MENDELIAN RANDOMIZATION AND SINGLE CELL DECONVOLUTION, TWO PROBLEMS IN STATISTICS GENETICS

Xuran Wang

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Nancy R. Zhang, Professor of Statistics

Graduate Group Chairperson

Charles L. Epstein, Thomas A. Scott Professor of Mathematics

Dissertation Committee

Nancy R. Zhang, Professor of Statistics

Dylan S. Small, Class of 1965 Wharton Professor of Statistics

Mingyao Li, Professor of Biostatistics

MENDELIAN RANDOMIZATION AND SINGLE CELL DECONVOLUTION, TWO PROBLEMS IN STATISTICS GENETICS

ACKNOWLEDGEMENT

First and foremost I want to thank my advisor Nancy Zhang. Without her support and guidance, I would not have entered the field of statistical genomics. She has taught me, both consciously and unconsciously, how applied statistics research is done. I appreciate all her contributions of time, ideas, and funding to make my PhD experience productive and stimulating. Her enthusiasm for research was contagious and motivational for me, even through the tough times in the PhD pursuit. The completion of my dissertation would not have been possible without the support and nurturing of Mingyao Li, who was the co-advisor for my major research project as well as my committee member. I'm extremely grateful for the excellent examples Mingyao and Nancy have provided as successful women researchers and professors. They mentored me on my research, pointed out and helped strengthen my weaknesses. I'm sure I will benefit from their invaluable words throughout my academic career. I would also like to extend my deepest gratitude to Dylan Small for being my co-advisor during my first two year of PhD study and research and my committee member. He not only taught me statistics, but mentored me with great patience during my transition period from a student to a beginner researcher. Nancy, Mingyao and Dylan have demonstrated how good research is done as well as what good researcher (and person) looks like.

I would like to extend my sincere thanks to Professor Katalin Susztak from Department of Medicine at University of Pennsylvania as well as Dr.Jihwan Park and Tong Zhou from Susztak's group, who kindly shared their data, constructive advice and valuable experiences during the deconvolution project. I very much appreciate Yan Che, who helped me process data in the Mendelian randomization project. I must also thank Dr.Charles Epstein, the graduate chair of Applied Mathematics and Computational Science program, for recruiting me. Otherwise, I would not have had the opportunity to join Penn and work with those wonderful professors.

Special thanks to the members of Nancy's group, Dr.Jingshu Wang, Chi-Yun Wu, Mo Huang, Zilu Zhou and Divyansh Agarwal, who have contributed immensely to my personal and professional time at Penn. The group has been a great source of friendship as well as good advice and collaboration. I'm grateful for those days and nights we spent together working in the office, and especially Mo, for helping me edit this thesis. I would also like to acknowledge friends of Nancy's group, Dr.Qingyuan Zhao and Professor Jian Ding for providing encouragements and stimulating motivations. Thanks should also go to my friends/colleagues at Penn: Lu Wang and Ruoqi Yu, whose professional and emotional support cannot be overestimated. I'd like to recognize the assistance that I received from the staff of the mathematics department, statistics department and Wharton computing group at Penn.

I'm deeply indebted to my parents for being my strongest support and providing endless and unconditional love. During the tough time of transition from an applied math student to a beginner researcher of statistical genomics, they never lost faith in me and believed that I can succeed even though sometimes I did not.

Lastly, I want to thank myself. With my sensitive and sentimental characteristics, I still made sensible choices. Worn by chronic pains, I never gave up.

Xuran Wang

University of Pennsylvania

April, 2019

ABSTRACT

MENDELIAN RANDOMIZATION AND SINGLE CELL DECONVOLUTION, TWO PROBLEMS IN STATISTICS GENETICS

Xuran Wang

Nancy R. Zhang

Finding interpretable targets within the genome for diseases is a primary goal of biomedical research. This thesis focuses on developing statistical models and methods for analysis of high throughput genomic and transcriptomic sequencing data with the goal of finding actionable targets of two types, disease-associated genes and disease-implicated cell types.

Traditional genome wide association studies(GWAS) focus on finding the association between genetic variants and diseases. However, GWAS results are often difficult to interpret, and they do not directly lead to an understanding of the true biological mechanism of diseases. Following GWAS findings, we can study the causal effect by Mendelian randomization(MR), which uses segregating genomic loci as instrumental variables to estimate the causal effect of a given exposure to disease outcome. In this thesis, we introduced the concept of "localizable exposures", which are exposures that can be localized, or mapped, to a specific region in the genome, such as the expression of a single gene or the methylation of a specific loci. With sequencing technology, allele specific reads are observable for localizable exposures, which allow their quantifications in an allele-specific manner. In the first part of this thesis, we present a new model, ASMR, uses allele-specific information for Mendelian randomization.

This thesis also develops methods for finding cell types implicated in disease through the joint analysis of bulk and single cell RNA sequencing data. Bulk tissue sequencing is often used to probe genes that have tissue-level expression changes between biological cohorts. However, tissue are usually a mixture of multiple distinct cell types and the tissue-level

changes are due to shifts of cell type proportions as well as cell type specific expression changes. Single-cell RNA sequencing (scRNA-seq) allows the investigation of the roles of individual cell types during disease initiation and development. We present MuSiC, a method that utilizes cell-type specific gene expression from single-cell RNA sequencing (RNA-seq) data to characterize cell type compositions from bulk RNA-seq data in complex tissues. When applied to pancreatic islet and whole kidney expression data in human, mouse, and rats, MuSiC outperforms existing methods, especially for tissues with closely related cell types. With MuSiC-estimated cell type proportions, we propose a reverse estimation procedure that can detect cell type specific differential expression, allowing for the elucidation of the roles of genes and cell types, as well as their interactions, on disease phenotypes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

PREFACE

This dissertation is submitted for the degree of Doctor of Philosophy at the University of Pennsylvania. The research described herein was conducted under the supervision of Professor Nancy R. Zhang and Professor Dylan S. Small in the Statistics Department, Professor Mingyao Li in the Biostatistics and Epidemiology Department, between May 2015 and April 2019.

This work is to the best of my knowledge original, except where acknowledgements and references are made to previous work. Neither this, nor any substantially similar dissertation has been or is being submitted for any other degree, diploma or other qualification at any other university.

Part of this work has been presented in the following publication:

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

CHAPTER 1 : Introduction

A central goal in biomedical research is to study the genome and find genes and other biological features that are responsible for diseases. Here we refer to such features on which downstream experiments can be performed to interpret, validate, quantify, and isolate their functional effects as actionable targets. In this thesis, we developed statistical models and computational methods for finding actionable targets using genetic and genomic data. The targets that we consider include the RNA expression of specific gene and the representation of specific cell types in tissue.

1.1. Allele specific Mendelian randomization

In genetics, Genome-wide association studies (GWAS) have been the stable of research for the last twenty years, since the sequencing of human genome allowed for the mapping to genes. The goal of GWAS is to find associations between genetic variants, such as single-nucleotide polymorphisms (SNPs), and observable traits such as diseases. The first successful GWAS published in 2002, found the association between *LTA* and myocardial infarction (Ozaki et al., 2002). Up until now, we have mapped over a hundred thousand genetic variants to various diseases, for example *DRD2* to schizophrenia, *PADI4* and *IL6R* to rheumatoid arthritis, and more (Visscher et al., 2017). Mapping diseases to genes and performing downstream analysis allow us to find treatments to diseases with gene knock-outs, gene editing and developing drugs.

The goal of GWAS is to find genes that are causally implicated in diseases by mapping diseases onto genetic variants. However, association between genetic variants and diseases is not enough to establish causal relationships or to explain the underlying biological mechanism. From the central dogma of biology, RNA is the intermediate molecule that leverages the effect of genetic variants to disease phenotypes. This motivates the study of expression quantitative traits loci (eQTL), defined as genomic loci that explain all or a fraction of the variation in the expression levels of mRNAs. The identification of disease associated

loci and their characterization as eQTLs allow us to quantify the causal effect of gene on disease phenotypes. The use of inherited genetic variation to study the causal effect of an exposure on a trait is called Mendelian randomization (MR). In this thesis, we consider the use of localizable exposures, which we define as exposures that can be localized to a specific region in the genome. For example, we can think of gene expression or epigenetic modification as an "exposure" in the terminology of epidemiologists, and since they can be directly mapping to a genome location, they are localizable exposures. With sequencing techniques, allele specific reads at heterozygous loci are observable, which allow the allele-specific quantifications of localizable exposures. We developed a method, ASMR, that incorporates the allele-specific information in a Mendelian randomization framework to estimate the causal effect of localizable exposure using maximum likelihood. A comparison of precision between two-stage-least-square (2SLS), a conventional estimation method for MR, and ASMR is presented with different strength of instrumental variables and confounder effects. We also illustrate how to find downstream targets of lncRNA using ASMR. This work is described in Chapter 2.

1.2. Cell type contributions to diseases

Unlike genetics, which focuses on segregating genetic variants and their affected genes, genomics is a relatively newer field that involves the whole genome and its quantitative behaviors. The technological evolution in whole genome RNA and DNA sequencing have allowed for genomic studies. A common type of analysis useful in genomics is differential expression analysis (Costa-Silva et al., 2017), which takes gene expression data to detects quantitative changes in expression levels between groups of samples such as between disease cohorts. Essentially, for each gene, statistical tests have been developed to decide whether an observed difference in expression between two cohorts is greater than the expected difference due to random variation. Large-scaled bulk tissue RNA and DNA sequencing datasets, such as Genotype-Tissue Expression (GTEx) Project (Carithers et al., 2015), have been generated, which allows tissue-specific investigation on diseases. However, bulk tissue sequencing data

reflects an average over all cells within the tissue, masking the contribution of individual cell types. It was difficult to sequence at the single cell level before 2009, because it was hard to isolate individual cells, and because the abundance of RNA in a single cell is too little to sequence. Recent advances in the techniques for isolating single cells and for amplifying their genetic material make possible the exploration of the transcriptome of single cells. With the birth of single-cell sequencing, researchers can study the transcriptomic heterogeneity of a single cell for thousands to millions of cells simultaneously.

"Cell type" is a classification used to distinguish between morphologically or phenotypically distinct cell forms. Usually complex tissues consist of cells of different cell types; for example in the human pancreas, there are endocrine cell types and exocrine cell types that jointly regulate glucose level. Cells of the same type show similar transcriptomic pattern, which can be captured by single cell RNA-seq data. Annotating cell types based on single-cell transcription profiles is a challenging problem and can be viewed as high-dimensional unsupervised clustering with high level of biological and technical noises. Currently, popular clustering methods include Seurat (Butler et al., 2018), TSCAN (Ji and Ji, 2016) and SC3 (Kiselev et al., 2017) (more in the recent review by Duò et al. (2018)). After clustering, the cell types are identified by their marker genes, which are genes that only express in a specific cell type. Based on cell types assigned in this way, we can study the inter- and intra-cell-type similarity and heterogeneity from single cell sequencing data and develop methods to study the role of cell types in diseases.

The observed differential expression in bulk tissue is a combined effect of cell type composition shifts and the expression shifts within cell types. Therefore, investigating the differences between groups of samples at the cell type levels can be framed as two problems, detecting cell type composition shifts and detecting cell type specific differentially expressed (DE) genes. Finding cell type specific DE genes is especially challenging because there is confounding from proportion shifts. One may ask, with the rapid adoption of of single cell sequencing, why not use single cell data to investigate cell-type level differences. However,

due to cell loss in the dissociation and isolation steps of single cell sequencing, proportions from single cell data do not reflect the true proportions in bulk tissue. Therefore, we and others(Avila Cobos et al., 2018) proposed the strategy of first estimating cell type proportions of bulk tissues with single cell expression as reference. Such estimation procedures are often referred as deconvolution. There are many existing deconvolution methods, such as CIBERSORT (Newman et al., 2015) and BSEQ-sc (Baron et al., 2016). However, they only use marker genes for estimation and ignore the cross-subject variations of gene expression when multi-subject single cell datasets are available. We developed a method, MuSiC, that takes the advantage of cross-subject variations of all genes in deconvolution without selecting marker genes. This is discussed in Chapter 3.

Now let's go back to the question of how to detect cell type specific differential expression from bulk tissue. Cell type specific differential expression testing starts from estimating cell type proportions by deconvolution guided with single cell reference. Deconvolution methods with pre-selected marker genes assume that the expression of marker genes are consistent across healthy and diseased status, which is not always true. For example, in pancreas islet, *INS* is the genes responsible for producing insulin and is a marker gene for beta cells. During the progress of type 2 diabetes, beta cells go through both loss of mass and lack of function, where there are less beta cells as well as lower *INS* expression in beta cells for diseased subjects. Although the expression of *INS* is higher in beta cells than other cell types, deconvolution with only marker genes like *INS* will mistaken the cell type specific expression changes for cell type proportion shifts. MuSiC eliminates the bias from marker genes by including all genes for deconvolution. Controlling the estimated proportion from MuSiC makes it possible to detect cell type specific DE genes between healthy and diseased status. We proposed a method for testing cell type specific DE genes by comparing two cell type models with and without disease indicator with a strategy of repetition partition of genes in to a set used for deconvolution and a separate set on which DE test is conducted. This is described in Chapter 4.

CHAPTER 2 : Allele specific Mendelian randomization

2.1. Introduction

A common goal of biomedical research is to elucidate the causal roles of genetic and epigenetic factors underlying complex human diseases. Mendelian randomization (MR) is a method for estimating the causal effect of an intermediate variable on an outcome of interest, in which inherited DNA variation are used as instrumental variables to overcome the effect of confounding and reverse causation. Mendelian randomization was adopted as early as 1986, when Katan (2004) used genetic marker *ApoE* as instrument for studying the causal effect of raised blood cholesterol on risk of cancers. Now it has been widely applied to epidemiology and integrative genetics models, more in the review paper by Burgess et al. (2017). For example, to measure the causal effect of alcohol consumption on the risk of coronary heart disease, the genetic variant *ALDH2*, which reduces alcohol consumption, has been used as an instrument (Smith and Ebrahim, 2004). In integrative genetics models, a large number of transcript abundances are measured with the ultimate aim of identifying causal relationships from a morass of expression quantitative trait loci (eQTL) using Mendelian randomization. For other examples, see recent review by Burgess and Thompson (2015).

