calling the CTS genes by the new proposed method (p-markers), we further compared them with CTS genes called by Wilcoxon rank-sum test (w-markers).

Table 3.1: Number of genes called as CTS genes with proposed method or Wilcoxon rank-sum test

Posterior Probability	Number of Samples showing DE	B cells	CD14+ Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A+ Monocytes	NK cells
pp = 0	0	3412	2558	2936	4072	729	1620	3324
	[1, 8]	752	2047	324	411	4343	3241	349
	[9, 16]	2	198	2	1	187	170	4
	[17, 24]	0	10	0	0	0	9	0
pp $\leq$ 0.95	0	361	0	615	789	1	0	1000
	[1, 8]	770	162	532	303	147	133	346
	[9, 16]	1	254	2	3	416	303	2
	[17, 24]	0	15	0	0	3	17	0
pp > 0.95	0	38	0	151	174	0	0	391
	[1, 8]	624	18	1135	408	11	13	660
	[9, 16]	178	347	300	45	257	272	90
	[17, 24]	93	622	233	25	137	453	65

The genes are categorized into three types: pp = 0 represents genes with negative mean LFC that discarded in the second step of EM algorithm; 0 < pp $\leq$ 0.95 represents genes failed to be identified as CTS genes in third step of EM algorithm; pp > 0.95 represents genes identified as CTS genes.

We found that some w-markers called in one or more samples are also called as p-markers (grey points in Figure 3.3(a), Table 3.1). The proportion of these w-

markers varies in different cell types (B cells: 36.98%, CD14+ Monocytes: 26.87%, CD4 T cells: 65.98%, CD8 T cells: 39.97%, Dendritic cells: 7.36%). The w-markers not called as p-markers are those with negative or small positive LFC defined as Equation 3.2 (blue and gold points in Figure 3.3(a)). In Figure 3.3(a), some genes are w-markers but with negative average LFC across samples, which indicates they show negative DE signals in some samples (blue points in Figure 3.3(a)). Since we are only interested in CTS genes with higher expression in cells from target cell type than from other cell types, these genes are failed to be called as p-markers. One example is gene CD74 in CD14+ Monocytes. In CD14+ Monocytes, CD74 has much higher expression than in CD4 T cells, NK cells, and CD8 T cells, but much lower expression values than in B cells and Dendritic cells (Figure 3.3(b)). It is more reasonable to define CD74 as CTS gene for B cells and Dendritic cells instead of CD14+ Monocytes in these samples. The significant Wilcoxon rank-sum test statistic of CD74 is due to much higher proportion of CD14+ Monocytes than B cells and Dendritic cells (Table B.1, Figure S1), which leads to higher rank for expression in cells of CD14+ Monocytes. Besides, there are some w-markers with very small positive average LFC across samples (gold in Figure 3.3(a)) that their DE signals are too weak to be called as p-markers.

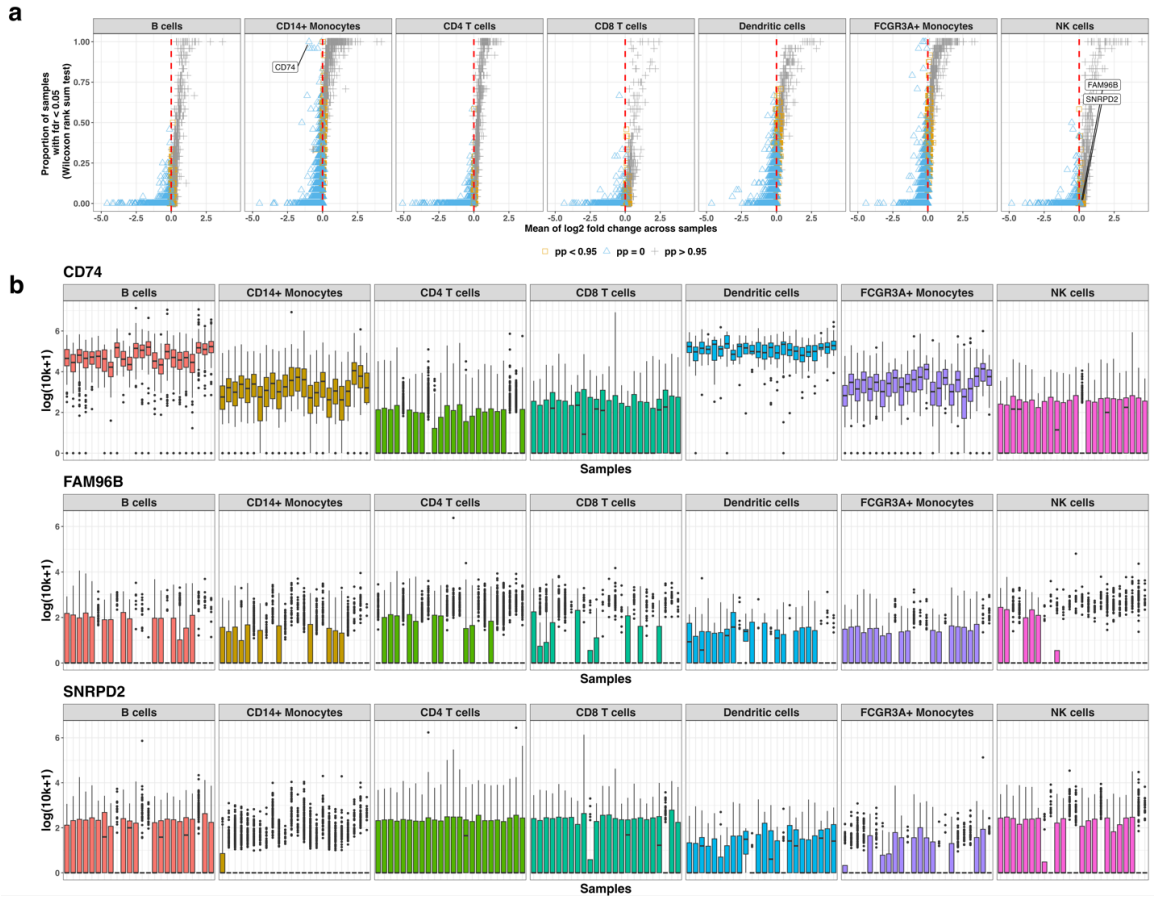

Figure 3.3: Comparison between CTS genes called by Wilcoxon rank-sum test (w-markers) and by the proposed method (p-markers). (a) Scatter plot of DE state of genes in target cell type. The y-axis is proportion of samples in which a gene being called DE (w-marker) by Wilcoxon rank-sum test with FDR < 0.05. The x-axis is the mean LFC defined in equation (2) across all twenty-four samples. Different colors represent posterior probability (pp) of genes to be p-markers (grey: pp > 0.95, is a p-marker; gold: pp < 0.95, not a p-marker and with positive LFC; blue: pp $= 0$, not a p-marker and with negative LFC). (b) Three example genes show difference between Wilcoxon rank-sum test and proposed method. The y-axis is the log transformed 10k counts. The x-axis represents samples. CD74 is w-marker in all samples, but not a p-marker. FAM96B is a p-marker with DE signal frequency 0.36, but not w-marker in any sample. SNRPD2 is a p-marker, but not w-marker in any sample.

There are also some genes identified as p-markers but not w-markers in B cells, CD8 T cells, CD4 T cells and NK cells (Table 3.1, Figure 3.3(a)). They are either with relative strong DE signal (but not strong enough to be tested by Wilcoxon rank-sum test) in a few samples, or with weak but consistent signal in most samples.

FAM96B and SNRPD2 are two example CTS genes in NK cells (Figure 3.3(b)). We performed enrichment analysis with these genes on Human gene atlas database (`http://biogps.org/downloads/`) (Wu et al., 2009; Su et al., 2004) stored in package *enrichR* (Kuleshov et al., 2016). In Table 3.2, we can observe that in cell type B cells, CD4 T cells, and NK cells, the most significant enriched terms are corresponding cell types. This indicates these genes, which are p-markers but not w-markers, may also serve as CTS genes. Discovery of such CTS genes are the result from pooling data from many samples together by the proposed model. Even though major interests usually lie in CTS genes with most significant DE signal.

Table 3.2: Statistically significant enriched terms from Human Gene Atlas with genes that are p-markers but not w-markers

Rank	B cells	CD4 T cells	CD8 T cells	NK cells
1	**CD19+ B cells (neg. sel.) (p.adjust: 2.79e-03)**	**CD4+ T cells (p.adjust: 7.57e-05)**	CD56 + NK cells (p.adjust: 1.42e-03)	**CD56+ NK cells (p.adjust: 8.33e-12)**
2	CD4+ T cells (p.adjust: 3.84e-02)	CD8+ T cells (p.adjust: 7.57e-05)	721 B lymphoblasts (p.adjust: 1.47e-03)	721 B lymphoblasts (p.adjust: 1.29e-06)
3	CD8+ T cells (p.adjust: 4.29e-02)	721 B lymphoblasts (p.adjust: 1.40e-03)	CD4+ T cells (p.adjust: 1.22e-02)	CD4 + T cells (p.adjust: 1.43e-04)
4			**CD8+ T cells (p.adjust: 2.70e-02)**	CD8+ T cells (p.adjust: 9.06e-04)
5			CD19+ B cells (neg. sel.) (p.adjust: 2.99e-02)	Heart (p.adjust: 4.27e-02)
6			Lymphoma burkitts (Raji) (p.adjust: 2.99e-02)	

Enrichment analysis was performed with genes which are called as p-markers (posterior probability > 0.95) but not w-markers for B cells (38 genes), CD4 T cells (151 genes), CD8 T cells (174 genes) or NK cells (394 genes) separately. There are total 87 terms corresponding to different cell types or tissues in the Human Gene Atlas database stored in package "enrichR". The table contains all enriched terms with adjust p-value smaller than 0.05.

We further investigated the genes which are both p-markers and w-markers by

comparing the estimated frequency showing DE from the proposed method with proportion of samples called DE by Wilcoxon rank-sum test. In Figure 3.4(a), we can observe that the estimated frequency of the proposed method is usually higher than proportion of samples called DE by Wilcoxon rank-sum test in B cells, CD4 T cells, CD8 T cells and NK cells. Meanwhile, the two metrics have much higher correlation in CD14+ Monocytes, FCGR3A+ Monocytes, and Dendritic cells. One example gene is NFATC1. It is a p-marker for CD4 T cells with estimated frequency equals to one, while being called DE in only three out of twenty-four samples by Wilcoxon rank-sum test. From Figure 3.4(b), we can observe the expression pattern of NFATC1 in cell types are similar across samples. Its weak DE signal leads to small power to call DE in all samples. Benefited from pooling samples in analysis, its estimated frequency showing DE is one with proposed method. At the same time, there are some genes with much lower estimated frequency showing DE with proposed method than proportion of samples called DE by Wilcoxon rank-sum test, such gene EIF4A1 in CD14+ Monocytes and gene CXCR4 in CD4 T cells (Figure 3.4(b)). In most samples, mean expression of gene EIF4A1 in CD14+ Monocytes is lower than in Dendritic cells. With definition of Equation 3.2, its estimated frequency showing DE is only 0.19. However, it is called DE in 19 out of 24 samples by Wilcoxon rank-sum test, which is due to high abundance of CD14+ Monocytes and low abundance of Dendritic cells (Table B.1), which can also explain gene CXCR4 in CD4 T cells.

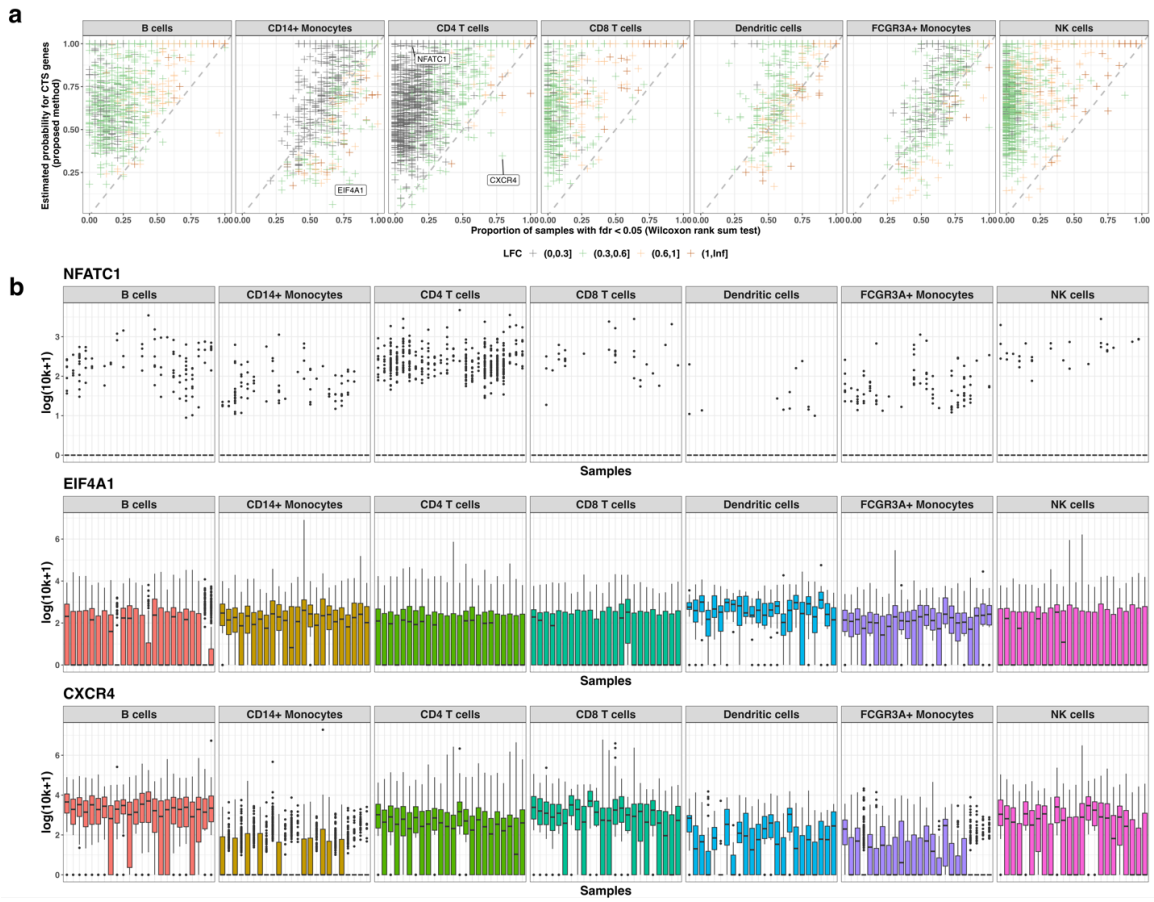

Figure 3.4: Comparison of estimated frequency showing DE signal in samples by proposed method with proportion of samples called DE with Wilcoxon rank-sum test for genes are both p-markers and w-markers. (a) Scatter plot of estimated frequency showing DE state by proposed method and Wilcoxon rank-sum test of genes in target cell type. The y-axis is the estimated frequency showing DE state among samples by proposed method. The x-axis is proportion of samples in which a gene being called DE (w-marker) by Wilcoxon rank-sum test with FDR < 0.05. Different colors represent estimated mean LFC among samples (grey: $0 < \text{LFC} \leq 0.30$; green: $0.30 < \text{LFC} \leq 0.6$; gold: $0.60 < \text{LFC} \leq 1.00$; brown: $\text{LFC} > 1.00$). (b) Three example genes show difference between Wilcoxon rank-sum test and proposed method. The y-axis is the log transformed 10k counts. The x-axis represents samples. NFATC1 is a p-marker with weak but consistent DE signal in samples but called DE in only 3 out of 24 samples by Wilcoxon rank-sum test for CD4 T cells. EIF4A1 is a p-marker with DE signal frequency 0.19 but called DE in 19 out of 24 samples by Wilcoxon rank-sum test for CD14+ Monocytes. CXCR4 is a p-marker with DE signal frequency 0.35 but called DE in 19 out of 24 samples by Wilcoxon rank-sum test for CD4 T cells.

3.3.4 Consistent CTS genes can improve performance of downstream analysis

Supervised cell type identification

After obtaining the CTS genes from PBMC Lupus data, we designed a simulation study based on real data to evaluate whether incorporating this (historical) information with proposed strategy can help to improve supervised cell type identification accuracy.

Among the twenty-four samples in PBMC Lupus data, we randomly select one sample as target, for which the cell types need to be identified. We randomly select another sample as reference, and all other samples were deemed historical samples, from which we identify consistent CTS genes across samples. We use three types CTS genes in the cell type identification: (1) "ref", CTS marker genes identified with existing methods (Bimod (McDavid et al., 2013), MAST (Finak et al., 2015) or Wilcoxon rank-sum test) from the reference sample; (2) "ref_hist", CTS marker genes identified with proposed strategy, which incorporate historical information with reference sample information; (3) "ref_target", overlap of CTS genes identified with existing methods in reference and target sample. In practice, since the cell types in target sample are unknown, the "ref_target" marker genes cannot be detected. We used them in this evaluation to serve as performance ceiling since using CTS genes in both target and reference should provide the most accurate results.

The cell typing analysis was performed with two popular methods: Seurat and SingleR (Aran et al., 2019). Seurat projects the PCA from the reference onto the target sample, while SingleR calculates the correlation between cells in target sample and cell type centroids of reference sample. There are totally six scenarios in one simulation (3 DE methods $\times$ 2 cell-typing methods). In each scenario, we tried different number of CTS genes (top 20, 50, 100, 200 and all CTS genes for each cell

type). If number of CTS gene for some cell type less than required number, then all of them are included for analysis. Number of "ref_target" CTS genes are less than "ref" CTS genes, because it is overlap of "ref" CTS genes and "target" CTS genes. We evaluated the cell-typing performance with two metrics – average Macro-F1 and average accuracy over $24 \times 23 = 532$ simulations.

In most scenarios, cell-typing with "ref_target" or "ref_hist" CTS genes has higher Macro-F1 score and accuracy than with "ref" CTS genes (Figure 3.5 and Table 3.3) on average. This confirms that CTS gene consistency between target and reference is crucial for cell typing. Compared to "ref" CTS genes, the improvement with historical CTS genes information ("ref_hist") is the most significant when using the top 20 and 50 CTS genes per cell type, which also indicates the top-ranking CTS genes selected by proposed method is more representative and informative. Besides, because number of "ref_target" CTS gene is smaller than the other two CTS gene types ("ref" and "ref_hist"), its performance is worse than "ref_hist" CTS genes when required CTS gene number is small (e.g., top 20, or top 50 per cell type). This implies importance of informative CTS gene number for cell typing analysis. Overall, the simulation results demonstrate that incorporating consistent CTS gene information from historical data can improve cell typing accuracy.

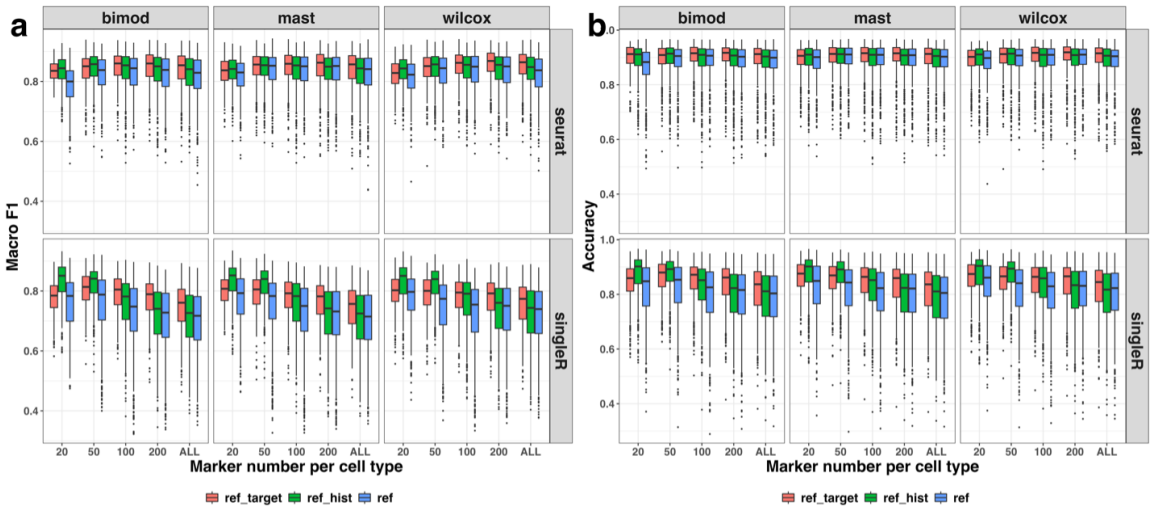

Figure 3.5: Accuracy evaluation of simulated cell typing with different types of CTS genes under various scenarios. Three types of CTS genes were in comparison: "ref", CTS genes identified in reference samples with Bimod, MAST or Wilcoxon rank-sum test; "ref_hist", CTS genes selected by proposed strategy incorporating historical marker information with reference sample information; "ref_target", overlap of CTS genes identified in both reference and target samples. Two cell typing methods were applied: Seurat and SingleR, which have different mechanisms for cell type annotation. The boxplot was generated based on totally $24 \times 23 = 532$ simulations. Two metrics were used: (a) Macro-F1 difference compared with "ref" marker of simulated cell typing, and (b) accuracy difference compared with "ref" marker of simulated cell typing. The x-axis is the number of marker selected in each cell type for cell typing analysis. Specifically, "ALL" represents all CTS marker genes are selected for analysis.

Table 3.3: Number of genes called as CTS genes with proposed method or Wilcoxon rank-sum test

Diff with ref (95% CI)		Seurat			SingleR		
Metrics	Marker number	Bimod	MAST	Wilcox	Bimod	MAST	Wilcox
Macro F1 Difference	20	**0.048 (0.044, 0.053)**	**0.017 (0.013,0.020)**	**0.022 (0.018,0.025)**	**0.073 (0.066,0.081)**	**0.060 (0.052,0.068)**	**0.051 (0.043,0.059)**
	50	**0.020 (0.017, 0.023)**	**0.004 (0.001, 0.007)**	**0.012 (0.009, 0.015)**	**0.068 (0.06, 0.076)**	**0.075 (0.067, 0.083)**	**0.081 (0.074, 0.088)**
	100	**0.006 (0.003, 0.01)**	-0.001 (-0.004, 0.001)	**0.003 (0.000, 0.005)**	**0.041 (0.036, 0.046)**	**0.040 (0.036, 0.045)**	**0.047 (0.043, 0.051)**
	200	**0.009 (0.006, 0.012)**	-0.002 (-0.005, 0.000)	**0.004 (0.001, 0.007)**	**0.017 (0.014, 0.02)**	**0.010 (0.007, 0.012)**	**0.012 (0.009, 0.016)**
	ALL	**0.011 (0.008, 0.014)**	**0.004 (0.002, 0.007)**	**0.013 (0.010, 0.016)**	**0.012 (0.01, 0.014)**	**0.011 (0.009, 0.014)**	**0.008 (0.005, 0.012)**
Accuracy Difference	20	**0.028 (0.024, 0.032)**	**0.006 (0.003, 0.009)**	**0.009 (0.006, 0.012)**	**0.056 (0.049, 0.063)**	**0.052 (0.044, 0.059)**	**0.036 (0.029, 0.042)**
	50	**0.007 (0.004, 0.011)**	-0.003 (-0.006, -0.001)	0.001 (-0.001, 0.004)	**0.054 (0.048, 0.061)**	**0.062 (0.055, 0.069)**	**0.062 (0.056, 0.069)**
	100	-0.006 (-0.009, -0.003)	-0.007 (-0.010, -0.004)	-0.007 (-0.010, -0.004)	**0.030 (0.026, 0.033)**	**0.031 (0.027, 0.034)**	**0.030 (0.027, 0.034)**
	200	-0.001 (-0.004, 0.001)	-0.004 (-0.006, -0.002)	-0.007 (-0.009, -0.004)	**0.012 (0.009, 0.014)**	**0.006 (0.003, 0.008)**	**0.003 (0.001, 0.005)**
	ALL	0.002 (-0.001, 0.004)	0.000 (-0.002, 0.003)	0.000 (-0.002, 0.003)	**0.008 (0.006, 0.01)**	**0.008 (0.007, 0.01)**	0.000 (-0.002, 0.002)

Cell typing accuracy difference between different types of CTS genes under various simulation scenarios. Average Macro-F1 and average accuracy differences and corresponding 95% confidence interval were calculated for "ref_hist" markers by being compared with "ref" markers for 532 simulations. Improvement in bold are statistically significant.

Bulk sample deconvolution

Since scRNA-seq data can provide gene expression information of different cell types in a tissue, it has been widely used as reference for bulk sample deconvolution. With well annotated cells, scRNA-seq data can be directly applied as input for some deconvolution methods like DWLS (Tsoucas et al., 2019) or pure profile of cell types can be summarized from scRNA-seq for methods like Cibersort (Newman et al., 2015) or non-negative least squares (NNLS) (Lawson and Hanson, 1995; Mullen and van

Stokkum, 2012).

Thus, we also designed a simulation to check whether incorporating the historical consistency information of CTS genes with proposed strategy can improve bulk sample deconvolution performance. Most of the settings in this simulation are similar as above cell typing analysis that three types CTS genes are compared with totally 532 simulations. The major difference is that we need to transform gene counts in target sample into pseudo bulk counts and summarize the proportion of cell types as golden standard. Besides, for methods like Cibersort or NNLS, we average gene expression within same cell type group and serve it as pure cell type profile for these methods. Three commonly used methods: Cibersort, NNLS, and DWLS are used. Performance is measured by rooted mean square deviance (RMSD) and Pearson correlation (Corr) between estimated and true proportions.

Generally, deconvolution with "ref_hist" genes can have better deconvolution performance (lower RMSD and higher correlation) than with "ref" marker genes in most scenarios. The performance improvement is most significant for NNLS method with 20 markers per cell type. For example, the RMSD with "ref_hist" is about 0.050 smaller than with "ref" markers and the correlation with "ref_hist" is about 0.124 greater than with "ref" markers on average (Table 3.4). Moderate decreasing of RMSD can be observed for methods Cibersort and DWLS and moderate increasing of Pearson correlation can be observed in method DWLS with makers generated by all three methods (Figure 3.6 and Table 3.4).

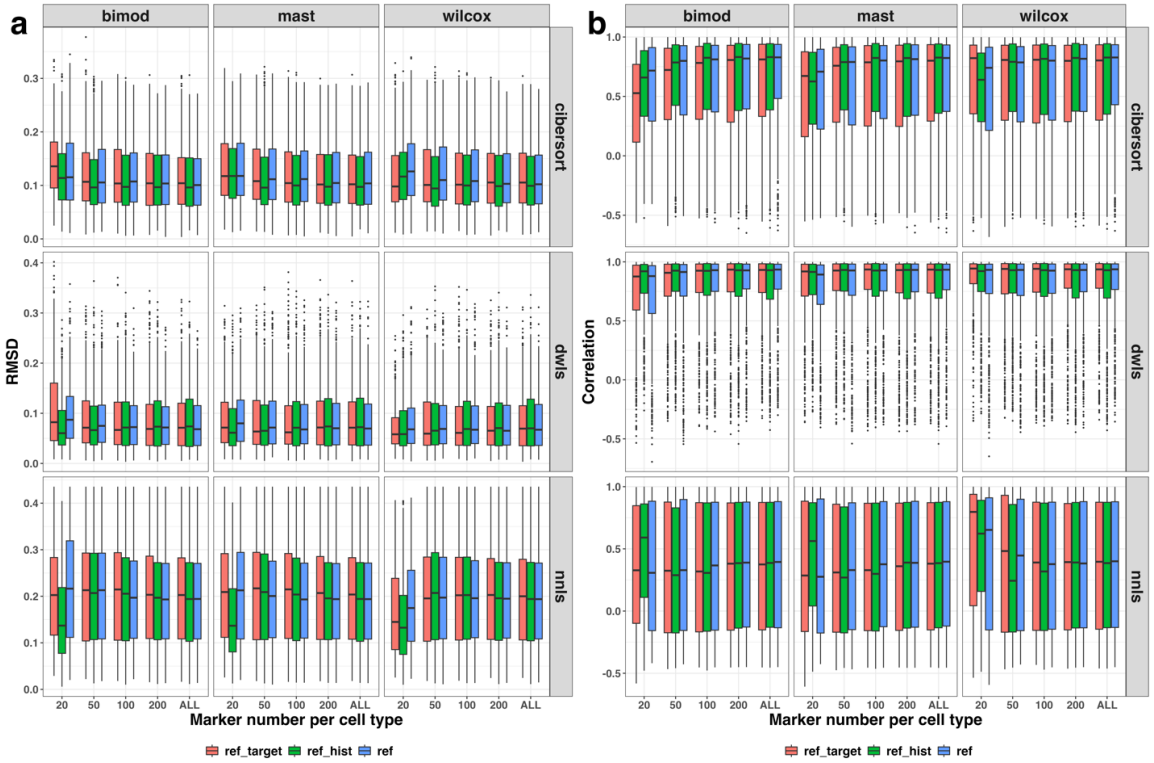

Figure 3.6: Accuracy evaluation of simulated bulk sample deconvolution with different types of CTS genes under various scenarios. Three types of CTS genes were in comparison: "ref", CTS genes identified in reference samples with Bimod, MAST or Wilcoxon rank-sum test; "ref_hist", CTS genes selected by proposed strategy incorporating historical marker information with reference sample information; "ref_target", overlap of CTS genes identified in both reference and target samples. Three commonly used cell typing methods were applied: Cibersort, DWLS, and NNLS. The boxplot was generated based on totally $24 \times 23 = 532$ simulations. Two metrics were used: (a) RMSD difference compared with "ref" marker of simulated deconvolution, and (b) Pearson correlation difference compared with "ref" marker of simulated deconvolution. The x-axis is the number of marker selected in each cell type for deconvolution analysis.

Table 3.4: Number of genes called as CTS genes with proposed method or Wilcoxon rank-sum test

Diff with ref (95% CI)		RMSD difference		Correlation difference	
		Marker number: 20	Marker number: 50	Marker number: 20	Marker number: 50
Cibersort	Wilcox	**-0.013 (-0.018, -0.008)**	**-0.009 (-0.012, -0.006)**	0.017 (-0.015, 0.049)	**0.022 (0.008, 0.037)**
	MAST	**-0.008 (-0.013, -0.004)**	**-0.010 (-0.013, -0.007)**	-0.012 (-0.04, 0.017)	**0.040 (0.026, 0.054)**
	Bimod	**-0.012 (-0.017, -0.007)**	**-0.008 (-0.011, -0.005)**	-0.001 (-0.029, 0.028)	0.01 (-0.006, 0.027)
NNLS	Wilcox	**-0.035 (-0.042, -0.028)**	0.004 (0.000, 0.008)	**0.080 (0.043, 0.117)**	-0.059 (-0.08, -0.039)
	MAST	**-0.050 (-0.057, -0.043)**	0.006 (0.003, 0.009)	**0.124 (0.087, 0.161)**	-0.041 (-0.055, -0.027)
	Bimod	**-0.059 (-0.067,-0.052)**	-0.001 (-0.005, 0.003)	**0.138 (0.099, 0.176)**	-0.038 (-0.059, -0.017)
DWLS	Wilcox	**-0.006 (-0.009, -0.002)**	-0.001 (-0.004, 0.003)	**0.020 (0.003, 0.037)**	0.010 (-0.007, 0.028)
	MAST	**-0.013 (-0.016, -0.009)**	**-0.005 (-0.008, -0.001)**	**0.042 (0.024, 0.060)**	**0.022 (0.008, 0.036)**
	Bimod	**-0.021 (-0.025, -0.017)**	**-0.004 (-0.007, -0.001)**	**0.095 (0.073, 0.117)**	**0.033 (0.015, 0.051)**

Bulk sample deconvolution difference between different types of CTS genes under various simulation scenarios. Average RMSD and average correlation differences and corresponding 95% confidence interval were calculated for "ref_hist" markers by being compared with "ref" markers for 532 simulations. Improvement in bold are statistically significant.

3.4 Discussion

In biological studies, CTS genes serve as cell type identities that it is not only required by scRNA-seq downstream analysis like cell type annotation, but also provides clues

to uncover hidden mechanisms under various conditions. Even though many methods have been developed to study CTS genes from scRNA-seq data, most of these work only study CTS genes in single dataset/sample that consistency of such CTS genes is not guaranteed in other datasets/samples. However, to serve as cell type identities, CTS genes are expected to be robust in different datasets/samples. Existing methods offering evaluation of CTS genes consistency across datasets/samples are limited and lack of rigorous statistics support.

In this work, we found that CTS genes identified through Wilcoxon rank-sum test or reported by public databases (PanglaoDB and GeneMarkeR) do not consistently appear in all samples of PBMC Lupus data. Inspired by this observation, we built a hierarchical model to identify CTS genes and evaluate their consistency across samples and proposed a strategy to make use this historical marker information. Finally, we applied our model and strategy on a data set containing multiple samples for performance evaluation.