To date, the "modifiable exposure" examined by Mendelian randomization studies have encompassed blood pressure, obesity, smoking, alcohol intake etc. Increasingly, with the ubiquitous adoption of high throughput sequencing technologies, the "exposure" of interest can now also focus in on the expression of a single gene or the methylation of a specific loci. We call such exposures **localizable exposures**, in that they can be localized, or mapped, to a specific region in the genome. High throughput sequencing has revolutionized the study of the transcriptome and the epigenome. One important benefit of sequencing is that it quantifies gene expression or epigenetic modification in an allele-specific manner, such that at heterozygous loci, the sequenced read indicates which locus it comes from. For

example, in RNA sequencing, this allows us to measure the number of transcript copies of each allele. This allele-level information has not, to our knowledge, been used in Mendelian randomization studies. We will develop a method for using such allele-level information and show that it can substantially boost estimation accuracy. Although this chapter focuses on gene expression, the methods we propose can be applied to other types of localizable exposures, such as methylation or splicing.

The measurement of allele-specific information in heterozygous individuals enables us to compare the difference between expressions of two paired alleles and consider expression level of different genotypes in a causal inference manner, where the causal effect is the difference between potential outcomes of the same subject. With the same idea, we use "dosage" to describe the difference between potential expressions of different alleles at the same loci, thus dosage measures the causal effect of changing the allele. The causal effect can not be measured directly because at each loci, at each haplotype, only one of the two potential outcomes can be observed. Even for heterozygous individuals, where the expression of both allele can be observed, the expression of each allele is a separate random variable which may be correlated due to the sharing of common environment, but cannot be considered to have the same potential outcomes.

The focus on localizable exposues has the additional benefit of allowing us to more easily satisfy the assumptions of Mendelian Randomization. The use of DNA-level variations as instrumental variables relies on two basic assumptions: First, the DNA variant must be stably maintained in our cells and not correlated with lifestyle and environment; this assumption is reasonable in most cases after proper adjustment for genetic ancestry. Second, the DNA variant must affect the outcome only through the exposure of interest, and this assumption is often violated, and very difficult to check due to pleiotropy. However, when the modifiable exposure is localizable, we can select to use only genetic variants that are physically close to the exposure on the genome. In genetics terminology, we limit our instruments to variants that affect the exposure in *cis*, as opposed to variants that act in

trans.

The classical approach for analyzing the causal effect in Mendelian randomization is two-stage least squares (2SLS). As long as the instrumental variable satisfies the assumptions for being a valid instrument, 2SLS gives consistent estimates of the causal effect of the exposure on outcome. In this chapter we propose an alternative method for estimating the causal effect of a localizable exposure on a quantitative phenotype. This chapter is organized as follows: In Section 2.2, we introduced the additive linear model and our notation. In Section 2.3, we described the model assumption and estimation procedure. Simulation results are shown in Section 2.4 to compare the powers of the proposed method to 2SLS. We illustrated the new framework to an example applicable in finding downstream regulatory targets of lncRNA in the data from the Geuvadis project (Lappalainen et al., 2013). A summary is given in Section 2.6. To simplify language, we use interchangeably localizable exposure and gene expression, since the latter is our primary example.

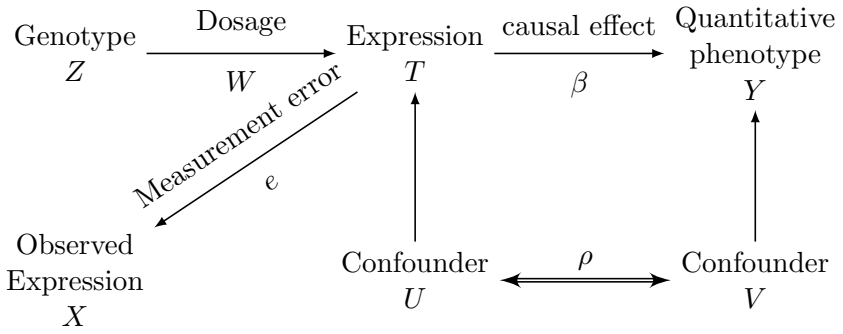
2.2. Model set up and notations

2.2.1. Model overview

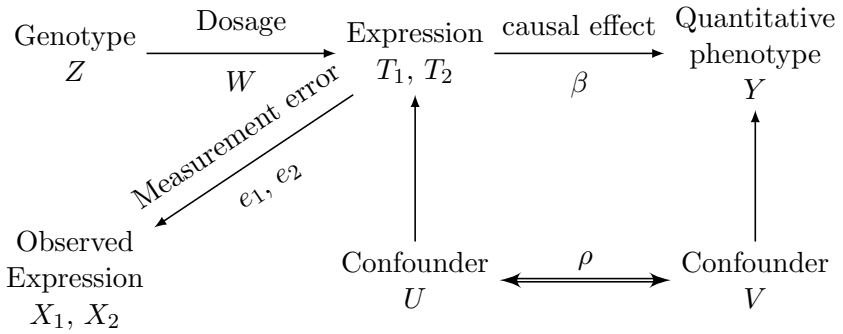
Figure 1 shows the relationship between the key variables of our model. Let T_i denote the total expression level for the gene of interest for individual i, $i \in \{1, 2, \ldots, n\}$. Y_i is the observed quantitative outcome for individual i and our goal is to estimate the causal effect of changing T_i on the outcome Y_i. Assume a simple linear model:

$$Y_i = \beta T_i + V_i \tag{2.1}$$

in which β represents the causal effect and V_i the total contribution of unobserved covariates and measurement errors. One could also extend the model to include possible observed covariates, but since we would be able to control for these by taking partial regression residuals, leading us back to (2.1), we keep things simple by ignoring observed covariates for now. We will use bold upper case un-subscripted letters to denote the vectors containing



(a) Homozygous



(b) Heterozygous

Figure 1: Diagram of allele specific Mendelian randomization model for (a) homozygous individuals and (b) heterozygous individuals.

the individual observations, $\mathbf{T} = (T_1, \ldots, T_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$.

We assume that within the transcribed region of the gene there is a SNP loci, which we call the tagging SNP, and that located outside the transcribed region in *cis* to the gene is a regulatory SNP. We also assume that there are only two alleles for each SNP and that the phase between these two SNPs is known. In other words, we assume that the haplotypes are known.

The observed total expression, denoted as X_i, differs from true expression level T_i by measurement error e_i. For individuals that are heterozygous at the tagging SNP, we observe the expressions of its two alleles,

$$\begin{aligned} X_{i1} = &\quad T_{i1} + e_{i1} \\ \text{and} \quad X_{i2} = &\quad T_{i2} + e_{i2}, \end{aligned} \tag{2.2}$$

as well as the observed total expression $X_i = X_{i1} + X_{i2}$; for homozygous individuals, only the total expression $X_i = T_i + e_i$ can be observed. Let $Z_{i1}, Z_{i2} \in \{0, 1\}$ be the two inherited alleles at the regulatory SNP for individual i. When both the tagging and the regulatory SNPs are heterozygous, let Z_{i1} and Z_{i2} correspond to the allele on the haplotype with expression T_{i1} and T_{i2}, respectively. We assume the additive model for the effect of the regulatory SNP on the expression of its linked allele,

$$T_{ij} = U_i + W_i^{(j)} Z_{ij} \quad j = 1, 2 \tag{2.3}$$

in which $j \in \{1, 2\}$ indexes the haplotype and $W_i^{(j)} Z_{ij}$ is the dosage effect of having the "1" haplotype on expression level T_{ij}. We assume independent *cis* regulatory effects, which means $W_i^{(1)}$ is independent with $W_i^{(2)}$.

2.2.2. Two stage least squares method for estimation β

A standardized regression of Y on X would yield a biased estimate for the causal effect β. This is due to existence of measurement error and correlation between U and V. To attain an unbiased estimate, the most widely used method is two-stage least squares (2SLS), which will be taken as the baseline for comparisons.

2SLS estimates the causal effect β by first forming a prediction for $\mathbf{X}$ based on the instrument $\mathbf{Z}$, and then regressing $\mathbf{Y}$ on $\mathbf{X}_Z$. In detail, let $\tilde{\mathbf{X}} = \mathbf{X} - \bar{X}$, $\tilde{\mathbf{Y}} = \mathbf{Y} - \bar{Y}$ and $\tilde{\mathbf{Z}} = \mathbf{Z} - \bar{Z}$, the projection of $\mathbf{X}$ on $\mathbf{Z}$ is

$$\mathbf{X}_Z = \bar{X} + (\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{X}}, \tag{2.4}$$

and the 2SLS estimate of β is

$$\hat{\beta}_{2SLS} = (\tilde{\mathbf{X}}_Z'\tilde{\mathbf{X}}_Z)^{-1}\tilde{\mathbf{X}}_Z'\tilde{\mathbf{Y}}. \tag{2.5}$$

The two-stage least squares estimate is a consistent estimator for β in our model, but ignores the information in the allele-specific measurements X_{i1}, X_{i2} for individuals that are heterozygous at the tagging SNP.

2.3. Estimation of causal effect in allele-specific model

The model described in Section 2.2.1 is a general model void of distribution assumptions. Here we add specific assumptions on the distribution of Z, U, V and W to allows maximum likelihood estimation of the causal parameter β.

2.3.1. Distribution assumptions

Assume genotype Z_{ij} follows Bernoulli distribution with minor allele frequency $P(Z_{ij} = 1) = p \in (0, 1)$. Without loss of generality, let $Z_{i1} = 1$ and $Z_{i2} = 0$ when individual i is heterozygous ($Z_i = 1$).

In (2.3), the independence of confounder U_i and dosage $W_i^{(j)}$ is assumed. This means that by changing the regulatory SNP's genotype Z_{ij}, the expression difference is independent of confounder effects contributing to the expression. This is a critical assumption for the instrumental variable to be a valid one. This assumption is not expected to always be true and even worse, it is usually hard to verify. With allele specific model, this assumption can be partially checked by $X_{i1} - X_{i2}$s and X_{i1}s from heterozygous individuals, see Figure 5 for examples. It is intuitive to consider that in (2.1), V_i, the total contribution of unobserved covariates and measurement errors on the outcome, is also independent of $W_i^{(j)}$. We assume that $W_i^{(j)}$ is i.i.d. normal with mean μ_w and variance σ_w^2, and that the joint distribution of (U_i, V_i) is i.i.d. bivariate normal with correlation ρ,

$$\begin{pmatrix} U_i \\ V_i \end{pmatrix} \sim N(\begin{pmatrix} \mu_u \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}).$$

Measurement error e in (2.2) is independent of the other variables. For each observation, measurement error follows $N(0, \sigma_e^2)$. That is, instead of observing T_{ij} or T_i, we observe $X_{ij} = T_{ij} + e_{ij}$ and $X_i = T_i + e_i$.

Given the observed distribution assumptions, the distribution for the observed data $\{Z_i, X_i, Y_i\}$ for homozygous and $\{Z_i, X_{i1}, X_{i2}, Y_i\}$ for heterozygous as follows:
For $(Z_{i1}, Z_{i2}) = (0, 0)$, i.e. homozygous individuals for the $(0, 0)$ haplotype,

$$(Y_i - \beta X_i,\ X_i)^T \sim N((\mu_v,\ 2\mu_u)^T, \Sigma_1); \tag{2.6}$$

for heterozygous individuals $(Z_{i1}, Z_{i2}) = (1, 0)$,

$$(Y_i - \beta X_i,\ X_i,\ X_{i1} - X_{i2})^T \sim N((\mu_v,\ 2\mu_u + \mu_w,\ \mu_w)^T, \Sigma_2); \tag{2.7}$$

for homozygous individual $(Z_{i1}, Z_{i2}) = (1, 1)$,

$$(Y_i - \beta X_i,\ X_i)^T \sim N((\mu_v,\ 2\mu_u + 2\mu_w)^T, \Sigma_3). \tag{2.8}$$

The covariance matrices in (2.6-2.8) are as follows,

$$\begin{aligned}
\Sigma_1 &= \begin{pmatrix} \sigma_v^2+\beta^2\sigma_e^2 & 2\rho\sigma_u\sigma_v-\beta\sigma_e^2 \\ 2\rho\sigma_u\sigma_v-\beta\sigma_e^2 & 4\sigma_u^2+\sigma_e^2 \end{pmatrix}, \\
\Sigma_2 &= \begin{pmatrix} \sigma_v^2+2\beta^2\sigma_e^2 & 2\rho\sigma_u\sigma_v-2\beta\sigma_e^2 & 0 \\ 2\rho\sigma_u\sigma_v-2\beta\sigma_e^2 & 4\sigma_u^2+\sigma_w^2+2\sigma_e^2 & \sigma_w^2 \\ 0 & \sigma_w^2 & \sigma_w^2+2\sigma_e^2 \end{pmatrix}, \\
\Sigma_3 &= \begin{pmatrix} \sigma_v^2 + \beta^2\sigma_e^2 & 2\rho\sigma_u\sigma_v-\beta\sigma_e^2 \\ 2\rho\sigma_u\sigma_v-\beta\sigma_e^2 & 4\sigma_u^2 + 2\sigma_w^2 + \sigma_e^2 \end{pmatrix}
\end{aligned} \tag{2.9}$$

2.3.2. Identifiability

Nine parameters are used to describe the distribution of the observed variables: 3 parameters (μ_u, μ_v, μ_w) for the mean of U, V, W; 5 parameters for the variances and covariance: $(\sigma_e^2, \sigma_v^2, \sigma_u^2, \sigma_w^2, \rho)$; and 1 parameter of the causal effect, β. There are 2 constraints: (i) variances are non-negative $\sigma_e^2 \geq 0$, $\sigma_v^2 \geq 0$, $\sigma_u^2 \geq 0$, $\sigma_w^2 \geq 0$; (ii) the correlation ρ is between -1 and 1.

From (2.6-2.8), the model is in the form of an exponential family. Therefore, to check identifiability we only need to evaluate the determinant of the Fisher information matrix R:

$$R = (R_{jk}), \quad R_{jk} = E[\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}] \tag{2.10}$$

where l is the log-likelihood function, $\boldsymbol{\theta} = (\mu_u, \mu_v, \mu_w, \sigma_v^2, \sigma_e^2, \sigma_u^2, \sigma_w^2, \rho_{uv}, \beta)$ and $\rho_{uv} = \rho\sigma_u\sigma_v$. If R is nonsingular in a convex set within the feasible parameter space, then every parameter point $\boldsymbol{\theta}$ in the feasible parameter space is globally identifiable (Rothenberg et al.,

1971).