The results showed that the proposed method can well evaluate consistency of CTS genes that it not only can identify CTS genes consistently appear in samples with large or small LFC, but also can successfully identify CTS genes only appear in partial samples. A simulated cell typing study showed us that the proposed strategy, which incorporates historical marker information with reference sample information, can help to improve cell typing accuracy. The most significant improvement appears in simulation with smaller number markers per cell type also imply the top-ranking CTS genes called by proposed strategy is more representative and informative. Based on the proposed model, if we want to study PBMC cell types, we can directly apply it on samples collected from different datasets (each data may have multiple samples). One limitation of this method is that the result may be dominated by the datasets with large sample size, since samples within same dataset show more similarity with each other. Thus, a sample size weighted estimation is needed for imbalance sampling

scenario. In the future, we plan to apply our proposed method on more scRNA-seq datasets and build up an interactive database which allows users deriving CTS genes information in various species or tissues easily. In addition, we plan to extend our model to incorporate covariates of sample information into analysis, which can help users to have deeper understanding of CTS genes in their study.

Chapter 4

Benchmark of Methods Designed for Rare Cell Population Identification in Single Cell RNA Sequencing Data

4.1 Introduction

In single cell RNA sequencing (scRNA-seq) data, cell types are identified by clustering cells based on their gene expression profile (Ianevski et al., 2022). Different clusters of cells represent different cell types and rare cell population (RCP) is a cluster of cells with extremely low abundance in one sample (Jiang et al., 2016; Andrews and Hemberg, 2018). While there is no restrictive upper limit of cell type abundance to define it as an RCP, most studies set their target RCP size with proportion within ranges from 0.1% to 15% (Andrews and Hemberg, 2018; Wegmann et al., 2019; Fa et al., 2021). These RCPs could be stem cells, short-lived progenitors, cancer stem cells, or circulating rumor cells. Thus, even though they have low abundance in a tissue, they can play important roles in biological development or disease progression (Orkin and Zon, 2008; Kreso and Dick, 2014; Plaks et al., 2013). For example, in skeletal muscle at homeostasis, muscle stem cells only account for less than 1% of total cells, and are essential for muscle homeostasis and repair (McKellar et al., 2021). In peripheral blood, the number of circulating tumor cells (CTCs) is reported to be between 0.1 and 10 cells per milliliter; however, they have been confirmed to be a prognostic cancer marker in breast cancers and prostate cancers (Enkhbat et al., 2021; Danila et al., 2007; Shaffer et al., 2007). Cancer stem cells (CSC) are reported to be preserved as a small population through self-renewal and to generate more differentiated progenies that constitute the bulk of the tumor mass (Suvà and Tirosh, 2020). In human melanoma, the frequency of CSCs is reported to be lower than 1 per million cells (Schatton et al., 2008).

To study the RCPs, a traditional way is to isolate them from bulk tissue via enrichment methods such as filtering and magnetic bead selection (Schreier et al., 2017). However, such method cannot guarantee the purity of isolated cells and highly relies on the prior knowledge of cell surface antigens for target RCP, which limits new cell type discovery (Zborowski and Chalmers, 2011). The development of scRNA-

seq provides researchers gene expression profile of each cell in studied tissue. It not only allows new RCP discovery, but also provides researchers possibility to study the relationship between target RCP and other cell types in the studied tissue under certain conditions. To correctly derive cell type information from scRNA-seq data, a crucial step is cells clustering. There are many methods having been developed (Satija et al., 2015; Kiselev et al., 2017) that can work well on cell types with large abundance. However, identification of RCPs brings additional challenges to these methods that RCPs contribute little to the global structure of a sample due to its small abundance. Furthermore, the genes that distinguish RCPs from other cell types may be only small proportion of all measured genes (DeMeo and Berger, 2021). Thus, computational methods specifically designed for RCP detection are needed to fully extract useful cell type information for scRNA-seq data.

Table 4.1: Summary of methods designed for RCP identification from scRNA-seq data

Key idea	Methods	Environment	Output info				Main characteristic
			RCP-cells	Marker	Final-clusters	Embeddings	
1	RaceID3 (Grün et al., 2015)	R	Yes	Yes	Yes	No	Find outlier cells from initial clustering result
	CellSIUS (Wegmann et al., 2019)	R	Yes	Yes	Yes	No	Find sub-clusters in initial clustering result
	SCISSORS (Leary et al., 2021)	R	Yes	No	Yes	No	Re-clustering initial clusters with high heterogeneity + Try different clustering parameter combinations + Report the one with lowest Silhouette score
2	Giniclust3 (Dong and Yuan, 2020)	Python	Yes	Yes	Yes	No	Clustering with RCP-related features measured by Gini index
	CIARA (Lubatti et al., 2022)	Python/R	Yes	Yes	Yes	No	Clustering with genes have their expression "highly localized"
	GapClust (Fa et al., 2021)	R	Yes	No	No	No	Identify RCP cells by detecting great distance change between their $k-1$-th and k-th neighbour
	SCA (DeMeo and Berger, 2021)	Python	No	No	No	Yes	Transform gene expression to information score - significance of local and background expression difference + Dimension reduction based on information score
	MicroCellClust (Gerniers et al., 2021)	R	Yes	Yes	No	No	Max-sum submatrix problem with constrains, bi-clustering
3	FiRE (Jindal et al., 2018)	Python/R	Yes	No	No	No	Rarity score for cells calculated by Sketching
	EDGE (Sun et al., 2020)	R	No	Yes	No	Yes	Cell similarity calculated by Sketching
	scAIDE (Xie et al., 2020)	Python	No	No	Yes	Yes	AE imputation + MDS dimesion reduction keep cell similarity + Random projection hashing - based k-means clustering
	DoRC (Chen et al., 2019)	Python	Yes	No	No	No	Rarity score for cells calculated by Isolation Forest
4	SCMER (Liang et al., 2021)	Python	No	Yes	No	No	Assume UMAP can capture RCP info + Select features keep the manifold

"Key idea" means the intuition behind methods: 1. RCP cells are mis-clustered into "raw" major clusters; 2. RCP is a small group of cells with certain genes highly expressed; 3. Under different transformations, similar cells always have similar transformed gene expression profiles; 4. Others. "Output info" means whether the method can provide corresponding information. "RCP-cells": whether the method provides RCP cell information in result; "Marker": whether the method provides RCP cells related markers; "Final-clusters": whether the method provides cluster information for all cell types in data; "Embeddings": whether method provides embedding info. "AE" is short for autoencoder and "MDS" is short for multidimensional scaling.

In scRNA-seq data analysis, there are multiple steps needed before deriving clustering result from a gene expression matrix of cells, such as feature selection, dimension reduction and similarity calculation between cells. Methods with different focuses on these steps have been developed to identify RCPs from scRNA-seq data. In general, based on the intuition behind these methods, most of them can be categorized into one of following three types:

RCP cells are mis-clustered into "raw" major clusters

Among existing methods, RaceID3 (Grün et al., 2015; Herman and Grün, 2018),

CellSIUS (Wegmann et al., 2019), and SCISSORS (Leary et al., 2021) assume that RCP cells are often mis-clustered with cells from major cell types (large abundance) with commonly used clustering methods (e.g., k-means, hierarchical clustering). Thus, these methods try to identify RCP cells from "raw"/"initial" clusters generated by a regular analysis pipeline.

For example, RaceID3 (Grün et al., 2015; Herman and Grün, 2018) identifies RCPs from outlier cells detected in clusters generated by k-medoids clustering. Within a cluster, an outlier cell must contain pre-defined number of "outlier" genes whose expression in this cell exceed their regular range in this cluster. This range is determined by a background distribution estimated from expression of all genes across cells in the data under the assumption that majority of genes do not exhibit cluster specific expression. Outlier cells with high transcriptome correlations will finally be merged into new clusters as RCPs.

CellSIUS (Wegmann et al., 2019) identifies RCPs by further refining given clusters. Within each cluster, it first identifies candidate genes with significant bimodal-expression (high expression in a small group of cells) and high cluster specificity (high expression only in target cluster). Then, cells are clustered into different subgroups based on gene sets consisted of highly correlated candidate genes selected in the first step. The final cluster assignment is a combination of all subgroups. Thus, new RCPs are sub-clusters of the original major clusters.

SCISSORS (Leary et al., 2021) assumes that a regular one-round clustering is not enough to identify RCPs, since RCP-specific features are often excluded due to a low overall variance in expression. Thus, RCP cells are mis-clustered together with cells from major cell types, which leads to a higher heterogeneity in these clusters. After initial clustering with conserved parameters, SCISSORS evaluates the heterogeneity of clusters with silhouette score. Clusters with high silhouette score are selected for re-clustering. In the re-clustering process, new highly variable genes (HVG) are re-

selected within selected clusters. After testing several user-defined combinations of clustering parameters, SCISSORS picks the one with lowest silhouette score as final output.

RCP is a small group of cells with certain genes highly expressed

There are also some works designing their methods by following the concept of RCP that it is a small group of cells with some specific genes highly expressed compared to cells from other cell types. So, such methods try to solve RCP identification problem by either identifying highly localized genes for clustering or directly identifying small-group cells showing high similarity with each other.

GiniClust3 (Dong and Yuan, 2020) identifies RCP with features selected by Gini index, which was originally developed to study social inequality (Gini, 1912). A high Gini index score indicates the wealth of a country is concentrated by a small number of individuals. Thus, the metric is particularly suitable for identifying rare cell-type-specific genes. Meanwhile, GiniClust3 identifies major cell populations with features selected by Fano factor (Grün et al., 2014). The two types information is combined by weighted consensus clustering algorithm (Li and Ding, 2008) to identify both common and rare cell types simultaneously.

To identify RCP-specific features, CIARA (Lubatti et al., 2022) ranks genes based on their enrichment in local neighborhoods defined from a K-nearest neighbors (KNN) graph. The "top-ranked" genes have the property of being "highly localized" in the gene expression space. A standard clustering algorithm with selected genes can be performed to identify RCP. Alternatively, cells with largest number of highly localized genes expressed in it and its KNN are defined as RCP cells.

SCA (DeMeo and Berger, 2021) is a dimension reduction method projecting the gene expression matrix to a linear subspace spanned by a set of bases vectors called *Shannon components*, which captures cells' variation caused by cell type difference.

These *Shannon components* are right eigenvectors of information score matrix transformed from the original gene expression matrix. In each cell, the score for a gene is the significance of expression difference between the cell's k nearest neighbors and the global background (a set of randomly picked k cells from entire data) measured by negative *log* transformed p-value of Wilcoxon rank-sum test. Higher score means a gene's local expression is far higher than expected by chance, which indicates it is a marker gene for the potential RCP.

GapClust (Fa et al., 2021) is inspired by the observation that a cell's distance (e.g., Euclidean distance) to cells from its same cell type is smaller than its distance to cells from other cell types. So, for a cell, a big difference ("gap") between distance to its $k-1$-th neighbour and distance to k-th neighbour indicates its $k-1$ nearest neighbours may come from its same cell type. In addition, smaller k implies a smaller cell type group (a.k.a RCP). Thus, GapClust first obtains K nearest neighbours for all cells, and confirms potential RCP size by checking the existence of "gap" at k (from 2 to $K-1$). For a candidate RCP size k, cell with the largest "gap" and its $k-1$ nearest neighbours will be considered as a candidate RCP.

MicroCellClust (Gerniers et al., 2021) transforms the RCP identification problem to a max-sum submatrix problem that given a gene expression matrix, simultaneously searching for subset of cells and subset of genes that maximize the sum of expression values within the selected submatrix. The entry of gene expression matrix is log transformed after adding pseudo count 0.1. Thus, the positive value represents the gene is expressed in the cell, while negative value represents genes are negligibly expressed or not expressed at all. Finally, MicroCellClust refines the objective function and adds constraints to search for rare and highly specific patterns of expressions within small subpopulation of cells.

Under different transformations, similar cells always have similar transformed gene expression profiles

Gene expression profiles of cells from same cell type are similar. So, after a transformation of gene expression, cells from same cell type are more likely to stay close with each other (a.k.a. have similar transformed gene expression) than cells from different cell types. After multiple trials of different gene expression transformations, cells share similar transformed profile in most of times have high probability that come from same cell type. Methods like FiRE (Jindal et al., 2018), EDGE (Sun et al., 2020), DoRC (Chen et al., 2019), and scAIDE (Xie et al., 2020) try to solve RCP identification problem by making use of this observation.

FiRE (Jindal et al., 2018) and EDGE (Sun et al., 2020) discover rare cells with Sketching technique (Wang et al., 2007), which randomly projects data points to a low-dimensional bit vector (hash code). Cells with similar gene expression patterns always tend to be projected to same hash code in a low-dimension space determined by randomly sampled genes. Thus, cells within same hash code show similarity in gene expression and populousness of a hash code is a measurement of the rareness of cells in it. After repeated Sketching, FiRE assigns the averaged populousness as robust rareness score to each cell, and EDGE performs dimension reduction on cells' similarity score matrix, in which the similarity score is the estimated probability two cells assigned into same hash code among all repeats.

scAIDE (Xie et al., 2020) first learns an autoencoder to embed the genes into 256 dimensions. After that, a random projection hashing based k-means algorithm is performed to identify RCPs. The random projection hashing technique can help to find suitable initial centers for sample containing cell types with imbalanced size (e.g., existence of RCP) by determining what cells are similar and should be merged to remove imbalance in data.

DoRC (Chen et al., 2019) treats RCP as anomalies in the whole scRNA-seq data.

It discovers rare cells with a method called Isolation Forest (Liu et al., 2008), which is a model-free algorithm widely applied in anomalies detection. In an Isolation Forest, cells are sub-sampled and processed in a tree structure based on random cuts in the values of randomly selected features. RCPs are those cells with the smaller path length in the tree (easy to be isolated). As a result, the aggregated lengths of the tree branches can be viewed as a measure of anomaly/rarity for each cell. To further distinguish cell types from RCP cells, a two-step procedure is proposed, in which Random Forest based similarity learning is first performed and followed by hierarchical clustering.

There are also some methods with novel design that cannot be grouped into any above three categories. For example, SCMER (Liang et al., 2021) hypothesizes that a manifold defined by pairwise cell similarity scores can sufficiently represent the complexity of the data, encoding both the global relationship between cell groups and the local relationship within cell groups. It selects an optimal set of features that can best preserve such a manifold of data. These features can sensitively delineate both common cell lineages and rare cellular states.

RCP identification is imperative for researchers to fully make use of scRNA-seq data, but so many methods may have confused users to choose one to apply. In some above introduced works, only partial methods were compared. Thus, in order to help users correctly choose a suitable tool for RCP analysis, we comprehensively evaluate the performance of above mentioned methods with simulated data. Different metrics are applied to evaluate different aspects of these methods.

4.2 Methods

4.2.1 Data simulation

To completely understand the performance of methods under different scenarios, we first designed a framework to simulate scRNA-seq data, in which the RCP group size, number of RCPs, and cell type specific (CTS) genes number can be well controlled.

Three different relationships between cell types

In the simulated data, cell types could be one of three types based on the their relationships with other cell types: "indep-ct", "sub-ct", and "transit-ct". "Indep-ct" cell types have no relationship with each other that their CTS genes are mutually selected from G genes. A "sub-ct" cell type means it is more similar to one "indep-ct" cell type than to other "indep-ct" cell types. In other words, it is a branch of its related "indep-ct" cell type. A "sub-ct" cell type shares same expression for CTS genes of its related "indep-ct" cell type and has additional smaller number of CTS genes for its own. So, under a lineage tree, a "sub-ct" cell type is under same node with its related "indep-ct" cell type. For example, we can treat CD8 cells is a "sub-ct" to CD4 cells. A "transit-ct" cell type is a cell type in transient state between two cell types. Thus, its CTS genes are same as the two cell types but the corresponding gene expression mean profile is between the two cell types. So, in the simulation process, we first simulate gene expression profiles for "indep-ct" cell types and then for "sub-ct" and "transit-ct" cell types.

Gene expression mean profile determination for "indep-ct" cell types

The data generation process starts with confirming the gene expression mean profile of a base cell type $\boldsymbol{\mu}_0 = (\mu_{10}, ..., \mu_{G0})^T$, where G is the total number of genes. The base cell type profile can be either estimated from a real data or arbitrarily specified

by users. This base cell type only serves for data generation, and does not appear in final data. In this benchmark work, we selected the K562 cell type in scRNA-seq provided in CellSIUS work (Wegmann et al., 2019) as base line cell type.

In next step, for "indep-ct" cell type $k \in \{1, ..., K\}$ in the final data, we determines its gene expression mean profile $\boldsymbol{\mu}_k$ by adding CTS (differential expression) signals $\boldsymbol{\delta}_k$ to the base profile $\boldsymbol{\mu}_0$ on randomly selected D_k genes.

$$\mu_{gk} = \mu_{g0} + \delta_{gk} \tag{4.1}$$

where $g = 1, ..., K$ and g-th gene is the one selected as CTS gene for cell type k. The number of CTS genes for k-th cell type D_k is determined by users. A larger D_k means less similarity (a.k.a. less differential expressed genes) between the cell type k and the base cell type. Usually, CTS genes are assumed to be uniquely high-expressed in target cell type. Thus, in our work, for simplicity, a gene g will be selected as CTS gene for only one cell type. The CTS signal δ_{gk} is related to μ_{g0}:

$$f(\delta_{gk}|\mu_{g0}) = p(\mu_{g0})\phi\left(\delta_{gk};\ h_1(\mu_{g0}), \sigma_1^2\right) + \left[1 - p(\mu_{g0})\right]\phi\left(\delta_{gk};\ h_2(\mu_{g0}), \sigma_2^2\right) \tag{4.2}$$

where $f(\delta_{gk})$ is probability density at point δ_{gk}, $\phi(x; m, \tau^2)$ is probability density at a point of x of normal distribution with mean m and variance τ^2. This Equation 4.2 is applied to capture the real data observation that CTS gene signal δ_{gk} is positively correlated with base cell type gene expression profile μ_{g0} in a non-linear form. In real data, given a certain gene expression μ_{g0}, the CTS signal δ_{gk} could be much larger than other genes with similar expression level. So, in Equation 4.2, the $f_1(\mu_{g0})$ is a non-linear function captures the relationship between CTS signal δ_{gk} and base gene expression μ_{g0} for most "regular" genes, while $f_2(\mu_{g0})$ is another non-linear function with same purpose but for "outlier" genes. $p(\mu_{g0})$ is a function describes how likely a gene shows irregular large CTS signal δ_{gk} given base gene expression μ_{g0}. The $f_1(\cdot)$

function is estimated by locally estimated scatter plot smoothing (LOESS) (Cleveland et al., 1992) with all CTS genes for one cell type. The function $f_2(\cdot)$ is estimated in a same way but with "outlier" genes, whose residuals exceed 95% quantile of a normal distribution with estimated $\hat{f}_1$ as mean and residuals variance $\hat{\sigma}_1^2$. The function $p(\cdot)$ is estimated by the proportion of outliers genes among all genes within each μ_{g0} intervals.

This is a major difference for our simulation method from other simulation methods used in RCP methods evaluation. One strategy is to apply package *Splatter* (Zappia et al., 2017), in which the fold change between two cell types is modeled to follow a log-normal distribution. This is a common way to simulate differential signal between two groups. However, this would lead to that a gene has small expression in base cell type also has small expression in the other cell type, even though the log2 fold change is 1 (e.g., gene expression in cell type A is 1, in cell type B is 2). In reality, some CTS genes in one cell type would have very high expression but no expression in other cell types, such as CD19 in Memory B-cell. At the same time, this strategy may also introduce abnormally large gene expression for one cell type if the expression in base cell type is large. Another strategy. which was used by CellSIUS (Wegmann et al., 2019), is to model the mean difference between two cell types to follow a log normal distribution and assume such CTS signal only appears in genes with extremely low expression in base cell type (less than 0.1). Obviously, such strategy would ignore differential signal for genes with large expression in base cell type.

Gene expression mean profile determination for "sub-ct" and "transit-ct" cell types

For "sub-ct" cell type k, its related "indep-ct" cell type $k^{'}$ is determined by users. The "sub-ct" cell type have same gene expression mean profile for CTS genes for "indep-

ct" cell type $k^{'}$. It also has its own D_k CTS genes that $D_k < D_{k'}$. The procedure to add the CTS signal is same as above described procedure for "indep-ct" cell type.

"Transit-ct" cell type does not have its own CTS genes, but its expression mean profile of CTS genes for its two related cell types is between the expression profile of the two cell types, which is realized by sampling with uniform distribution with left and right bound as the gene expression of the two related cell types.

Gene expression generation for each cell

In scRNA-seq data, due to cells are sequenced one by one, they have different library sizes. So, after deriving the gene expression mean profile for k-th cell type $\boldsymbol{\mu}_k$ ($k = 1, ..., K$), the mean expression of c-th cell from this cell type is scaled by the library size factor:

$$\mu_{gkc} = \mu_{gk} \times s_{kc} \tag{4.3}$$

where μ_{gkc} is the mean expression of g-th gene in c-th cell from k-th cell type, and s_{kc} is the library size factor for c-th cell of k-th cell type. The library size factor is simulated follow a log-normal distribution as used in CellSIUS work (Wegmann et al., 2019):

$$log2(s_{kc}) \sim N(0, 0.5^2) \tag{4.4}$$

where $k = 1, ..., K$ and $c = 1, ..., C_k$.

We assume gene expression of a cell follows a negative binomial distribution:

$$X_{gkc} \sim NB(\mu_{gkc}, \lambda_{gkc}) \tag{4.5}$$

where μ_{gkc} is the mean, $\lambda_{gkc} = \frac{\mu_{gkc}^2}{\sigma_{gkc}^2 - \mu_{gkc}}$ is the dispersion and σ_{gkc}^2 is the variance.

In addition, due to the observation that gene expression variance is correlated with expression mean, a second-order polynomial was fit to the sample variance σ^2_{gkc} as a function of the mean μ_{gkc} in logarithmic space on K562 data as suggested in CellSIUS work (Wegmann et al., 2019). The estimated function is:

$$log2(\sigma^2_{gkc}) = 0.310 + 1.193 \times log2(\mu_{gkc}) + 0.049 \times \left[log2(\mu_{gkc})\right]^2 \tag{4.6}$$

4.2.2 Benchmark and evaluation

In this work, we totally compared twelve methods (GapClust (Andrews et al., 2016), GiniClust3 (Dong and Yuan, 2020), DoRC (Chen et al., 2019), FiRE (Jindal et al., 2018), SCA (DeMeo and Berger, 2021), CellSIUS (Wegmann et al., 2019), RaceID (Grün et al., 2015), EDGE (Sun et al., 2020), SCMER (Liang et al., 2021), scAIDE (Xie et al., 2020), MicroCellClust (Gerniers et al., 2021), and CIARA (Lubatti et al., 2022)) designed for RCP detection with Seurat (Satija et al., 2015), which is one of the most commonly used packages for scRNA-seq analysis. Since these methods are specifically developed for RCP detection, we focuses on the evaluation of their performance in RCP detection. Thus, we want to compare them on simulated data to answer three questions:

1. How is the performance in data with single RCP group? (simplest situation)
2. How is the performance in data with multiple RCP group? (more complex but more realistic situation)
3. What is the computation efficiency?

To answer above questions, we designed different scenarios and used different metrics for evaluation. We selected the K562 cell type data provided in CellSIUS work (Wegmann et al., 2019) as base cell type and estimated its mean expression profile $\boldsymbol{\mu}_0$ and the variance-mean relation shown in Equation 4.6. We also estimated

the h_1, σ_1^2, h_2, σ_2^2, and p in Equation 4.2 by pooling all positively differential expressed gene in other cell types (e.g., Jurkat, A549) compared to K562 together for analysis. The pooling process can increase robustness for estimation by increasing available data.

One RCP group in data

In this scenario, we assigned two major cell types that each has 500 cells and one RCP with size varies between 5, 10, 20 cells in the simulated data. The RCP group can be "indep-ct", "sub-ct", or "transit-ct". There are totally 5000 genes and the number of CTS genes for each major cell type is 200. The number of CTS genes for RCP group varies between 50, 100, 200 for "indep-ct" RCP; between 50, 100 for "sub-ct" RCP. For scenario in which RCP is "transit-ct", the CTS gene number for two major cell types varies between 50, 100, 200. So, for "indep-ct" RCP there are totally nine scenarios (RCP CTS gene number: 50, 100, 200 $\times$ RCP size: 5, 10, 20); for "sub-ct" RCP, there are totally six scenarios (RCP CTS gene number: 50, 100 $\times$ RCP size: 5, 10, 20), for "transit-ct" RCP there are totally nine scenarios (Major CTS gene number: 50, 100, 200 $\times$ RCP size: 5, 10, 20).

The metrics we used is precision, recall, Matthew's correlation coefficient (MCC), and Macro-F1, which are commonly used metrics for binary classification accuracy evaluation. In the comparison process, one major challenge is that the output of the thirteen methods cannot be compared directly that some of them provide RCP cells index directly (e.g., GapClust), some of them only provides embedding or selected features that can help to identify RCPs (e.g., SCA, SCMER), while others provide clustering result without providing specific RCP information for each cell type (e.g., Seurat, SCISSORS).

To make sure these methods generating different types of results can be compared, for methods that are neither provide RCP cell index nor clustering results, we first

used the features or embeddings they provide to generate clustering results. The clustering process follows regular pipeline in Scanpy (Wolf et al., 2018) (python environment) or Seurat (R environment) that selected features are passed for dimension reduction by PCA, and final clustering is performed with Leiden algorithm (Traag et al., 2019) with different resolutions (0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 1.0, 1.2, 1.4). For example, the SCMER provides features that can help to identify RCP cells, then we used the selected features instead of highly variable features (HVG) for dimension reduction and clustering. Secondly, for methods only provide clustering results without specifying the RCP cell index, we used every cluster they generated as a RCP group (keep others as a major group) and calculated corresponding evaluation metrics. The cluster with highest Macro-F1 is finally treated as RCP group generated by the method and its corresponding metrics are kept to represent this method's performance. In addition, if the clustering method was tried with multiple resolutions, then the resolution with largest Macro-F1 will be treated for the method's final result. Most of the methods are applied by following their tutorials and using their recommended parameter values. Some modifications are described in Appendix section C.1.

Multiple RCP groups in data

In this scenario, we wanted to answer following two specific questions:

1. Can methods find all RCP cells from different RCP groups?

2. Can methods distinguish cells from different RCP groups?

We assigned two major cell types that each has 500 cells and three RCP groups in the simulated data. The three different RCP groups have same group size (varies between 5, 10 or 20 cells per group) and CTS gene number (varies between 50, 100 or 200 genes).

To answer the first question, we continue to use the precision, recall, MCC, and Macro-F1 metrics and re-group the three RCP cell types into one "RCP" group and two major cell types into one "major" group (two groups in total). For the second question, we added two metrics normalized mutual information (NMI) and adjusted rand index (ARI) and only re-group the two major cell types into one "major" group (four groups in total). If one method cannot distinguish the three RCP groups, then the NMI and ARI would be lower than the method that can distinguish them well.

Same as above, for methods that are neither provide RCP cell index nor clustering results, we first used the features or embeddings they provide to generate clustering results. The clustering process follows regular pipeline in Scanpy (Wolf et al., 2018) (Python environment) or Seurat (R environment) that selected features are passed for dimension reduction by PCA, and final clustering is performed with Leiden algorithm (Traag et al., 2019) with different resolutions (0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 1.0, 1.2, 1.4).

The difference is that for methods only provide clustering results without RCP cell index, we identify the clusters best matching each RCP group in simulated data (a.k.a cluster with highest Macro-F1). Same cluster could be corresponding to two or three RCP groups. This means all RCP groups in simulated data are in same cluster generated by the method (a.k.a the method cannot distinguish the RCP groups). Next, for each RCP group, its corresponding cluster must be with precision greater than 0.1, otherwise, we thought this RCP cannot be identified by the method and drop it. This is because a low precision means the cluster has too many cells from other cell types and it contains cells from target RCP group may be just due to its large group size. Theoretically, a higher precision threshold should be applied (e.g., 0.5 that more than 50% cells in the cluster come from the target cell type). However, by observation, we found that some method can identify all cells from the three RCP groups but cannot distinguish them. This leads to a cluster containing all RCP cells

and a low precision for each RCP group (precision is about 0.33). To avoid excluding such result, we set a relative smaller threshold 0.1. The final remained clusters are marked as RCP groups for calculating the six metrics as other methods providing RCP cells index information.

So, after above processing, for methods that do not provide RCP cell index, we can pick the clusters mostly consisted of RCP cells as the RCP groups identified by the clustering methods. Then for the first question, we combined the identified RCP clusters into one group, the remained cluster into another group and compared the result with true labels (also merged into two groups). Similarly, for the second question, we kept the RCP clusters and combined the non-RCP clusters into one group and compared the result with true labels (which merged into four groups: one major group + three RCP groups).

Computation efficiency

We compared the computation time of each method by applying them on a data set with 5020 cells (two major cell types with 2500 cells , one RCP group: 20 cells) and 5000 genes. The result is an average of three repeated simulations. The computation environment is Linux system with 2.80 GHz CPU and 100G RAM. To be specific, for methods like scAIDE, which expects GPU for computation, the result in comparison may not be representative and fare for them.

4.3 Results

4.3.1 Synthetic data can well capture differential signal pattern in real data

Modeling the fold change of gene expression between two cell types to follow a log-normal distribution is a commonly used strategy to simulate gene expression profiles

for different cell types, such as *Splatter* (Zappia et al., 2017). The modeling method can easily control differential signal by tuning the mean and variance parameters of the normal distribution part. However, by exploring real data, we found that this way is not perfect to describe the relationship between differential signal of gene and the base gene expression level. In Figure 4.1 (left and middle column), we can observe that the log2 mean expression difference between two cell types in *Splatter* simulated data is highly linear-correlated with log2 mean expression of the base cell type. However, such linear-correlation can hardly be seen in real data (Figure 4.1, right column, PBMC68K and CellSIUS human cell lines) that a gene expression with low expression in base cell type can have very high differential expression in the other cell type. For example, in *Splatter* simulated data, given a gene with mean expression equals to $2^{-5} = 0.03125$, then its differential expression is still around $2^{-5} = 0.03125$ and hardly greater than $2^{0} = 1$. This leads to the expression in the compared cell type is still small ($0.0625 \sim 1.03125$). However, in real data, the differential expression could be greater than 3, which would lead to the expression in compared cell type can be greater than 8. Thus, to make sure the simulated data can well represent real data, a new simulation strategy is desired.

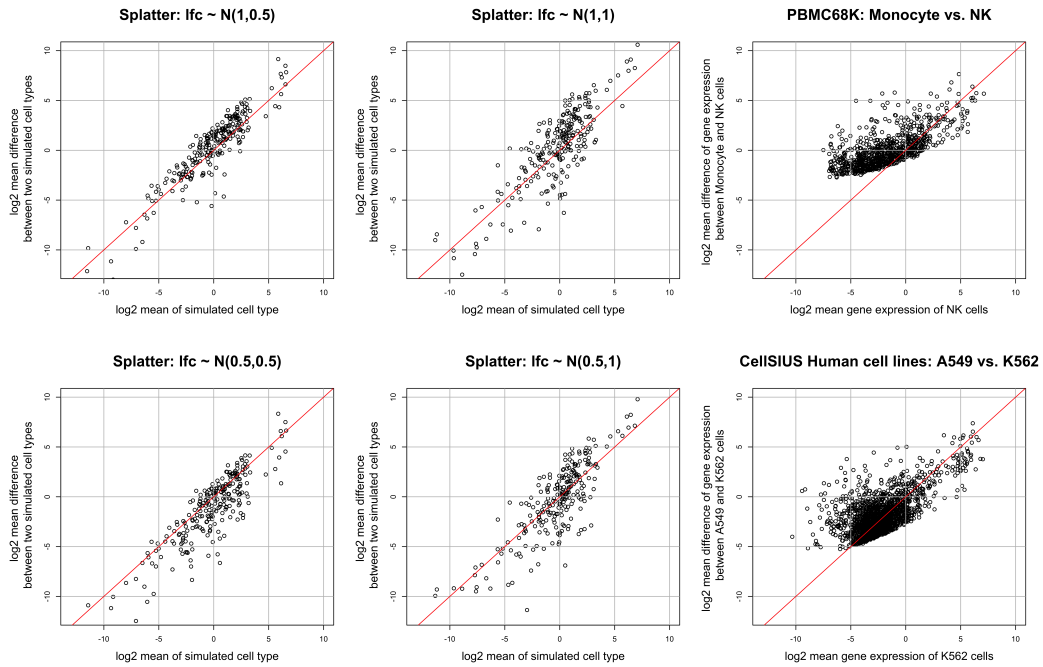

Figure 4.1: Scatter plot showing the relation of differential signal and mean expression in base line cell type in *Splatter* simulated data and real data. The x-axis is the log2 transformed mean expression of CTS genes in base line cell type. The y-axis is the log2 transformed mean expression difference of CTS genes between of the other cell type compared to the base line cell type. A point represents a gene. The left and middle columns show data generated by *Splatter* with different mean and variance of log2 fold change. The right column shows the relationship in real data (up: PBMC-68K, Monocyte vs. NK; down: CellSIUS human cell lines: A549 vs. K562).

In Figure 4.2, we can observe that the relationship between simulated differential signal and mean expression profile of base cell type for CTS genes of the other cell type (Figure 4.2, right panel) is very similar to what it is in the real data (Figure 4.2, left panel). This indicates that the simulation strategy proposed in Methods 4.2.1 can well capture the relationship between differential signal and mean expression of CTS genes, which ensures the simulated data can better serve for RCP identification evaluation, since the differential signal in CTS genes plays a key role in RCP identification.

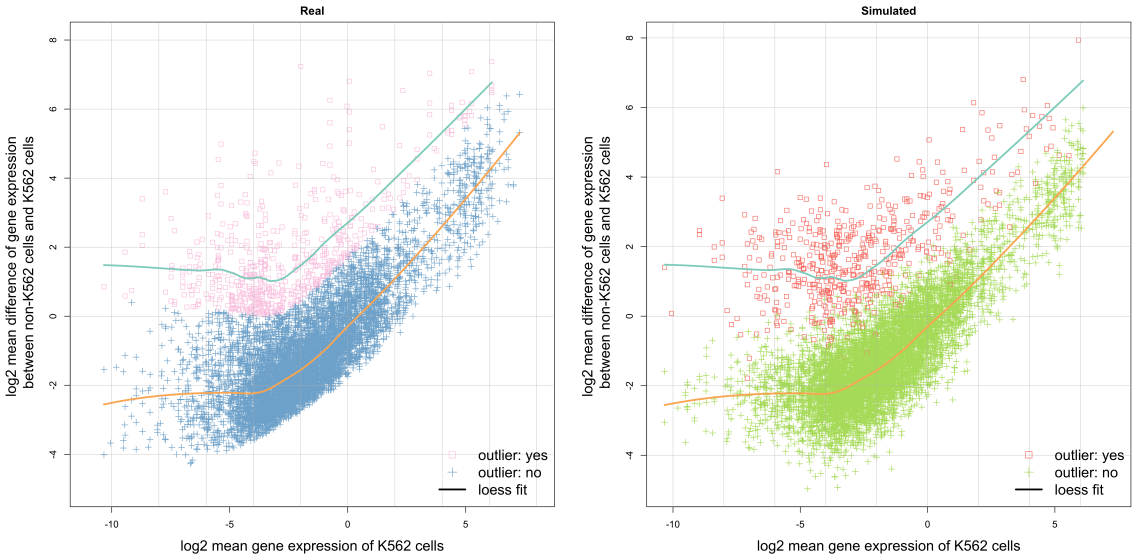

Figure 4.2: Relation between differential signal and mean expression of CTS genes in real data and simulated data.The left panel is from real data - PBMC 68K, where K562 serves as base line cell type. The right panel is from data simulated by proposed strategy. The square/cross points represent the differential signal is "outlier"/"regular" that defined in Methods 4.2.1. The two solid lines depict estimated mean differential signal given mean expression in base line cell type (orange: $h_1(\cdot)$) and (blue: $h_2(\cdot)$) by LOESS fit.

Figure 4.3 shows five simulated cell population (two cell types with large abundance: "major 1" and "major 2"; three RCP groups: "indep-rcp", "transit-rcp" and "sub-rcp") in UMAP (Becht et al., 2019). The "sub-rcp" is a "sub-ct" cell type related to "major 1" and in the UMAP map, it is close to the "major 1" population. The "transit-rcp" is a "transit-ct" cell type related to two large "indep-ct" cell type "major 1" and "major 2". In the UMAP map, the "transit-rcp" is located between the two major cell types; meanwhile, it also has cells blended in the two major cell types. Thus, the proposed simulation strategy can well simulate different relations between cell types.

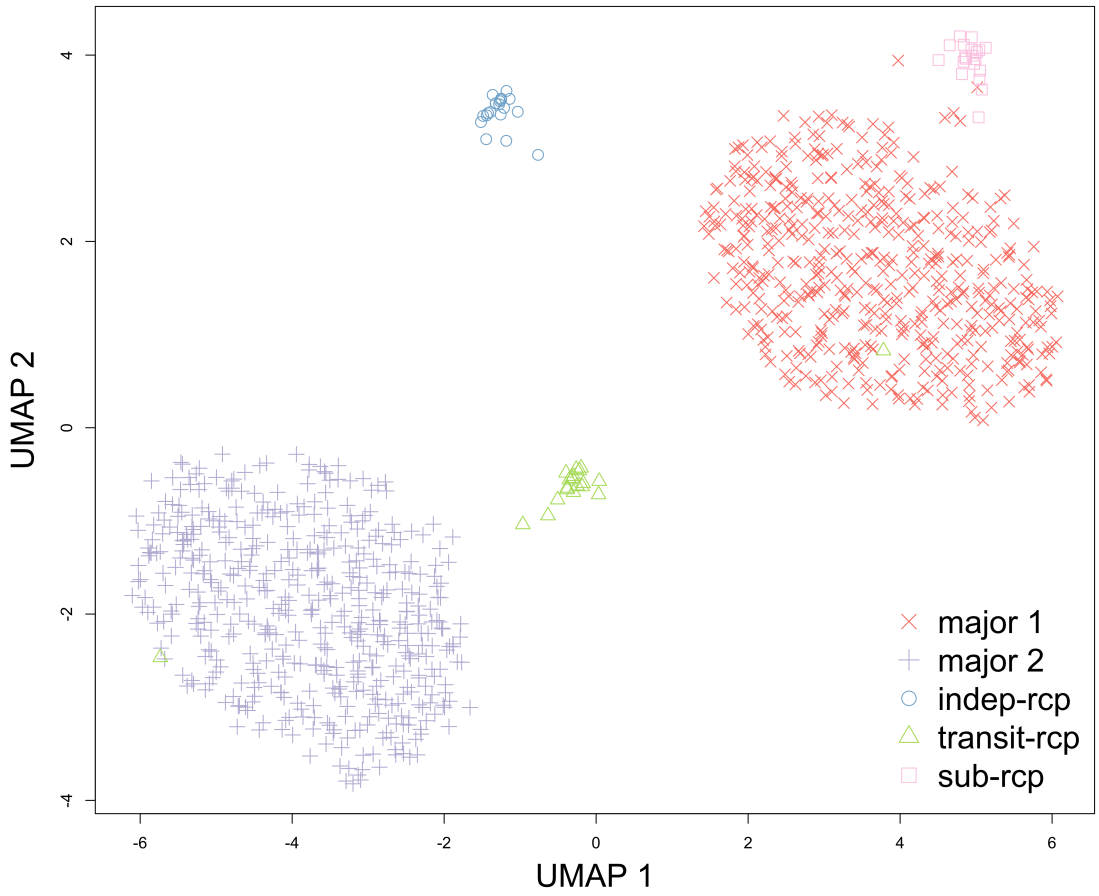

Figure 4.3: Illustration of simulated cell population in UMAP. The "major 1", "major 2" and "indep-rcp" are three "indep-ct" cell types with different abundance. The "sub-rcp" is a "sub-ct" cell type related to "major 1". The "transit-rcp" is a "transit-ct" cell type related to "major 1" and "major 2" cell types.

4.3.2 Performance of methods when only one RCP exists

Overall most methods have their performance decrease with the decrease of RCP size and decrease of CTS gene number in RCP cell type (Figure 4.4). For example, the Macro-F1 of SCMER is 0.81 when RCP size is 20 and CTS gene number is 200. However it drops to 0.205 when RCP size decreases to 20 and CTS gene number is still 200. Similarly, the metric drops to 0.662 when the CTS gene number drops to 50 and RCP size is kept at 20. Such trend follows expectation since more RCP cell number

or CTS gene number means greater signal, which would make the RCP identification much easier for these methods. One exception is MicroCellClust (MCC1) that given CTS gene number, it performs best when RCP size is 10. For example, when CTS gene number is 200, the Macro-F1 of MicroCellClust (MCC1) is 0.429 for RCP size 10, 0.194 for RCP size 20, and 0.253 for RCP size 5.

Even though Seurat is not specifically designed for RCP detection, it surprisingly performs well in identifying RCP with size 20 for all three levels CTS gene number (high: 200, middle: 100, low: 50) that its precision, recall, Macro-F1 and MCC are all greater than 0.9 (Figure 4.4). Seurat also performs well when RCP size is 10 with CTS gene number equal or greater than 100. In the remaining scenarios that either with small RCP size or small CTS gene number, Seurat cannot accurately identify out RCP cells. For example, when RCP size is 5 and CTS gene number is 50, the recall of Seurat is 0.76, but the precision is only 0.114, which implies that the reported cluster contains 76% RCP cells, but meanwhile it also contains many cells from two major cell types. Another surprising finding is that some methods designed for RCP identification performs worse than Seurat in almost all scenarios, such as EDGE, MicroCellClust (MCC1), RaceID, GiniClust3, scAIDE, FiRE, and DoRC. These methods have very low precision values, which indicates that they cannot distinguish the RCP cells from cells of major cell types. For example, in the easiest scenario (RCP size: 20, CTS gene number: 200), recall of EDGE is 0.58, but its precision is 0.484; in contrast, both the recall and precision of Seurat are 1. Methods like CIARA, CellSIUS, and SCMER have similar performance as Seurat that they can identify RCP cells when RCP size or CTS gene number is large enough (e.g., RCP size 20 and CTS gene number 200), but cannot work well when RCP sizes drops to 10 or CTS gene number drops to 50. For example, Macro-F1 of CIARA is 0.997 when RCP size is 20 and CTS gene number is 200; but it drops to 0.53 when RCP size is 5 and CTS gene is still 200.

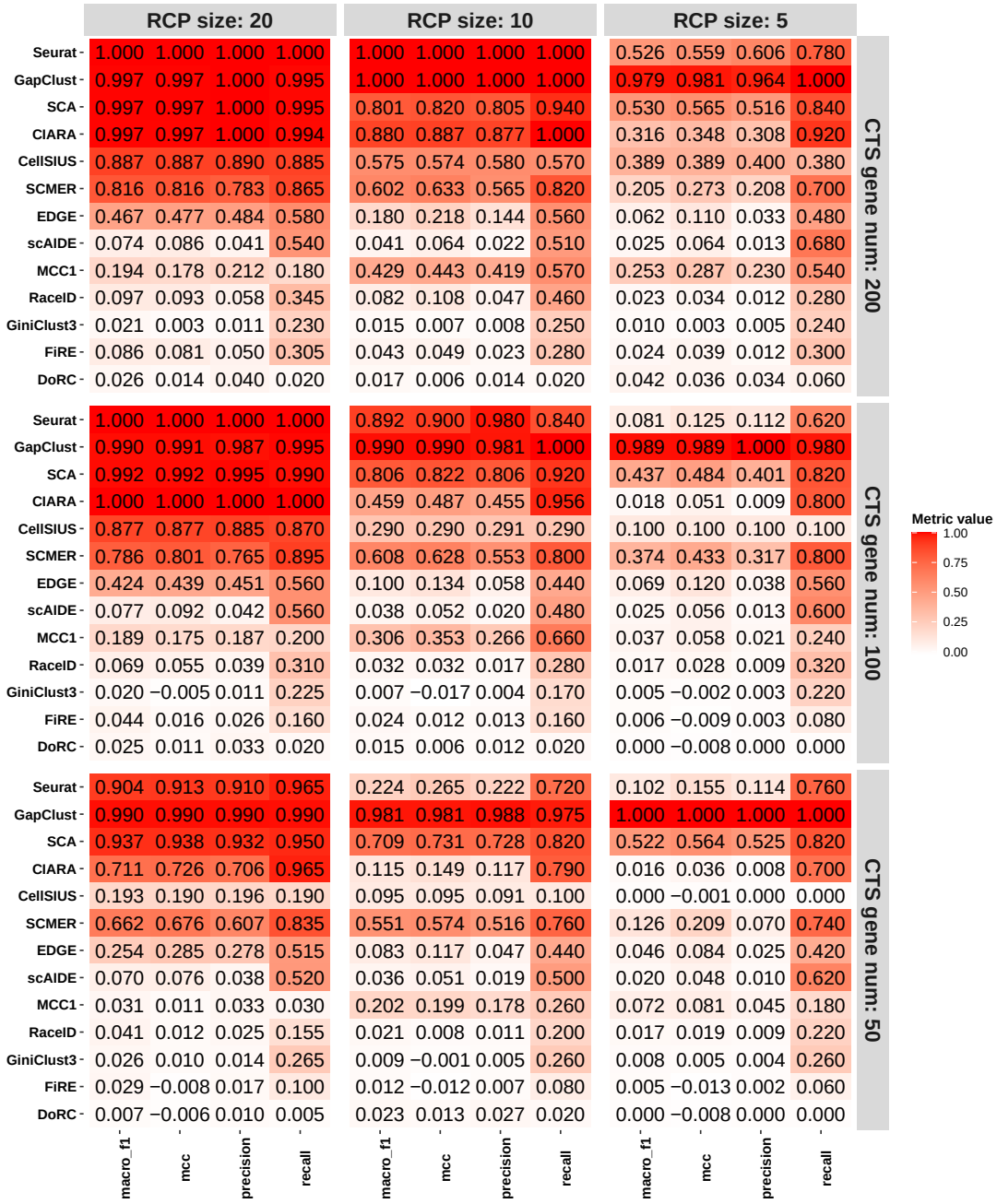

Figure 4.4: Evaluation of methods performance in RCP identification when single RCP group exist. Only one RCP group and two major cell types exist in the data. The cell number in major cell type is 500, and CTS gene number of major cell type is 200. The cell number of RCP varies between 5, 10 20, and the CTS gene number of RCP varies between 50, 100, 200. The result is averaged by 10 simulations

GapClust and SCA are two methods can still accurately identify RCP cells even in scenarios that Seurat cannot work. The recall of GapClust is greater than 0.98

and its precision is greater than 0.96 in all scenarios. SCA has similar performance as GapClust when RCP size is 20. However, when RCP size is less than 20, its precision is lower than GapClust, even though its recall is still high. For example when RCP size is 5 and CTS gene number is 50, the recall of SCA is 0.82 but its precision is only 0.525. This indicates that when SCA can identify four out of five RCP cells, it also falsely report four cells from major cell types as RCP cells.

When the RCP group is a "sub-ct" cell type of "major 1" cell type, only Seurat can accurately identify RCP cells when the RCP size is 20 and CTS gene number is 100 that its accuracy is 0.855 and precision is 1 (Figure C.1). GapClust failed to report any RCP cells (thus not shown in Figure C.1) and SCA's performance is also very poor that the Macro-F1 is only 0.107. CIARA has recall equals to 0.98 and precision equals to 0.514, which means that it can identify most of the RCP cells with some false positive cells. The recall of CellSIUS is 0.595, and precision of CellSIUS is 0.6, which implies CellSIUS can only identify partial RCP cells together with some false positive cells. All other methods cannot identify RCP cells well that their recall is below 0.4 and precision is below 0.1. We can observe similar result when the RCP group is a "transit-ct" cell type between "major 1" and "major 2" that only Seurat can identify most RCP cells accurately (recall: 0.9, precision: 1) when RCP size is 20 and CTS gene number is 200 (C.2). CIARA can also identify most RCP cells (recall: 0.94) but falsely report some cells in major cell types as RCP cells (precision: 0.711). The SCA has a lower recall (0.775) but a slightly higher precision (0.766) than CIARA. Such decrease in performance is due to less and weaker differential signal between RCP cells and major cell types.

In conclusion, when there is only one RCP group ("indep-ct") in data, GapClust is the best choice while SCA is the second choice to replace the use of Seurat. If the RCP is "sub-ct" or "transit-ct", no methods can accurately identify them in different scenarios.

4.3.3 Performance of methods when multiple RCPs exist

In real studies, if the studied tissue has complex composition, there could be multiple RCP groups in the data. Thus, in this section, we expanded the number of RCP groups to three. We expect to evaluate the selected methods in two aspects: (1) whether the method can identify all cells from the RCP groups; (2) whether the method can distinguish the cells from the three RCP groups.

To answer the first question, we combine cells from three RCP groups as one group and cells from two major cell types as one group. Then the question becomes a binary classification problem. We can continue to use the four metrics: Macro-F1, MCC, precision, and recall for evaluation. We can observe that Seurat still has good performance when RCP size is 20 or 10. For example, when RCP size is 10 and CTS gene number is 50, the recall of Seurat is 0.897 and precision is 0.979, which means Seurat can accurately identify about 27 out of 30 RCP cells. When the RCP size is 5 and CTS gene number is 200, recall of Seurat is 0.547 and precision is 1. This indicates Seurat can still identify partial RCP cells. But when the CTS gene number drops to 100 or 50, both recall and precision of Seurat are very low, which means Seurat cannot identify RCP in these two scenarios.

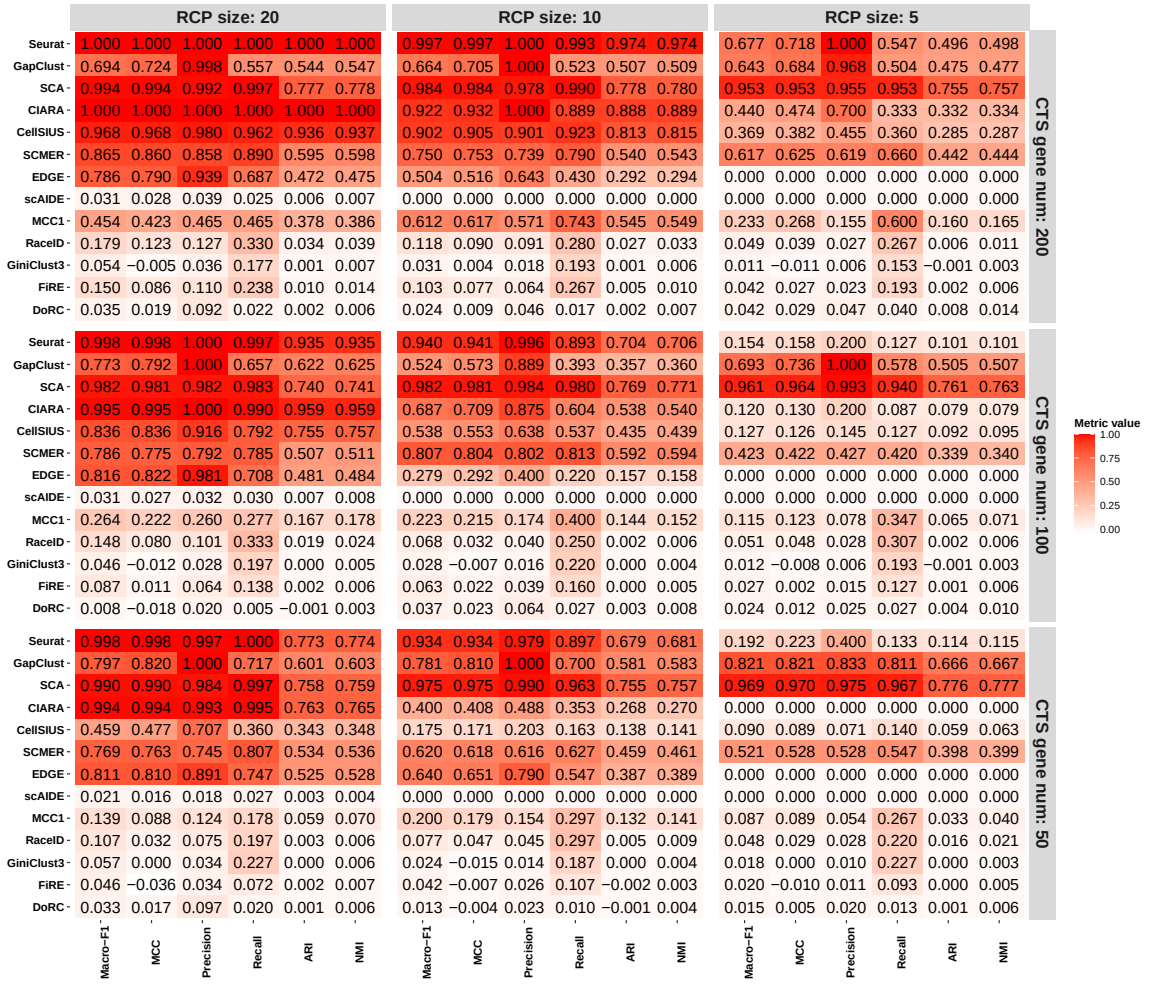

Figure 4.5: Evaluation of methods performance in RCP identification when multiple RCP groups exist. There are three RCP groups with same size and CTS gene number and two major cell types in the data. The cell number in major cell type is 500, and CTS gene number of major cell type is 200. The cell number of RCP groups varies between 5, 10 20, and the CTS gene number of RCP varies between 50, 100, 200. The result is averaged by 10 simulations

Similar as single RCP scenario, methods like EDGE, MicroCellClust (MCC1), RaceID, GiniClust3, scAIDE, FiRE, and DoRC performs worse than Seurat in all scenarios. For example, in the easiest scenario (RCP size: 20, CTS gene number: 200), Macro-F1 is only 0.454, while Seurat is 1. Specifically, for methods EDGE and scAIDE, all metrics are 0 in some scenarios (e.g., RCP size: 5, CTS gene number: 200). This is because the values are arbitrarily assigned as 0, since all clusters they generated have very low precision (smaller than 0.1 defined in Methods 4.2.2) for

any of the three RCP groups. Methods like CIARA, CellSIUS and SCMER can also identify most RCP cells as Seurat in scenarios that RCP size or CTS gene number is large. For example, when RCP size is 10 and CTS gene number is 200, CIARA's recall and precision are 0.889 and 1, CellSIUS's recall and precision are 0.923 and 0.901, SCMER's recall and precision are 0.79 and 0.739. In other scenarios, these two methods' performance decreases that they can only identify partial RCP cells and falsely report cells from major cell type as RCP cells. For example, when RCP size is 5 and CTS gene number is 200, the recall of SCMER is 0.66 and precision is 0.619.

Different from scenarios that only single RCP exists in data, even though the precision of GapClust is still very high, the recall of GapClust ranges from 0.393 to 0.717 when multiple RCP groups in data. This is because GapClust can only accurately identify one or two RCP groups. In contrast, SCA can accurately identify all cells from the three RCP groups in all scenarios that its recall and precision are 0.967 and 0.975 when RCP size is 5 and CTS gene number is 50.

To answer the second question, we combined cells in the two major cell types into one group and keep the three RCP groups. Then we used NMI and ARI to evaluate whether the methods can well distinguish the cells from three RCP groups. Higher NMI and ARI (close to 1) value indicates the methods can well separate cells from the three RCP groups. However, due to the RCP size is much smaller than the major cell types (500 cells per group), when ARI is around or lower than 0.8 , it means cells from the three RCP cannot be separated correctly. From Figure 4.5 we can observe that the ARI and NMI for GapClust and SCA are all lower than 0.8, which means that the two methods cannot correctly separate cells from three RCP groups. Only Seurat and CIARA can successfully distinguish cells from the three RCP groups in following three scenarios: RCP size is 20, CTS gene number is 200; RCP size is 20, CTS gene number is 100; and RCP size is 10, CTS gene number is 200.

Overall, SCA is the first choice to identify RCP cells when there are multiple RCP

groups are in the data. However, further examination needs to be performed for SCA result to separate RCP cells from different group. GapClust is another method in consideration since it can correctly identify partial RCP cells. A potential usage is to apply GapClust repeatedly that in each round remove cells that have been already identified as RCP.

4.3.4 Computation efficiency

We compared the computation time of methods in a simulated data with 5000 cells and 5000 genes. The tasks were run on Linux environment with 2.80 GHz CPU and 100G RAM. The evaluation is for the whole analysis pipeline for each method, which means that for feature-selection or dimension reduction methods, the computation time also includes their downstream analysis steps like clustering with Leiden algorithm. Besides, for method scAIDE, which needs GPU, the computation time for it is not fare since the computation was run only on CPU.

From Figure 4.6 we can observe that GapClust and GiniClust3 are the top two fastest methods that they take less than 1 minute to complete the computation. The following methods are DoRC, FiRE, Seurat, SCA, CellSIUS and RaceID that they take less than 10 min to complete the computation. EDGE, SCMER, and scAIDE take more than 60 minutes to complete the computation. The slowest method is MicroCellClust (MCC1) and CIARA that they take over 400 minutes to complete the computation. Since CIARA performs test for each background-filtered genes, its computation time depends on how many genes remained after filtering. In our evaluation, we make the background filtering threshold less rigorous than default one to make sure less genes will miss the formal test selection step. Thus, the CIARA computation time could be smaller than the time reported in Figure 4.6 if users provide less genes for its test step. Besides, in the second generation of MicroCellClust (MCC2), the authors improved its computation speed. However, this method requires

output of FiRE or DoRC. So, we did not include it for comparison.

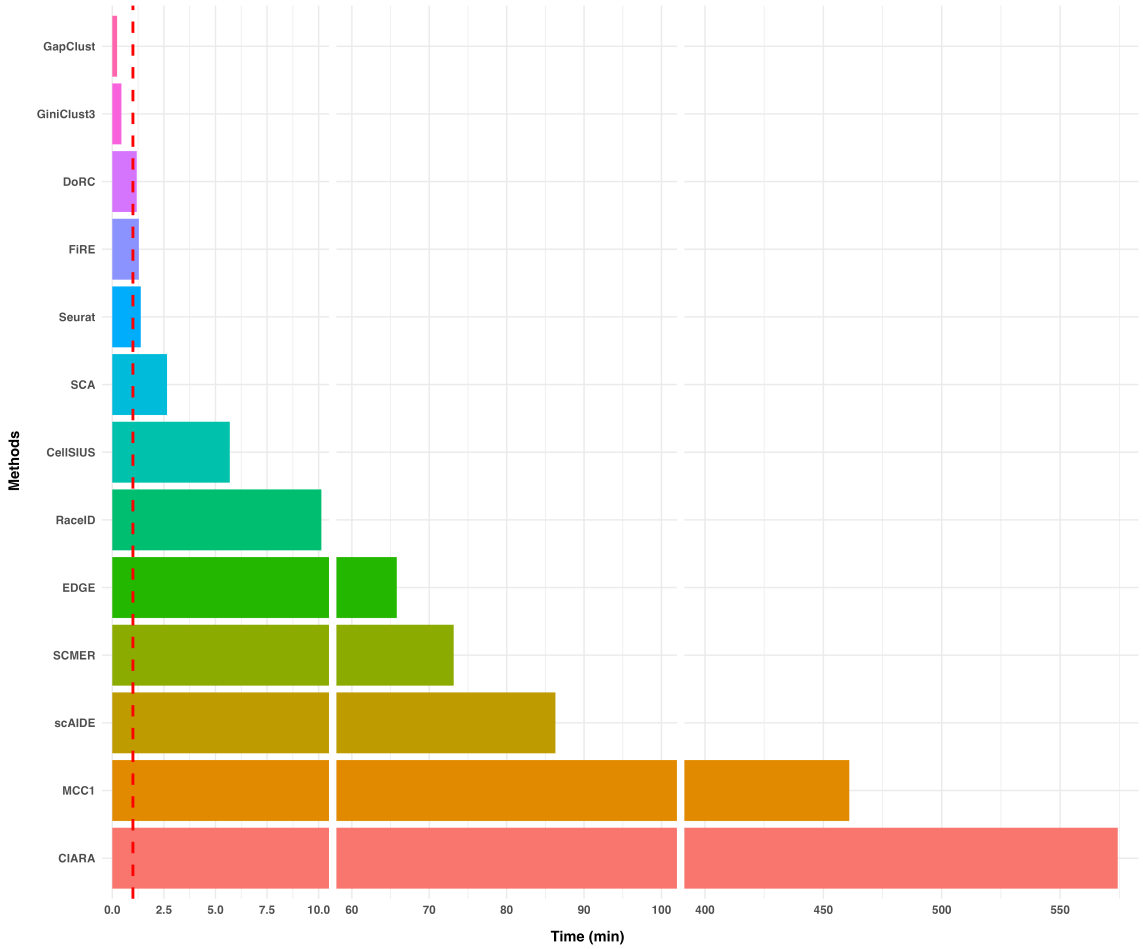

Figure 4.6: Computation time of methods for RCP identification in data with 5000 cells and 5000 genes. The result is an average of three simulations. The red dashed line represents 1 minute. The time unit is minute.

4.4 Discussion

In this work, we evaluated performance and computation time of twelve methods designed for RCP identification from scRNA-seq data and compared them with Seurat, which is one of the most commonly used package for scRNA-seq analysis. We found that even though Seurat is not specifically designed for RCP identification, it can work better than many methods when RCP group size is large enough (e.g., 20 cells). There are two methods that work well in scenarios when Seurat cannot work

- GapClust and SCA. The two methods also have fast computation speed.

One limitation of this work is that all the evaluations are based on simulated data, since there may be some unknown characteristics of real data that can affect RCP identification not considered in our data simulation procedure. Thus, in the future real data should be included to provide more comprehensive evaluation. In addition, in our work only unsupervised methods are included for discussion. There are several supervised methods like CAMLU (Li et al., 2022b) designed for outlier or RCP identification. So, another work direction is to collect these supervised methods and provide comprehensive evaluation for them. Last but not least, with development of scRNA-seq technique, more and more population level scRNA-seq data will be available, thus how these methods' performance after integrating these data into one matrix and what integration methods can best serve for RCP identification task is also an interesting question.

Chapter 5

Summary and future research plan

5.1 Summary

Aiming at deciphering the cell type specific activities from high-throughput omics data, this dissertation contains three parts. In the first part, we reported the observation in real data that DE/DM states in cell types are correlated. Inspired by this observation, we designed a novel statistical model incorporating such correlation with a cell type hierarchy to improve the power and accuracy of cell type specific differential analysis in bulk omics data. The comprehensive simulation analysis demonstrates that our designed model CeDAR has better performance than existing methods (especially for low abundance cell type) when DE/DM state correlation between cell types exists. Meanwhile, the simulation results also demonstrate the robustness of CeDAR to different DE/DM correlation pattern among cell types, estimated proportion and mis-specified tree structures. Better performance of CeDAR compared to other methods is also shown in multiple real data analyses.

In the second part, we reported the observation that cell type specific (CTS) genes may show their DE signal (differential expression between target cell type and all other cell types) inconsistently among subject samples of scRNA-seq data. We designed a hierarchical model to describe the observation. Through this model, we can identify CTS genes, evaluate their consistency among subject samples and measure the DE signal strength at population level. The proposed method was applied on a PBMC data set and it can identify CTS genes with different characteristics. We also compared our method with the strategy performing Wilcoxon rank-sum test on each subject sample, and found that our proposed method can identify some CTS genes that missed by Wilcoxon rank-sum test and filtered some genes that called by Wilcoxon rank-sum test but violates our definition of CTS genes. Lastly, we also designed strategies to make use of the CTS genes information identified with our method from historical data for cell typing. The analysis shows that once the historical data is representative for the target sample, our identified CTS genes can

greatly improve cell typing accuracy than genes identified by Wilcoxon rank-sum test.

In the third part, we collected and summarized twelve unsupervised methods for rare cell population (RCP) identification. We then compared these methods with Seurat, most common used package in scRNA-seq analysis, to answer three interested questions: (1) Can these methods identify RCP cells when only one RCP group exists in data? (2) Can these methods identify and distinguish RCP cells when multiple RCP groups exist in data? (3) How is the computation performance of these methods? All the analyses were performed on simulated data generated by our proposed simulation strategy, which can better mimic the relationship between differential expression signal and baseline expression level. The final result indicates that only SCA and GapClust have better performance than Seurat in extreme scenarios (rare cell population size: 5 cells and cell type specific gene number: 50 genes). Besides, SCA and GapClust also have good computation performance.

5.2 Future research plan

In the future, we first plan to expand our developed methods for more general application. In the first project, we assumed normal distribution for microarray data and derived the likelihood. In biological study, bulk RNA-sequencing data is another popular applied method, in which negative binomial distribution is often applied for statistical analysis. Thus, we will generalize our CeDAR function to allow analysis for both microarray and sequencing data. In the second project, a key implicit assumption is that the population level scRNA-seq data we analyzed are homogeneous. However, in practice, there are many data collecting subject samples with different group information. Thus, we plan to generalize our method with regression model to account for these available group information (confounders). In the last project, we only benchmarked unsupervised methods for RCP identification. There are also

some supervised method designed for this task, such as CAMLU (Li et al., 2022b) and scSynO (Bej et al., 2021). Thus, a complete comparison that includes these methods is expected.

Besides above work extensions, to better understand cell type specificity in biological activities, we plan to develop new models to identify RCP on population-level scRNA-seq data. When we are looking for some RCP in scRNA-seq data under certain condition, we more expect it to appear in a group of subject samples instead of uniquely appearing in one subject sample caused by technical or subject-specific factor. Thus, we want to develop a method to identify two types of RCP from population-level scRNA-seq data simultaneously: pop-RCP, which appears in many subject samples; and subject-RCP, which only appears in one subject sample. The first advantage of this method is that by pooling information from multiple subject sample, we can have greater power to identify RCP with extremely low abundance in single sample (e.g., only 1 or 2 cells in every sample). The second advantage is that the detected pop-RCP means this RCP consistently appear in most subject samples within a studied group, which means its appearance is less likely due to subject-specific factors. Thus, such detected RCP is worth of further studying in a systematic way. The third advantage is that if we can identify subject-RCP, it means some special event is happening in the analyzed subject (e.g., cancer stem cell identified in a subject sample collected from "healthy" population). Such identified subject-RCP can be used for disease diagnosis. During this process, one challenge is that how to make subject samples from different sources comparable.