Let $p_1 = \sigma_v^2 + \beta^2\sigma_e^2$, $p_2 = 2\rho\sigma_u\sigma_v - \beta\sigma_e^2$, and $p_3 = 4\sigma_u^2 + \sigma_e^2$ and replace the parameters $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^* = (\mu_u, \mu_v, \mu_w, p_1, p_2, p_3, \sigma_e^2, \sigma_w^2, \beta)$. The Jacobian matrix between variance covariance parameters $(p_1, p_2, p_3, \sigma_e^2, \sigma_w^2, \beta)$ and $(\sigma_v^2, \sigma_e^2, \sigma_u^2, \sigma_w^2, \rho_{uv}, \beta)$ is

$$\frac{\partial(p_1, p_2, p_3, \sigma_e^2, \sigma_w^2, \beta)}{\partial(\sigma_v^2, \sigma_e^2, \sigma_u^2, \sigma_w^2, \rho_{uv}, \beta)} = \begin{bmatrix} 1 & \beta^2 & 0 & 0 & 0 & 2\beta\sigma_e^2 \\ 0 & -\beta & 0 & 0 & 2 & -\sigma_e^2 \\ 0 & 1 & 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.11}$$

The Jacobian matrix is non-singular with determinant 8. This means that if $\boldsymbol{\theta}^*$ is identifiable, then so is $\boldsymbol{\theta}$.

Rewriting the distribution (2.6-2.8) using the parameters $\boldsymbol{\theta}^*$, for homozygous $(Z_{i1}, Z_{i2}) = (0, 0)$ individuals,

$$(Y_i - \beta X_i,\ X_i)^T \sim N((\mu_v,\ 2\mu_u)^T, \Sigma_1^*), \tag{2.12}$$

for heterozygous $(Z_{i1}, Z_{i2}) = (1, 0)$ individuals,

$$(Y_i - \beta X_i,\ X_i,\ X_{i1} - X_{i2})^T \sim N((\mu_v,\ 2\mu_u + \mu_w,\ \mu_w)^T, \Sigma_2^*), \tag{2.13}$$

for homozygous $(Z_{i1}, Z_{i2}) = (1, 1)$ individuals,

$$(Y_i - \beta X_i,\ X_i)^T \sim N((\mu_v,\ 2\mu_u + 2\mu_w)^T, \Sigma_3^*), \tag{2.14}$$

with covariance matrices

$$\Sigma_1^* = \begin{pmatrix} p_1 & p_2 \\ p_2 & p_3 \end{pmatrix};$$
$$\Sigma_2^* = \begin{pmatrix} p_1+\beta^2\sigma_e^2 & p_2-\beta\sigma_e^2 & 0 \\ p_2-\beta\sigma_e^2 & p_3+\sigma_w^2+\sigma_e^2 & \sigma_w^2 \\ 0 & \sigma_w^2 & \sigma_w^2+2\sigma_e^2 \end{pmatrix}; \tag{2.15}$$
$$\Sigma_3^* = \begin{pmatrix} p_1 & p_2 \\ p_2 & p_3+2\sigma_w^2 \end{pmatrix}.$$

The observed data can be divided by genotypes into 3 groups. The observations $(X_i, X_{i1}, X_{i2}, Z_i, Y_i)$ are independent across individuals, and individuals within same group follow the same distribution (2.12-2.14). Therefore the log-likelihood function of the whole population can be written as the sum of the log-likelihood functions for each group: $l^* = l_1^* + l_2^* + l_3^*$.

The Fisher information matrix, expressed in terms of the new parameters $\boldsymbol{\theta}^*$, is

$$R^* = (R_{jk}^*), \quad R_{jk}^* = E[\frac{\partial l^*}{\partial \theta_j^*}\frac{\partial l^*}{\partial \theta_k^*}] \tag{2.16}$$

$$\begin{aligned} R_{jk}^* =& E[\frac{\partial l^*}{\partial \theta_j^*}\frac{\partial l^*}{\partial \theta_k^*}] \\ =& E[\frac{\partial(l_1^*+l_2^*+l_3^*)}{\partial \theta_j^*}\cdot\frac{\partial(l_1^*+l_2^*+l_3^*)}{\partial \theta_k^*}] \\ =& \sum_{s=1}^{3} E[\frac{\partial l_s^*}{\partial \theta_j^*}\frac{\partial l_s^*}{\partial \theta_k^*}] + (\frac{\partial l_1^*}{\partial \theta_j^*}\cdot\frac{\partial l_2^*}{\partial \theta_k^*} + \frac{\partial l_1^*}{\partial \theta_k^*}\cdot\frac{\partial l_2^*}{\partial \theta_j^*}) \\ &+ (\frac{\partial l_2^*}{\partial \theta_j^*}\cdot\frac{\partial l_3^*}{\partial \theta_k^*} + \frac{\partial l_2^*}{\partial \theta_k^*}\cdot\frac{\partial l_3^*}{\partial \theta_j^*}) + (\frac{\partial l_3^*}{\partial \theta_j^*}\cdot\frac{\partial l_1^*}{\partial \theta_k^*} + \frac{\partial l_3^*}{\partial \theta_k^*}\cdot\frac{\partial l_1^*}{\partial \theta_j^*}). \end{aligned} \tag{2.17}$$

Note that $E[\frac{\partial l_s^*}{\partial \theta_j^*}\cdot\frac{\partial l_t^*}{\partial \theta_k^*}] = E[\frac{\partial l_s^*}{\partial \theta_j^*}]\cdot E[\frac{\partial l_t^*}{\partial \theta_k^*}]$ when $s \neq t$. $E[\frac{\partial l_s^*}{\partial \theta_j^*}] = 0$ for all $s = 1, 2, 3$ and

$j = 1, \ldots, 9$. Therefore

$$R^*_{jk} = E[\frac{\partial l^*}{\partial \theta^*_j}\frac{\partial l^*}{\partial \theta^*_k}] = \sum_{s=1}^{3} E[\frac{\partial l^*_s}{\partial \theta^*_j}\frac{\partial l^*_s}{\partial \theta^*_k}]$$

Let $R^{*(s)}, s = 1, 2, 3$ denote the Fisher information matrix for each genotype group, $R^* = R^{*(1)} + R^{*(2)} + R^{*(3)}$. Individuals in the same group have independent identical distribution, and consequently, the same log-likelihood function. Take the first group ($Z_i = 0$) as example:

$$l^*_1 = \sum_{i:Z_i=0} l^{*(i)}_1$$

where $l^{*(i)}_1$ is the log-likelihood function of individual i in the first group.

$$\begin{aligned} R^{*(1)}_{jk} = & \quad E[(\sum_{i:Z_i=0}\frac{\partial l^{*(i)}_1}{\partial \theta^*_j})(\sum_{i:Z_i=0}\frac{\partial l^{*(i)}_1}{\partial \theta^*_k})] \\ = & \quad \sum_{i:Z_i=0} E[\frac{\partial l^{*(i)}_1}{\partial \theta^*_j}\cdot\frac{\partial l^{*(i)}_1}{\partial \theta^*_k}] = N_1 \cdot E[\frac{\partial l^{*(i)}_1}{\partial \theta^*_j}\cdot\frac{\partial l^{*(i)}_1}{\partial \theta^*_k}]. \end{aligned}$$

We denote N_1, N_2 and N_3 as the sample size of genotype $Z_i = 0$, $Z_i = 1$, and $Z_i = 2$ correspondingly. $N_1 + N_2 + N_3 = N$.

Proposition 2.1 *Let G^*_1 denote the Fisher information matrix for each individual in the first group and $R^{*(1)} = N_1 G^*_1$. Similarly, G^*_2 and G^*_3 are the Fisher information matrix for individuals in second and third group.*

(a) $\text{rank}(G^*_1) = 5$, $\det(G^*_1) = 0$*; the identifiable parameters are:* $(\mu_v, \mu_u, p_1, p_2, p_3)$*;*

(b) $\text{rank}(G^*_3) = 5$, $\det(G^*_3) = 0$*; the identifiable parameters are:* $(\mu_v, \mu_u + \mu_w, p_1, p_2, p_3 + \sigma^2_w)$*;*

(c) $\text{rank}(G^*_2) = 9$ *with*

$$\det(G^*_2) = \frac{2^{19}\sigma^4_w}{\det(\Sigma^*_2)^5};$$

$\det(G^*_2) \neq 0$ *when* $\sigma^2_w \neq 0$ *and* $\Sigma^*_2 \neq 0$*;*

(d) $\text{rank}(N_1G_1^* + N_3G_3^*) = 8$, $\det(N_1G_1^* + N_3G_3^*) = 0$ *for all* $N_1, N_3 \in \mathbf{N}$*; the identifiable parameters are:* $\tilde{\theta} = (\mu_v, \mu_u, \mu_w, p_1, p_2, p_3, \sigma_w^2, \beta)$*; Let* $\tilde{G_1}$ *and* $\tilde{G_3}$ *be the Fisher information matrix of homozygous observations corresponding to parameter* $\tilde{\boldsymbol{\theta}}$*. And the Fisher information matrix with only homozygous individuals is as follows:*

$$\begin{aligned}\det \quad & [N_1\tilde{G_1} + N_3\tilde{G_3}] \\ & = \frac{2^{21}N_1^3N_3^3(N_1+N_3)p_1^2((N_1+N_3)p_1\sigma_w^2 + \mu_w^2(N_3\det(\Sigma_1^*) + N_1\det(\Sigma_3^*)))}{\det(\Sigma_1^*)^4\det(\Sigma_3^*)^4}\end{aligned}$$

(e) $\text{rank}(R^*) = 9$, $\det(R^*) = \det(N_1G_1^* + N_2G_2^* + N_3G_3^*) \neq 0$*, if* $N_2 \neq 0$ *because the numerator of* $\det(R^*)$ *is proportionate to* N_2*.*

Thus, as long as both homozygous genotypes are observed ($N_1N_3 \neq 0$), and the gene has a non-zero effect on either mean or variance of expression level (σ_w^2 and μ_w is not zero at the same time), the model is identifiable for $\tilde{\boldsymbol{\theta}}$. Although we can not estimate σ_e^2 from $\tilde{\boldsymbol{\theta}}$, β, as well as p_1, p_2 p_3, σ_w^2 and $\boldsymbol{\mu}$, are identifiable. All parameters are identifiable when $N_2 > 0$, that is, if we can observe heterozygous individuals.

2.3.3. Maximum likelihood estimation

The full log-likelihood function follows exactly the joint distribution (2.6-2.8). Let $\boldsymbol{\mu} = (\mu_u, \mu_v, \mu_w)$, $\boldsymbol{\sigma} = (\sigma_e^2, \sigma_v^2, \sigma_u^2, \sigma_w^2, \rho)$, the log-likelihood is

$$\begin{aligned} l_1(\boldsymbol{\mu}, \boldsymbol{\sigma}, \beta, X, Y, |Z=0) = -\frac{1}{2}\sum_{Z_i=0}\Big\{ & \log(|\Sigma_1|) \\ & - \frac{1}{2}\begin{pmatrix} Y_i - \beta X_i - \mu_v \\ X_i - 2\mu_u \end{pmatrix}^T \Sigma_1^{-1} \begin{pmatrix} Y_i - \beta X_i - \mu_v \\ X_i - 2\mu_u \end{pmatrix}\Big\}, \end{aligned} \tag{2.18}$$

$$
\begin{aligned}
l_2(\boldsymbol{\mu}, \boldsymbol{\sigma}, \beta, X, Y, |Z=1) = -\frac{1}{2} \sum_{Z_i=1} \Big\{ & \log(|\Sigma_2|) \\
& - \frac{1}{2} \begin{pmatrix} Y_i - \beta X_i - \mu_v \\ X_i - 2\mu_u - \mu_w \\ X_{i1} - X_{i2} - \mu_w \end{pmatrix}^T \Sigma_2^{-1} \begin{pmatrix} Y_i - \beta X_i - \mu_v \\ X_i - 2\mu_u - \mu_w \\ X_{i1} - X_{i2} - \mu_w \end{pmatrix} \Big\},
\end{aligned} \tag{2.19}
$$

$$
\begin{aligned}
l_3(\boldsymbol{\mu}, \boldsymbol{\sigma}, \beta, X, Y, |Z=2) = -\frac{1}{2} \sum_{Z_i=2} \Big\{ & \log(|\Sigma_3|) \\
& - \frac{1}{2} \begin{pmatrix} Y_i - \beta X_i - \mu_v \\ X_i - 2\mu_u - 2\mu_w \end{pmatrix}^T \Sigma_3^{-1} \begin{pmatrix} Y_i - \beta X_i - \mu_v \\ X_i - 2\mu_u - 2\mu_w \end{pmatrix} \Big\},
\end{aligned} \tag{2.20}
$$

where Σ_1, Σ_2 and Σ_3 are defined in (2.9). We used the `optim` function directly from `R` to get the estimated parameters with maximum likelihood. Here, we define this method as ASMR, refer to Allele Specific Mendelian Randomization estimation.

2.4. Simulation study

The variance of true expression level T_i is composed of the variance due to genotype Z_i as $W_i^{(1)} Z_{i1} + W_i^{(2)} Z_{i2}$ and the variance of U_i, according to (2.3). Compare the variance accounts for genotype Z_i,

$$
\begin{aligned}
\mathrm{Var}[W_i^{(1)} Z_{i1} + W_i^{(2)} Z_{i2}] &= 2\mathrm{Var}[W_i^{(1)} Z_{i1}] \\
&= 2(p\sigma_w^2 + p(1-p)\mu_w^2).
\end{aligned} \tag{2.21}
$$

with the variance of observed expression level X_i,

$$
\begin{aligned}
\mathrm{Var}[X_i] &= E[\mathrm{Var}[X_i|Z_i]] + \mathrm{Var}[E[X_i|Z_i]] \\
&= 4\sigma_u^2 + \sigma_e^2 + 2(p\sigma_w^2 + p(1-p)(\mu_w^2 + \sigma_e^2)).
\end{aligned} \tag{2.22}
$$

The variance of the confounder U (σ_u), which contributes to the expression, is taken as the standard to measure other parameters: μ_w/σ_u, σ_w/σ_u, and σ_e/σ_u. Therefore in every simulation, we are going to fix $\sigma_u = 1$. We are also going to fix $\mu_u = 1$ and $\mu_v = 0$ while changing other parameters.

To measure the strength of instruments, we used the concentration parameter by Stock et al. (2002), defined as:

$$\mu_c^2 = \frac{E[(W_i^{(1)} Z_{i1} + W_i^{(2)} Z_{i2})^2]}{4\sigma_u^2} = p\big(\frac{\mu_w^2}{\sigma_u^2} + \frac{\sigma_w^2}{2\sigma_u^2}\big). \tag{2.23}$$

When μ_c^2 is greater than 1.82, the instrument can be considered as strong.