Appendix A

Appendix for Chapter 2

A.1 Evaluation of CeDAR method

Given simulated mixture gene expression and known cell type proportion in each sample, TOAST was run first to provide cell type specific DE inference result. Then tree structure depicting cell types correlation was estimated by following estimation procedure Methods section 2.2.3 for CeDAR-M based on the TOAST results. Genes with FDR smaller than 0.01 in any cell type were selected for tree structure estimation. For CeDAR-S method, an arbitrarily defined single layer tree was used directly. After deriving estimated tree structure, prior probability on each node of the tree was estimated by following proposed estimation procedures in Method section 2.2.3 with the threshold set as 0.01. In TCA, variances in the model are learned by maximum likelihood estimation. In csSAM, all results are based on 200 permutations. To evaluate the accuracy of each method (CeDAR-S, CeDAR-M, TOAST, TCA, csSAM and CellDMC), the threshold-averaged ROC curve (Fawcett, 2006), area under receiver operating characteristic curve (AUC-ROC), and area under precision-recall curves (AUC-PR) were calculated based on 50 simulations by using R package *ROCR* (Sing et al., 2005). To evaluate the FDR control of these methods, observed FDR was

calculated at cut off (TOAST, TCA, csSAM, CellDMC: estimated FDR; CeDAR-S, CeDAR-M: posterior probability of non-DEG) 0.05. In addition, another metric - Matthews correlation coefficient (MCC), which is used to measure quality of binary classification, was calculated at cut off (TOAST, TCA, csSAM, CellDMC: estimated FDR; CeDAR-S, CeDAR-M: posterior probability of non-DEG) 0.05. Both reported observed FDR and MCC results are average of 50 simulations.

A.2 Cell-type-specific differential methylation in brain

We downloaded the processed Illumina 450k data, which contains both bulk brain tissue samples and pure sorted neuronal and glia samples for a number of individuals with sex information, from GEO with accession number GSE41826 (Guintivano et al., 2013). We used function *dmpFinder* in R package *minfi* (Aryee et al., 2014; Maksimovic et al., 2012; Fortin et al., 2014; Fortin and Hansen, 2015; Fortin et al., 2017; Triche Jr et al., 2013; Andrews et al., 2016) to call DMCs by performing two group comparison between the fourteen healthy males and fifteen healthy females in each cell type based on pure cell type profiles. True DM sites were defined when corresponding FDR is smaller than 0.01; non-DM sites were defined when corresponding FDR is greater than 0.8. The DNA methylation reference used for estimating mixture proportions of bulk samples is the mean profile of each cell type. The top 1000 sites with largest variance among bulk samples were used to estimate mixture proportion with *EpiDISH* (Zheng et al., 2018a,b; Teschendorff et al., 2017; Newman et al., 2015; Houseman et al., 2012) in RPC mode. Threshold used in CeDAR-S to estimate prior probability on each node is $pval = 10^{-5}$. Result of csSAM is based on 200 permutations.

A.3 Cell-type-specific differential methylation in whole blood

We downloaded the processed Illumina MethylationEPIC data with the whole blood profiles, as well as the cell-sorted CD4, CD8, B cells, Monocytes, Granulocytes profiles for 30 individuals (GSE166844 (Hannon et al., 2021)). In the QC step, any site with detection p-value greater than 0.01 in any sample was removed. There are 757,133 sites kept and 45,083 sites removed. We used function *dmpFinder* in R package *minfi* to call DMCs by performing two group comparison between twelve healthy males and eighteen healthy females in each cell type based on pure cell type profiles. True DM sites were defined when corresponding FDR is smaller than 0.01; non-DM sites were defined when corresponding FDR is greater than 0.8. The DNA methylation reference used for estimating mixture proportions of bulk samples is the mean profile of each cell type. Cell type specific markers were selected with two sample t-test by setting target cell type samples as one group, and all other samples corresponding to remaining cell types as the other group. Sites with FDR smaller than 0.05 and beta value with a 0.2 difference greater than any other cell types were selected as markers. We selected 10 markers per cell type with the largest variance among bulk samples to estimate mixture proportion with *EpiDISH* in RPC mode. Then tree structure depicting cell types correlation was estimated by following estimation procedure Methods section 2.2.3 for CeDAR-M based on the TOAST results. Genes with FDR smaller than 0.01 in any cell type were selected for tree structure estimation. Threshold used in CeDAR-S and CeDAR-M to estimate prior probability is $pval = 10^{-5}$. Result of csSAM is based on 200 permutations.

A.4 Cell-type-specific differential methylation in RA EWAS study

We downloaded raw Illumina 450K data with the peripheral blood lymphocytes profile for 332 normal individuals and 354 rheumatoid arthritis (RA) patients (GSE42861 (Liu et al., 2013; Kular et al., 2018)). Any probe with detection p-value greater than the threshold 10^{-16} was treated as missing value. Samples with call rate $<$ 95% and probes with call rate $<$ 90% were excluded. Probes located on chromosome X and chromosome Y were removed. We also dropped probes containing a SNP at the CpG interrogation and/or at the single nucleotide extension. Two samples without smoking status information were removed. Normalization was completed by *Funnorm* method (Fortin et al., 2014) in *minfi*. Missing values were imputed by function *impute.knn* in R package *impute* (Hastie et al., 2021). Finally, beta value was calculated for cell type specific DM analysis. We estimated cell type fractions of six major immune cell types (B cells, CD4, CD8, NK, and Monocytes) by using *EpiDISH* in RCP mode with a DNAm reference consisting of 333 immune cell type-specific DMCs (Teschendorff et al., 2017). In the cell type specific DM analysis, both disease state (RA vs. normal) and age are assumed to have cell type specific effects, while smoking status and gender were treated as global confounders (have same effect on all cell types). In TCA, variances in the model are learned by maximum likelihood estimation. Same as simulation settings, for TOAST, TCA, csSAM and CellDMC, probes with FDR $<$ 0.05 were reported as DMC; for CeDAR-S and CeDAR-M, probes with posterior probability of DM $>$ 0.95 were reported as DMC. Enrichment analysis were performed with *gometh* function in package *missMethyl* (Phipson et al., 2015) for KEGG (Kanehisa and Goto, 2000; Kanehisa, 2019; Kanehisa et al., 2021) pathways.

A.5 Additional real data analysis showing DE/DM state correlations among cell types

We obtained three additional datasets from GEO database, which measure gene expression/DNA methylation profiles of different cell types from samples of different groups. The first data set (GEO accession number GSE149050 (Panwar et al., 2021)) contains gene expression profile (raw counts) from RNA-seq for six major circulating immune cell types (T cells, B cells, polymorphonuclear neutrophils, conventional dendritic cells, plasmacytoid dendritic cells, classical Monocytes) from blood of healthy subjects and Systemic Lupus Erythematosus (SLE) patients with high expressed type I interferon – related genes. The second dataset (GSE59250 (Absher et al., 2013)) contains DNA methylation profiles (normalized beta value) measured by Illumina HumanMethylation450 for cell types (CD4, CD8, and Monocytes) of SLE patients and controls. The third dataset (GSE131525 (Speake et al., 2019)) contains gene expression profile (raw counts) from RNA-seq for cell types (CD4, CD8, B cells and Monocytes) of SLE patients and healthy subjects. For DNA methylation data (GSE59250), we used function *dmpFinder* in R package *minfi* to call DM for SLE vs. control comparison. CpG site with q-value less than 0.05 are deemed differentially methylated sites. For the gene expression data (GSE149050, GSE131525), we used *DEseq2* (Love et al., 2014) to call DE for SLE vs. control comparison. DE genes are defined as genes with false discovery rate (FDR) less than 0.05. Then, we evaluated the pairwise correlation among cell types in terms of their DE/DM status, using both Pearson correlation coefficient (PCC) of log transformed p-values from the DE/DM tests for all features, and the odds ratio (OR) of being DE/DM from the cell types.

The pairwise scatter plots for the comparisons are shown in Figure A.1. In data GSE149050 (Figure A.1(a)), the p-values from all cell types are statistically significant positive that the smallest PCC is 0.36 between B cells and polymorphonuclear

neutrophils (PMN) and the largest PCC is 0.63 between classical Monocytes (cMo) and conventional dendritic cells (cDC)/PMN. Besides, the ORs for being DE between any two cell types are also statistically greater than 1 that smallest OR is 3.0 between B cells and PMN and the largest OR is 24 between classical Monocytes and plasmacytoid dendritic cells. In data GSE59250 (Figure A.1(b)), even though the PCCs are smaller than those in GSE149050 that the largest value is 0.25 between CD4 and B cells, the ORs are all statistically significantly greater than 1 that smallest value is 26 between CD4 and Monocytes, and largest value is 180 between B cells and Monocytes. The results of the two data sets indicate existence of DE/DM state correlation among cell types. In addition, in data GSE131525 (Figure A.1(c)), we can observe that between CD4 and CD8 both PCC (0.65) and OR (37) are greatly larger than other pairs of cell types (remaining largest PCC is 0.35, largest OR is 7.1), which implies a cell type hierarchy of DE/DM state correlation. Overall, these results demonstrate that there are strong correlations among cell types in terms of their DE/DM status.

A.6 Additional simulation analysis evaluating impact of data noise on observed FDR for CeDAR method

To illustrate the effect of data noise on observed FDR from CeDAR, we performed simulation with different data noise levels (extremely low: 0.01, low: 0.1, normal: 1, high: 2). In the settings, normal level (noise level 1) is the setting we used in our reported simulations. We modify the noise level by multiplying 0.01, 0.1 or 2 to the standard deviation of both cell type specific gene expression and bulk gene expression.

We first performed the simulation on two cell types with proportion ratio 9 : 1.

As can be seen from Table A.5, when noise level is low (0.01, 0.1), the FDR of cell type 2 with true prior is still smaller with estimated prior (0.024 vs. 0.039, 0.025 vs. 0.037). But when noise level is larger (1, 2), we can observe larger FDR in cell type 2 with true prior (0.083 vs. 0.047, 0.225 vs. 0.126).

We then performed the simulation on six cell types with true/estimated prior probability and tree structure on the four different noise levels (extremely low: 0.01, low: 0.1, normal: 1, high: 2). Same conclusion can be derived from Table A.6 that when data noise is small (noise level 0.01, 0.1), CeDAR with true prior probability has lower FDR than CeDAR with estimated prior probability (e.g., in cell type 2, 0.066 vs. 0.089, 0.069 vs. 0.080). When data noise is larger (noise level 1, 2), CeDAR with true prior probability has higher FDR than CeDAR with estimated prior probability (e.g., in cell type 2, 0.165 vs. 0.073, 0.345 vs. 0.206).

Overall, the FDR difference between CeDAR with true prior probability and estimated probability is related with data noise. When data noise is large, CeDAR with estimated prior prob has smaller FDR and when data noise is small, CeDAR with true prior prob has smaller FDR.

A.7 Additional simulation analysis evaluating impact of mis-specified tree structures as input of CeDAR-M

To evaluate impact of mis-specified tree structure as input for CeDAR-M, we designed additional simulation with either correct or mis-specified tree structure as input for CeDAR-M. Correct tree structure means applying the tree structure generating DE state in simulation data as input of CeDAR-M, while mis-specified tree structure means applying tree structures with cell types arbitrarily switched under nodes. The

simulation is performed with six cell types under different sample sizes per group (50, 100, 200). The evaluation process is similar as process described in Section A.1, except that all the tree structures are pre-specified without estimation.

We performed simulation with six cell types (proportions of cell type 1-6: 0.63, 0.10, 0.11, 0.06, 0.06 and 0.05). The "correct" tree structure and "mis-specified" tree structures are shown in the top row of Figure A.6 and A.7 (Correct: "tree 1"; Mis-specified: "tree 2", "tree 3", "tree 4" and "tree 5"). In "correct" tree structure, cell type 1 and 2 are set under same node, while cell type 4, 5, 6 are set under same node with cell type 3 but with different DE state correlation level. In "mis-specified" tree structures, we switch cell type 2 with cell type 3/4 ("tree 2"/ "tree 5") to check impact of a cell type mis-clustered with small proportion cell types Besides, we also switch cell type 4 with cell type 5/6, which decreases DE state correlation between cell type 4 and cell type 3. Such mis-specification is common during estimation process because small proportion providing less information for accurate clustering.

The simulation result (Figure A.6, A.7, and Table A.7) shows that using "mis-specified" tree structures as input of CeDAR-M has small impact on csDE inference compared to using "correct" tree structure. For cell types with large proportion (e.g., cell type 1 with mean proportion 0.63), we can barely observe difference of ROC curves and box plot of observed FDR between correct and mis-specified tree structures. When cell type 1 is clustered with cell type 3 or cell type 4 ("tree 2"/ "tree5"), compared to "correct" tree structure the decrease of AUC-ROC, AUC-PR, MCC and increase of observed FDR are extremely small. For example, with sample size 100 per group, the AUC-ROC for cell type 1 with "tree 1" vs."tree 2" is 0.989 vs. 0.987, and the observed FDR is 0.068 vs. 0.073. For small proportion cell types that are mis-clustered with other weak correlated cell types that have small proportions (e.g., cell type 2 in "tree 2" and "tree 5", cell type 4 in "tree 2"), we can observe decrease of AUC-ROC and inflation of observed FDR compared to the result with

"tree 1". For example, with sample size 100 per group, the AUC-ROC for cell type 2 with "tree 1" vs. "tree 2" is 0.919 vs. 0.871, and the observed FDR is 0.070 vs. 0.102. Such change is because cell types with relatively small proportions (cell type 4, 5, 6) cannot provide accurate information as cell type 1. Besides, for cell type 2, "tree 3" and "tree 4" have similar performance in AUC-ROC and observed FDR as "tree 1", which indicates that when cell type 2 is correctly clustered with large proportion cell type 1, the mis-specified tree structure in other sibling nodes have little impact on it. For small proportion cell types that are mis-clustered with other cell types under same non-root node (e.g., cell type 4, 6 in "tree 3" and cell type 4, 5 in "tree 4"), we can observe that the change of AUC-ROC and observed FDR is small compared to "tree 1". For example, with sample size 100 per group, the AUC-ROC for cell type 4 with "tree 1" vs. "tree 4" is 0.850. vs. 0.847, and for cell type 5 is 0.829 vs. 0.830; the observed FDR for cell type 4 with "tree 1" vs. "tree 4" is 0.097 vs. 0.092, and for cell type 5 is 0.129 vs. 0.138. In addition, with increasing sample size, CeDAR-M performance with "mis-specified" tree can improve. For example, from sample size 50 to 200, the AUC-ROC of cell type 2 in "tree 2" increases from 0.852 to 0.894 and the observed FDR decreases from 0.175 to 0.075. Overall, using "mis-specified" tree structure as input has little impact on cell types with large proportion, cell types that are correctly clustered in sibling nodes, or cell types that are mis-clustered with other cell types in same non-root node. The main impact of "mis-specified" tree structure (decrease of AUC-ROC, inflation of observed FDR) is observed for small proportion cell types that are clustered with other weak correlated (under different non-root node) cell types with small proportion. Meanwhile, with increasing sample size, the performance of CeDAR-M with "mis-specified" tree structure can be improved.

A.8 Additional real data analyses

We applied CeDAR for three more real data analyses and compared it with other methods (TOAST, TCA, csSAM and CellDMC). The first two analyses were performed separately on Down syndrome (DS) methylation data (GSE74486 (Mendioroz et al., 2015)) and Systemic Lupus Erythematosus (SLE) methylation data (GSE118144 (Yeung et al., 2019)), which contain both bulk samples and pure cell type samples. We identified cell type specific differential methylation sites from pure cell type samples and use them as gold standard to benchmark the result of csDM analysis on bulk samples. In the third analysis, we performed csDM analysis on two DNA methylation data (GSE42861 and GSE40279 (Hannum et al., 2013)) and examined whether methods in comparison can identify seven reported smoking associated cell type specific probes.

A.8.1 Cell-type-specific differential methylation in Down syndrome study

The DS methylation data (GSE74886) contains both bulk samples of frontal cortex grey matter (14 DS vs. 8 normal) and pure cell type samples of glia and neuron cells derived by FACS from DS subjects and healthy control subjects. We first performed two-group comparison (DS vs. normal) separately for glia and neuron samples to identify csDMCs serving for gold standard by using dmpFinder function in *minfi* package with default settings. We defined sites with FDR < 0.01 as true DM; FDR > 0.8 as non-DM in the two cell types. Among all 390,089 sites, there are 8099 and 12,438 true DM sites identified in glia and neuron respectively. The two cell types share 1284 common true DM sites. We estimated the mixture proportions for each bulk sample by using EpiDISH with RPC-mode, in which the mean profile of each cell type is used as reference and top 1000 sites with largest variance among bulk samples

were used for deconvolution. The estimated mixture proportions and the whole-tissue DNA methylation data were used as inputs for TOAST, TCA, csSAM, CellDMC and CeDAR-S. Result of csSAM is generated based on 200 permutations. Threshold used in CeDAR-S to estimate prior probability on each node is $pval = 10^{-5}$. Accuracy was measured by true discovery rate (TDR) in top ranked sites. The TDR curves in Figure A.10 show that CeDAR-S has significantly higher accuracy among the top CpG sites than all other methods in both glia and neuron that the differences of TDR between CeDAR-S and TOAST among top ranked 5000 sites in both cell types are more than 20%.

A.8.2 Cell-type-specific differential methylation in Systemic Lupus Erythematosus study

The SLE methylation data (GSE118114) contains both bulk samples of whole blood (16 SLE vs. 13 normal) and pure cell type samples of neutrophils, CD8, CD4, and B cells from SLE patients and healthy control subjects. We performed two-group comparison (SLE vs. normal) separately for neutrophils, CD8, CD4 and B cells to identify csDMCs serving for gold standard by using *dmpFinder* function in *minfi* package with default settings. We defined sites with FDR < 0.01 as true DM; FDR > 0.8 as non-DM in the four cell types. Among all the 662,741 sites, there are 5425 (neutrophils), 6886 (CD4), 59 (CD8), 25 (B cells) true DM sites identified. We estimated the mixture proportions for each bulk sample by using *EpiDISH* with RPC-mode, in which the reference is a DNAm reference consisting of 333 immune cell type-specific DMCs (Teschendorff et al., 2017) and sites in both reference data and bulk samples were kept for deconvolution. The estimated mixture proportions and the whole-tissue DNA methylation data were used as inputs for TOAST, TCA, csSAM, CellDMC, CeDAR-S and CeDAR-M. Result of csSAM is generated based on 200 permutations. Sites with p-value smaller than 0.01 in any cell type were selected for tree structure

estimation. Threshold used in CeDAR-S and CeDAR-M to estimate prior probability on each node is $pval = 10^{-5}$. Accuracy was measured by true discovery rate (TDR) in top ranked sites. The TDR curves in Figure A.11 show that both CeDAR-S and CeDAR-M have higher accuracies than all other methods. For example, in cell type neutrophils, which has largest mean proportion (0.67), we can see that CeDAR-S and CeDAR-M have higher TDR curve than other methods. Meanwhile, in low abundant cell types, like CD4 (mean proportion: 0.064), the performances of all methods are not good, but only CeDAR-S and CeDAR-M can identify some true DM sites among top ranked 5000 sites. This can also be observed in cell type CD8 and B cells, which only have 59 and 25 true DM sites respectively.

A.8.3 Cell-type-specific differential methylation analysis for smoking associated DNA methylation sites

Su et al. (2016) reported five smoking associated Myeloid-specific DM sites, which are cg05575921, cg21566642, cg09935388, cg06126421, and cg03636183; and two smoking associated Lymphoid-specific DM sites, which are cg19859270 and cg09099830. We performed csDM analysis on two DNA methylation data (Liu's data: GSE42861 and Hannum's data: GSE40279) to check whether CeDAR and other methods can identify these csDMCs. In the analysis we compare CeDAR-S with TOAST, TCA, csSAM and CellDMC.

For Liu's data, after preprocessing described in Section A.5, proportions of seven blood cell types (B cells, CD4, CD8, NK, Monocytes, neutrophils, and eosinophils) were first estimated by *EpiDISH* with RPC-mode, which using DNAm reference consisting of 333 immune cell type specific DMCs as reference. Then proportion of lymphoid is the summation of estimated proportions of B cells, CD4, CD8, and NK cells. Similarly, proportion of Myeloid is the summation of estimated proportions of Monocytes, neutrophils, and eosinophils. We defined smoking status as

binary variable (never vs. smoke) that never-smokers are in "never" group, while ex-smokers, occasional-smokers and current-smokers are in "smoke" group. In the cell type specific DM (csDM) analysis, disease state, age, and smoking status are assumed to have cell type specific effect, and gender is treated as global confounder (have same effect on all cell types). For Hannum's data, pre-processed data was derived online from figshare (`https://figshare.com/articles/online_resource/CompCellDMCtoTCA/12922322/1`) and the proportion estimation process is same as Liu's data. Similarly, we defined smoking status as binary variable (never vs. smoke) that never-smokers are in "never" group, while ex-smokers and current-smokers are in "smoke" group. In the csDM analysis, age and smoking status are assumed to have cell type specific effect, and plate is treated as global confounder. In both analyses, threshold used in CeDAR-S to estimate prior probability on each node is $pval = 10^{-5}$. Same as simulation settings, for TOAST, TCA, csSAM and CellDMC, probes with FDR < 0.05 were reported as DMC; for CeDAR-S, probes with posterior probability of DM > 0.95 were reported as DMC. As can be seen from Figure A.12, CeDAR-S can identify more smoking associated DNA methylation sites reported by Su et al. than other four methods in both Liu's data and Hannum's data. In Liu's data, while TOAST, TCA, csSAM and CellDMC can identify four myeloid-specific sites (cg05575921, cg21566642, cg06126421, and cg03636183) but zero lymphoid-specific sites, CeDAR-S can identify all myeloid-specific sites and one more lymphoid-specific site (cg19859270). Meanwhile, CeDAR-S identified a myeloid-specific site (cg03636183) in lymphoid cells. In Hannum's data, while TOAST, TCA and CellDMC can only identify one myeloid-specific site (cg05575921) and csSAM cannot identify any site, CeDAR-S can identify four out of five myeloid-specific sites (cg05575921, cg09935388, cg06126421, and cg03636183) and one lymphoid-specific site (cg19859270) with one myeloid-specific (cg21566642) site identified in lymphoid cells.

Overall, all the three analyses demonstrate that incorporating DM state correlation among cell types can improve accuracy and power in csDM analysis.

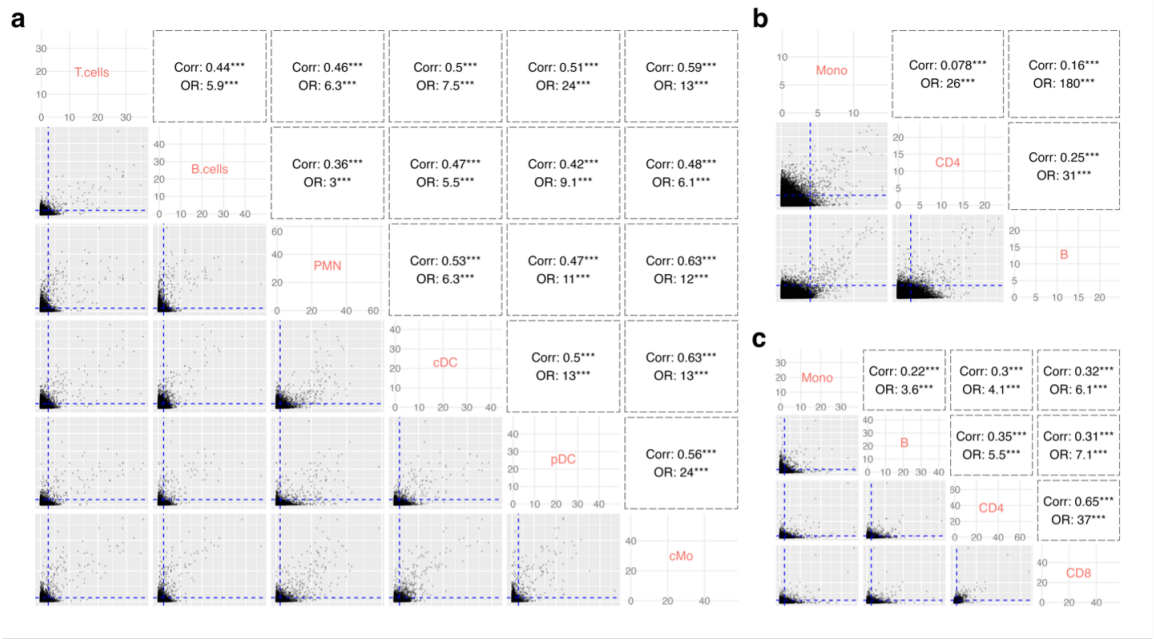

Figure A.1: Correlations among cell types from cell type specific differential analysis. (a) cell type specific differential expression analysis on data GSE149050 (healthy controls vs. SLE patients with high expressed type I interferon - related genes) in major circulating immune cell types (T cells, B cells, Polymorphonuclear Neutrophils (PMNs), conventional dendritic cells (cDC), plasmacytoid dendritic cells (pDC), classical Monocytes (cMo)); (b) cell type specific differential methylation analysis on data GSE59250 (lupus patients vs. controls) in cell types (CD14 Monocytes, CD4, and B cells); (c) cell type specific differential expression analysis on data GSE131525 (SLE patients vs. healthy controls) in cell types (Monocytes, CD8, CD4, and B cells). DE/DM tests were applied for each feature in each cell type. X-axis and Y-axis represent $-log10$ transformed p-value from DE/DM tests in corresponding cell types. Each point represents a gene or CpG site. Dashed blue lines represent the thresholds used to define DEG/DMC in each cell type. Pearson correlation coefficients (PCC) of transformed p-values and odds ratio (OR) of differential state are tested for their significance. $***$ represents p-value < 0.01.

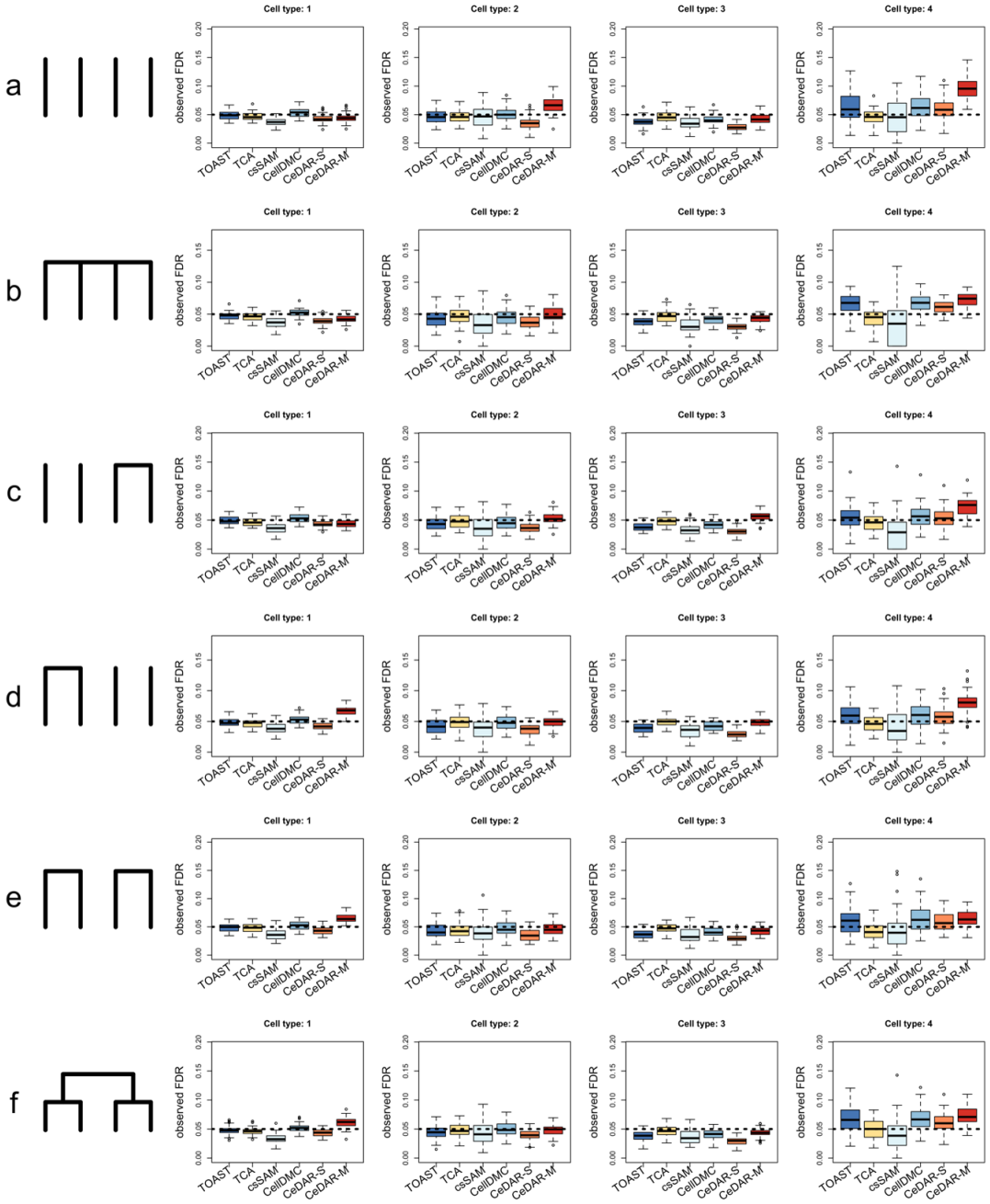

Figure A.2: Observed FDR under different DE patterns (strong correlation). DE genes were defined with rule: FDR < 0.05 (TOAST, TCA, csSAM, CellDMC); posterior probability of DE > 0.95 (CeDAR-M, CeDAR-S). Observed FDR of 50 simulations were summarized in box plot.

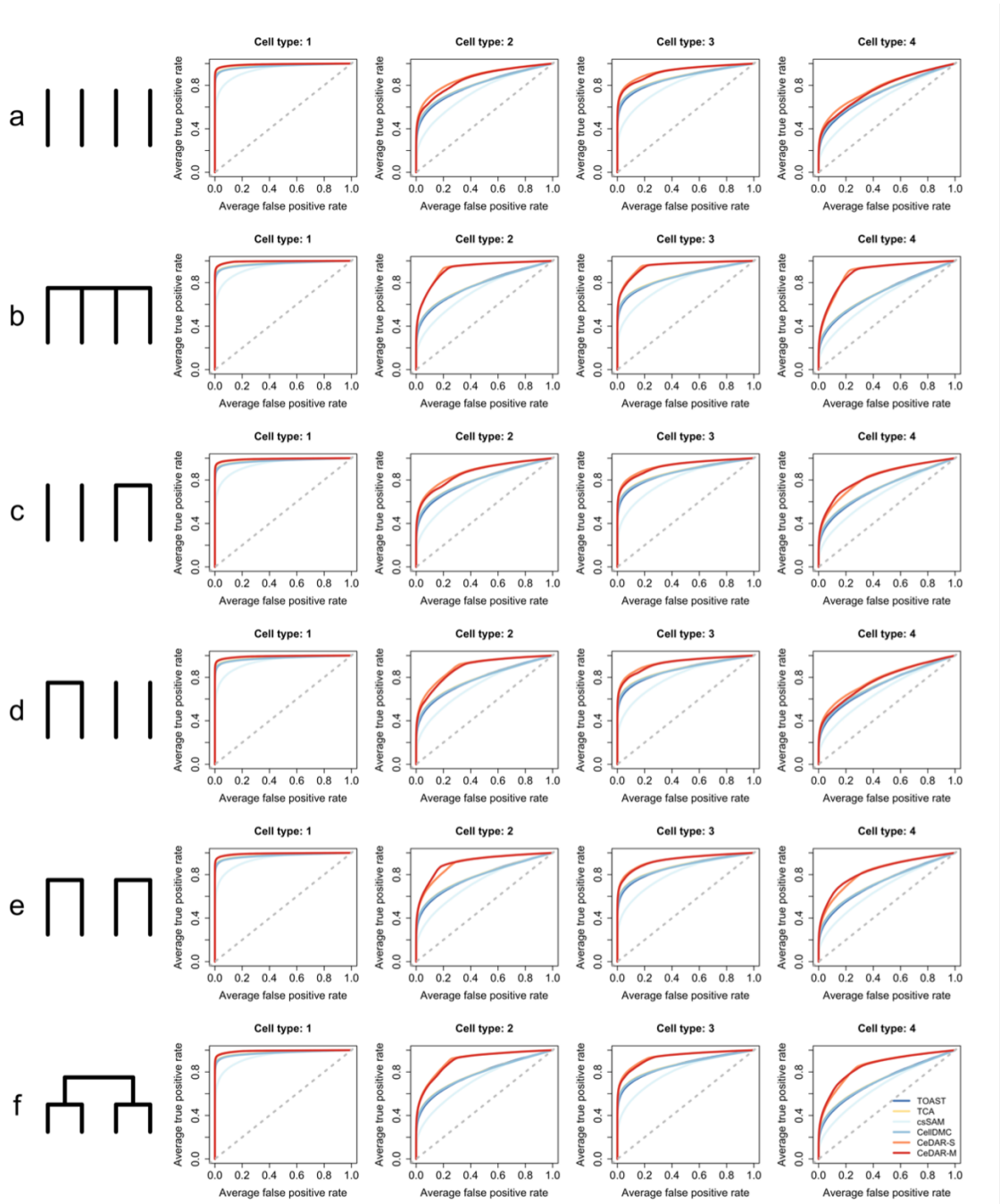

Figure A.3: ROC curves under different DE patterns (weak correlation). The simulation mimics a two-group comparison based on bulk microarray gene expression - a mixture of four common blood immune cell types (1: Neutrophils, 2: Monocytes, 3: CD4, 4: CD8) under six different DE patterns: (a) all cell types are independent; (b) all cell types are correlated under a single layer tree structure; (c) only cell types 3 and 4 are correlated; (d) only cell types 1 and 2 are correlated; (e) cell types 1 and 2 are correlated, and cell types 3 and 4 are correlated; (f) all cell types are correlated under a multiple-layer tree structure). Methods under comparison include TOAST, TCA, csSAM, CellDMC, CeDAR-S and CeDAR-M. Reported ROC curves are average results from 50 simulations.