The classical estimation method, 2SLS estimation, is used to evaluate the performance of ASMR proposed in Section 2.3. Theoretically, 2SLS estimation is unbiased with asymptotic variance

$$\mathrm{Var}[\hat{\beta}_{2SLS}] = \frac{1}{n} M_{XZ}^{-1} E[X_i Z_i'] E[Z_i Z_i']^{-1} E[(Y_i - \beta X_i)^2 Z_i Z_i'] E[Z_i Z_i']^{-1} E[X_i Z_i']' M_{XZ}^{-1} \tag{2.24}$$

in which $M_{XZ} = E[X_i Z_i'] E[Z_i Z_i']^{-1} E[X_i Z_i']'$. $\mathrm{Var}[\hat{\beta}_{2SLS}]$ does not depends on σ_w^2, the variance of the dosage.

The performance of ASMR and 2SLS are measured by the medians of estimated causal effect and the median absolute difference between true and estimated causal effect across $nsim = 1000$ simulations of $n = 1000$ individuals. Data is generated from distribution (2.6-2.8). In each simulation setting, we varied the confounding effect ρ from 0 (no confounding) to 1 (fully confounded). Across simulation settings, we change the strength of instruments by varying the mean and variance of dosage, μ_w and σ_w^2 and the results are shown in Figure 2-3.

We changed the strength of instrument by varying the mean of dosage μ_w (Figure 2a and b) while fix other parameters: true causal effect $\beta = 1$, minor allele frequency $p = 0.1$,

variance of dosage $\sigma_w^2 = 1$ and the variance of confounders $\sigma_v^2 = 1$, $\sigma_u^2 = 1$. The strength is considered strong when $\mu_w = 6$ and $\mu_w = 2$, and weak when $\mu_w = 0.5$. We also examined median strength IV with $\mu_w = 1$.

Comparing the median of estimated causal effect and true causal effect, in general, 2SLS and ASMR methods are both approximately unbiased across the range of the confounder strength ρ. The estimates are more accurate across ρ with stronger instruments (Figure 2b). The stronger the instrument is, the smaller difference between true and estimated causal effect for both 2SLS and ASMR methods (Figure 2a). When the instrument is weak ($\mu_w = 0.5$), ASMR improved the estimation accuracy significantly compared with 2SLS. The accuracy of ASMR estimation with $\mu_w = 0.5$ is similar with 2SLS with $\mu_w = 1$. With increasing instrument strength ($\mu_w = 1, 2$ and 6), the difference of accuracy between ASMR and 2SLS becomes smaller. When the instrument is a strong instrument ($\mu_w = 6$), there is almost no difference between ASMR and 2SLS. If the strength of instrument is too weak ($\mu_w = 0.2$) to explain the gene's expression level X or even invalid ($\mu_w = 0$), the estimation is unacceptable with large amount of outliers for both 2SLS and ASMR (Figure 17). In conclusion, when varying the strength of instrument by the mean of dosage, ASMR gives more accurate estimates of causal effect than 2SLS.

The strength of instrument is determined not only by the mean of dosage, but also by the variance of dosage σ_w^2 (Equation (2.23)), which determines how much a genotype can explain the true expression level. Two estimation methods are compared by the median of estimated causal effect $\hat{\beta}$ in Figure 3a. 2SLS and ASMR both give unbiased estimation across different confounder effects when the correlation ρ increases from 0 to 1. We also compared the median absolute difference $|\hat{\beta} - \beta|$ of two methods (Figure 3b). The median absolute differences do not show any change for 2SLS when σ_w^2 increases from 0.5 to 6. This is because the variance of 2SLS estimation is invariant regarding to σ_w^2, wee equation (2.24). However, the median absolute difference for ASMR estimation gets smaller as σ_w^2 gets bigger. When $\sigma_w^2 = 0$, the median absolute difference of ASMR is slightly smaller than

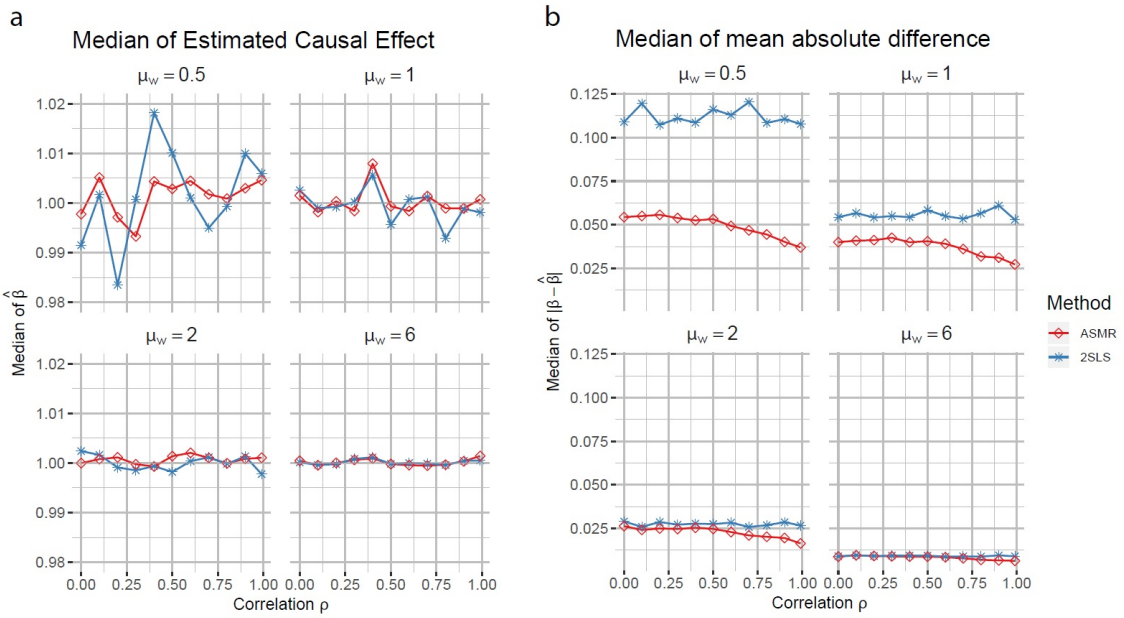


Figure 2: Simulation results when changing instrument strength by the mean of dosage μ_w. Under each setting, we simulated 1000 times with the range of confounder strength $\rho \in [0, 1]$.

2SLS while when $\sigma_w^2 = 6$, the median absolute difference of ASMR is about 0.01, which is much lower than that of 2SLS, which is around 0.06. If the variance is smaller, $\sigma_w^2 = 0.2$, or even no variation for dosage $\sigma_w^2 = 0$, there are barely no difference between ASMR and 2SLS estimates (Figure 3).

In short, the results from simulations show that when the instruments are too weak to explain the gene's expression level or when the instruments are very strong, ASMR performs similarly to 2SLS. When the instruments are weak and can only partially explain the gene's expression level, the ASMR has higher power than 2SLS.

2.5. Real data example: Finding downstream targets of lincRNA

Long intergenic non-coding RNAs (lincRNAs) have gained widespread attention in recent years as a potentially new and crucial layer of biological regulation. lincRNAs of all kinds have been implicated in a range of developmental processes and diseases, but knowledge of the mechanisms by which they act is still limited. Rinn and Chang (2012) hypothesized

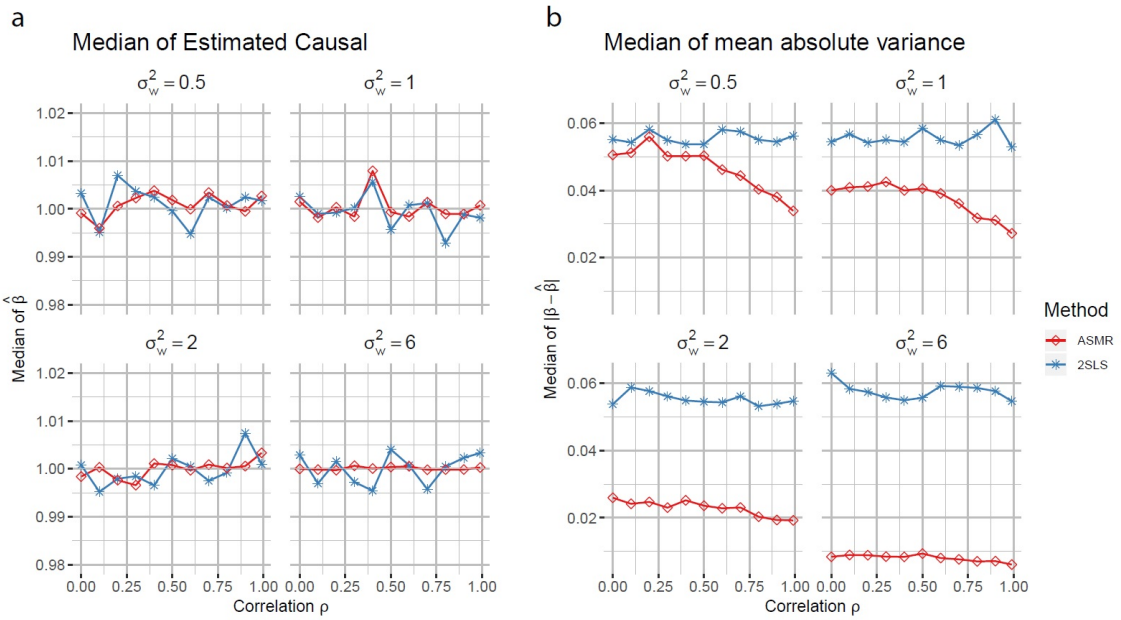


Figure 3: Simulation results when changing instrument strength by the variance of dosage σ_w^2. Under each setting, we simulated 1000 times with the range of confounder strength $\rho \in [0, 1]$.

that lincRNAs regulate the transcription of other genes and their mRNA expression. The causal relations between lincRNAs and mRNAs from coding-genes have been examined in Mendelian randomization by McDowell et al. (2016) using 2SLS, where they found significant pairs of lincRNAs and genes. For more information, see recent review by Mattick (2018). In this section, we applied the allele specific model to real data of lincRNA and mRNA expression, estimated causal effects by ASMR and 2SLS. We also tested the significant pairs of lincRNAs and genes from estimated causal effects with empirical variances and false discovery rate (FDR) control.

2.5.1. Data processing

The raw data is from Geuvadis samples (Lappalainen et al., 2013), filtered by Hardy-Weinberg equilibrium and read strand bias. There are 87 individuals available in this dataset. Each individual has 1513 lincRNAs expressions with their tagging SNPs' genotypes from 1000 Genome Project (Consortium et al., 2015) and the expression levels of genes,

measured by Fragments Per Kilobase of transcript per Million (FPKM). The expressions of lincRNAs are measured by read counts and the allele specific information of the lincRNA expressions are available for heterozygous individuals with respect to the tagging SNPs. The total expressions are obtained by adding allele specific expressions together. Gene expressions are measured by FPKM on 26144 non-zero genes.

We selected desirable SNPs before estimating causal effect to save computing time. SNPs that: (i) can observe 3 genotypes across individuals; (ii) have over 30% of heterozygous; (iii) have at least one non-zero heterozygous read counts; (iv) have at least one non-zero homozygous read counts; will be chosen. After filtering, there are 321 SNPs left.

Based on concentration parameter (equation (2.23)) and simulation results, with fixed σ_u, larger μ_w and σ_w lead to more accurate estimation of causal effect β. Before estimating causal effect from full likelihood model, we first estimate σ_u, μ_w and σ_w by maximizing likelihood of model (2.3). We selected genes that are strong instruments based the ratio of estimated σ_w/σ_u and μ_w/σ_u (Figure 4). There are 6 SNPs that shows both large μ_w/σ_u ratio and σ_w/σ_u ratio and are considered as strong instruments. The allele specific information provided by heterozygous individuals allows us to partially check the independence between confounder U and dosage W by looking at the the allele specific expression from different genotypes. Assume X_{i1} and X_{i2} are the observed expressions for genotype $Z_{i1} = 0$ and $Z_{i2} = 1$ of individual i. $X_{i1} = U_i + e_{i1}$ and $X_{i2} - X_{i1} = W_i^{(2)} + e_{i2} - e_{i1}$ can be viewed as an approximation of confounder U_i and dosage $W_i^{(2)}$. We can checked the correlation between X_{i1} and $X_{i2} - X_{i1}$ for the independence of confounder U and dosage W from heterozygous individuals. As an example, we checked SNP rs11061295 in Figure 5 and the Spearman correlation between X_{i1} and X_{i2} is 0.0096.

2.5.2. Estimation of causal effect

We estimate the causal effect as well as other parameters by ASMR from log-likelihoods (equations (2.6 - 2.8)). As noted earlier, 2SLS estimation is a well-studied, unbiased method

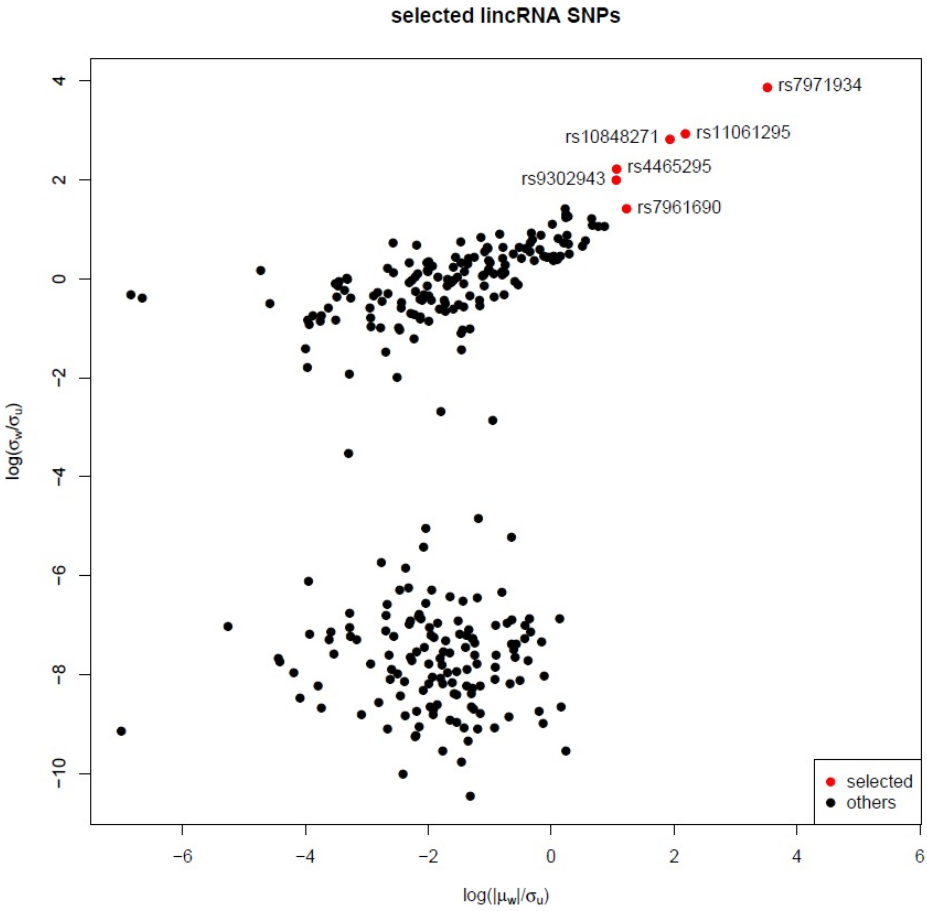


Figure 4: Estimated $\log(|\mu_w|/\sigma_u)$ and $\log(\sigma_w/\sigma_u)$ from first stage likelihood of model (2.3). The red dots are selected lincRNA SNPs with high concentration parameter. Those SNPs are strong instruments and used for next step causal effect analysis. This step selected 6 SNPS: rs4465295, rs7971934, rs7961690, rs11061295, rs10848271 and rs9302943.