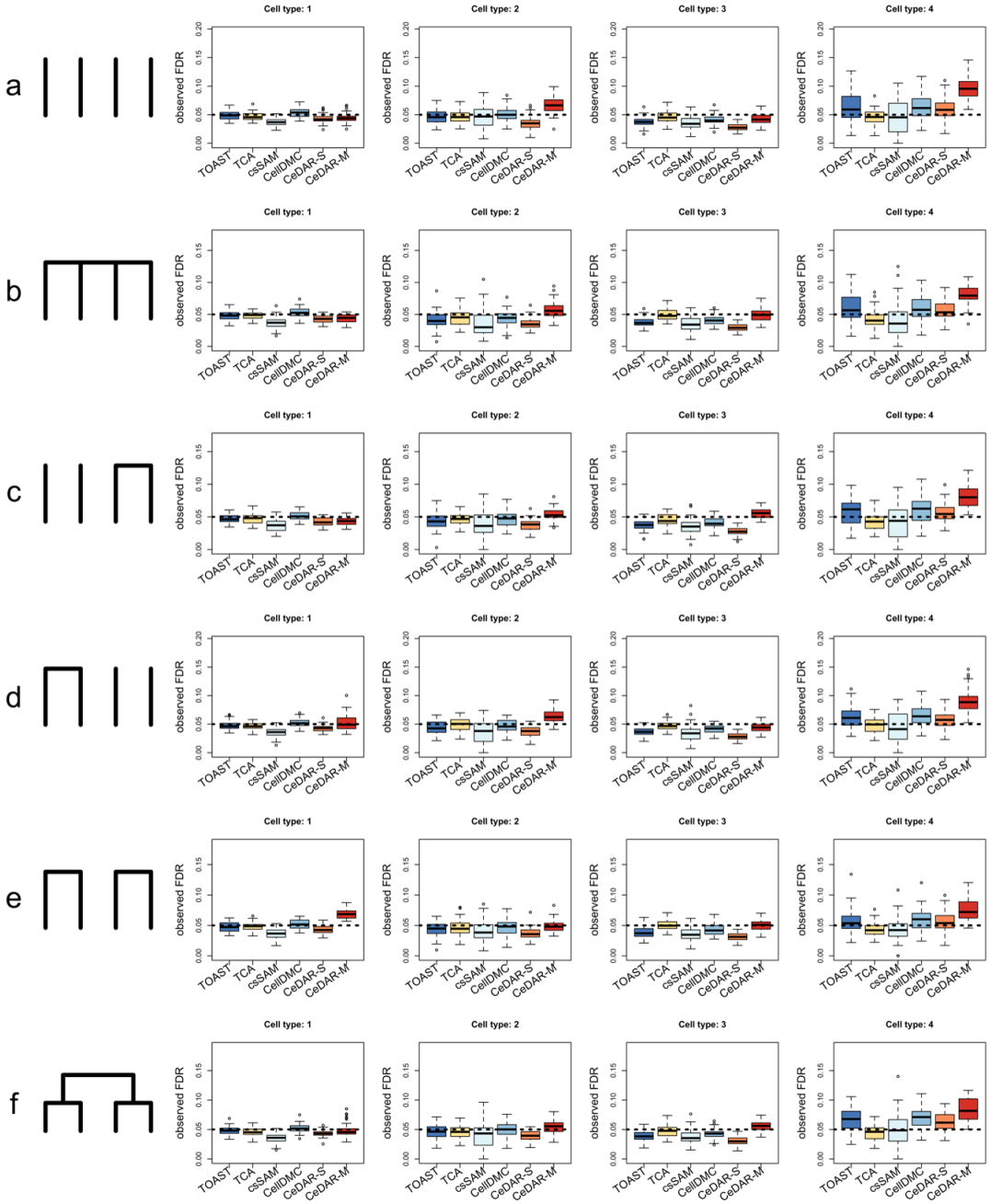

Figure A.4: Observed FDR under different DE patterns (weak correlation). DE genes were defined with rule: FDR < 0.05 (TOAST, TCA, csSAM and CellDMC); posterior probability of DE > 0.95 (CeDAR-M, CeDAR-S). Observed FDR of 50 simulations were summarized in box plot.

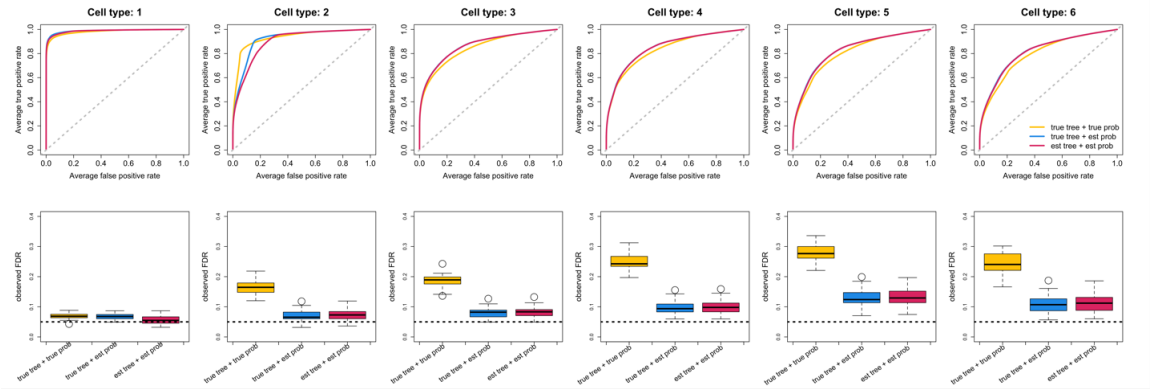

Figure A.5: Evaluation of effect on csDE detection performance by using estimated tree structure and estimated prior probability for each node on estimated tree. The upper panel shows ROC curves of csDE analysis by CeDAR-M with true tree + true prior probability (gold), true tree + estimated prior probability (blue), and estimated tree + estimated prior probability (red). The lower panel shows observed FDR of using true/estimated tree structures and prior probabilities. DE genes were defined with rule: posterior probability of DE > 0.95. Observed FDR of 50 simulations were summarized in box plot.

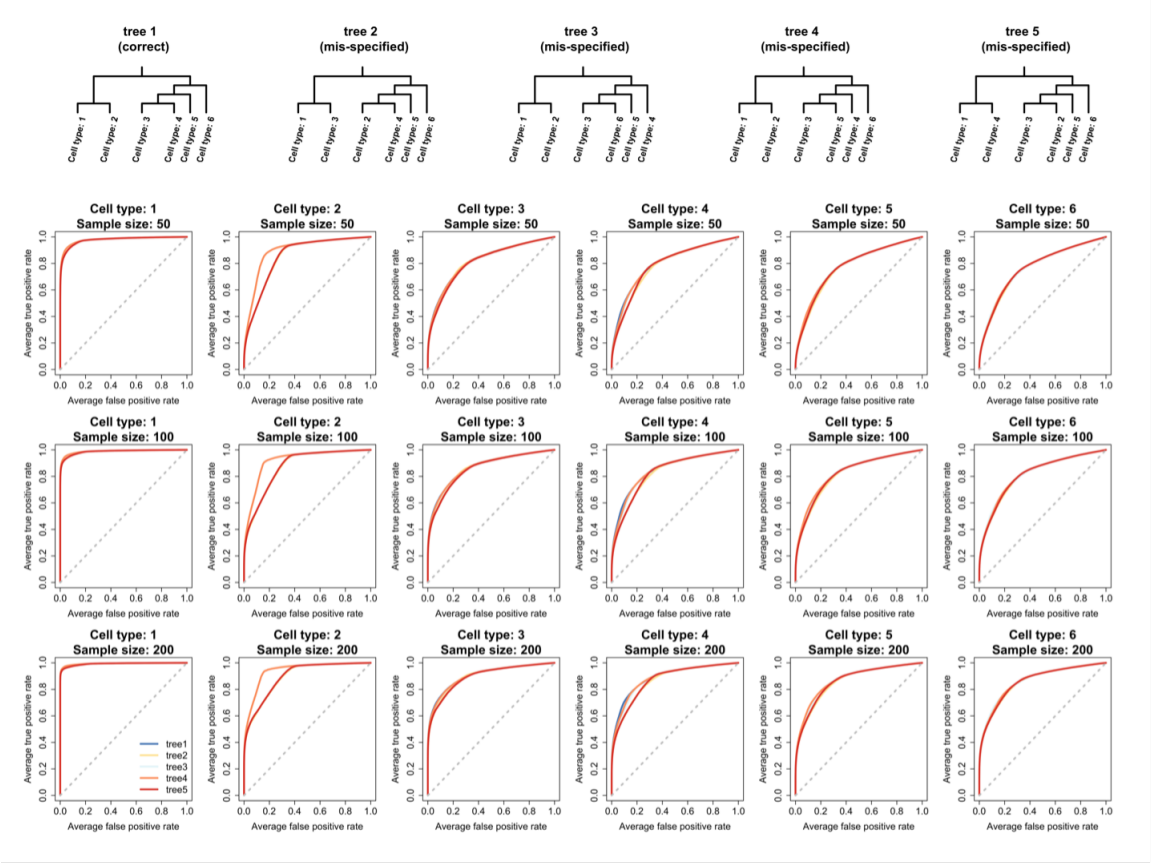

Figure A.6: ROC curves with correct/mis-specified tree structure as input of CeDAR-M for cell type specific differential expression analysis. The simulation mimics a two-group comparison based on bulk microarray gene expression – a mixture of six common blood immune cell types (1: Neutrophils, 2: Monocytes, 3: CD4, 4: CD8, 5: B cells, 6: NK cells) with different sample sizes per group (50, 100, and 200). "tree 1" is the correct tree structure used to generated simulation data; "tree 2", "tree 3", "tree 4" and "tree 5" are mis-specified tree structures by switching cell type 2 with cell type 3, and by switching cell type 4 with cell type 2/5/6, which are used for evaluating impact of mis-specified tree structure. Reported ROC curves are average results from 50 simulations.

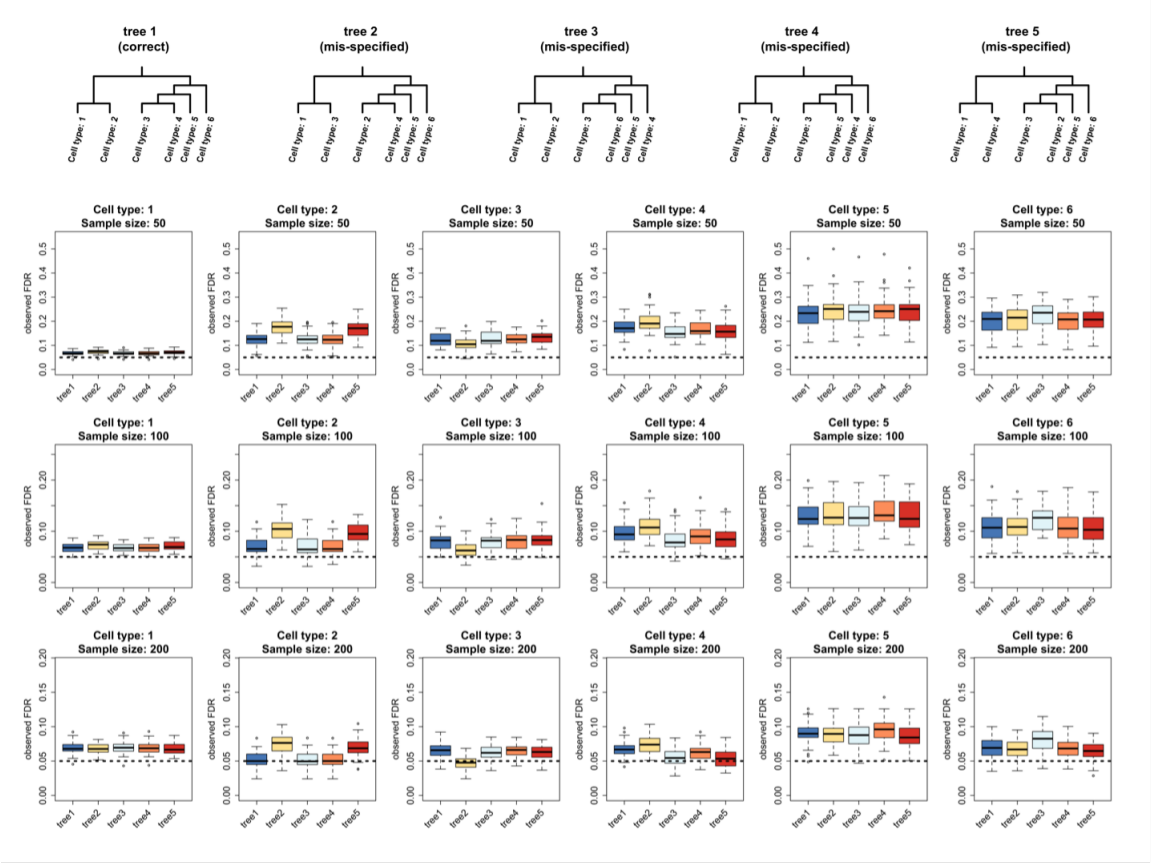

Figure A.7: Observed FDR with correct/mis-specified tree structure as input of CeDAR-M for cell type specific differential expression analysis. The simulation mimics a two-group comparison based on bulk microarray gene expression – a mixture of six common blood immune cell types (1: Neutrophils, 2: Monocytes, 3: CD4, 4: CD8, 5: B cells, 6: NK cells) with different sample sizes per group (50, 100, and 200). "tree 1" is the correct tree structure used to generated simulation data; "tree 2", "tree 3", "tree 4" and "tree 5" are mis-specified tree structures by switching cell type 2 with cell type 3, and by switching cell type 4 with cell type 2/5/6, which are used for evaluating impact of mis-specified tree structure. Reported observed FDR values are average results from 50 simulations.

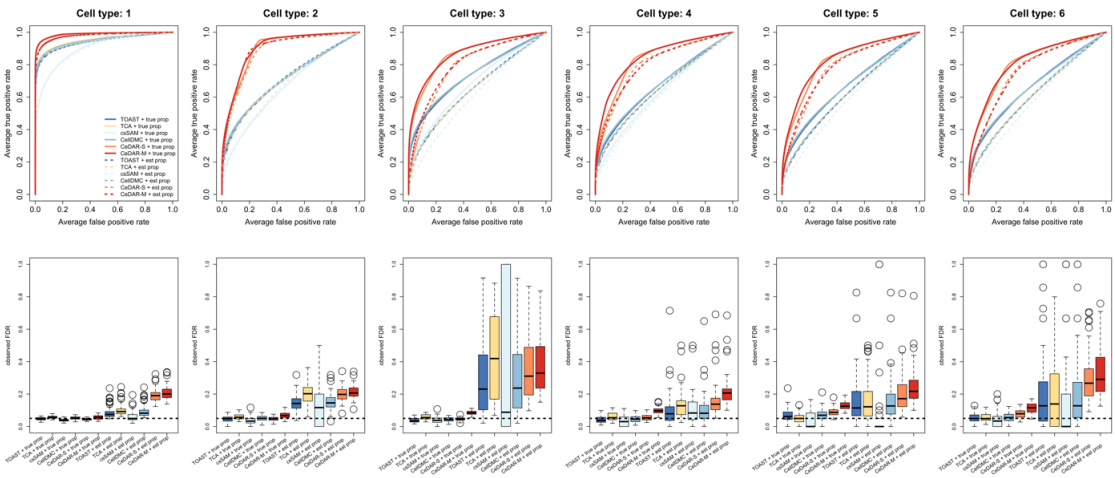

Figure A.8: Evaluation of effect on csDE detection performance by using estimated proportion. The upper panel shows ROC curves of six methods with either true proportion (solid line) or estimated proportion (dashed line). The lower panel shows observed FDR of six methods with either true proportion (left six) or estimated proportion (right six). DE genes were defined with rule: FDR < 0.05 (TOAST, TCA, csSAM and CellDMC); posterior probability of DE > 0.95 (CeDAR-M, CeDAR-S). Observed FDR of 50 simulations were summarized in box plot.

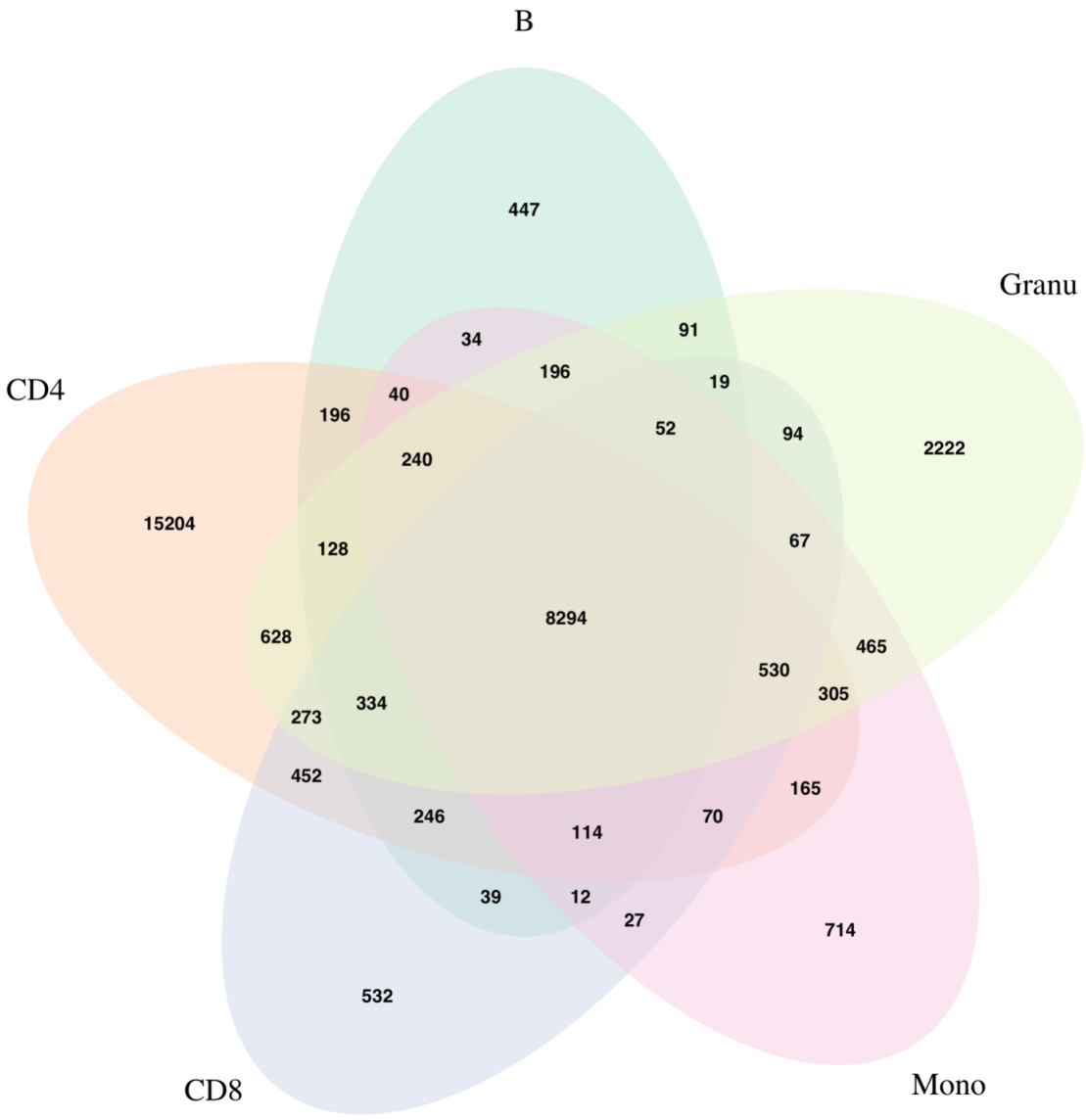

Figure A.9: Overlap of DMCs detected in pure cell types for data set GSE166844. DMCs in five cell types (Granulocytes, Monocytes, CD4, CD8, and B cells) were defined with rule FDR < 0.01.

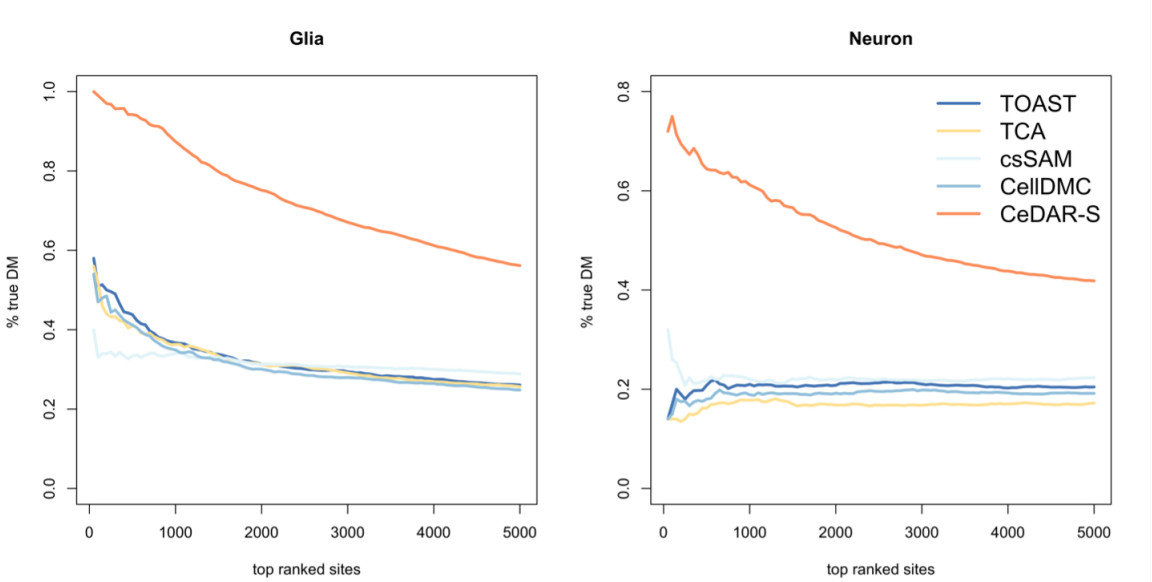

Figure A.10: Accuracy of detecting csDM associated with Down syndrome (DS) in human frontal cortex grey matter methylation data. The human frontal cortex grey matter methylation dataset (GEO accession number: GSE74486) contains both bulk samples from frontal cortex grey matter and pure cell type samples of glia and neuron cells derived by FACS. The csDM sites associated with disease DS were identified between 14 DS and 8 normal bulk samples using TOAST, TCA, csSAM, CellDMC, and CeDAR-S. The accuracy was evaluated by TDR curves.

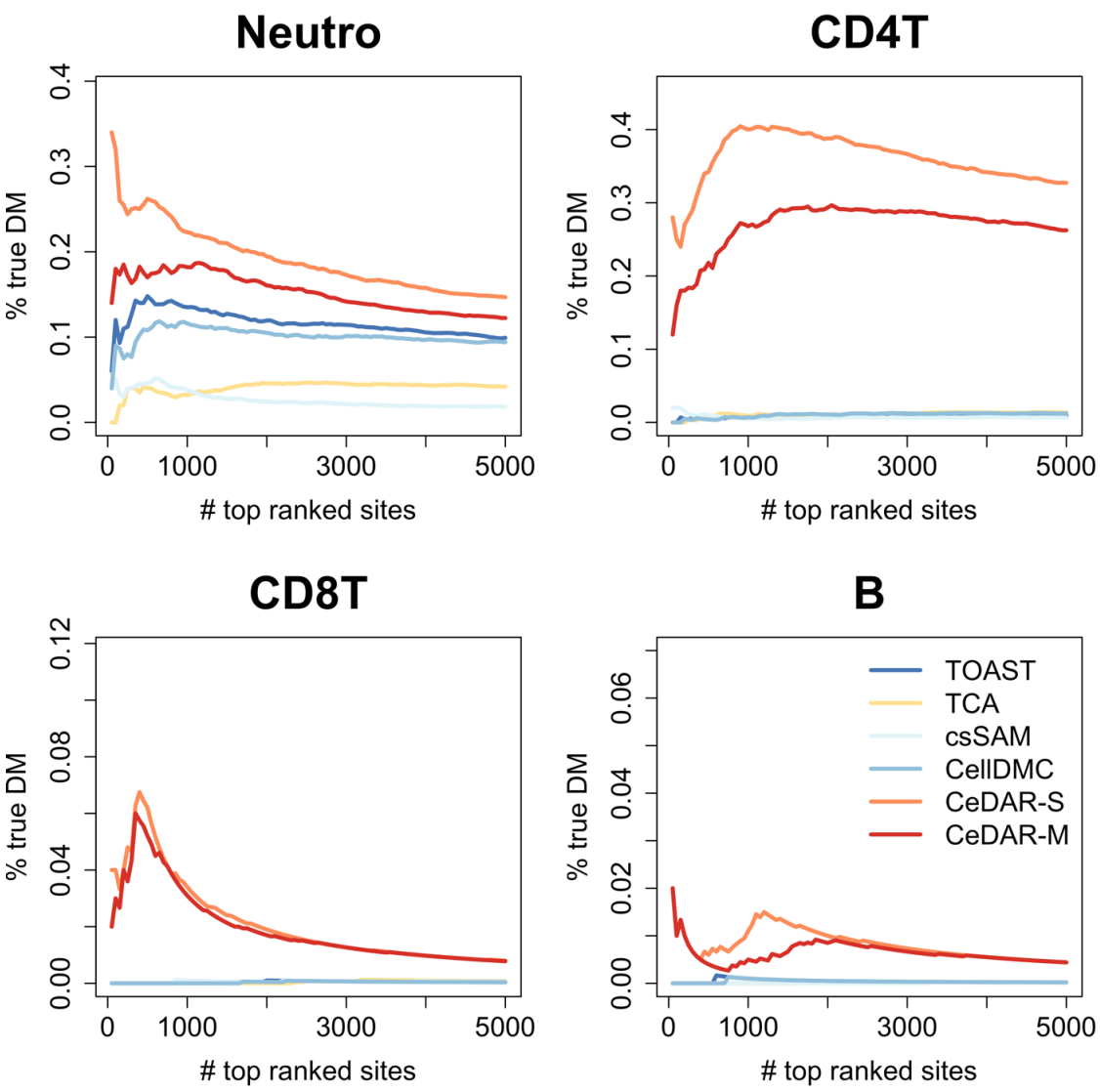

Figure A.11: Accuracy of detecting csDM associated with Systemic Lupus Erythematosus (SLE) in human whole blood methylation data. The human whole blood methylation dataset (GEO accession number: GSE118144) contains both bulk samples from whole blood and pure cell type samples of neutrophils, CD8, CD4, and B cells derived by FACS. The csDM sites associated with disease SLE were identified between 16 SLE and 13 normal bulk samples using TOAST, TCA, csSAM, CellDMC, CeDAR-S and CeDAR-M. The accuracy was evaluated by TDR curves.

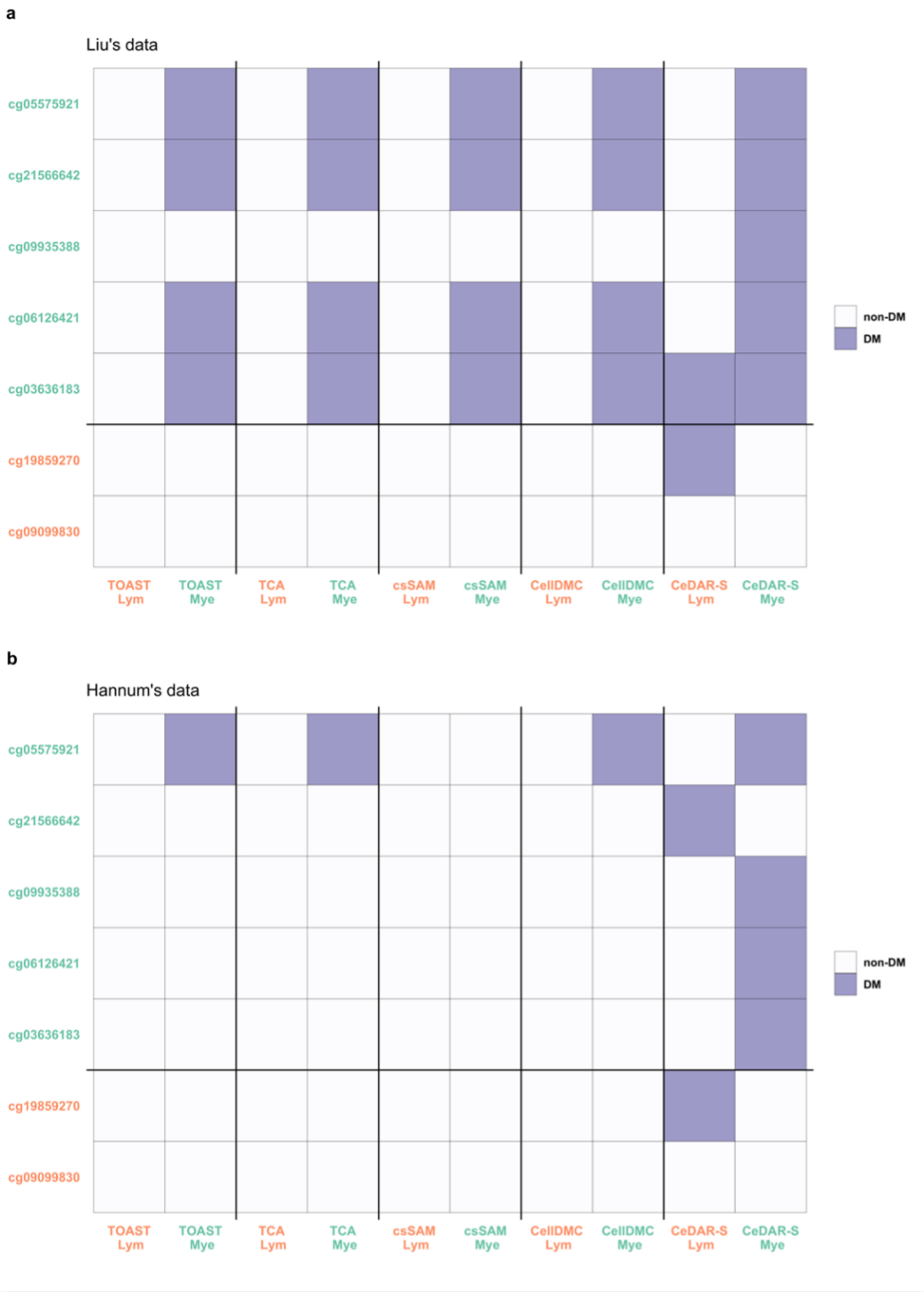

Figure A.12: Cell type specific DMC result associated with smoking status for blood methylation data. Examination of TOAST, TCA and CeDAR-S in identifying cs-DMCs of Lymphoid (Lym) and myeloid (Mye) cells in (a) Liu's DNA methylation data (GSE42861), and (b) Hannum's DNA methylation data (GSE40279). Five smoking associated Mye-specific DMCs (cg05575921, cg21566642, cg09935388, cg06126421, and cg03636183) and two Lym-specific DMCs (cg19859270 and cg09099830) used for evaluation were reported by Su et al. The csDMCs were called by FDR < 0.05 for TOAST, TCA, csSAM and CellDMC; by posterior probability of DM >0.95 for CeDAR-S.

Table A.1: Evaluation of different methods with correlated DE states among cell types under various sample sizes per group for cell type specific differential expression analyses.