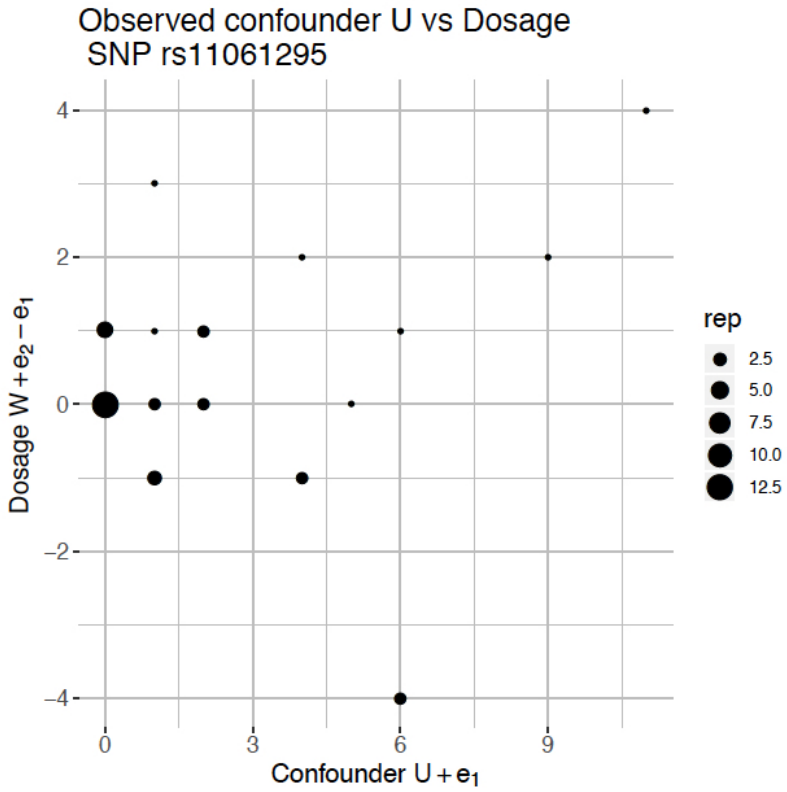


Figure 5: Scatter plot of X_{i1} vs. $X_{i2} - X_{i1}$ for SNP rs11061295 with heterozygous individuals. The Spearman correlation is 0.0096.

and it is compared with ASMR results. To detect reliable causal relation between the localizable exposure and gene expression, we compare the causal effect estimated by 2SLS ($\hat{\beta}_{2SLS}$) and ASMR ($\hat{\beta}_{ASMR}$) after filtering out ASMR results that does converge. In addition to $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{ASMR}$, we also calculated the standard deviation of $\hat{\beta}_{ASMR}$ by the Fisher information matrix (see Section 2.3.2), denoted as $sd(\hat{\beta}_{ASMR})$, under the null hypothesis, where there is no causal effect between lincRNA expression and gene expression. We also calculated the standard deviation of $\hat{\beta}_{2SLS}$ from equation (2.24), denoted as $sd(\hat{\beta}_{2SLS})$. The z-values from ASMR and 2SLS are calculated by $z_{ASMR} = \hat{\beta}_{ASMR}/sd(\hat{\beta}_{ASMR})$and $z_{2SLS} = \hat{\beta}_{2SLS}/sd(\hat{\beta}_{2SLS})$ respectively. If we assume standard normality for z_{2SLS} and z_{ASMR}, we will reject the genes that have z-values greater than qnorm(0.975) or less than qnorm(0.025) for the null hypothesis and consider those genes with real causal effect from ASMR or from 2SLS. Since we are testing 26144 genes at once, we need to control for false discovery rate (FDR) for multiple testing. We calculated the empirical null distributions for the z-values from both ASMR and 2SLS and adjusted for multiple testing. Then, for both methods, we select significant genes with reject level $\alpha = 0.05$. Across 6 candidate SNPs, we focus on genes that are have large z-values from ASMR or 2SLS.

Here we also take SNP rs11061295 as an example. The allele specific information allows us to partially check the independence between the After adjusted by empirical null, there are 296 genes that are significant differ from zero from ASMR and 1975 genes from 2SLS where 22 of them are also significant from ASMR. We first plot the z-values from estimated causal effects and standard deviations by 2SLS and ASMR (Figure 6). After adjusted by empirical null, we selected 20 genes that are uniquely selected by ASMR or 2SLS. From the analysis and simulations above, we are expected to select more significant genes from ASMR because it estimates causal effect β more accurately with stronger instruments. Since both 2SLS and ASMR are unbiased methods for estimating causal effect, we are also expected to see a overlap between significant genes selected by ASMR and 2SLS. However, in this example, we did not see both patterns. Take a closer look at the expression of SNP as well as the selected significant genes. From the scales of the z-value histograms from ASMR and 2SLS,

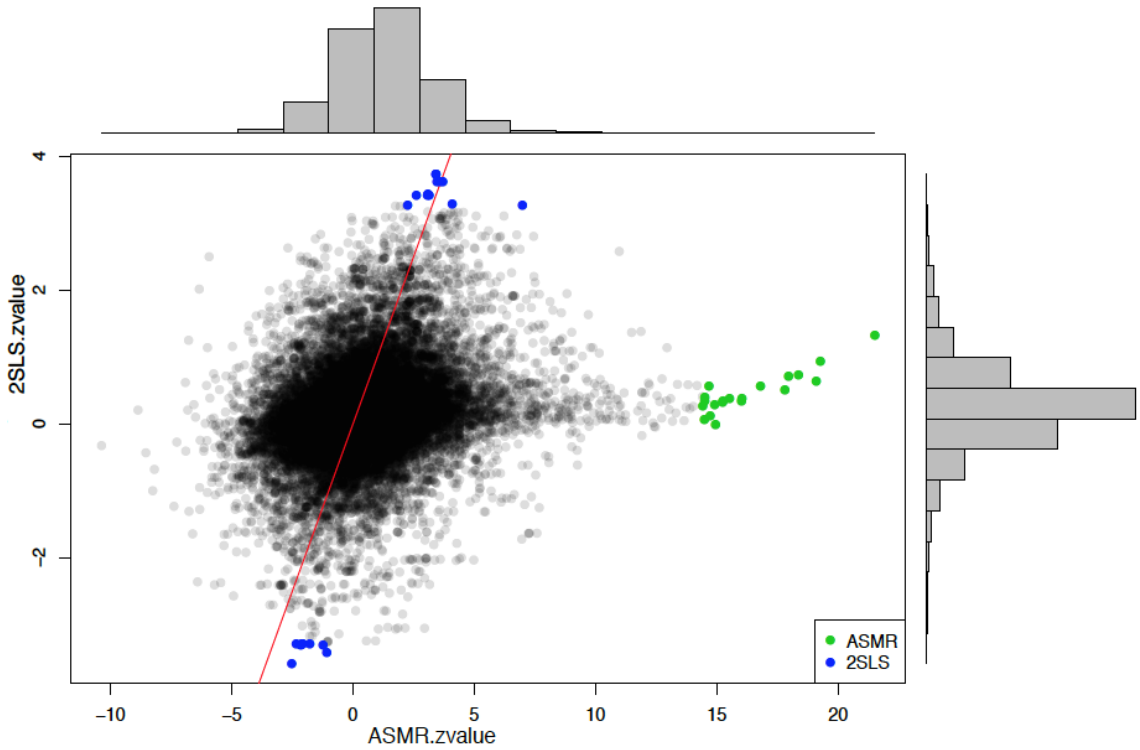


Figure 6: Scatter plot with histograms of the z-values from ASMR and 2SLS. The dots represent the genes and the x-axis and y-axis are the z-values from ASMR and 2SLS, respectively. The red line is the $x = y$ line. We highlighted the top 20 genes that are of high z-values from ASMR and are only significant from ASMR by green and top 20 genes from 2SLS by blue.

the z-values from ASMR with empirical null distribution with larger variance. Even though the 2SLS significant genes have similar z-values from ASMR and 2SLS, the large dispersion from ASMR empirical null distribution leads to the overlook from ASMR. The significant genes from ASMR are with extremely high z-values while the 2SLS provides z-values that are close to zero. This is not because that the actual estimated causal effect from ASMR $\hat{\beta}_{ASMR}$ is larger than that from 2SLS $\hat{\beta}_{2SLS}$, in fact, they are similar, but because that the calculated standard deviation from ASMR are extremely small.

2.6. Conclusion

Mendelian randomization (MR) has been widely studied for estimate causal effect between gene's expression level and phenotype of interest. Two-stage least square regression, with DNA-variants as instruments, is the classical method for MR. In this article, we look for a new method, ASMR, that can accommodate the development of sequencing techniques that produced allele-specific data. Maximum likelihood estimation, i.e., estimating every parameters including the one of interests under the normal assumption, has better estimation power than classical two stage least squares (2SLS) that ignores allele specific information. This new method incorporates the allele-specific expression in heterozygous individuals by separating expression X_{i1} and X_{i2} in distribution. It would be of interest for future study to consider the properties of maximum likelihood estimation in the non-normal setting.

CHAPTER 3 : Bulk tissue deconvolution with single cell RNA sequencing

3.1. Introduction

Bulk tissue RNA-seq is a widely adopted method to understand genome-wide transcriptomic variations in different conditions such as disease states. Bulk RNA-seq measures the average expression of genes, which is the sum of cell type-specific gene expression weighted by cell type proportions. Knowledge of cell type composition and their proportions in intact tissues is important, because certain cell types are more vulnerable for disease than others. Characterizing the variation of cell type composition across subjects can identify cellular targets of disease, and adjusting for these variations can clarify downstream analysis.

The rapid development of single-cell RNA-seq (scRNA-seq) technologies have enabled cell type-specific transcriptome profiling. Although cell type composition and proportions are obtainable from scRNA-seq, scRNA-seq is still costly, prohibiting its application in clinical studies that involve a large number of subjects. Furthermore, scRNA-seq is not well suited to characterizing cell type proportions in a solid tissue, because the cell dissociation step is biased towards certain cell types (Park et al., 2018).

Computational methods have been developed to deconvolve cell type proportions using cell type-specific gene expression references (Park et al., 2018). CIBERSORT (Newman et al., 2015), based on support vector regression, is a widely used method designed for microarray data. More recently, BSEQ-sc (Baron et al., 2016) extended CIBERSORT to allow the use of scRNA-seq gene expression as a reference. TIMER (Li et al., 2016), developed for cancer data, focuses on the quantification of immune cell infiltration. These methods rely on pre-selected cell type-specific marker genes, and thus are sensitive to the choice of significance threshold. More importantly, these methods ignore cross-subject heterogeneity in cell type-specific gene expression as well as within-cell type stochasticity of single-cell gene expression, both of which cannot be ignored based on our analysis of multiple scRNA-seq datasets (Figure 20a).

Here we introduce a new MUlti-Subject SIngle Cell deconvolution (MuSiC) method (code available) that utilizes cross-subject scRNA-seq to estimate cell type proportions in bulk RNA-seq data. Through comprehensive benchmark evaluations, and applications to pancreatic islet and whole kidney expression data in human, mouse, and rats, we show that MuSiC outperformed existing methods, especially for tissues with closely related cell types.

3.2. Methods

3.2.1. Method overview

An overview of MuSiC is shown in Figure 7. MuSiC starts with multi-subject scRNA-seq data, and assumes that the cells for each subject have been classified into a set of fixed cell types that are shared across subjects. MuSiC deconvolves bulk RNA-seq samples to obtain the proportions of these cell types in each sample. A key concept in MuSiC is marker gene consistency. We show that, when using scRNA-seq data as a reference for cell type deconvolution, two fundamental types of consistency must be considered: cross-subject and cross-cell, in which the first is to guard against bias in subject selection, and the second is to guard against bias in cell capture in scRNA-seq. By incorporating both types of consistency, MuSiC allows for scRNA-seq datasets to serve as effective references for independent bulk RNA-seq datasets involving different individuals.

Rather than pre-selecting marker genes from scRNA-seq based only on mean expression, MuSiC gives weight to each gene, allowing for the use of a larger set of genes in deconvolution. The weighting scheme prioritizes consistent genes across subjects: up-weighing genes with low cross-subject variance (informative genes) and down-weighing genes with high cross-subject variance (non-informative genes). This requirement on cross-subject consistency is critical for transferring cell type-specific gene expression information from one dataset to another.