Cell type	Metrics	Sample size: 50						Sample size: 100						Sample size: 200					
		TOAST	TCA	csSAM	CellDMC	CeDAR-S	CeDAR-M	TOAST	TCA	csSAM	CellDMC	CeDAR-S	CeDAR-M	TOAST	TCA	csSAM	CellDMC	CeDAR-S	CeDAR-M
1	ROC-AUC	0.897	0.897	0.841	0.898	0.979	0.977	0.950	0.952	0.905	0.949	0.988	0.988	0.975	0.978	0.944	0.974	0.993	0.994
	PR-AUC	0.732	0.736	0.528	0.746	0.921	0.913	0.872	0.882	0.700	0.878	0.958	0.958	0.940	0.947	0.818	0.941	0.976	0.980
	MCC	0.600	0.638	0.315	0.632	0.829	0.814	0.797	0.815	0.537	0.812	0.901	0.896	0.891	0.903	0.703	0.897	0.936	0.929
	FDR	0.049	0.093	0.036	0.054	0.050	0.055	0.047	0.060	0.040	0.052	0.045	0.056	0.050	0.050	0.035	0.054	0.042	0.071
2	ROC-AUC	0.642	0.640	0.608	0.641	0.877	0.861	0.704	0.706	0.662	0.704	0.897	0.899	0.771	0.775	0.721	0.771	0.916	0.940
	PR-AUC	0.264	0.269	0.198	0.275	0.487	0.458	0.381	0.396	0.270	0.395	0.575	0.585	0.513	0.534	0.365	0.528	0.668	0.723
	MCC	0.160	0.187	0.084	0.180	0.274	0.286	0.303	0.332	0.173	0.327	0.388	0.402	0.446	0.478	0.274	0.470	0.514	0.538
	FDR	0.048	0.099	0.033	0.049	0.089	0.139	0.046	0.059	0.043	0.048	0.055	0.076	0.045	0.050	0.041	0.049	0.038	0.055
3	ROC-AUC	0.664	0.664	0.625	0.666	0.821	0.818	0.728	0.732	0.675	0.731	0.865	0.866	0.795	0.800	0.737	0.797	0.901	0.905
	PR-AUC	0.304	0.310	0.227	0.317	0.477	0.483	0.430	0.448	0.301	0.448	0.595	0.614	0.565	0.588	0.402	0.582	0.699	0.723
	MCC	0.202	0.232	0.137	0.219	0.320	0.355	0.351	0.384	0.224	0.376	0.444	0.489	0.496	0.531	0.325	0.520	0.570	0.617
	FDR	0.042	0.091	0.039	0.048	0.080	0.132	0.038	0.057	0.037	0.043	0.043	0.081	0.039	0.048	0.042	0.043	0.032	0.065
4	ROC-AUC	0.616	0.615	0.587	0.616	0.797	0.800	0.670	0.672	0.628	0.672	0.837	0.847	0.734	0.739	0.681	0.736	0.873	0.887
	PR-AUC	0.226	0.229	0.180	0.235	0.379	0.404	0.323	0.337	0.236	0.338	0.480	0.526	0.448	0.469	0.317	0.465	0.590	0.642
	MCC	0.135	0.151	0.086	0.142	0.216	0.252	0.246	0.272	0.159	0.265	0.320	0.368	0.380	0.414	0.249	0.404	0.444	0.493
	FDR	0.047	0.095	0.043	0.050	0.131	0.188	0.044	0.062	0.041	0.048	0.063	0.100	0.042	0.051	0.039	0.045	0.042	0.072
5	ROC-AUC	0.605	0.605	0.580	0.606	0.778	0.776	0.656	0.659	0.617	0.659	0.820	0.826	0.722	0.727	0.671	0.724	0.860	0.870
	PR-AUC	0.205	0.209	0.164	0.213	0.342	0.351	0.293	0.309	0.213	0.308	0.441	0.466	0.416	0.443	0.292	0.435	0.557	0.587
	MCC	0.107	0.118	0.037	0.114	0.191	0.215	0.213	0.233	0.091	0.232	0.291	0.321	0.350	0.381	0.180	0.374	0.417	0.448
	FDR	0.095	0.119	0.046	0.107	0.184	0.243	0.070	0.062	0.037	0.068	0.095	0.132	0.058	0.047	0.039	0.062	0.059	0.086
6	ROC-AUC	0.609	0.607	0.580	0.610	0.771	0.766	0.671	0.673	0.626	0.673	0.820	0.820	0.733	0.736	0.679	0.735	0.860	0.863
	PR-AUC	0.204	0.205	0.161	0.212	0.338	0.335	0.313	0.327	0.220	0.330	0.460	0.468	0.436	0.457	0.297	0.454	0.571	0.583
	MCC	0.087	0.099	0.060	0.097	0.182	0.198	0.216	0.240	0.115	0.240	0.305	0.327	0.364	0.392	0.203	0.390	0.434	0.452
	FDR	0.077	0.170	0.042	0.086	0.183	0.247	0.054	0.055	0.037	0.056	0.074	0.102	0.048	0.053	0.034	0.051	0.046	0.065

The reported metrics (AUC-ROC, AUC-PR, MCC, and observed FDR) are average of 50 simulations. For TOAST, TCA, csSAM, and CellDMC, the MCC and observed FDR were derived by calling DE with estimated FDR < 0.05; for CeDAR-S and CeDAR-M, the MCC and observed FDR were derived by calling DE with estimated posterior probability of DE > 0.95.

Table A.2: Evaluation of different methods under various DE state patterns for cell type specific differential expression analyses (Corresponding to Figure 2.4 and Figure A.2: strong correlation).

Fig 2.4 panel	Methods	Cell type 1				Cell type 2				Cell type 3				Cell type 4			
		ROC-AUC	PR-AUC	MCC	FDR	ROC-AUC	PR-AUC	MCC	FDR	ROC-AUC	PR-AUC	MCC	FDR	ROC-AUC	PR-AUC	MCC	FDR
a	TOAST	0.976	0.943	0.896	0.049	0.803	0.569	0.489	0.046	0.868	0.713	0.647	0.037	0.732	0.434	0.362	0.065
	TCA	0.978	0.949	0.907	0.047	0.809	0.594	0.523	0.046	0.874	0.736	0.680	0.047	0.739	0.464	0.397	0.046
	csSAM	0.945	0.821	0.706	0.037	0.740	0.392	0.293	0.044	0.807	0.526	0.435	0.036	0.675	0.297	0.182	0.044
	CellDMC	0.975	0.944	0.901	0.053	0.806	0.588	0.517	0.050	0.870	0.729	0.673	0.041	0.736	0.456	0.393	0.065
	CeDAR-S	0.992	0.976	0.937	0.043	0.875	0.681	0.566	0.036	0.932	0.820	0.724	0.028	0.782	0.514	0.430	0.061
	CeDAR-M	0.991	0.974	0.933	0.045	0.861	0.650	0.575	0.067	0.925	0.804	0.728	0.043	0.768	0.488	0.439	0.096
b	TOAST	0.970	0.930	0.882	0.047	0.761	0.491	0.421	0.042	0.836	0.651	0.589	0.038	0.698	0.373	0.307	0.065
	TCA	0.972	0.937	0.894	0.046	0.763	0.510	0.448	0.048	0.839	0.669	0.618	0.047	0.701	0.393	0.332	0.043
	csSAM	0.940	0.810	0.694	0.037	0.717	0.353	0.256	0.036	0.789	0.494	0.405	0.033	0.660	0.269	0.148	0.037
	CellDMC	0.969	0.931	0.887	0.052	0.760	0.504	0.444	0.046	0.834	0.662	0.610	0.043	0.697	0.386	0.329	0.067
	CeDAR-S	0.997	0.984	0.934	0.040	0.961	0.756	0.508	0.038	0.976	0.860	0.674	0.030	0.947	0.661	0.388	0.061
	CeDAR-M	0.996	0.983	0.932	0.043	0.963	0.778	0.541	0.049	0.976	0.866	0.700	0.043	0.953	0.703	0.425	0.073
c	TOAST	0.977	0.944	0.895	0.049	0.796	0.558	0.479	0.043	0.868	0.711	0.645	0.039	0.721	0.412	0.337	0.055
	TCA	0.979	0.950	0.907	0.047	0.802	0.583	0.514	0.047	0.873	0.733	0.676	0.048	0.725	0.436	0.366	0.046
	csSAM	0.945	0.820	0.704	0.036	0.733	0.384	0.284	0.038	0.807	0.526	0.435	0.034	0.671	0.287	0.177	0.031
	CellDMC	0.976	0.945	0.900	0.053	0.799	0.578	0.509	0.047	0.869	0.726	0.669	0.042	0.723	0.429	0.363	0.057
	CeDAR-S	0.992	0.978	0.939	0.043	0.878	0.680	0.558	0.037	0.931	0.817	0.721	0.030	0.867	0.557	0.402	0.054
	CeDAR-M	0.992	0.976	0.935	0.044	0.869	0.662	0.566	0.053	0.930	0.819	0.745	0.056	0.895	0.638	0.448	0.073
d	TOAST	0.974	0.939	0.891	0.048	0.767	0.508	0.437	0.043	0.870	0.713	0.646	0.039	0.731	0.433	0.361	0.059
	TCA	0.976	0.946	0.902	0.047	0.769	0.525	0.465	0.048	0.875	0.737	0.679	0.049	0.739	0.461	0.392	0.047
	csSAM	0.944	0.817	0.700	0.039	0.719	0.362	0.271	0.038	0.807	0.527	0.435	0.035	0.675	0.297	0.188	0.041
	CellDMC	0.974	0.940	0.896	0.053	0.767	0.521	0.461	0.048	0.872	0.730	0.672	0.042	0.736	0.454	0.389	0.061
	CeDAR-S	0.992	0.974	0.933	0.042	0.922	0.673	0.505	0.037	0.939	0.831	0.727	0.030	0.787	0.519	0.431	0.057
	CeDAR-M	0.993	0.980	0.931	0.068	0.950	0.743	0.532	0.048	0.932	0.809	0.727	0.048	0.778	0.499	0.436	0.080
e	TOAST	0.974	0.940	0.893	0.048	0.768	0.510	0.439	0.043	0.868	0.712	0.646	0.037	0.719	0.409	0.338	0.062
	TCA	0.976	0.945	0.903	0.048	0.770	0.527	0.468	0.044	0.874	0.735	0.677	0.048	0.724	0.434	0.367	0.044
	csSAM	0.944	0.819	0.705	0.037	0.721	0.366	0.278	0.039	0.808	0.526	0.432	0.035	0.667	0.283	0.168	0.041
	CellDMC	0.974	0.941	0.897	0.053	0.767	0.522	0.463	0.048	0.870	0.728	0.670	0.041	0.721	0.427	0.361	0.064
	CeDAR-S	0.993	0.977	0.935	0.043	0.932	0.695	0.512	0.037	0.938	0.829	0.726	0.030	0.881	0.578	0.405	0.059
	CeDAR-M	0.994	0.982	0.933	0.065	0.954	0.755	0.537	0.046	0.940	0.836	0.749	0.043	0.907	0.665	0.449	0.065
f	TOAST	0.973	0.936	0.889	0.047	0.762	0.497	0.428	0.045	0.852	0.682	0.619	0.038	0.711	0.394	0.325	0.067
	TCA	0.975	0.942	0.900	0.046	0.765	0.515	0.454	0.048	0.856	0.701	0.649	0.047	0.714	0.417	0.353	0.048
	csSAM	0.944	0.817	0.701	0.033	0.716	0.355	0.259	0.044	0.799	0.511	0.423	0.036	0.666	0.277	0.156	0.040
	CellDMC	0.972	0.937	0.893	0.052	0.761	0.509	0.450	0.049	0.852	0.695	0.641	0.041	0.711	0.409	0.348	0.070
	CeDAR-S	0.995	0.978	0.932	0.043	0.946	0.717	0.504	0.040	0.956	0.835	0.699	0.029	0.914	0.613	0.401	0.061
	CeDAR-M	0.995	0.982	0.931	0.061	0.957	0.754	0.522	0.048	0.955	0.838	0.721	0.044	0.928	0.677	0.435	0.072

There are six different DE patterns corresponding to six panels in Figure 2.4 (a: all cell types are independent; b: all cell types are correlated under a single layer tree structure; c: only cell types 3 and 4 are correlated; d: only cell types 1 and 2 are correlated; e: cell types 1 and 2 are correlated, and cell types 3 and 4 are correlated; f: all cell types are correlated under a multiple-layer tree structure) The reported metrics (AUC-ROC, AUC-PR, MCC, and observed FDR) are average of 50 simulations. For TOAST, TCA, csSAM, and CellDMC the observed FDR was derived by calling DE with estimated FDR < 0.05; for CeDAR-S and CeDAR-M, the MCC and observed FDR was derived by calling DE with estimated posterior probability of DE > 0.95.

Table A.3: Evaluation of different methods under various DE state patterns for cell type specific differential expression analyses (corresponding to Figure A.3 and Figure A.4: weak correlation).

Fig A.3 panel	Methods	Cell type 1				Cell type 2				Cell type 3				Cell type 4			
		ROC-AUC	PR-AUC	MCC	FDR	ROC-AUC	PR-AUC	MCC	FDR	ROC-AUC	PR-AUC	MCC	FDR	ROC-AUC	PR-AUC	MCC	FDR
a	TOAST	0.976	0.943	0.896	0.049	0.803	0.569	0.489	0.046	0.868	0.713	0.647	0.037	0.732	0.434	0.362	0.065
	TCA	0.978	0.949	0.907	0.047	0.809	0.594	0.523	0.046	0.874	0.736	0.680	0.047	0.739	0.464	0.397	0.046
	csSAM	0.945	0.821	0.706	0.037	0.74	0.392	0.293	0.044	0.807	0.526	0.435	0.036	0.675	0.297	0.182	0.044
	CellDMC	0.975	0.944	0.901	0.053	0.806	0.588	0.517	0.05	0.870	0.729	0.673	0.041	0.736	0.456	0.393	0.065
	CeDAR-S	0.992	0.976	0.937	0.043	0.875	0.681	0.566	0.036	0.932	0.820	0.724	0.028	0.782	0.514	0.43	0.061
	CeDAR-M	0.991	0.974	0.933	0.045	0.861	0.650	0.575	0.067	0.925	0.804	0.728	0.043	0.768	0.488	0.439	0.096
b	TOAST	0.973	0.936	0.887	0.048	0.778	0.524	0.453	0.041	0.854	0.686	0.621	0.037	0.713	0.396	0.329	0.059
	TCA	0.975	0.942	0.898	0.049	0.781	0.544	0.481	0.045	0.857	0.704	0.651	0.049	0.716	0.420	0.359	0.043
	csSAM	0.943	0.814	0.697	0.037	0.725	0.370	0.273	0.037	0.800	0.514	0.424	0.035	0.666	0.279	0.169	0.04
	CellDMC	0.972	0.937	0.892	0.053	0.778	0.539	0.477	0.044	0.853	0.698	0.643	0.041	0.713	0.412	0.355	0.059
	CeDAR-S	0.994	0.977	0.932	0.043	0.926	0.708	0.528	0.036	0.954	0.829	0.698	0.030	0.895	0.589	0.401	0.056
	CeDAR-M	0.994	0.976	0.929	0.044	0.924	0.707	0.549	0.056	0.951	0.821	0.713	0.049	0.894	0.599	0.426	0.08
c	TOAST	0.976	0.942	0.895	0.047	0.797	0.560	0.484	0.044	0.868	0.711	0.647	0.038	0.726	0.423	0.352	0.057
	TCA	0.977	0.948	0.906	0.047	0.803	0.586	0.517	0.047	0.874	0.733	0.678	0.046	0.732	0.450	0.383	0.043
	csSAM	0.945	0.820	0.702	0.037	0.734	0.386	0.286	0.039	0.809	0.529	0.438	0.035	0.673	0.292	0.177	0.041
	CellDMC	0.975	0.943	0.900	0.052	0.800	0.580	0.511	0.047	0.870	0.726	0.671	0.041	0.729	0.442	0.378	0.06
	CeDAR-S	0.992	0.976	0.937	0.043	0.876	0.679	0.561	0.037	0.932	0.818	0.721	0.028	0.824	0.532	0.416	0.055
	CeDAR-M	0.991	0.974	0.934	0.043	0.867	0.662	0.568	0.053	0.926	0.804	0.733	0.056	0.831	0.554	0.441	0.082
d	TOAST	0.975	0.940	0.893	0.048	0.781	0.529	0.454	0.044	0.867	0.712	0.649	0.037	0.732	0.433	0.361	0.062
	TCA	0.977	0.947	0.905	0.046	0.786	0.551	0.486	0.048	0.873	0.735	0.681	0.047	0.740	0.462	0.394	0.047
	csSAM	0.944	0.819	0.703	0.036	0.727	0.372	0.276	0.037	0.806	0.527	0.438	0.035	0.674	0.296	0.184	0.042
	CellDMC	0.974	0.942	0.898	0.053	0.783	0.546	0.481	0.047	0.869	0.728	0.673	0.042	0.737	0.454	0.389	0.064
	CeDAR-S	0.992	0.974	0.935	0.042	0.896	0.668	0.527	0.037	0.935	0.824	0.727	0.028	0.784	0.515	0.430	0.058
	CeDAR-M	0.991	0.973	0.930	0.052	0.885	0.644	0.536	0.064	0.929	0.808	0.728	0.044	0.771	0.491	0.438	0.09
e	TOAST	0.975	0.941	0.894	0.047	0.783	0.535	0.461	0.044	0.870	0.712	0.648	0.039	0.727	0.421	0.352	0.058
	TCA	0.977	0.946	0.904	0.048	0.788	0.557	0.492	0.045	0.874	0.734	0.677	0.050	0.732	0.447	0.381	0.043
	csSAM	0.943	0.817	0.701	0.036	0.728	0.374	0.276	0.041	0.810	0.532	0.439	0.035	0.674	0.293	0.184	0.043
	CellDMC	0.974	0.942	0.899	0.052	0.784	0.552	0.487	0.046	0.871	0.727	0.671	0.043	0.730	0.440	0.377	0.059
	CeDAR-S	0.992	0.975	0.935	0.043	0.901	0.680	0.534	0.036	0.934	0.822	0.722	0.031	0.828	0.536	0.418	0.056
	CeDAR-M	0.992	0.975	0.926	0.069	0.911	0.700	0.548	0.049	0.932	0.813	0.731	0.050	0.837	0.565	0.439	0.075
f	TOAST	0.974	0.940	0.893	0.048	0.781	0.530	0.457	0.046	0.861	0.699	0.636	0.039	0.723	0.414	0.346	0.067
	TCA	0.976	0.946	0.904	0.047	0.785	0.553	0.489	0.046	0.864	0.719	0.665	0.047	0.728	0.440	0.376	0.045
	csSAM	0.944	0.820	0.704	0.036	0.726	0.373	0.278	0.042	0.804	0.520	0.432	0.037	0.673	0.287	0.171	0.052
	CellDMC	0.974	0.941	0.898	0.052	0.782	0.547	0.483	0.049	0.862	0.713	0.658	0.043	0.725	0.431	0.371	0.068
	CeDAR-S	0.992	0.975	0.935	0.043	0.909	0.686	0.530	0.040	0.940	0.820	0.710	0.030	0.853	0.552	0.412	0.063
	CeDAR-M	0.992	0.974	0.930	0.049	0.904	0.677	0.540	0.054	0.936	0.808	0.723	0.055	0.856	0.568	0.436	0.084

There are six different DE patterns corresponding to six panels in Figure A.3 (a: all cell types are independent; b: all cell types are correlated under a single layer tree structure; c: only cell types 3 and 4 are correlated; d: only cell types 1 and 2 are correlated; e: cell types 1 and 2 are correlated, and cell types 3 and 4 are correlated; f: all cell types are correlated under a multiple-layer tree structure) The reported metrics (AUC-ROC, AUC-PR, MCC, and observed FDR) are average of 50 simulations. For TOAST, TCA, csSAM, and CellDMC the observed FDR was derived by calling DE with estimated FDR < 0.05; for CeDAR-S and CeDAR-M, the MCC and observed FDR was derived by calling DE with estimated posterior probability of DE >0.95.

Table A.4: Evaluation of CeDAR with true/estimated tree structure and true/estimated prior probability of nodes on the tree as input for cell type specific differential analyses.

Tree structure type		True	True	Estimated
Prior probability type		True	Estimated	Estimated
Cell type 1	ROC-AUC	0.982	0.989	0.988
	PR-AUC	0.940	0.961	0.956
	MCC	0.872	0.899	0.894
	FDR	0.069	0.068	0.057
Cell type 2	ROC-AUC	0.933	0.919	0.898
	PR-AUC	0.720	0.630	0.586
	MCC	0.512	0.413	0.408
	FDR	0.165	0.070	0.073
Cell type 3	ROC-AUC	0.849	0.868	0.868
	PR-AUC	0.592	0.620	0.618
	MCC	0.509	0.495	0.495
	FDR	0.186	0.080	0.083
Cell type 4	ROC-AUC	0.833	0.850	0.849
	PR-AUC	0.513	0.532	0.529
	MCC	0.406	0.367	0.367
	FDR	0.250	0.097	0.100
Cell type 5	ROC-AUC	0.808	0.829	0.828
	PR-AUC	0.441	0.473	0.470
	MCC	0.347	0.324	0.324
	FDR	0.280	0.129	0.133
Cell type 6	ROC-AUC	0.795	0.818	0.817
	PR-AUC	0.426	0.460	0.457
	MCC	0.341	0.321	0.322
	FDR	0.243	0.109	0.112

True tree structure/prior probability represents using parameters generating simulation data as CeDAR input. Estimated tree structure/prior probability represents using tree structure/prior probability that are estimated from estimation procedure described in Methods section 2.2.3. The reported metrics (AUC-ROC, AUC-PR, MCC, and observed FDR) are average of 50 simulations. For TOAST, TCA, csSAM, and CellDMC, the MCC and observed FDR was derived by calling DE with estimated FDR < 0.05; for CeDAR-S and CeDAR-M, the MCC and observed FDR was derived by calling DE with estimated posterior probability of DE >0.95.

Table A.5: Observed FDR of CeDAR-S with estimated/true prior probability as input on simulated data with different noise level (two cell types).

Noise level	Estimated prior probability		True prior probability	
	FDR in Cell type 1	FDR in Cell type 2	FDR in Cell type 1	FDR in Cell type 2
0.01	0.043	0.039	0.026	0.024
0.1	0.043	0.037	0.026	0.025
1	0.063	0.047	0.024	0.083
2	0.046	0.126	0.024	0.225

True prior probability represents using parameters generating simulation data as CeDAR input. Estimated tree prior probability represents using estimated prior probability that are estimated from estimation procedure described in Methods section 2.2.3. The reported observed FDR is average of 50 simulations. The observed FDR was derived by calling DE with estimated posterior probability of DE > 0.95. Noise level 1 is the parameter setting we used in other simulations. For other noise levels (extremely low 0.01, low: 0.1, high: 2), we multiply 0.01, 0.1 or 2 to the standard deviation of both cell type specific gene expression and bulk expression. Sample size is 100 per group.

Table A.6: Observed FDR of CeDAR-M with estimated/true prior probability as input on simulated data with different noise level (six cell types)

Prior prob and tree structure type	Noise level	FDR in Cell type 1	FDR in Cell type 2	FDR in Cell type 3	FDR in Cell type 4	FDR in Cell type 5	FDR in Cell type 6
Estimated	0.01	0.097	0.089	0.105	0.108	0.096	0.073
	0.1	0.095	0.080	0.101	0.095	0.089	0.069
	1	0.057	0.073	0.083	0.100	0.133	0.112
	2	0.042	0.206	0.168	0.292	0.395	0.374
True	0.01	0.068	0.066	0.086	0.090	0.076	0.056
	0.1	0.068	0.069	0.089	0.095	0.082	0.062
	1	0.069	0.165	0.186	0.250	0.280	0.243
	2	0.107	0.345	0.408	0.528	0.587	0.555

True tree structure/prior probability represents using parameters generating simulation data as CeDAR input. Estimated tree structure/prior probability represents using prior probability that are estimated from estimation procedure described in Methods section 2.2.3. The reported observed FDR is average of 50 simulations. The observed FDR was derived by calling DE with estimated posterior probability of DE > 0.95. Noise level 1 is the parameter setting we used in other simulations. For other noise levels (extremely low: 0.01, low: 0.1, high: 2), we multiply 0.01, 0.1 or 2 to the standard deviation of both cell type specific gene expression and bulk expression. Sample size is 100 per group.

Table A.7: Evaluation of CeDAR-M with correct/mis-specified tree structure as input for cell type specific differential expression analyses from different methods.

Cell type	Tree type	Sample size: 50					Sample size: 100					Sample size: 200				
		Tree 1 (Correct)	Tree 2 (Mis-specified)	Tree 3 (Mis-specified)	Tree 4 (Mis-specified)	Tree 5 (Mis-specified)	Tree 1 (Correct)	Tree 2 (Mis-specified)	Tree 3 (Mis-specified)	Tree 4 (Mis-specified)	Tree 5 (Mis-specified)	Tree 1 (Correct)	Tree 2 (Mis-specified)	Tree 3 (Mis-specified)	Tree 4 (Mis-specified)	Tree 5 (Mis-specified)
1	ROC-AUC	0.979	0.977	0.979	0.979	0.976	0.989	0.987	0.989	0.989	0.987	0.994	0.992	0.994	0.994	0.992
	PR-AUC	0.923	0.911	0.923	0.923	0.908	0.961	0.952	0.961	0.961	0.951	0.980	0.974	0.980	0.980	0.973
	MCC	0.834	0.822	0.834	0.833	0.816	0.899	0.888	0.899	0.899	0.886	0.930	0.925	0.930	0.930	0.923
	FDR	0.067	0.074	0.067	0.067	0.071	0.068	0.073	0.068	0.068	0.072	0.069	0.068	0.069	0.069	0.068
2	ROC-AUC	0.897	0.852	0.897	0.897	0.854	0.919	0.871	0.920	0.919	0.873	0.940	0.894	0.940	0.940	0.896
	PR-AUC	0.531	0.437	0.532	0.531	0.439	0.630	0.532	0.630	0.630	0.536	0.723	0.630	0.723	0.723	0.635
	MCC	0.284	0.285	0.285	0.285	0.285	0.413	0.412	0.413	0.413	0.411	0.534	0.527	0.534	0.534	0.526
	FDR	0.128	0.175	0.127	0.127	0.168	0.070	0.102	0.070	0.070	0.097	0.052	0.075	0.052	0.052	0.070
3	ROC-AUC	0.821	0.816	0.819	0.820	0.812	0.868	0.862	0.866	0.867	0.858	0.905	0.899	0.903	0.904	0.895
	PR-AUC	0.493	0.468	0.484	0.489	0.464	0.620	0.587	0.609	0.614	0.585	0.725	0.694	0.715	0.719	0.692
	MCC	0.363	0.330	0.357	0.360	0.347	0.495	0.453	0.486	0.490	0.472	0.617	0.576	0.607	0.612	0.592
	FDR	0.124	0.105	0.127	0.124	0.135	0.080	0.063	0.080	0.081	0.083	0.066	0.047	0.063	0.065	0.063
4	ROC-AUC	0.802	0.783	0.799	0.801	0.789	0.850	0.828	0.845	0.847	0.831	0.889	0.867	0.883	0.886	0.869
	PR-AUC	0.413	0.364	0.393	0.401	0.366	0.532	0.472	0.505	0.516	0.469	0.643	0.583	0.614	0.626	0.579
	MCC	0.257	0.238	0.238	0.246	0.223	0.367	0.345	0.343	0.353	0.326	0.489	0.465	0.464	0.474	0.446
	FDR	0.175	0.199	0.154	0.166	0.162	0.097	0.111	0.084	0.092	0.087	0.067	0.074	0.054	0.061	0.054
5	ROC-AUC	0.780	0.771	0.780	0.781	0.775	0.829	0.819	0.829	0.830	0.823	0.870	0.859	0.869	0.871	0.863
	PR-AUC	0.352	0.332	0.351	0.359	0.340	0.473	0.449	0.471	0.481	0.457	0.584	0.560	0.581	0.593	0.565
	MCC	0.201	0.195	0.200	0.209	0.196	0.324	0.316	0.321	0.332	0.315	0.445	0.437	0.442	0.455	0.436
	FDR	0.237	0.247	0.239	0.247	0.243	0.129	0.134	0.128	0.138	0.130	0.090	0.089	0.087	0.096	0.086
6	ROC-AUC	0.771	0.766	0.770	0.771	0.768	0.818	0.813	0.818	0.818	0.815	0.863	0.859	0.864	0.863	0.860
	PR-AUC	0.348	0.339	0.353	0.348	0.342	0.460	0.449	0.469	0.460	0.452	0.587	0.577	0.597	0.587	0.579
	MCC	0.201	0.198	0.211	0.201	0.198	0.321	0.317	0.334	0.321	0.316	0.456	0.451	0.470	0.456	0.449
	FDR	0.202	0.208	0.226	0.200	0.204	0.109	0.109	0.126	0.109	0.106	0.067	0.066	0.081	0.067	0.065

The simulation mimics a two-group comparison based on bulk microarray gene expression – a mixture of six common blood immune cell types (1: Neutrophils, 2: Monocytes, 3: CD4, 4: CD8, 5: B cells, 6: NK cells) with different sample sizes per group (50, 100, and 200). "tree 1" is the correct tree structure used to generated simulation data; "tree 2", "tree 3", "tree 4" and "tree 5" are mis-specified tree structures by switching cell type 2 with cell type 3, and by switching cell type 4 with cell type 2/5/6, which were used for evaluating impact of mis-specified tree structure. The reported metrics (AUC-ROC, AUC-PR, MCC, and observed FDR) are average of 50 simulations. The MCC and observed FDR were derived by calling DE with estimated posterior probability of DE > 0.95.

Table A.8: Evaluation of different methods with true/estimated cell type composition as input for cell type specific differential expression analyses from different methods.

Proportion type		True						Estimated					
Methods		TOAST	TCA	csSAM	CellDMC	CeDAR-S	CeDAR-M	TOAST	TCA	csSAM	CellDMC	CeDAR-S	CeDAR-M
Cell type 1	ROC-AUC	0.948	0.951	0.902	0.948	0.988	0.988	0.940	0.943	0.899	0.940	0.978	0.978
	PR-AUC	0.870	0.881	0.697	0.877	0.959	0.958	0.859	0.868	0.703	0.865	0.926	0.926
	MCC	0.796	0.816	0.535	0.812	0.901	0.896	0.793	0.803	0.576	0.802	0.822	0.815
	FDR	0.047	0.059	0.038	0.052	0.044	0.057	0.089	0.104	0.062	0.097	0.194	0.211
Cell type 2	ROC-AUC	0.708	0.710	0.665	0.708	0.898	0.902	0.708	0.704	0.674	0.708	0.884	0.897
	PR-AUC	0.389	0.405	0.276	0.404	0.583	0.596	0.342	0.336	0.244	0.352	0.519	0.548
	MCC	0.308	0.338	0.176	0.332	0.395	0.411	0.187	0.222	0.064	0.210	0.331	0.355
	FDR	0.047	0.061	0.036	0.051	0.047	0.069	0.147	0.203	0.124	0.154	0.199	0.214
Cell type 3	ROC-AUC	0.734	0.738	0.681	0.737	0.865	0.867	0.647	0.643	0.615	0.648	0.801	0.807
	PR-AUC	0.438	0.457	0.310	0.456	0.597	0.618	0.259	0.249	0.192	0.268	0.371	0.401
	MCC	0.358	0.393	0.228	0.383	0.450	0.497	0.170	0.182	0.043	0.184	0.241	0.287
	FDR	0.039	0.059	0.039	0.043	0.046	0.083	0.304	0.443	0.370	0.304	0.371	0.382
Cell type 4	ROC-AUC	0.672	0.675	0.630	0.675	0.836	0.848	0.650	0.651	0.619	0.651	0.804	0.802
	PR-AUC	0.329	0.344	0.240	0.343	0.483	0.531	0.281	0.287	0.211	0.292	0.413	0.425
	MCC	0.251	0.278	0.160	0.270	0.325	0.371	0.198	0.219	0.123	0.213	0.272	0.306
	FDR	0.041	0.061	0.035	0.046	0.056	0.099	0.102	0.136	0.105	0.108	0.168	0.223
Cell type 5	ROC-AUC	0.662	0.664	0.621	0.663	0.822	0.828	0.615	0.616	0.589	0.615	0.782	0.781
	PR-AUC	0.301	0.317	0.217	0.315	0.449	0.473	0.216	0.223	0.167	0.224	0.342	0.353
	MCC	0.220	0.239	0.095	0.239	0.296	0.324	0.108	0.122	0.027	0.118	0.191	0.211
	FDR	0.069	0.052	0.040	0.071	0.089	0.127	0.162	0.167	0.093	0.161	0.208	0.248
Cell type 6	ROC-AUC	0.667	0.668	0.624	0.669	0.815	0.815	0.613	0.612	0.587	0.613	0.772	0.770
	PR-AUC	0.309	0.320	0.217	0.324	0.452	0.458	0.208	0.210	0.158	0.215	0.321	0.321
	MCC	0.213	0.233	0.112	0.234	0.298	0.319	0.087	0.094	0.030	0.097	0.168	0.179
	FDR	0.055	0.056	0.040	0.058	0.079	0.114	0.216	0.212	0.165	0.215	0.306	0.336

True proportion represents using cell type compositions generating simulation data as input; Estimated proportion represents using estimated cell type compositions (by *ged* function of CellDMC package) as input. The reported metrics (AUC-ROC, AUC-PR, MCC, and observed FDR) are average of 50 simulations. For TOAST, TCA, csSAM, and CellDMC, the MCC and observed FDR were derived by calling DE with estimated FDR < 0.05; for CeDAR-S and CeDAR-M, the MCC and observed FDR were derived by calling DE with estimated posterior probability of DE > 0.95.

Table A.9: Computation time of various methods with different number of cell types and different sample sizes.

Methods	Cell type number	Sample size: 50	Sample size: 100	Sample size: 200
TCA	4	531.495	603.709	751.259
csSAM	4	38.241	64.792	111.977
CellDMC	4	19.674	20.257	22.137
TOAST	4	0.161	0.395	1.353
CeDAR-M	4	3.611	10.369	33.879
TCA	6	618.769	679.791	876.800
csSAM	6	39.104	66.050	122.227
CellDMC	6	23.207	24.466	26.644
TOAST	6	0.154	0.409	1.383
CeDAR-M	6	10.877	36.759	130.927
TCA	8	870.97	757.761	989.960
csSAM	8	41.432	73.086	124.592
CellDMC	8	26.804	28.785	31.872
TOAST	8	0.176	0.424	1.405
CeDAR-M	8	50.959	152.417	524.238

TOAST, TCA, csSAM, CellDMC and CeDAR-M were evaluated for 12,402 genes with different number of cell types (4, 6, 8) and different sample sizes per group (50, 100, 200). Simulation was run on Linux with 2.80 GHz CPU and 8G RAM. Reported time (in seconds) is average of five simulations.