Solid tissues often contain closely related cell types, and correlation of gene expression between these cell types leads to collinearity, making it difficult to resolve their relative

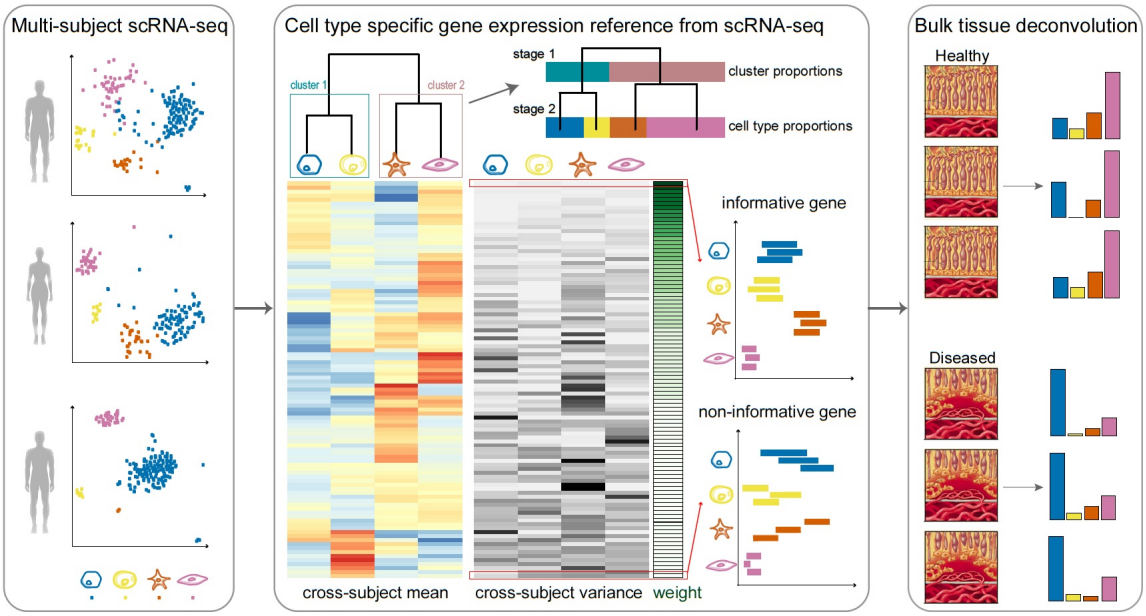


Figure 7: Overview of MuSiC framework.
Music starts from scRNA-seq data from multiple subjects, classified into cell types (shown in different colors), and constructs a hierarchical clustering tree reflecting the similarity between cell types. Based on this tree, the user can determine the stages of recursive estimation and which cell types to group together at each stage. MuSiC then determines the group-consistent genes and calculates cross-subject mean (red to blue) and cross-subject variance (black to white) for these genes in each cell type. MuSiC up-weighs genes with low cross-subject variance and down-weighs genes with high cross-subject variance. In the example shown, deconvolution is performed in two stages, only cluster proportions are estimated for the first stage. Constrained by these cluster proportions, the second stage estimates cell type proportions, illustrated by the length of the bar with different colors. The deconvolved cell type proportions can then be compared across disease cohorts.

proportions in bulk data. To deal with collinearity, MuSiC employs a tree-guided procedure that recursively zooms in on closely related cell types. Briefly, we first group similar cell types into the same cluster and estimate cluster proportions, then recursively repeat this procedure within each cluster (Figure 7). At each recursion stage, we only use genes that have low within-cluster variance, a.k.a. the cross-cell consistent genes. This is critical as the mean expression estimates of genes with high variance are affected by the pervasive bias in cell capture of scRNA-seq experiments, and thus cannot serve as reliable reference.

3.2.2. MuSiC model set-up

In this section, we derive the relation ship between gene expression in bulk tissue and cell type-specific gene expression in single cells. This relationship forms the basis of our deconvolution procedure. For gene g, let X_{jg} be the total number of mRNA molecules in subject j of the given tissue, which is composed of K cell types. Then $X_{jg} = \sum_{k=1}^{K} \sum_{c \in C_j^k} X_{jgc}$, where X_{jgc} is the number of mRNA molecules of gene g in cell c of subject j, and C_j^k is the set of cell index for cell type k in subject j with $m_j^k = |C_j^k|$ being the total number of cells in this set. The relative abundance of gene g in subject j for cell type k is

$$\theta_{jg}^k = \frac{\sum_{c \in C_j^k} X_{jgc}}{\sum_{c \in C_j^k} \sum_{g'=1}^{G} X_{jg'c}}. \tag{3.1}$$

We can show that

$$X_{jg} = \sum_{k=1}^{K} m_j^k S_j^k \theta_{jg}^k = m_j \sum_{k=1}^{K} p_j^k S_j^k \theta_{jg}^k, \tag{3.2}$$

where for subject j, $S_j^k = \frac{\sum_{c \in C_j^k} \sum_{g'=1}^{G} X_{jg'c}}{m_j^k}$ is the average number of total mRNA molecules for cells of cell type k (also referred as "cell type" below), $m_j = \sum_{k=1}^{K} m_j^k$ is the total number of cells in the bulk tissue, and $p_j^k = \frac{m_j^k}{m_j}$ is the proportion of cells from cell types. Let $Y_{jg} = \frac{X_{jg}}{\sum_{g'=1}^{G} X_{jg'}}$ be the relative abundance of gene g in the bulk tissue of subject j. Equation (3.2) implies

$$Y_{jg} \propto \sum_{k=1}^{K} p_j^k S_j^k \theta_{jg}^k. \tag{3.3}$$

Thus, across G genes in subject j, we have

$$\begin{bmatrix} Y_{j1} \\ \vdots \\ Y_{jg} \end{bmatrix} \propto \begin{bmatrix} \theta_{j1}^1 & \cdots & \theta_{j1}^K \\ \vdots & \ddots & \vdots \\ \theta_{jG}^1 & \cdots & \theta_{jg}^K \end{bmatrix} \cdot \begin{bmatrix} S_j^1 & & 0 \\ & \ddots & \\ 0 & & S_j^k \end{bmatrix} \cdot \begin{bmatrix} p_j^1 \\ \vdots \\ p_j^K \end{bmatrix} \tag{3.4}$$

The goal of MuSiC is to estimate p_j^k using data from scRNA-seq and bulk RNA-seq.

3.2.3. Model assumptions

If scRNA-seq were available for subject j, we would be able to obtain the cell size factor S_j^k (or the relative values of S_j^k, see below) and cell type-specific relative abundance θ_{jg}^k. With bulk RNA-seq data in subject j, we get the bulk tissue relative abundance Y_{jg}, and, if θ_{jg}^k and S_j^k were known, we would be able to perform a regression to estimate p_j^k. However, since scRNA-seq is still costly, most studies cannot afford the sequencing of a large number of individuals using scRNA-seq. To make deconvolution possible for a broader range of studies, it is desirable to utilize cell type-specific gene expression from other studies or from a smaller set of individuals in the same study. This is feasible under the following three assumptions:

Assumption 3.1 *Individuals with scRNA-seq and bulk RNA-seq are from the same population, with their cell-type specific relative abundances θ_{jg}^k in equation* (3.1) *following the same distribution with mean θ_g^k and variance σ_{gk}^2,*

$$\theta_{jg}^k \sim F(\theta_g^k, \sigma_{gk}^2). \tag{3.5}$$

Here, $F(\cdot, \cdot)$ represents a general distributional function, which is not assumed to be of any particular form. Under this assumption, deconvolution can use available single-cell data from other subjects or even subjects from other studies as reference.

Assumption 3.2 *The ratio of cell size S_k^j across cell types are the same across subjects*

and studies:

$$\frac{S_j^k}{S_j^{k'}} = \frac{S_{j'}^k}{S_{j'}^{k'}} \quad \textit{for all subjects } j, j' \in \{1, \ldots, N\} \textit{ and cell types } k, k' \in \{1, \ldots, K\}. \tag{3.6}$$

This second assumption allows us to replace S_j^k by a common value S^k across subjects. We want to emphasize that we assume the ratio, and not the absolute value, of cell size to be constant across subjects and studies, because to utilize the common value S^k, we need a constant scalar in equation (3.8) as shown below.

In practice, we do not observe the actual cell sizes S_j^k, since (1) for non-UMI data we observe read counts, not molecule counts and (2) for each cell we observe library size, not cell size. Let $\tilde{X}_{jg}$ and $\tilde{X}_{jgc}$ denote the read counts for a bulk sample and for a specific cell c in the sample, respectively. Let $\tilde{S}_j^k = \frac{\sum_{c \in C_j^k} \sum_{g'=1}^{G} \tilde{X}_{jg'c}}{m_j^k}$ denote the average library size of cell type k for subject j. We define the efficiency of cell type k for subject j as $\gamma_j^k = \tilde{S}_j^k / S_j^k$. We assume

Assumption 3.3 *The ration of average library size is the same across cell type regardless of subjects and studies*

$$\frac{\tilde{S}_j^k}{\tilde{S}_j^{k'}} = \frac{\tilde{S}_{j'}^k}{\tilde{S}_{j'}^{k'}} \quad \textit{for all } j, j' \in \{1, \cdots, N\} \textit{ and } k, k' \in \{1, \cdots, K\}. \tag{3.7}$$

Combined with assumption 3.2, equation (3.7) is equivalent to assuming that the ratio of efficiency between cell types is conserved across subjects and studies

$$\frac{\gamma_j^k}{\gamma_j^{k'}} = \frac{\gamma_{j'}^k}{\gamma_{j'}^{k'}} \quad \text{for all } j, j' \in \{1, \cdots, N\} \text{ and } k, k' \in \{1, \cdots, K\}.$$

This assumption seems plausible, since although efficiency varies across cell types and subjects, its ratio between cell types should be less variable. Assumption 3.2 and 3.3 allow us to use the common value of library size $\tilde{S}^k$ across subjects in the read count setting.

Assumption 3.1-3.3 enable us to recover the trend of cell type proportion change across subjects, as shown in Section 3.3, but does not enable the recovery of absolute cell type proportions.

To recover absolute cell type proportions, a stronger version of Assumption 3.3 is needed, which we called

Assumption 3.4 *The ratio of average library size is equal to the ratio of average cell size, for all pairs of cell types and across all subjects and studies*

$$\frac{\tilde{S}_j^k}{\tilde{S}_j^{k'}} = \frac{S_j^k}{S_j^{k'}} = \frac{S_{j'}^k}{S_{j'}^{k'}} = \frac{\tilde{S}_{j'}^k}{\tilde{S}_{j'}^{k'}} \quad \textit{for all } j, j' \in \{1, \cdots, N\} \textit{ and } k, k' \in \{1, \cdots, K\}.$$

Given Assumption 3.2 and 3.4 is equivalent to assume that the efficiency γ_j^k is the same across cell types, subjects and studies

$$\gamma_j^k = \gamma_{j'}^{k'} \quad \text{for all } j, j' \in \{1, \cdots, N\} \text{ and } k, k' \in \{1, \cdots, K\}.$$

This stronger assumption indicates that we can safely interchange the ratio of library size with the ratio of cell size to estimate cell type proportion.When this assumption is not satisfied, we can estimate the fraction of RNA molecules from each cell type, represented by $p_j^k \times S_j^k$, but the estimate of cell type proportion, p_j^k, will be biased.

3.2.4. Cell type proportion estimation

To estimate cell type proportions $\mathbf{p}_j = \{p_j^k, k = 1, \cdots, K\}$, we need to consider two constraints: (C1) Non-negativity: $p_j^k \geq 0$ for all j, k; (C2) Sum-to-one: $\sum_{k=1}^K p_j^k = 1$ for all j. Because the bulk tissue and single-cell relationship derived in equation (3.5) is a "proportional to" relationship, to satisfy the (C2) constraint, we need a normalizing constant C_j so that

$$Y_j g = C_j \big(\sum_{k=1}^{K} p_{jk} S_k \theta_{jg}^k + \epsilon_{jg} \big), \tag{3.8}$$

where $\epsilon_{jg} \sim N(0, \delta_{jg}^2)$ represents bulk tissue RNA-seq gene expression measurement noise. When cell type proportions $\mathbf{p}_j$ and subject-specific relative abundances $\boldsymbol{\theta}_{jg} = \{\theta_{jg}^k, k = 1, \ldots, K\}$ are known, the variance of bulk tissue gene expression measurement is

$$\text{Var}[Y_{jg}|\mathbf{p}_j, \boldsymbol{\theta}_{jg}] = C_j^2 \delta_{jg}^2. \tag{3.9}$$

Given only cell type proportions, the variance is

$$\begin{aligned} \text{Var}[Y_{jg}|\mathbf{p}_j] &= E[\text{Var}[Y_{jg}|\mathbf{p}_j, \boldsymbol{\theta}_{jg}]] + \text{Var}[E[Y_{jg}|\mathbf{p}_j, \boldsymbol{\theta}_{jg}]] \\ &= C_j^2 \delta_{jg}^2 + \text{Var}\big[C_j \sum_{k=1}^{K} p_{jk} S_k \theta_{jg}^k\big] \\ &= C_j^2 \delta_{jg}^2 + C_j^2 \cdot \sum_{k=1}^{K} p_j g^2 S_k^2 \text{Var}[\theta_{jg}^k] = C_j^2 \delta_{jg}^2 + C_j^2 \sum_{k=1}^{K} p_{jk}^2 S_k^2 \sigma_{gk}^2 \\ &= \frac{1}{w_{jg}^2}. \end{aligned} \tag{3.10}$$

Because of the heteroscedasticity of gene expression over genes, including the weight w_{jg} can improve estimates. Since δ_{jg}^2 is unknown, we will estimate the weight $w_j g$ iteratively, initialized by NNLS. MuSiC is robust and converges to the same value even with different starting points (Appendix A.2.7, Figure 24).

Given that bulk and single-cell expression data are generated via different protocols, it may also be necessary to consider gene-specific protocol bias. We note that the difference between the grand average of the single-cell and bulk expression profiles does not necessarily reflect bias between protocols, because the difference between cell type proportions of single-cell and bulk expression data can also lead to expression differences of marker genes even in the absence of protocol bias. To address potential protocol bias between bulk and single-cell expression data, we add a gene- and subject-specific intercept in equation (3.8), that is $Y_{jg} = C_j \cdot (\alpha_j g + \sum_{k=1}^{K} p_{jk} S_k \theta_{jg}^k + \epsilon_{jg}$. After adjusting for the protocol bias, MuSiC can detect significant biological signals across protocols (Figure 19, Table 6).

MuSiC is a weighted non-negative least squares regression (W-NNLS), which does not require pre-selected marker genes. Indeed, the iterative estimation procedure automatically imposes more weight on informative genes and less weight on non-informative genes. Because it is a linear regression-based method, genes showing less cross cell type variations will have low leverage, thus having less influence on the regression, whereas the most influential genes are those with high weight and high leverage. To illustrate this point, we also performed benchmarking experiments to show that applying MuSiC using all genes gives more accurate results than applying MuSiC using pre-selected marker genes, thus demonstrating that MuSiCs weighting scheme makes marker gene pre-selection unnecessary (Figure 20c, Figure 21). MuSiC can also deal with batch effect with its weighting scheme. When batch effect is present, the variance of relative abundance will generally increase for all cell types. This means that the batch effect with be absorbed in σ_{kg}, meaning that MuSiC not only up-weighs cross-subject consistent genes, but also cross-batch consistent genes. Thus, by down-weighting cross-batch variable genes, MuSiC effectively deals with batch effects.

The weighting scheme in MuSiC enables automatic selection of marker genes for deconvolution, as supported by our findings from the pancreas and kidney data (marker genes are highlighted with colors in Tables 11-13). However, we note that some of the top-ranked genes are not necessarily marker genes. This is because genes in MuSiC are weighed by the combined effect of cross-subject variation and cross-cell-type variation, which are very different concepts. The cross-subject variation measures the consistency of genes across subjects while the cross-cell-type variation measures the cell type specificity of genes. The top ranked non-markers genes for the analyses in Results tend to be consistently expressed across subjects, and are usually highly expressed. Although they are not exclusively expressed in a particular cell type, they are differentially expressed across cell types, thus offering power to differentiate different cell types. We believe that MuSiC benefit from these genes and hence yield more accurate cell type proportions than methods that only use marker genes in deconvolution.

3.2.5. Recursive tree-guided deconvolution for closely related cell types

Complex solid tissues often include closely related cell types with similar gene expression levels. Correlation in gene expression can lead to collinearity, making it difficult to reliably estimate cell type proportions, especially for less frequent and rare cell types. Although the collinearity problem can be improved by selecting marker genes through support vector regression, as is done in CIBERSORT[ref3] and BSEQ-sc[ref4], these approaches still have limited power to resolve similar cell types. In MuSiC, we introduce a recursive tree-guided deconvolution procedure based on a cell type similarity tree, which can be easily obtained through hierarchical clustering. In stage 1 of this procedure, cell types in the design matrix are divided into high-level clusters by hierarchical clustering with closely related cell types clustered together. Proportion for these cell type clusters are estimated using genes with small intra-cluster variance (cluster-consistent genes) using the above described W-NNLS. In stage 2, for cell types in each cluster, the cell type proportions are estimated using W-NNLS with genes displaying small intra-cell type variance, subject to the constraint on the pre-estimated cluster proportions. If necessary, more than 2 stages of recursion can be applied, with each stage separating the cell types within each large cluster into finer clusters, and using cluster-consistent genes to do W-NNLS subject to the constraint that fixes higher-level cluster proportions.

To illustrate this recursive tree-guided deconvolution procedure, we start with a simple case with four cell types and G genes. Let X_1,X_2,X_3,X_4 represent cell type-specific expression in the design matrix, obtained from scRNA-seq, and let Y be the gene expression vector in the bulk RNA-seq data. The relationship of bulk and single-cell data can be written as

$$\begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} = \begin{pmatrix} X_1^{(1)} & X_2^{(1)} & X_3^{(1)} & X_4^{(1)} \\ X_1^{(2)} & X_2^{(2)} & X_3^{(2)} & X_4^{(2)} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} + \begin{pmatrix} \epsilon^{(1)} \\ \epsilon^{(1)}, \end{pmatrix} \tag{3.11}$$

where the superscripts $^{(1)}$ and $^{(2)}$ indicate two sets of genes. Suppose the four cell types are grouped into two clusters, (X_1, X_2) and (X_3, X_4). The first set of genes are those showing small intra-cluster variance in gene expression, that is, $X_1^{(1)} \approx X_2^{(1)}$ and $X_3^{(1)} \approx X_4^{(1)}$, whereas the second set of genes are the remaining genes.

Stage 1 Estimate cluster proportions $\pi_1 = p_1 + p_2$ and $\pi_2 = p_3 + p_4$,

$$Y^(1) = X_1^{(1)}\pi_1 + X_3^{(1)}\pi_2 + \epsilon^{(1)}. \tag{3.12}$$

The cluster proportions, $\hat{\pi}_1$ and $\hat{\pi}_2$, are estimated by W-NNLS using intra-cluster homogenous genes.

Stage 2 : Estimate cell type proportions (p_1, p_2, p_3, p_4),

$$Y^{(2)} = X_1^{(2)}p_1 + X_2^{(2)}p_2 + X_3^{(2)}p_3 + X_4^{(2)}p_4 + \epsilon^{(2)}. \tag{3.13}$$

The cell type proportions are estimated by W-NNLS using the remaining genes subject the constraint that

$$\hat{p}_1 + \hat{p}_2 = \hat{\pi}_1, \quad \text{and} \quad \hat{p}_3 + \hat{p}_4 = \hat{\pi}_2. \tag{3.14}$$

3.2.6. Interconversion of different gene expression measures

MuSiC links bulk and single-cell gene expression by mRNA molecule counts. There are many measures of mRNA abundance, such as read counts, UMI counts, RPKM and TPM. As molecule counts are not observed in real studies, we approximate the molecule counts by read counts and estimate cell type proportions based on assumptions 3.1 - 3.3. The interconversion between other gene expression measures and read count determines if MuSiC can utilize other measures as the input for deconvolution. One step in MuSiC estimation is the use of average library size as a proportional measure of average cell size for a given cell type, which is absent in normalized measurements of mRNA abundance such as RPKM and TPM. For RPKM, we would need the average library size for each cell type to be provided,

or the average cell size for each cell type to be obtained from other sources. Cell type proportions cannot be estimated by MuSiC with TPM information alone. Below, we derive the relationships of various types of gene expression measures in detail.

Let L_g denote the length of gene g, and the corresponding RPKMs of bulk and single-cell data are denoted by $\hat{X}_{jg}$ and $\hat{X}_{jgc}$, respectively. For simplicity, we omit the 10^3 scalar for now. By definition,

$$\hat{X}_{jg} = \frac{\tilde{X}_{jg}/L_g}{\sum_{g'=1}^{G} \tilde{X}_{jg'}}, \quad \tilde{X}_{jgc} = \frac{\tilde{X}_{jgc}/L_g}{\sum_{g'=1}^{G} \tilde{X}_{jg'c}}, \tag{3.15}$$

where $\tilde{X}_{jg}$ and $\tilde{X}_{jgc}$ denote the bulk and single cell read counts, respectively.

Based on the model set-up described earlier, we can show that the relationship between bulk and single-cell RPKMs is

$$\hat{X}_{jg} \propto \frac{\tilde{X}_{jg}}{L_g} = \sum_{k=1}^{K} \sum_{c \in C_j^k} \Big(\frac{\tilde{X}_{jgc}/L_g}{\sum_{g'=1}^{G} \tilde{X}_{jg'c}} \cdot \sum_{g'=1}^{G} \tilde{X}_{jg'c} \Big) = \sum_{k=1}^{K} \sum_{c \in C_j^k} \hat{X}_{jgc} \tilde{S}_{jc} \tag{3.16}$$

where $\tilde{S}_{jc}$ is the library size of cell c. Equation (3.16) can be further approximated by

$$\hat{X}_{jg} \propto \sum_{k=1}^{K} \sum_{c \in C_j^k} \hat{X}_{jgc} \tilde{S}_{jc} \approx \sum_{k=1}^{K} m_j^k \hat{\theta}_{jg}^k \tilde{S}_j^k = m_j \sum_{k=1}^{K} p_j^k \hat{\theta}_{jg}^k \tilde{S}_j^k \tag{3.17}$$

where $\hat{\theta}_{jg}^k = \sum_{c \in C_j^k} \hat{X}_{jgc}/m_j^k$ is the average RPKM of gene g in subject j for cell type k.

To utilize multi-subject information, we assume $\hat{\theta}_{jg}^k$ follows the same assumption as Assumption 3.1, that is, individuals with scRNA-seq and bulk RNA-seq are from the same population, with their cell-type specific average RPKM $\hat{\theta}_{jg}^k$ following the same distribution with mean $\hat{\theta}_g^k$ and variance $\hat{\sigma}_{gk}^2$,

$$\hat{\theta}_{jg}^k \sim \tilde{F}(\hat{\theta}_g^k, \hat{\sigma}_{jg}^2). \tag{3.18}$$

Assumption 3.2 states that the ratio of average library size is consistent across subjects and studies, which justifies the use of $\tilde{S}_j^k$ from other studies if these quantities are not available

for the same data set. The linear relation between bulk RPKM and average cell-type specific single cell RPKM is approximated by formula (3.17). Since this is an approximation, MuSiC estimates using RPKM may not be as accurate as those using read or UMI count. In our test of MuSiC using RPKM values for the pancreatic islets bulk mixture experiment, we found that it is not as accurate as MuSiC estimates using read count, but still higher than NNLS, BSEQ-sc, and CIBERSORT (Figure 23d).

Another widely used normalized mRNA measure is TPM. Let $\hat{Z}_{jg}$ and $\hat{Z}_{jgc}$ denote the bulk and single-cells TPM values, respectively. By definition, . Let Z_{jg} and Z_{jgc} be the gene length normalized read count in bulk and single cell, that is, $Z_jg = \tilde{X}_{jg}/L_g$ and $Z_{jgc} = \tilde{X}_{jgc}/L_g$. The link between bulk and single-cell TPMs is

$$\hat{Z}_{jg} \propto Z_{jg} = \sum_{k=1}^{K} \sum_{c \in C_j^k} Z_{jgc} = \sum_{k=1}^{K} \sum_{c \in C_j^k} \Big(\frac{Z_{jgc}}{\sum_{g'=1}^{G} Z_{jg'c}} \cdot \sum_{g'=1}^{G} Z_{jg'c} \Big) = \sum_{k=1}^{K} \sum_{c \in C_j^k} \hat{Z}_{jgc} \hat{S}_{jc}, \quad (3.19)$$

where $\hat{S}_{jc}$ is the summation of normalized read counts in cell c for subject j. Equation (3.19) suggests that it is difficult to make assumptions or approximations to express relative abundance as a function of TPM.

3.2.7. Construction of benchmark datasets and evaluation metrics

To evaluate MuSiC and compare with other deconvolution methods, we need bulk RNA-seq data with known cell type proportions. Therefore, we construct artificial bulk tissue data from a scRNA-seq dataset in which the bulk data is obtained by summing up gene counts from all cells in the same subject. Relative abundance is calculated by equation (3.1). The true cell type proportions in the artificial bulk data can be directly obtained from the scRNA-seq data and this allows us to use this artificially constructed bulk data as a benchmark dataset to evaluate the performance of different deconvolution methods (A.2.2). Denote the true cell type proportions by $\mathbf{p}$ and the estimated proportions by $\hat{\mathbf{p}}$. Deconvolution methods are evaluated by the following metrics.

(i) Pearson correlation, $R = \text{Cor}[\mathbf{p}, \hat{\mathbf{p}}]$;

(ii) Root mean squared deviation, $\text{RMSD} = \sqrt{\text{avg}(\mathbf{p} - \hat{\mathbf{p}})^2}$;

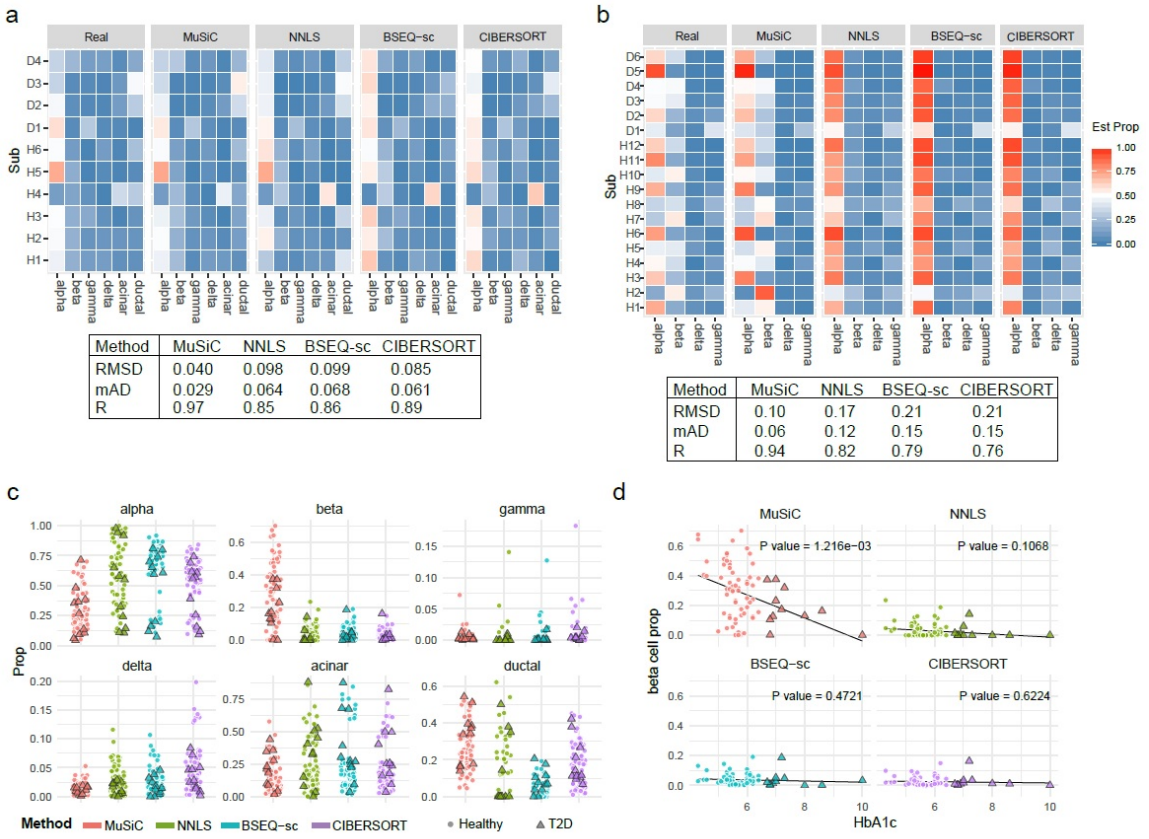
(iii) Mean absolute deviation, $\text{mAD} = \text{avg}(|\mathbf{p} - \hat{\mathbf{p}}|)$.