Appendix B

Appendix for Chapter 3

B.1 A more general framework for different types marker identification

Different studies may focus on different types of CTS genes markers. In our study we mainly focus on CTS genes that uniquely have higher expression in one cell type, but lower level in other cell types (one vs. others) with Equation 3.2. In practice, we found that this expression can help us to find such CTS genes, but it can still provide some genes that have high expression in more than one cell type. This is because in the contrast group – "other" cell types, the expression is an average of all remained cell types. Thus, even there is another cell type has high expression, the mean expression of "other" group can still be low if most other cell types have extremely low expression. In addition, some studies may interest in CTS marker genes have high expression in more than one cell type.

To avoid the above described problem and generalize our methods for different requirements, we proposed the following summary statistic to study CTS genes in samples. Suppose we want to study CTS genes that have higher expression in cell

types belong to set $C \subset \{1, 2, ..., K\}$:

$$Y_{giC} = min\{\bar{X}_{gik}\}_{k \in C} - max\{\bar{X}_{gik'}\}_{k' \notin C} \tag{B.1}$$

The variance of Y_{giC} can be derived by bootstrapping, in which a new sample simulated by randomly draw cells from each cell type with fixed cell type composition.

B.2 Standard error calculation for estimated log2 fold change in one sample

In the proposed method (Equation 3.1 and Equation 3.2), we know $\bar{X}_{gik}$ is the arithmetic mean expression of normalized 10k counts for g-th gene of k-th cell type in i-th sample and Y_{gik} is log2 fold change for g-th gene ($g = 1, ..., G$) of k-th cell type ($k = 1, ..., K$) in i-th sample ($i = 1, ..., N$).

Since K cell type are assumed to be independent, the joint distribution of the estimators for mean expression in K cell types is:

$$\begin{bmatrix} \bar{X}_{gik1} \\ \vdots \\ \bar{X}_{giK} \end{bmatrix} \sim AN_K \left(\begin{bmatrix} \mu_{gi1} \\ \vdots \\ \mu_{giK} \end{bmatrix}, \begin{bmatrix} \frac{\omega^2_{gi1}}{C_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\omega^2_{giK}}{C_K} \end{bmatrix} \right) \tag{B.2}$$

For the following context, we treat k-th cell type as the target cell type for which we want to study its CTS genes. Based on property of multivariate normal distribution, we have:

$$\begin{bmatrix} \bar{X}_{gik} \\ \frac{\sum_{k' \neq k} \bar{X}_{gik'}}{K-1} \end{bmatrix} \sim AN_2 \left(\begin{bmatrix} \mu_{gik} \\ \frac{\sum_{k' \neq k} \mu_{gik'}}{K-1} \end{bmatrix}, \begin{bmatrix} \frac{\omega^2_{gik}}{C_k} & 0 \\ 0 & \frac{\sum_{k' \neq k} \frac{\omega^2_{gik'}}{C_{k'}}}{(K-1)^2} \end{bmatrix} \right) \tag{B.3}$$

The $log2$ transformed average expression (a pseudo-count 1 is added) based on Delta method is:

$$\begin{bmatrix} log2(\bar{X}_{gik}+1) \\ log2(\frac{\sum_{k'\neq k}\bar{X}_{gik'}}{K-1}+1) \end{bmatrix} \sim \tag{B.4}$$
$$AN_2\left(\begin{bmatrix} log2(\mu_{gik}+1) \\ log2(\frac{\sum_{k'\neq k}\mu_{gik'}}{K-1}+1) \end{bmatrix}, \begin{bmatrix} \frac{\omega^2_{gik}/C_k}{\Big(ln(2)\times(\mu_{gik}+1)\Big)^2} & 0 \\ 0 & \sum_{k'\neq k}\frac{\omega^2_{gik'}/C_{k'}}{\Big(ln(2)\times(\sum_{k'\neq k}\mu_{gik'}+K-1)\Big)^2} \end{bmatrix}\right)$$

Then we can derive the form of the square of standard error for LFC - Y_{gik}:

$$\sigma^2_{gik} = \frac{\omega^2_{gik}/C_k}{\Big(ln(2)\times(\mu_{gik}+1)\Big)^2} + \sum_{k'\neq k}\frac{\omega^2_{gik'}/C_{k'}}{\big(ln(2)\times(\sum_{k'\neq k}\mu_{gik'}+K-1)\big)^2} \tag{B.5}$$

The estimate of σ^2_{gik} can be derived by plugged in the estimates of μ_{gik} and ω^2_{gik} for g-th gene of k-th cell type in i-th sample.

B.3 EM algorithm details

B.3.1 Details in step 1

For g-th gene ($g = 1, ..., G$), we assume $D_{gk} = 1$, denote $\mathbf{\Theta} = \{q_{gk}, m_{gk}, \tau^2_{gk}\}$ and define $\mathbf{\Theta}_t$ as the parameters derived at t-th iteration. In addition, we define: $p_{0gik,t} = \phi(Y_{gik}; 0, \sigma^2_{gik} + \tau^2_{gk})$ and $p_{1gik,t} = \phi(Y_{gik}; m_{gk,t}, \sigma^2_{gik} + \tau^2_{gk})$.

Based on the complete likelihood shown in Equation 3.4, we can easily derive the

log-likelihood as following:

$$
\begin{aligned}
l(\boldsymbol{\Theta}) = & \sum_{g=1}^{G} \Bigg\{ (1 - D_{gk}) log(1 - \pi_k) + D_{gk} log(\pi_k) \\
& + \sum_{i=1}^{N} \Bigg[(D_{gk} - D_{gk} Z_{gik}) log(1 - q_{gk}) + D_{gk} Z_{gik} log(q_{gk}) \\
& \quad - \frac{log(\tau_{gk}^2)}{2} - \frac{\Delta_{gik}^2}{2\tau_{gk}^2} - \frac{D_{gk} Z_{gik} m_{gk}^2}{2\tau_{gk}^2} + \frac{D_{gk} Z_{gik} m_{gk}^2 \Delta_{gik}}{\tau_{gk}^2} \Bigg] \Bigg\} + Constant
\end{aligned}
\tag{B.6}
$$

E-step in Step 1

First, we can have conditional expectation of random variable Z_{gik}:

$$
\begin{aligned}
E\{Z_{gik} | \boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk} = 1\} &= P(Z_{gik} = 1 | \boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk} = 1) \\
&= \frac{p_{1gik,t} \times q_{gk,t}}{p_{1gik,t} \times q_{gk,t} + p_{0gik,t} \times (1 - q_{gk,t})} \equiv a_{gik,t}^*
\end{aligned}
\tag{B.7}
$$

Second, based on the proposed model in Equation 3.3, we can derive the distribution of Δ_{gik} conditioned on $\boldsymbol{Y}_{gk}, \boldsymbol{Z}_{gk}, \boldsymbol{\Theta}_t$:

$$
\begin{aligned}
\Delta_{gik} | Y_{gik}, Z_{gik} = 0, \boldsymbol{\Theta}_t &\sim N(\frac{Y_{gik}/\sigma_{gik}^2}{1/\sigma_{gik}^2 + 1/\tau_{gk}^2}, \frac{1}{1/\sigma_{gik}^2 + 1/\tau_{gk}^2}) \\
\Delta_{gik} | Y_{gik}, Z_{gik} = 1, \boldsymbol{\Theta}_t &\sim N(\frac{Y_{gik}/\sigma_{gik}^2 + m_{gk}/\tau_{gk}^2}{1/\sigma_{gik}^2 + 1/\tau_{gk}^2}, \frac{1}{1/\sigma_{gik}^2 + 1/\tau_{gk}^2})
\end{aligned}
\tag{B.8}
$$

Then we can derive the expectation of missing variable Δ_{gik}:

$$
E\{\Delta_{gik} | \boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk} = 1\} = E\left\{E\{\Delta_{gik} | Z_{gik}, \boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk} = 1\} | \boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk} = 1\right\}
\tag{B.9}
$$

$$
= a_{gik}^* \frac{Y_{gik}/\sigma_{gik}^2 + m_{gk}/\tau_{gk}^2}{1/\sigma_{gik}^2 + 1/\tau_{gk}^2} + (1 - a_{gik}^* \frac{y_{gik}/\sigma_{gik}^2}{1/\sigma_{gik}^2 + 1/\tau_{gk}^2})
$$

Similarly, we can have:

$$E\{Z_{gik}\Delta_{gik}|\boldsymbol{Y}_{gk},\boldsymbol{\Theta}_t,D_{gk}=1\}=a^*_{gik}\frac{Y_{gik}/\sigma^2_{gik}+m_{gk}/\tau^2_{gk}}{1/\sigma^2_{gik}+1/\tau^2_{gk}} \tag{B.10}$$

$$E\{\Delta^2_{gik}|\boldsymbol{Y}_{gk},\boldsymbol{\Theta}_t,D_{gk}=1\}=a^*_{gik}\left[\frac{Y_{gik}/\sigma^2_{gik}+m_{gk}/\tau^2_{gk}}{1/\sigma^2_{gik}+1/\tau^2_{gk}}+\frac{1}{1/\sigma^2_{gik}+1/\tau^2_{gk}}\right] \tag{B.11}$$

$$+(1-a^*_{gik})\left[\frac{Y_{gik}/\sigma^2_{gik}}{1/\sigma^2_{gik}+1/\tau^2_{gk}}+\frac{1}{1/\sigma^2_{gik}+1/\tau^2_{gk}}\right]$$

M-step in Step 1

In this step, we maximize the "Q function" (the expected complete data log-likelihood with respect to $\boldsymbol{\Theta}$) shown in following Equation(16) to obtain $\boldsymbol{\Theta}_{t+1}$.

$$\begin{aligned}Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}_t)&=E\{l(\boldsymbol{\Theta})|\boldsymbol{\Theta}_t\} \\ &=\sum_{i=1}^{N}\left[(1-a^*_{gik,t})log(1-q_{gk})+a^*_{gik,t}log(q_{gk})\right] \\ &+\sum_{i=1}^{N}\left[-\frac{log(\tau^2_{gk})}{2}-\frac{E\{\Delta^2_{gik}|\boldsymbol{Y}_{gk},\boldsymbol{\Theta}_t,D_{gk}=1\}}{2\tau^2_{gk}}-\frac{a^*_{gik,t}m^2_{gk}}{2\tau^2_{gk}}\right. \\ &\left.+\frac{E\{Z_{gik}\Delta_{gik}|\boldsymbol{Y}_{gk},\boldsymbol{\Theta}_t,D_{gk}=1\}}{\sum_{i=1}^{N}a^*_{gik,t}}\right]\end{aligned} \tag{B.12}$$

Then by solving $\frac{\partial Q}{\partial m_{gk}}=0$, we can update m_{gk}:

$$m_{gk,t+1}=\frac{\sum_{i=1}^{N}E\{Z_{gik}\Delta_{gik}|\boldsymbol{Y}_{gk},\boldsymbol{Q}_t,D_{gk}=1\}}{\sum_{i=1}^{N}a^*_{gik,t}} \tag{B.13}$$

Similarly, by solving $\frac{\partial Q}{\partial \tau_{gk}^2} = 0$, we can update τ_{gk}^2:

$$\tau_{gk,t+1}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} E\{\Delta_{gik}^2 | \boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk}=1\} - 2m_{gk,t+1}E\{Z_{gik}\Delta_{gik}|\boldsymbol{Y}_{gk}, \boldsymbol{\Theta}_t, D_{gk}=1\} + m_{gk,t+1}^2 a_{gik,t}^*\right) \tag{B.14}$$

By solving $\frac{\partial Q}{\partial q_{gk}} = 0$, we can update q_{gk}:

$$q_{gk,t+1} = \frac{\sum_{i=1}^{N} a_{gik,t}^*}{N} \tag{B.15}$$

B.3.2 Details in step 3

In step 3, our purpose is to estimate parameter π_k by fixing $\boldsymbol{q}_k = \{q_{1k}, ..., q_{Gk}\}$, $\boldsymbol{m}_k = \{m_{1k}, ..., m_{Gk}\}$, $\boldsymbol{\tau}_k^2 = \{\tau_{1k}^2, ..., \tau_{Gk}^2\}$ with corresponding estimates derived in step 1. Since $\boldsymbol{q}_k$, $\boldsymbol{m}_k$, and $\boldsymbol{\tau}_k^2$ are known, the complete data log-likelihood is part of Equation B.6, which is shown in following Equation B.16:

$$l(\pi_k) = \sum_{g=1}^{G}\big\{(1-D_{gk})log(1-\pi_k) + D_{gk}log(\pi_k)\big\} + Constant \tag{B.16}$$

In the E-step, we can derive the expectation of variable D_{gk} for g-th gene $g = 1, ..., G$:

$$E\{D_{gk}|\boldsymbol{Y}_{gk}, \pi_{k,t}\} = \frac{P(\boldsymbol{Y}_{gk}|D_{gk}=1, \pi_{k,t})\pi_{k,t}}{P(\boldsymbol{Y}_{gk}|D_{gk}=1, \pi_{k,t})\pi_{k,t} + P(\boldsymbol{Y}_{gk}|D_{gk}=0, \pi_{k,t})(1-\pi_{k,t})} \tag{B.17}$$

where $P(\boldsymbol{Y}_{gk}|D_{gk}=1, \pi_{k,t}) = \prod_{i=1}^{N} P(Y_{gik}|D_{gk}=1, \pi_{k,t}) = \prod_{i=1}^{N} P(Y_{gik}|D_{gk}=1, \pi_{k,t}) = \prod_{i=1}^{N}\{q_{g,t}\phi(Y_{gik}; m_{gk,t}, \sigma_{gik}^2+\tau_{gk}^2)+(1-q_{g,t})\phi(Y_{gik}; 0, \sigma_{gik}^2+\tau_{gk}^2)\}$, $P(\boldsymbol{Y}_{gk}|D_{gk}=0, \pi_{k,t}) = \prod_{i=1}^{N}\phi(Y_{gik}; 0, \sigma_{gik}^2+\tau_{gk}^2)$. Then in the M-step, we look for the π_k value that

maximize the "Q function" in Equation(21):

$$Q(\pi_k|\pi_{k,t}) = E\{l(\pi_k)|\pi_{k,t}\} = \sum_{g=1}^{G}\big\{(1-b_{gk,t})log(1-\pi_k) + b_{gk,t}log(\pi_k)\big\} \tag{B.18}$$

By solving $\frac{\partial Q}{\partial \pi_k} = 0$, we can update π_k:

$$\pi_{k,t+1} = \frac{\sum_{g=1}^{G} b_{gk,t}}{G} \tag{B.19}$$

Table B.1: Cell type composition of samples in PBMC Lupus data.

Sample	B cells	CD14+ Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A+ Monocytes	NK cells
1043_ctrl	0.057	0.263	0.477	0.073	0.015	0.056	0.059
1079_ctrl	0.094	0.193	0.427	0.045	0.014	0.116	0.113
1085_ctrl	0.096	0.087	0.639	0.075	0.013	0.022	0.069
1154_ctrl	0.048	0.256	0.219	0.328	0.015	0.034	0.099
1249_ctrl	0.065	0.188	0.404	0.130	0.025	0.021	0.168
1493_ctrl	0.274	0.103	0.329	0.080	0.012	0.064	0.138
1511_ctrl	0.103	0.105	0.426	0.108	0.017	0.041	0.201
1598_ctrl	0.166	0.066	0.530	0.098	0.020	0.053	0.069
101_ctrl	0.126	0.238	0.346	0.087	0.025	0.094	0.083
107_ctrl	0.086	0.409	0.309	0.039	0.016	0.063	0.078
1015_ctrl	0.167	0.287	0.312	0.066	0.011	0.093	0.065
1016_ctrl	0.073	0.215	0.246	0.313	0.011	0.067	0.075
1039_ctrl	0.063	0.293	0.440	0.060	0.021	0.079	0.044
1244_ctrl	0.062	0.223	0.573	0.031	0.025	0.030	0.056
1256_ctrl	0.103	0.186	0.478	0.063	0.012	0.033	0.125
1488_ctrl	0.108	0.147	0.587	0.029	0.021	0.050	0.058
101_stim	0.118	0.218	0.327	0.092	0.030	0.124	0.090
107_stim	0.107	0.323	0.359	0.030	0.026	0.069	0.087
1015_stim	0.146	0.287	0.316	0.057	0.013	0.097	0.084
1016_stim	0.068	0.204	0.217	0.305	0.010	0.078	0.119
1039_stim	0.066	0.273	0.460	0.0530	0.029	0.078	0.041
1244_stim	0.065	0.199	0.588	0.022	0.022	0.024	0.079
1256_stim	0.103	0.188	0.474	0.056	0.016	0.049	0.114
1488_stim	0.111	0.137	0.590	0.017	0.022	0.057	0.067

Table B.2: Number of genes showing DE signals (called by Wilcoxon rank-sum test) in different number of samples in PBMC Lupus data.

Number of samples showing DE signal	B cells	CD14+ Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A+ Monocytes	NK cells
24	22	167	96	6	10	71	26
23	10	83	19	3	14	59	5
22	12	65	19	1	12	48	5
21	7	63	19	3	23	35	4
20	6	65	17	4	12	60	7
19	7	58	17	1	24	62	8
18	14	71	14	5	23	60	6
17	15	75	32	2	22	84	4
16	12	108	24	5	43	63	7
15	13	91	21	6	62	66	5
14	19	94	36	6	76	80	10
13	24	102	35	3	87	77	14
12	22	70	43	3	111	81	13
11	24	100	32	5	141	118	12
10	37	124	47	10	160	132	13
9	30	110	66	11	180	128	22
8	54	135	69	16	249	133	27
7	88	150	87	28	328	183	36
6	92	153	140	30	395	237	38
5	107	184	131	36	534	301	44
4	171	196	192	48	640	349	65
3	249	282	257	83	724	499	100
2	445	402	376	182	810	701	229
1	940	725	740	699	821	984	816
0	3811	2558	3702	5035	730	1620	4715

Table B.3: Number of genes showing DE signals (called by Wilcoxon rank-sum test) in different number of samples in PBMC Lupus data.

Posterior Probability	B cells	CD14+ Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A+ Monocytes	NK cells
pp = 0	4166	4813	3263	4484	5259	5040	3677
$0 <$ pp ≤ 0.95	1132	431	1149	1095	567	453	1348
pp > 0.95	933	987	1819	652	405	738	1206

The genes are categorized into three types: pp = 0 represents genes with negative mean LFC that discarded in the second step of EM algorithm; $0 <$ pp ≤ 0.95 represents genes failed to be identified as CTS genes in third step of EM algorithm; pp > 0.95 represents genes identified as CTS genes.

Table B.4: Number of CTS genes falling in different categories of frequency to show DE and LFC level.

Freq (q_{gk})	LFC (m_{gk})	B cells	CD14+ Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A+ Monocytes	NK cells
0-0.25	0-0.5	1	11	20	8	0	1	4
	0.5-0.1	5	9	0	1	1	13	5
	> 1	0	2	0	0	1	0	0
0.25-0.5	0-0.5	136	67	308	109	16	54	147
	0.5-0.1	21	26	6	25	27	14	29
	> 1	2	10	1	1	6	2	2
0.5-0.75	0-0.5	273	153	605	168	69	124	350
	0.5-0.1	88	26	15	72	59	18	116
	> 1	1	11	2	7	17	6	8
0.75-1	0-0.5	182	437	661	141	89	285	240
	0.5-0.1	195	149	176	96	81	144	247
	> 1	29	86	25	24	39	77	58

The CTS genes identified with posterior probability greater than 0.95 are categorized into different categories. "Freq (q_{gk})" is the probability to show DE signal in a random picked sample estimated with proposed method. "LFC (m_{gk})" is the mean of LFC in samples estimated with proposed method.

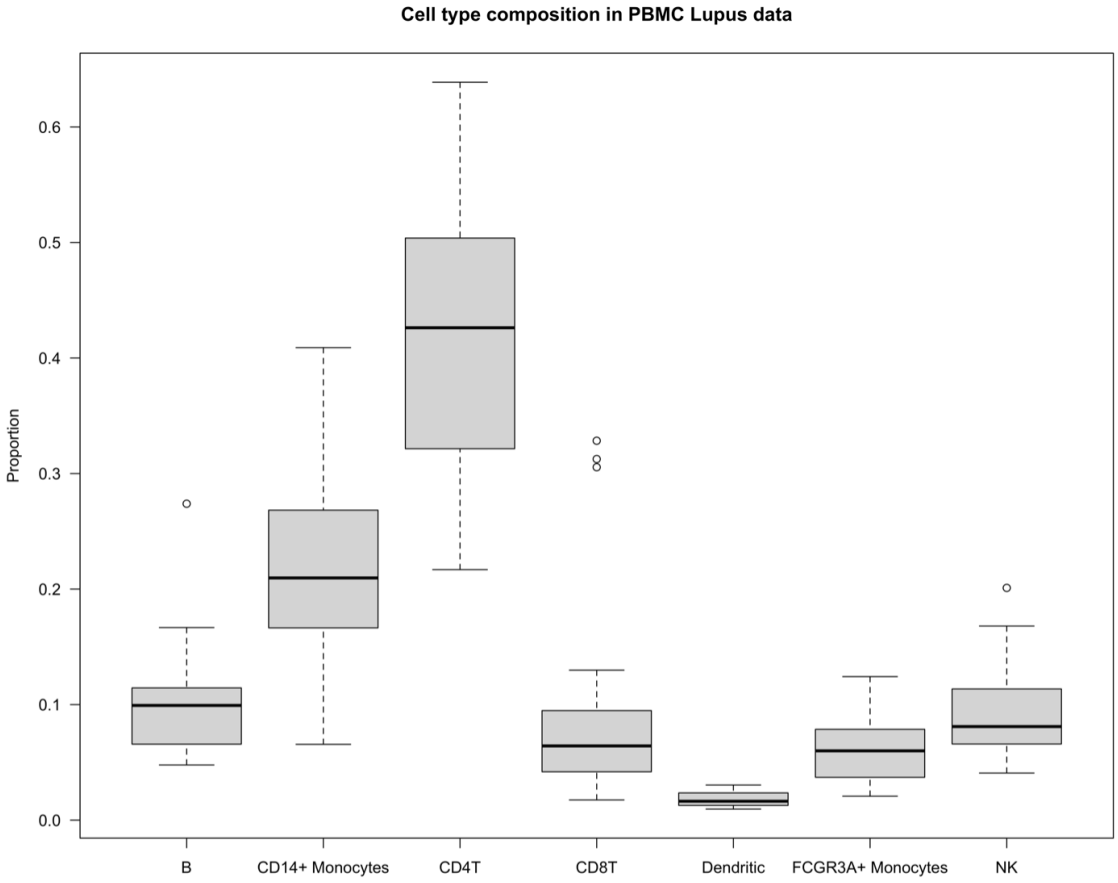

Figure B.1: Cell type composition of samples in PBMC Lupus data. There are 24 samples in the data set. The y-axis is the cell type proportion, and the x-axis is the cell type.

Appendix C

Appendix for Chapter 4

C.1 Details of benchmark pipeline for each method

C.1.1 Seurat

Functions used in this analysis are all from Seurat package. Raw data was first normalized to 10k with function *NormalizeData*. Then top 2000 most variable genes were selected by function *FindVariableFeatures*. Before dimension reduction with PCA, data was scaled with function *ScaleData*. The first 30 PCA components were kept for computing shared nearest neighbor (SNN) graph. The final clustering was performed by Leiden algorithm with different resolutions (0.01, 0.05, 0,1, 0.3, 0.5, 0.7, 1.0, 1.2, 1.4) on the constructed SNN graph.

C.1.2 RaceID

Genes with no expression in all cells were filtered from raw data. We applied this less rigorous criteria to keep more genes that are potential related to RCP cells. A K-medoids clustering was performed on the filtered raw data in which the distance between cells is calculated by Pearson coefficient correlation. The number of clusters (value of K) is provided as the number of major cell types (two in our scenarios). All

parameters required for outlier identification with function *findoutliers* are default value.

In RaceID output, there are two clustering results that one is first round K-medoids clustering result (stored in *clustering*\$*kpart*), the other one is the final clustering result with additional clusters (stored in *cpart*) representing identified outliers. Thus, in evaluation of binary classification performance, we merged all these outliers into one group and treated them as RCP cells identified by RaceID. In evaluation of performance for distinguishing multiple RCP groups, these outlier clusters are kept (not merged) for ARI and NMI calculation.

C.1.3 CellSIUS

In analysis with CellSIUS, the normalization was performed with *SCRAN* package (L Lun et al., 2016), which was recommended by CellSIUS work (Wegmann et al., 2019). CellSIUS requires a "rough" clustering result as input and identifies outlier cells from these clusters. It provides two default clustering methods - hierarchical clustering ("hlucst") and "igraph" clustering. In the analysis, we tried the two clustering methods separately.

For the parameters used in main function *CellSIUS*, we set "min_n_cells" equals to 3 instead of default value 10. Because this specifies the minimum number of cells per mode, a value 3 allows CellSIUS to deal with scenarios in which RCP size is 5 cells. We tried three different combinations of parameters "min_fc" and "fc_between_cutoff": $(2, 1)$, $(1, 0.5)$, and $(0.5, 0.25)$. The default combination is the first one - "min_fc" equals to 2 and "fc_between_cutoff" equals to 1. "min_fc" represents the minimum difference in mean (log2) between the two modes of the gene expression distribution and "fc_between_cutoff" represents minimum difference (log2) in gene expression between cells in the sub-cluster and all other cells. The higher, the more cluster-specific is the gene signature. In the tutorial, "fc_between_cutoff" is required not to be set

higher than "min_fc". We tried the other two less rigorous thresholds to avoid outlier genes mis-filtered. So in total, for a single data, there are six runs with combination of different clustering method ("hclust" and "igraph") and three different sets of fold change thresholds. We picked the combination with best result to report as final CellSIUS result.

In CellSIUS result, sub-clusters (a.k.a the RCP) of a major cluster (identified with "hclust" or "igraph") will be marked with underscore symbol. Thus, in evaluation of binary classification performance, we merged all these sub-clusters into one group and treated them as RCP cells identified by CellSIUS. In evaluation of performance for distinguishing multiple RCP groups, these sub-clusters are kept (not merged) for ARI and NMI calculation.

C.1.4 EDGE

In analysis of EDGE, the raw data is normalized by median normalization and transformed with log2 by adding pseudo-count one. All parameters are set with default value except the hash table size and number of weak learners. The optimal hash table size suggested by EDGE is 1,017,881, however this requires large computation resources. Thus we used the EDGE recommended hash table size 101,107, which balance the computation resources and the accuracy. The number of weak learners is 10,000, which is a recommended value. After deriving the EDGE embeddings, we used it to construct SNN graph and derived the clustering result with Leiden algorithm with resolutions (0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 1.2, 1.4).

C.1.5 GapClust

In analysis of GapClust, the raw data is normalized by median normalization. The number of neighbors to consider and the upper limit of minor cluster size is set as 200.

C.1.6 FiRE

In the analysis of FiRE, raw data was pre-processed by function *ranger_preprocess*, in which data is normalized by median normalization and top 1000 most variable genes are kept for following analysis. Same as EDGE, the number of weak learner is 10,000, and the hash table size is 101,107. The other parameters are all default values. In FiRE, the default threshold used for identify RCP cells is that FiRE score $\leq q3 + 1.5 \times IQR$, where $q3$ and IQR denote the third quartile and the interquartile range (75th percentile − 25th percentile). We also tried two less rigorous thresholds: $q3 + 1.0 \times IQR$ and $q3 + 0.5 \times IQR$ to detect RCP cells. The threshold with best result was used as final FiRE result.

C.1.7 CIARA

In the analysis CIARA, all pre-processes and clustering steps follow its online tutorial and all parameter values are the same as the example analysis shown in `https://github.com/ScialdoneLab/CIARA` except parameters in the function used for pre-selecting genes that have higher expression a small group cells (*get_background_full*). In the default setting, genes expression greater than 1 in at least 3 cells and at most 20 cells are kept for next CIARA test step. This step only keeps genes uniquely expressed in certain cell types. However, any genes with higher expression than the other group could be marker gene for RCP group. So, we changed the parameters values to keep genes only have expression greater than m, where $m = 1, 2, ..., 14$, in at least 3 cells and at most 50 cells.

C.1.8 MicroCellClust (MCC1)

In the analysis of MicroCellClust, raw data was first normalized to 10K and then log10 transformed with pseudo count 0.1. In this way, small and negligible gene expression

can be transformed to negative value. Then genes expressed in more than 25% cells were filtered out. All other values of parameters are same as recommended by the author. The number of RCP groups in data is provided to MicroCellClust that it will generate corresponding number of RCP clusters by repeated running.

C.1.9 SCISSORS

In analysis of SCISSORS, a regular clustering procedure with Seurat was first performed. In this procedure, top n highly variable genes were selected first, where n is the sum of CTS genes in all cell types; top 30 principle components used for non-linear dimension reduction; the other parameter values follow default settings. The generated clusters with size greater than 50 cells are further submitted to the second round clustering. In the second round clustering procedure, top n highly variable genes were selected first, where n is the sum of CTS genes in all cell types; top 30 principle components used for non-linear dimension reduction. In addition, the k value for k-nearest neighborhood varies is set as 3, 5, 10, 25 or 50 and the clustering resolution is 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 1.0, 1.2, and 1.4. Thus, there are $5 \times 9 = 45$ combinations tried in the second round clustering.

C.1.10 SCMER

In the analysis of SCMER, raw data was first normalized to 10K and log transformed with pseudo count 1 by *Scanpy* (Wolf et al., 2018). The highly variable genes were selected by function *highly_variable_genes* with parameter "max_mean" equals to 10. Genes were selected by function *UmapL1*, in which the parameter lasso was set as 0.001 and other parameters were set as their default values. After deriving the selected genes, regular clustering pipeline (scale data, PCA dimension reduction, construct nearest neighborhood, clustering) used in Scanpy tutorial was applied with default settings. The clustering was completed by Leiden algorithm with resolutions: 0.01,

0.05, 0.1, 0.3, 0.5, 0.7, 1.0, 1.2, and 1.4.

C.1.11 scAIDE

In the analysis of scAIDE, raw data was normalized to 10K and then log transformed with pseudo count 1. The autoencoder-imputed distance-preserved embedding was derived by function *AIDE* with default settings. The final RPHKMeans clustering was performed on the previous derived embedding with cluster number set as the total number of cell types in the data.

C.1.12 GiniClust3

In analysis of GiniClust3, genes have no expression in all cells were filtered out. We followed instruction of Giniclust and Ginlcust2 that data does not need to be normalized. In the *calGini* function, the p-value threshold is set as 0.001 and the minimum Gini value is set as 0.2. In the *clusterGini* function, the Gini neighbor was set as 5. The next Fano clustering and Concensus clustering follow default settings.

We used the rare clustering result, which is generated by Gini clustering, for evaluation. In the rare clustering result, all clusters with label greater than "0" are thought as identified RCP clusters. Thus, in evaluation of binary classification performance, we merged all these RCP clusters into one group and treated them as RCP cells identified by Giniclust3. In evaluation of performance for distinguishing multiple RCP groups, these RCP clusters are kept (not merged) for ARI and NMI calculation.

C.1.13 SCA

In analysis of SCA, raw data was first normalized to 10K and log transformed with pseudo count 1. All parameters follow default to derive the top 50 Shannon com-

ponents. The nearest neighborhood was constructed with the 50 components and clustering was completed by Leiden algorithm with resolution: 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 1.0, 1.2, and 1.4.

C.1.14 DoRC

In the analysis of DoRC, raw data was pre-processed by function *ranger_preprocess*, in which data is normalized by median normalization and top 1000 most variable genes are kept for following analysis. The RCP identification process follows default settings of function *dorc*.

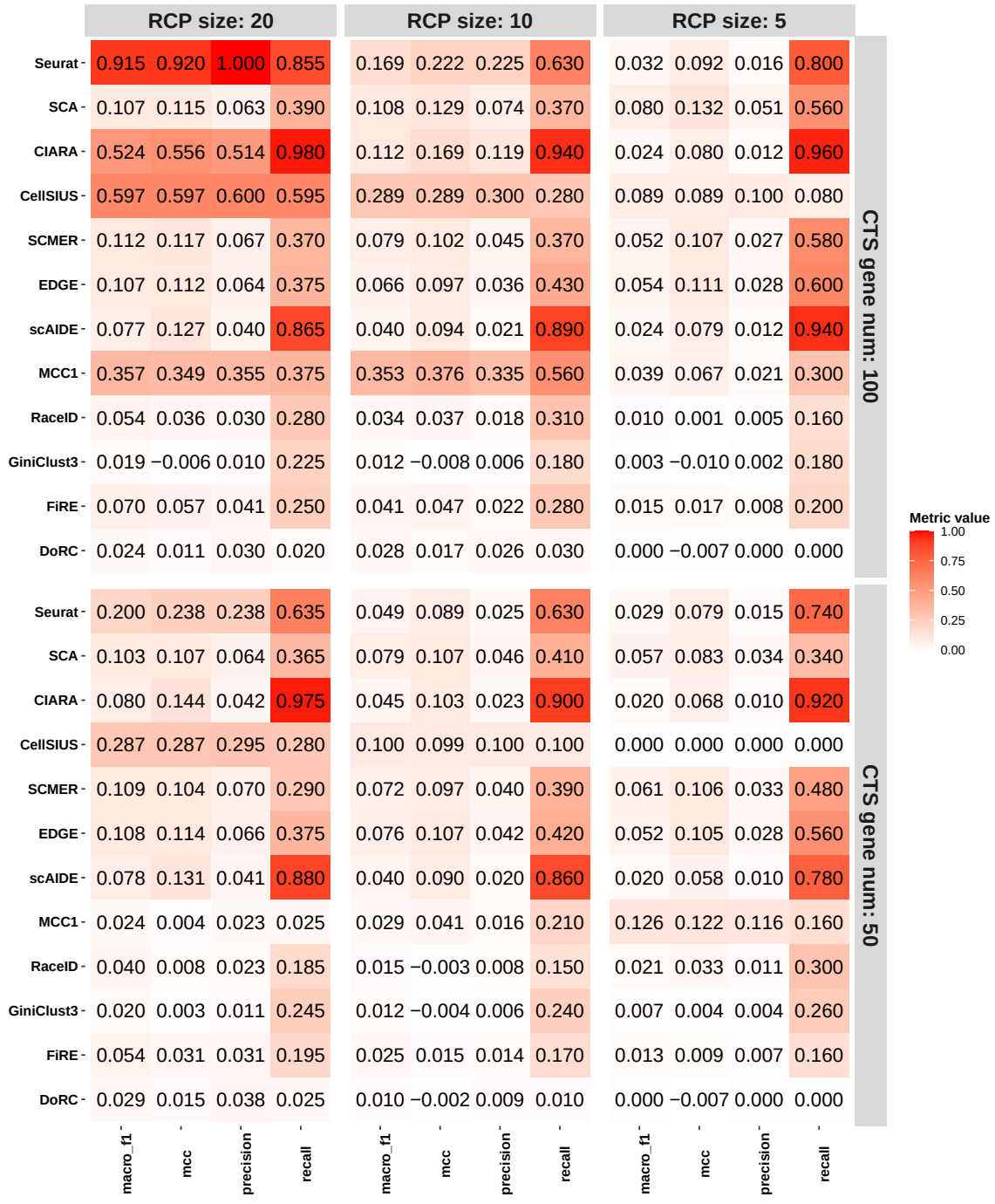

Figure C.1: Evaluation of methods performance in RCP identification when single RCP group ("sub-ct" cell type) exist. Only one RCP group and two major cell types exist in the data. The cell number in major cell type is 500, and CTS gene number of major cell type is 200. The cell number of RCP varies between 5, 10 20, and the CTS gene number of RCP varies between 50, 100. The result is averaged by 10 simulations. GapClust is not included because it fails to report any RCP cells.

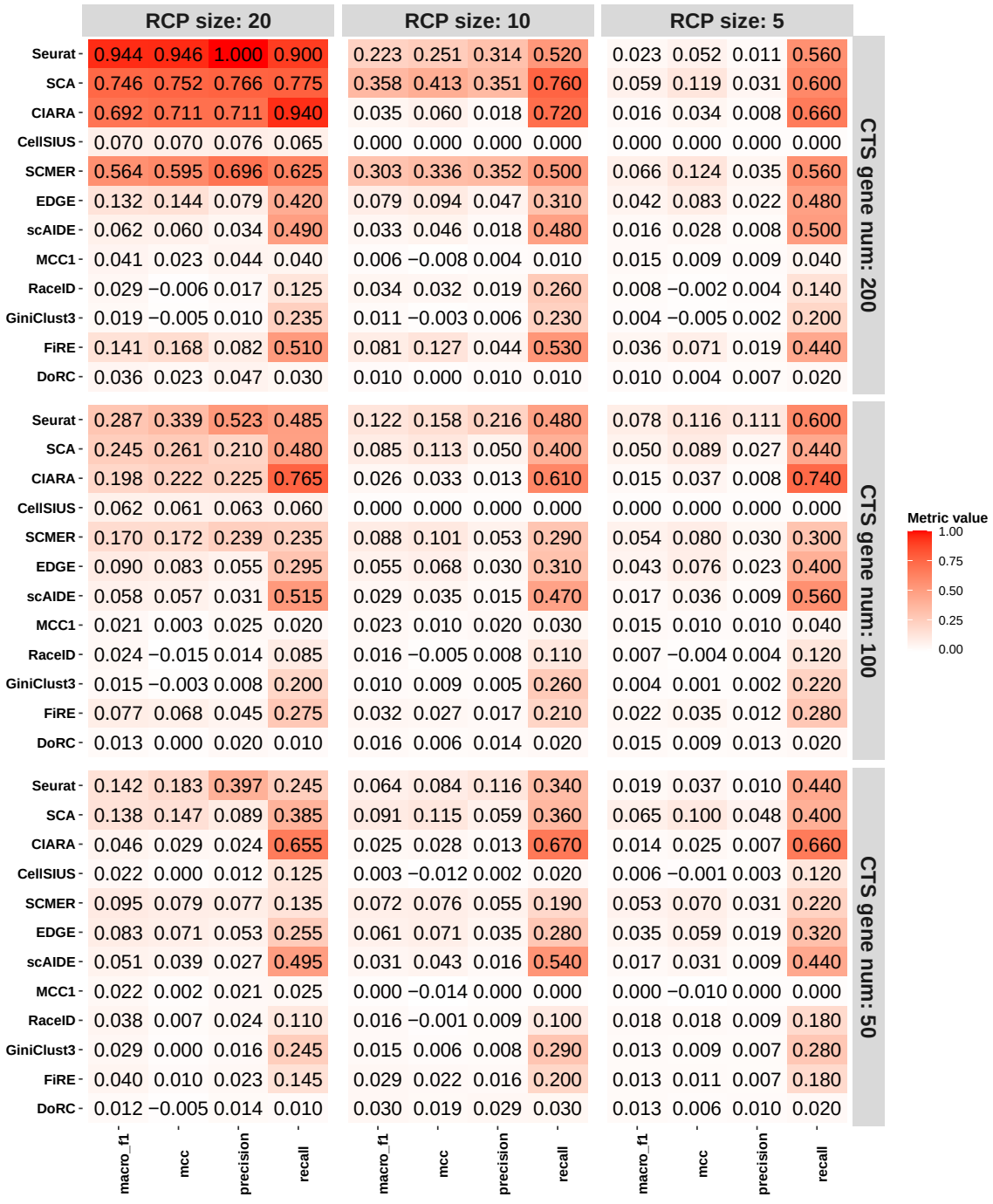

Figure C.2: Evaluation of methods performance in RCP identification when single RCP group exist ("transit-ct" cell type). Only one RCP group and two major cell types exist in the data. The cell number in major cell type is 500, and CTS gene number of major cell type is 200. The cell number of RCP varies between 5, 10 20, and the CTS gene number of RCP varies between 50, 100, 200. The result is averaged by 10 simulations. GapClust is not included because it fails to report any RCP cells.

Bibliography

Abbas, A. R., Baldwin, D., Ma, Y., Ouyang, W., Gurney, A., Martin, F., Fong, S., van Lookeren Campagne, M., Godowski, P., Williams, P., et al. (2005). Immune response in silico (iris): immune-specific genes identified from a compendium of microarray expression data. *Genes & Immunity*, 6(4):319–331.

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7):e6098.

Absher, D. M., Li, X., Waite, L. L., Gibson, A., Roberts, K., Edberg, J., Chatham, W. W., and Kimberly, R. P. (2013). Genome-wide dna methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to cd4+ t-cell populations. *PLoS genetics*, 9(8):e1003678.

Aevermann, B., Zhang, Y., Novotny, M., Keshk, M., Bakken, T., Miller, J., Hodge, R., Lelieveldt, B., Lein, E., and Scheuermann, R. H. (2021). A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell rna sequencing. *Genome research*, 31(10):1767–1780.

Alix-Panabières, C. and Pantel, K. (2014). Challenges in circulating tumour cell research. *Nature Reviews Cancer*, 14(9):623–631.

Andreatta, M., Berenstein, A. J., and Carmona, S. J. (2022). scgate: marker-based purification of cell types from heterogeneous single-cell rna-seq datasets. *Bioinformatics*, 38(9):2642–2644.

Andrews, S. V., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Fallin, M. D. (2016). "gap hunting" to characterize clustered probe signals in illumina methylation array data. *Epigenetics & chromatin*, 9(1):1–21.

Andrews, T. S. and Hemberg, M. (2018). Identifying cell populations with scrnaseq. *Molecular aspects of medicine*, 59:114–122.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44.

Bej, S., Galow, A.-M., David, R., Wolfien, M., and Wolkenhauer, O. (2021). Automated annotation of rare-cell types from single-cell rna-sequencing data through synthetic oversampling. *BMC bioinformatics*, 22(1):1–17.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., et al. (2013). An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., et al. (2006). High-throughput dna methylation profiling using universal bead arrays. *Genome research*, 16(3):383–393.

Chemin, K., Gerstner, C., and Malmström, V. (2019). Effector functions of cd4+ t cells at the site of local autoimmune inflammation—lessons from rheumatoid arthritis. *Frontiers in immunology*, 10:353.

Chen, X., Wu, F.-X., Chen, J., and Li, M. (2019). Dorc: Discovery of rare cells from ultra-large scrna-seq data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 111–116. IEEE.

Chen, Y., Lun, A. T., and Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Research*, 5.

Cici, D., Corrado, A., Rotondo, C., and Cantatore, F. P. (2019). Wnt signaling and biological therapy in rheumatoid arthritis and spondyloarthritis. *International journal of molecular sciences*, 20(22):5552.

Cleveland, W., Grosse, E., and Shyu, W. (1992). Local regression models. chapter 8 in statistical models in s (jm chambers and tj hastie eds.), 608 p. *Wadsworth & Brooks/Cole, Pacific Grove, CA*.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Danila, D. C., Heller, G., Gignac, G. A., Gonzalez-Espinoza, R., Anand, A., Tanaka, E., Lilja, H., Schwartz, L., Larson, S., Fleisher, M., et al. (2007). Circulating

tumor cell number and prognosis in progressive castration-resistant prostate cancer. *Clinical cancer research*, 13(23):7053–7058.

De Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F. C. (2019). Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic acids research*, 47(16):e95–e95.

DeMeo, B. and Berger, B. (2021). Discovering rare cell types through information-based dimensionality reduction. *bioRxiv*, pages 2021–01.

Deng, Y., Bao, F., Dai, Q., Wu, L. F., and Altschuler, S. J. (2019). Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nature methods*, 16(4):311–314.

Diez-Silva, M., Dao, M., Han, J., Lim, C.-T., and Suresh, S. (2010). Shape and biomechanical characteristics of human red blood cells in health and disease. *MRS bulletin*, 35(5):382–388.

Dong, R. and Yuan, G.-C. (2020). Giniclust3: a fast and memory-efficient tool for rare cell type identification. *BMC bioinformatics*, 21(1):1–7.

Dörner, T. and Burmester, G. R. (2003). The role of b cells in rheumatoid arthritis: mechanisms and therapeutic targets. *Current opinion in rheumatology*, 15(3):246–252.

Dumitrascu, B., Villar, S., Mixon, D. G., and Engelhardt, B. E. (2021). Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature communications*, 12(1):1186.

Enkhbat, M., Liu, Y.-C., Kim, J., Xu, Y., Yin, Z., Liu, T.-M., Deng, C.-X., Zou, C., Xie, X., Li, X., et al. (2021). Expansion of rare cancer cells into tumoroids for therapeutic regimen and cancer therapy. *Advanced Therapeutics*, 4(7):2100017.

Fa, B., Wei, T., Zhou, Y., Johnston, L., Yuan, X., Ma, Y., Zhang, Y., and Yu, Z. (2021). Gapclust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. *Nature Communications*, 12(1):4197.

Fan, X., Dong, J., Zhong, S., Wei, Y., Wu, Q., Yan, L., Yong, J., Sun, L., Wang, X., Zhao, Y., et al. (2018). Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell rna-seq analysis. *Cell research*, 28(7):730–745.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69.

Feng, Z., Ren, X., Fang, Y., Yin, Y., Huang, C., Zhao, Y., and Wang, Y. (2020). sctim: seeking cell-type-indicative marker from single cell rna-seq data by consensus optimization. *Bioinformatics*, 36(8):2474–2485.

Feng, Z.-Y. and Wang, Y. (2018). Elf: extract landmark features by optimizing topology maintenance, redundancy, and specificity. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(2):411–421.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13.

Fischer, S. and Gillis, J. (2021). How many markers are needed to robustly determine a cell's type? *Iscience*, 24(11):103292.

Fortin, J.-P. and Hansen, K. D. (2015). Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16(1):1–23.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M., and Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology*, 15(11):1–17.

Fortin, J.-P., Triche Jr, T. J., and Hansen, K. D. (2017). Preprocessing, normalization and integration of the illumina humanmethylationepic array with minfi. *Bioinformatics*, 33(4):558–560.

Franzén, O., Gan, L.-M., and Björkegren, J. L. (2019). Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019.

Gaujoux, R. and Seoighe, C. (2013). Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29(17):2211–2212.

Georges, P. C. and Janmey, P. A. (2005). Cell type-specific response to growth on soft materials. *Journal of applied physiology*, 98(4):1547–1553.

Gerniers, A., Bricard, O., and Dupont, P. (2021). Microcellclust: mining rare and highly specific subpopulations from single-cell expression data. *Bioinformatics*, 37(19):3220–3227.

Gerniers, A. and Dupont, P. (2022). Microcellclust 2: a hybrid approach for multivariate rare cell mining in large-scale single-cell data. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 148–153. IEEE.

Gini, C. (1912). Variabilità e mutabilità reprinted in memorie di metodologica statis-

tica ed e pizetti and t salvemini (rome: Libreria eredi virgilio veschi) go to reference in article.

Grubman, A., Chew, G., Ouyang, J. F., Sun, G., Choo, X. Y., McLean, C., Simmons, R. K., Buckberry, S., Vargas-Landin, D. B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature neuroscience*, 22(12):2087–2097.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and Van Oudenaarden, A. (2015). Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.

Gu, J., Barrera, J., Yun, Y., Murphy, S. K., Beach, T. G., Woltjer, R. L., Serrano, G. E., Kantor, B., and Chiba-Falek, O. (2021). Cell-type specific changes in dna methylation of snca intron 1 in synucleinopathy brains. *Frontiers in neuroscience*, page 493.

Guintivano, J., Aryee, M. J., and Kaminsky, Z. A. (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302.

Guo, H. and Li, J. (2021). scsorter: assigning cells to known cell types according to marker genes. *Genome biology*, 22(1):1–18.

Hannon, E., Mansell, G., Walker, E., Nabais, M. F., Burrage, J., Kepa, A., Best-Lane, J., Rose, A., Heck, S., Moffitt, T. E., et al. (2021). Assessing the co-variability of dna methylation across peripheral cells and tissues: Implications for the interpretation of findings in epigenetic epidemiology. *PLoS genetics*, 17(3):e1009443.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2021). impute: impute: Imputation for microarray data. *R package version 1.68.0*.

Herman, J. S. and Grün, D. (2018). Fateid infers cell fate bias in multipotent progenitors from single-cell rna-seq data. *Nature methods*, 15(5):379–386.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13:1–16.

Hu, J., Li, X., Hu, G., Lyu, Y., Susztak, K., and Li, M. (2020). Iterative transfer learning with neural network for clustering and cell type classification in single-cell rna-seq analysis. *Nature machine intelligence*, 2(10):607–618.

Ianevski, A., Giri, A. K., and Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature communications*, 13(1):1246.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209.

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting immune circuits by linking crispr-pooled screens with single-cell rna-seq. *Cell*, 167(7):1883–1896.

Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome biology*, 17(1):1–13.

Jin, C., Chen, M., Lin, D.-Y., and Sun, W. (2021). Cell-type-aware analysis of rna-seq data. *Nature computational science*, 1(4):253–261.

Jindal, A., Gupta, P., and Sengupta, D. (2018). Discovery of rare cells from voluminous single cell expression data. *Nature communications*, 9(1):4719.

Julià, A., Absher, D., López-Lasanta, M., Palau, N., Pluma, A., Waite Jones, L., Glossop, J. R., Farrell, W. E., Myers, R. M., and Marsal, S. (2017). Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in b cells. *Human Molecular Genetics*, 26(14):2803–2811.

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). Kegg: integrating viruses and cellular organisms. *Nucleic acids research*, 49(D1):D545–D551.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kang, D. W., Park, M.-K., Oh, H.-J., Lee, D.-G., Park, S.-H., Choi, K.-Y., Cho, M.-L., and Min, D. S. (2013). Phospholipase d1 has a pivotal role in interleukin-1β-

driven chronic autoimmune arthritis through regulation of nf-κb, hypoxia-inducible factor 1α, and foxo3a. *Molecular and cellular biology*, 33(14):2760–2772.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., et al. (2018). Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94.

Kim, H. J., Wang, K., Chen, C., Lin, Y., Tam, P. P., Lin, D. M., Yang, J. Y., and Yang, P. (2021). Uncovering cell identity through differential stability with cepo. *Nature Computational Science*, 1(12):784–790.

Kim, J.-H., Ho, S. B., Montgomery, C. K., and Kim, Y. S. (1990). Cell lineage markers in human pancreatic cancer. *Cancer*, 66(10):2134–2143.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., et al. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620.

Kreso, A. and Dick, J. E. (2014). Evolution of the cancer stem cell model. *Cell stem cell*, 14(3):275–291.

Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L., and Luthi-Carter, R. (2011). Population-specific expression analysis (psea) reveals molecular changes in diseased brain. *Nature methods*, 8(11):945–947.

Kular, L., Liu, Y., Ruhrmann, S., Zheleznyakova, G., Marabita, F., Gomez-Cabrero, D., James, T., Ewing, E., Lindén, M., Górnikiewicz, B., et al. (2018). Dna methyla-

tion as a mediator of hla-drb1* 15: 01 and a protective variant in multiple sclerosis. *Nature communications*, 9(1):2397.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97.

L Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*. SIAM.

Leary, J., Xu, Y., Morrison, A., Jin, C., Shen, E. C., Su, Y., Rashid, N., Yeh, J. J., and Peng, X. L. (2021). Sub-cluster identification through semi-supervised optimization of rare-cell silhouettes (scissors) in single-cell sequencing. *bioRxiv*, pages 2021–10.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043.

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X., and Zhang, Z. (2020a). Scibet as a portable and fast single cell type identifier. *Nature communications*, 11(1):1818.

Li, D., Ding, J., and Bar-Joseph, Z. (2022a). Unsupervised cell functional annotation for single-cell rna-seq. *Genome Research*, 32(9):1765–1775.

Li, T. and Ding, C. (2008). Weighted consensus clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 798–809. SIAM.

Li, Z., Guo, Z., Cheng, Y., Jin, P., and Wu, H. (2020b). Robust partial reference-free cell composition estimation from tissue expression. *Bioinformatics*, 36(11):3431–3438.

Li, Z., Wang, Y., Ganan-Gomez, I., Colla, S., and Do, K.-A. (2022b). A machine learning-based method for automatically identifying novel cells in annotating single-cell rna-seq data. *Bioinformatics*, 38(21):4885–4892.

Li, Z. and Wu, H. (2019). Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome biology*, 20(1):1–17.

Li, Z., Wu, Z., Jin, P., and Wu, H. (2019). Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics*, 35(20):3898–3905.

Liang, S., Mohanty, V., Dou, J., Miao, Q., Huang, Y., Müftüoğlu, M., Ding, L., Peng, W., and Chen, K. (2021). Single-cell manifold-preserving feature selection for detecting rare cell populations. *Nature Computational Science*, 1(5):374–384.

Limbach, M., Saare, M., Tserel, L., Kisand, K., Eglit, T., Sauer, S., Axelsson, T., Syvänen, A.-C., Metspalu, A., Milani, L., et al. (2016). Epigenetic profiling in cd4+ and cd8+ t cells from graves' disease patients reveals changes in genes associated with t cell receptor signaling. *Journal of autoimmunity*, 67:46–56.

Linsley, P. S., Speake, C., Whalen, E., and Chaussabel, D. (2014). Copy number loss of the interferon gene cluster in melanomas is linked to reduced t cell infiltrate and poor patient prognosis. *PloS one*, 9(10):e109760.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013). Epigenome-wide

association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2):142–147.

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

Lubatti, G., Stock, M., Iturbe, A., Segura, M. L. R. T., Tyser, R., Theis, F. J., Srinivas, S., Torres-Padilla, M.-E., and Scialdone, A. (2022). Ciara: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell rna seq data. *bioRxiv*, pages 2022–08.

Ludwig, P. E., Reddy, V., and Varacallo, M. (2022). Neuroanatomy, neurons. In *StatPearls [Internet]*. StatPearls Publishing.

Luo, X., Yang, C., and Wei, Y. (2019). Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nature communications*, 10(1):3113.

Maceyka, M. and Spiegel, S. (2014). Sphingolipid metabolites in inflammatory disease. *Nature*, 510(7503):58–67.

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

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome biology*, 13:1–12.

Marston, B., Palanichamy, A., and Anolik, J. H. (2010). B cells in the pathogenesis and treatment of rheumatoid arthritis. *Current opinion in rheumatology*, 22(3):307.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297.

McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, 29(4):461–467.

McKellar, D. W., Walter, L. D., Song, L. T., Mantri, M., Wang, M. F., De Vlaminck, I., and Cosgrove, B. D. (2021). Large-scale integration of single-cell transcriptomic data captures transitional progenitor states in mouse skeletal muscle regeneration. *Communications biology*, 4(1):1280.

Mendioroz, M., Do, C., Jiang, X., Liu, C., Darbary, H. K., Lang, C. F., Lin, J., Thomas, A., Abu-Amero, S., Stanier, P., et al. (2015). Trans effects of chromosome aneuploidies on dna methylation patterns in human down syndrome and mouse models. *Genome biology*, 16(1):1–26.

Miao, Z., Moreno, P., Huang, N., Papatheodorou, I., Brazma, A., and Teichmann, S. A. (2020). Putative cell type discovery from single-cell gene expression data. *Nature methods*, 17(6):621–628.

Micheel, C. M., Nass, S. J., Omenn, G. S., et al. (2012). Omics-based clinical discovery: Science, technology, and applications. In *Evolution of Translational Omics: Lessons Learned and the Path Forward*. National Academies Press (US).

Montaño, C. M., Irizarry, R. A., Kaufmann, W. E., Talbot, K., Gur, R. E., Feinberg, A. P., and Taub, M. A. (2013). Measuring cell-type specific differential methylation in human brain tissue. *Genome biology*, 14(8):1–9.

Moonen, S., Koper, M. J., Van Schoor, E., Schaeverbeke, J. M., Vandenberghe, R., von Arnim, C. A., Tousseyn, T., De Strooper, B., and Thal, D. R. (2023). Pyroptosis in alzheimer's disease: Cell type-specific activation in microglia, astrocytes and neurons. *Acta Neuropathologica*, 145(2):175–195.

Mullen, K. M. and van Stokkum, I. H. M. (2012). Nnls: The lawson-hanson algorithm for non-negative least squares. *R package version 1.4*.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782.

Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644.

Paisley, B. M. and Liu, Y. (2021). Genemarker: a database and user interface for scrna-seq marker genes. *Frontiers in Genetics*, page 1906.

Panwar, B., Schmiedel, B. J., Liang, S., White, B., Rodriguez, E., Kalunian, K., McKnight, A. J., Soloff, R., Seumois, G., Vijayanand, P., et al. (2021). Multi–cell type gene coexpression network analysis reveals coordinated interferon response and cross–cell type correlations in systemic lupus erythematosus. *Genome Research*, 31(4):659–676.

Park, J., Shrestha, R., Qiu, C., Kondo, A., Huang, S., Werth, M., Li, M., Barasch, J., and Suszták, K. (2018). Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*, 360(6390):758–763.

Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.-S., et al. (2019). Single-cell rna-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research*, 29(9):725–738.

Phipson, B., Maksimovic, J., and Oshlack, A. (2015). missmethyl: an r package for analyzing data from illumina's humanmethylation450 platform. *Bioinformatics*, 32(2):286–288.

Plaks, V., Koopman, C. D., and Werb, Z. (2013). Circulating tumor cells. *Science*, 341(6151):1186–1188.

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

Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., Rosset, S., Sankararaman, S., and Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications*, 10(1):3417.

Rauluseviciute, I., Drabløs, F., and Rye, M. B. (2019). Dna methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. *Clinical epigenetics*, 11(1):1–13.

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

San Segundo-Val, I. and Sanz-Lozano, C. S. (2016). Introduction to the gene expression analysis. *Molecular genetics of asthma*, pages 29–43.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502.

Saul, D., Leite Barros, L., Wixom, A. Q., Gellhaus, B., Gibbons, H. R., Faubion, W. A., and Kosinsky, R. L. (2022). Cell type-specific induction of inflammation-associated genes in crohn's disease and colorectal cancer. *International journal of molecular sciences*, 23(6):3082.

Schatton, T., Murphy, G. F., Frank, N. Y., Yamaura, K., Waaga-Gasser, A. M., Gasser, M., Zhan, Q., Jordan, S., Duncan, L. M., Weishaupt, C., et al. (2008). Identification of cells initiating human melanomas. *Nature*, 451(7176):345–349.

Schreier, S., Sawaisorn, P., Udomsangpetch, R., and Triampo, W. (2017). Advances in rare cell isolation: an optimization and evaluation study. *Journal of Translational Medicine*, 15:1–16.

Shaffer, D. R., Leversha, M. A., Danila, D. C., Lin, O., Gonzalez-Espinoza, R., Gu, B., Anand, A., Smith, K., Maslak, P., Doyle, G. V., et al. (2007). Circulating tumor cell analysis in patients with progressive castration-resistant prostate cancer. *Clinical cancer research*, 13(7):2023–2029.

Shelef, M. A., Bennin, D. A., Yasmin, N., Warner, T. F., Ludwig, T., Beggs, H. E., and Huttenlocher, A. (2014). Focal adhesion kinase is required for synovial fibroblast invasion, but not murine inflammatory arthritis. *Arthritis research & therapy*, 16:1–10.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M.,

Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type–specific gene expression differences in complex tissues. *Nature methods*, 7(4):287–289.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941.

Smith, E. A. and Hodges, H. C. (2019). The spatial and genomic hierarchy of tumor ecosystems revealed by single-cell technologies. *Trends in cancer*, 5(7):411–425.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical issues in cdna microarray data analysis. *Functional genomics: methods and protocols*, pages 111–136.

Speake, C., Skinner, S. O., Berel, D., Whalen, E., Dufort, M. J., Young, W. C., Odegard, J. M., Pesenacker, A. M., Gorus, F. K., James, E. A., et al. (2019). A composite immune signature parallels disease progression across t1d subjects. *JCI insight*, 4(23).

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, 101(16):6062–6067.

Su, D., Wang, X., Campbell, M. R., Porter, D. K., Pittman, G. S., Bennett, B. D., Wan, M., Englert, N. A., Crowl, C. L., Gimple, R. N., et al. (2016). Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. *PloS one*, 11(12):e0166486.

Su, K., Yu, T., and Wu, H. (2021). Accurate feature selection improves single-cell rna-seq cell clustering. *Briefings in Bioinformatics*, 22(5):bbab034.

Sun, X., Liu, Y., and An, L. (2020). Ensemble dimensionality reduction and feature gene extraction for single-cell rna-seq data. *Nature communications*, 11(1):5853.

Suvà, M. L. and Tirosh, I. (2020). The glioma stem cell model in the era of single-cell genomics. *Cancer cell*, 37(5):630–636.

Swanson, C. D., Akama-Garren, E. H., Stein, E. A., Petralia, J. D., Ruiz, P. J., Edalati, A., Lindstrom, T. M., and Robinson, W. H. (2012). Inhibition of epidermal growth factor receptor tyrosine kinase ameliorates collagen-induced arthritis. *The Journal of Immunology*, 188(7):3513–3521.

Teh, A. L., Pan, H., Lin, X., Lim, Y. I., Patro, C. P. K., Cheong, C. Y., Gong, M., MacIsaac, J. L., Kwoh, C.-K., Meaney, M. J., et al. (2016). Comparison of methyl-capture sequencing vs. infinium 450k methylation array for methylome analysis in clinical samples. *Epigenetics*, 11(1):36–48.

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC bioinformatics*, 18(1):1–14.

Teschendorff, A. E. and Zheng, S. C. (2017). Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5):757–768.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

Triche Jr, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund,

K. D. (2013). Low-level processing of illumina infinium dna methylation beadarrays. *Nucleic acids research*, 41(7):e90–e90.

Tserel, L., Kolde, R., Limbach, M., Tretyakov, K., Kasela, S., Kisand, K., Saare, M., Vilo, J., Metspalu, A., Milani, L., et al. (2015). Age-related profiling of dna methylation in cd8+ t cells reveals changes in immune response and transcriptional regulator genes. *Scientific reports*, 5(1):13107.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.-C. (2019). Accurate estimation of cell-type composition from gene expression data. *Nature communications*, 10(1):2975.

Tsoucas, D. and Yuan, G.-C. (2018). Giniclust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome biology*, 19:1–13.

van Loosdregt, J., Rossetti, M., Spreafico, R., Moshref, M., Olmer, M., Williams, G. W., Kumar, P., Copeland, D., Pischel, K., Lotz, M., et al. (2016). Increased autophagy in cd4+ t cells of rheumatoid arthritis patients results in t-cell hyperactivation and apoptosis resistance. *European journal of immunology*, 46(12):2862–2870.

Vasilopoulos, Y., Gkretsi, V., Armaka, M., Aidinis, V., and Kollias, G. (2007). Actin cytoskeleton dynamics linked to synovial fibroblast activation as a novel pathogenic principle in tnf-driven arthritis. *Annals of the rheumatic diseases*, 66(suppl 3):iii23–iii28.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., and Kriegstein, A. R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, 364(6441):685–689.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods*, 14(4):414–416.

Wang, J., Roeder, K., and Devlin, B. (2021). Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. *Genome research*, 31(10):1807–1818.

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019a). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380.

Wang, Y., Lloyd, K. A., Melas, I., Zhou, D., Thyagarajan, R., Lindqvist, J., Hansson, M., Svärd, A., Mathsson-Alm, L., Kastbom, A., et al. (2019b). Rheumatoid arthritis patients display b-cell dysregulation already in the naïve repertoire consistent with defects in b-cell tolerance. *Scientific reports*, 9(1):19995.

Wang, Z., Dong, W., Josephson, W., Lv, Q., Charikar, M., and Li, K. (2007). Sizing sketches: a rank-based analysis for similarity search. In *Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 157–168.

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

Wegmann, R., Neri, M., Schuierer, S., Bilican, B., Hartkopf, H., Nigsch, F., Mapa, F., Waldt, A., Cuttat, R., Salick, M. R., et al. (2019). Cellsius provides sensitive and specific detection of rare cell populations from complex single-cell rna-seq data. *Genome biology*, 20(1):1–21.

Westra, H.-J., Arends, D., Esko, T., Peters, M. J., Schurmann, C., Schramm, K.,

Kettunen, J., Yaghootkar, H., Fairfax, B. P., Andiappan, A. K., et al. (2015). Cell specific eqtl analysis without sorting cells. *PLoS genetics*, 11(5):e1005223.

Wilhelm-Benartzi, C. S., Koestler, D. C., Karagas, M. R., Flanagan, J. M., Christensen, B. C., Kelsey, K. T., Marsit, C. J., Houseman, E. A., and Brown, R. (2013). Review of processing and analysis methods for dna methylation array data. *British journal of cancer*, 109(6):1394–1402.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W., et al. (2009). Biogps: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology*, 10(11):1–8.

Wu, Z. and Wu, H. (2020). Accounting for cell type hierarchy in evaluating single cell rna-seq clustering. *Genome biology*, 21(1):1–14.

Xie, K., Huang, Y., Zeng, F., Liu, Z., and Chen, T. (2020). scaide: clustering of large-scale single-cell rna-seq data reveals putative and rare cell types. *NAR genomics and bioinformatics*, 2(4):lqaa082.

Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R. S., Aldridge, A. I., Ament, S. A., Bartlett, A., Behrens, M. M., Van den Berge, K., et al. (2021a). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110.

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., et al. (2021b). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.

Yeung, K. S., Lee, T. L., Mok, M. Y., Mak, C. C. Y., Yang, W., Chong, P. C. Y., Lee, P. P. W., Ho, M. H. K., Choufani, S., Lau, C. S., et al. (2019). Cell lineage-specific genome-wide dna methylation analysis of patients with paediatric-onset systemic lupus erythematosus. *Epigenetics*, 14(4):341–351.

Yoo, H. J., Hwang, W. C., and Min, D. S. (2020). Targeting of phospholipase d1 ameliorates collagen-induced arthritis via modulation of treg and th17 cell imbalance and suppression of osteoclastogenesis. *International Journal of Molecular Sciences*, 21(9):3230.

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

Zborowski, M. and Chalmers, J. J. (2011). Rare cell separation and analysis by magnetic sorting.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019a). Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728.

Zhang, Z., Luo, D., Zhong, X., Choi, J. H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E. W., Modrusan, Z., et al. (2019b). Scina: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes*, 10(7):531.

Zhao, W., Dovas, A., Spinazzi, E. F., Levitin, H. M., Banu, M. A., Upadhyayula, P., Sudhakar, T., Marie, T., Otten, M. L., Sisti, M. B., et al. (2021). Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell rna-seq. *Genome Medicine*, 13(1):82.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049.

Zheng, S. C., Breeze, C. E., Beck, S., and Teschendorff, A. E. (2018a). Identification of differentially methylated cell types in epigenome-wide association studies. *Nature methods*, 15(12):1059–1066.

Zheng, S. C., Webster, A. P., Dong, D., Feber, A., Graham, D. G., Sullivan, R., Jevons, S., Lovat, L. B., Beck, S., Widschwendter, M., et al. (2018b). A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics*, 10(7):925–940.