3.3. Results of deconvolution

3.3.1. Application to pancreatic islets in human

To demonstrate and evaluate MuSiC, we started with a well-studied tissue, the islets of Langerhans, which are clusters of endocrine cells within the pancreas that are essential for blood glucose homeostasis. Pancreatic islets contain five endocrine cell types (α, β, δ, ϵ, and γ), of which β cells, which secrete insulin, are gradually lost during type 2 diabetes (T2D). We applied MuSiC to bulk pancreatic islet RNA-seq samples from 89 donors from Fadista et al. (2014), to estimate cell type proportions and to characterize their associations with hemoglobin A1c (HbA1c) level, an important biomarker for T2D. We were motivated to re-analyze this data because, as shown in Figure 8 and in Baron et al. (2016), existing methods failed to recover the correct β cell proportions, which should be around 50-60%, and also failed to recover their expected negative relationship with HbA1c level. As reference, we experimented with scRNA-seq data from two sources: 6 healthy and 4 T2D adult donors from Segerstolpe et al. (2016), and 12 healthy and 6 T2D adult donors from Xin et al. (2016). All bulk and single-cell datasets in this analysis are summarized in Table 1.

First, to systematically benchmark, we applied MuSiC and three other methods (Non-negative least squares (NNLS), CIBERSORT, and BSEQ-sc) to artificial bulk RNA-seq data constructed by simply summing the scRNA-seq read counts across cells for each single-cell sequenced subject. In this case, true cell type proportions are known, which allows the evaluation of accuracy. More details on artificial bulk construction are described in the Appendix A.2.2. Figure 8a, 20c and Figure 21b show the estimation results when the artificial bulk and the single-cell reference data are from the same study, either both from



Method	MuSiC	NNLS	BSEQ-sc	CIBERSORT
RMSD	0.040	0.098	0.099	0.085
mAD	0.029	0.064	0.068	0.061
R	0.97	0.85	0.86	0.89

Method	MuSiC	NNLS	BSEQ-sc	CIBERSORT
RMSD	0.10	0.17	0.21	0.21
mAD	0.06	0.12	0.15	0.15
R	0.94	0.82	0.79	0.76

Figure 8: Pancreatic islet cell type composition in healthy and T2D human samples. **a** and **b** Benchmarking of deconvolution accuracy on bulk data constructed by combining together scRNA-seq samples. **a.** The bulk data is constructed for 10 subjects from Segerstolpe et al. (2016) while the single cell reference is taken from the same dataset. The cell type proportions of healthy subjects are estimated by leave-one-out single cell reference. The subject names are relabeled; the table shows average root mean square error (RMSD), mean absolute deviation (mAD), and Pearson correlation (R) across all samples and cell types. **b.** The bulk data is constructed for 18 subjects from Xin et al. (2016) while the single cell reference is 6 healthy subjects from Segerstolpe et al. (2016). **c.** Jitter plots of estimated cell type proportions for Fadista et al. (2014) subjects, color-coded by deconvolution method. Of the 89 subjects from Fadista et al. (2014), only the 77 that have recorded HbA1c level are plotted, and T2D subjects are denoted as triangles while non-diabetic subjects are denoted as dots. **d.** HbA1c vs beta cell type proportions estimated by each of 4 methods. The reported p-values are from single variable regression β cell proportion HbA1c. Multivariable regression results are reported in Table 6. Figure 19 shows the deconvolution results of Fadista et al. (2014) with the inDrop data from Baron et al. (2016) as single-cell reference. The corresponding multivariable regression results are shown in Table 6.

Name	Accession number	Data type	Protocol	# sample	# cells	# genes
Segerstolpe et al. (2016)	E-MTAB-5061	Single cell	Smart-seq2	10 (6H + 4 T2D)	2209	25453
Segerstolpe et al. (2016)	E-MTAB-5060	Single cell	Smart-seq2	7 (3H + 4 T2D)	NA	25453
Xin et al. (2016)	GSE81608	Bulk		18 (12H + 6 T2D)	1492	39849
Baron et al. (2016)	GSE81433	Single cell	InDrop	3H	7729	17434
Fadista et al. (2014)	GSE50244	Bulk		89	NA	56638

Table 1: Pancreatic islet datasets
For data type column, "Single cell" means "single cell RNA-seq" and "Bulk" represents "bulk RNA-seq". For # sample column, "H" means "healthy" and "T2D" means "type 2 diabetes".

Segerstolpe et al. (2016) or both from Xin et al. (2016). MuSiC achieves improved accuracy over existing procedures. Figure 8b and Figure 21a show the estimation results when the artificial bulk and the single-cell reference data are from different studies. This is a more challenging but more realistic scenario, since library preparation protocols vary across labs and bulk deconvolution analyses are often performed using single-cell reference generated by others. MuSiC still maintains high accuracy, while other methods perform substantially worse. Further comparisons show that, unlike existing methods that rely on pre-selected marker genes, MuSiC gives accurate results when the cell type composition in the bulk data is substantially different from that of the single cell reference (Figure 21c and Appendix A.2.3), and when the bulk tissue contains minority cell types that are missing in the reference (Figure 22 and Appendix A.2.4). MuSiCs ability to transfer knowledge across data sources is derived from its consideration of marker gene consistency.

We now turn to the deconvolution of bulk RNA-seq data from Fadista et al. (2014). We first used the scRNA-seq data from Segerstolpe et al. (2016) as reference for all methods. MuSiC recovers the expected $\approx 50 - 60\%$ β cell proportion for the healthy subjects (Cabr-

era et al., 2006), whereas other methods grossly overestimate the proportion of α cells and underestimate the proportion of β cells. Furthermore, MuSiC detects a significant association of β cell proportion with HbA1c level (p-value = 0.00126, Figure 8d). Based on clinical standard, HbA1c level $< 6.0\%$ is classified as normal, and $> 6.5\%$ is classified as diabetic. After adjusting for age, gender and body mass index, MuSiC estimates suggest that a 0.5% increase in HbA1c level, representing the magnitude of increase from normal to the diabetes cutoff, corresponds to a drop of $3.07\% \pm 2.49\%$ in β cell proportion (Table 5). The scRNA-seq data from Segerstolpe et al. (2016) was generated by the Smart-seq2 protocol. Similar results are obtained when using the InDrop scRNA-seq data from Baron et al. (2016) as reference. MuSiC detects the significant association of β cell proportion with HbA1c level with and without adjustment for covariates (Figure 19, Table 6). The weight ordered gene list for pancreatic islet analysis are provided in Table 11.

3.3.2. Application to kidney in mouse and rats

As a second tissue example, we used the kidney, a complex organ consisting of several anatomically distinct segments each playing critical roles in the filtration and re-absorption of electrolytes and small molecules of the blood. Chronic kidney disease (CKD), the gradual loss of kidney function, is increasingly recognized as a major health problem, affecting 10-16% of the global adult population. We aim to characterize how kidney cell type composition changes during CKD. Fibrosis is the histologic hallmark common to all CKD models, and hence, we analyzed the bulk RNA-seq data from three mouse models for renal fibrosis: unilateral ureteric obstruction induced by surgical ligation of the ureter (UUO, Arvaniti et al. (2016)), toxic precipitation in the tubules induced by high dose folic acid injection (FA, Craciun et al. (2016)), or genetic alteration by transgenic expression of genetic risk variant *APOL1* in podocytes (*APOL1* transgenic mice12). As reference, we used the mouse kidney specific scRNA-seq data from Park et al. (2018). Details of all datasets are summarized in Table 2. We systematically benchmarked all methods on artificial bulk experiments performed using the Park et al. (2018) scRNA-seq data, finding similar trends as those in

Figure 8a-b (Figure 25a-b).

Name	Accession number	Data type	Protocol	# sample	# cells	# genes
Park et al. (2018)	GSE107585	Single cell	10x	7H	43745	16273
Beckerman et al. (2017)	GSE81492	Bulk		10 (6 control + 4 *APOL1*)	NA	19033
Lee et al. (2015)	GSE56743	Bulk		118 replicates (14 segments)	NA	10903
Craciun et al. (2016)	GSE65267	Bulk		18 replicates (6 time points)	NA	25219
Arvaniti et al. (2016)	GSE79443	Bulk		10 replicates (Sham + 2 time points)	NA	38683

Table 2: Mouse/Rat kidney datasets
For data type column, "Single cell" means "single cell RNA-seq" and "Bulk" represents "bulk RNA-seq". For # sample column, "H" means "healthy" and "T2D" means "type 2 diabetes".

Hierarchical clustering of the cell types in the single cell reference reveals that, apart from neutrophils and podocytes, kidney cells fall into two large groups: Immune cell types (macrophages, fibroblasts, T lymphocytes, B lymphocytes, and natural killer cells) and kidney-specific cell types (proximal tubule, distal convolved tubule, loop of Henle, two cell types forming the collecting ducts, and endothelial cells). Of these, proximal tubule (PT) is the dominant cell type in kidney, and the proportion of PT cells is known to decrease with CKD progression. MuSiC finds this decrease in all three mouse models (Figure 9b-d). Other methods also detect this association for the *APOL1* and UUO mouse models, but showed ambiguous results for the FA model.

Distal convolved tubule cells (DCT) are known to be the second most numerous cell type in kidney, with an expected proportion of 10-20%. Yet, CIBERSORT did not detect DCT in any of the three bulk datasets; BSEQ-sc missed it in two datasets and grossly over-estimated its proportion in the third dataset at the cost of a grossly underestimated PT proportion. This is due to the high similarity between DCT and PT, observable in Figure 9a. Through

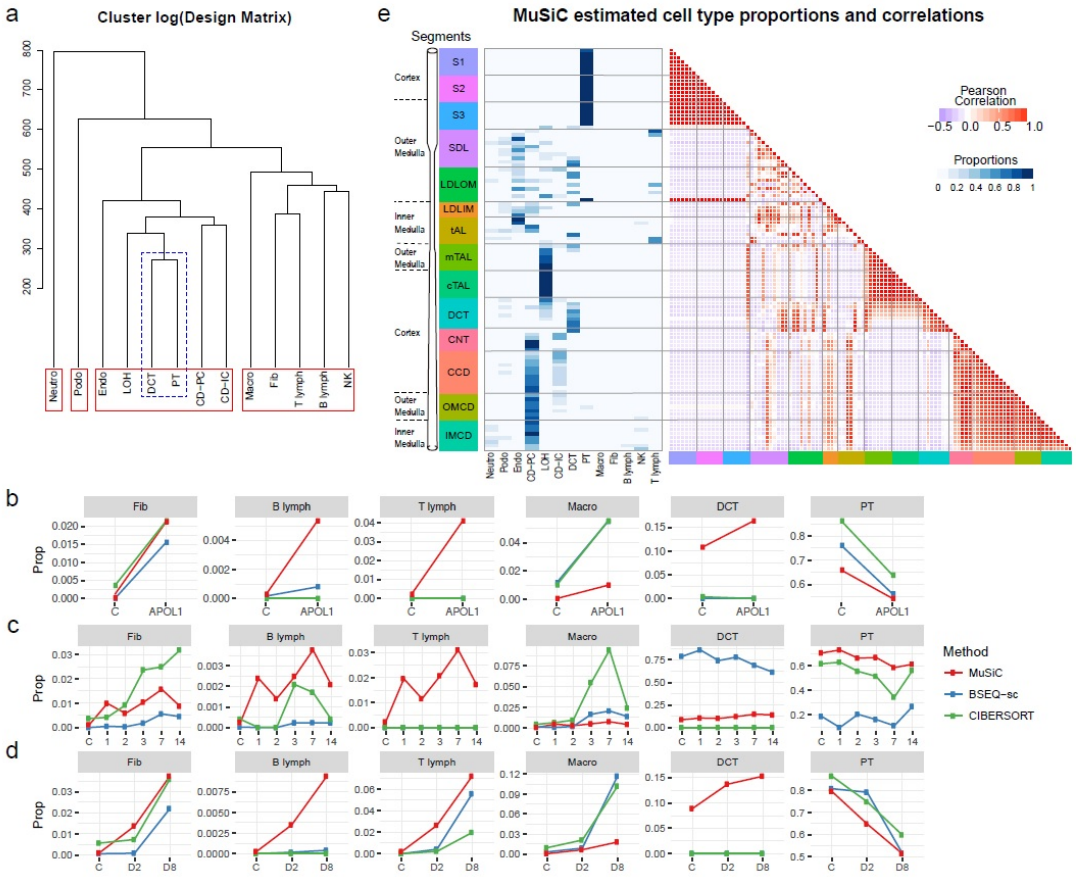


Figure 9: Cell type composition in kidney of mouse CKD models and rat.
a. Cluster dendrogram showing similarity between 13 cell types that were confidently characterized in Park et al. (2018). Abbreviations: Neutro: neutrophils, Podo: podocytes, Endo: endothelials, LOH: loop of Henle, DCT: distal convolved tubule, PT: proximal tubule, CD-PT: collecting duct principal cell, CD-IC: CD intercalated cell, Macro: macrophages, Fib: fibroblasts, NK: natural killers. **b**, **c** and **d.** Average estimated proportions for 6 cell types in bulk RNA-seq samples taken from 3 different studies, each study based on a different mouse model for chronic kidney disease. Results from three different deconvolution methods (MuSiC, BSEQ-sc and CIBERSORT) are shown by different colors. Supplementary Figure 27a-c show complete estimation results of all 13 cell types. **b.** Bulk samples are from Beckerman et al. (2017), who sequenced 6 control and 4 *APOL1* mice. **c.** Bulk data are from Craciun et al. (2016), where samples are taken before (C) and at 1, 2, 3, 7, 14 days after administering folic acid. Line plot shows cell type proportion changes over time (days), averaged over 3 replicates at each time point. **d.** Bulk data are from Beckerman et al. (2017), where samples are taken from mice after Sham operation (C), 2 days after UUO operation (D2), and 8 days after UUO operation (D8). The average proportions at each time point are plotted. **e.** MuSiC estimated cell type proportions of rat renal tubule segments. The estimated cell type proportions (left) and the proportions correlations between samples (right) are shown as heatmap. Segment names are color coded and aligned according to their physical positions along the renal tubule. Figure 26a-c show NNLS, BSEQ-sc and CIBERSORT results. See Table 10 for full segment names.

its tree-guided recursive algorithm, MuSiC first estimates the combined proportion of kidney cell types versus immune cell types using consistent genes for these two large groups, and then zooms in and deconvolves the kidney cell types using genes re-selected for each kidney cell type. This allows MuSiC to successfully separate PT and DCT cells in all three bulk datasets, recovering a consistent DCT proportion between 8-20%, matching expectations. Interestingly, unlike for PT, the proportion of DCT cells show a consistent increase with disease progression across all three mouse models. This may seem counter-intuitive given that loss of kidney function is expected to be associated with the loss of kidney cell types. But given the substantial drop of the dominant PT cell type, the proportion of DCT cells relative to the whole may increase, even if its absolute count drops.

Next, we consider immune cells, which are known to play a central role in the pathogenesis of CKD. MuSiC found the largest immune sub-type to be macrophage, and all methods detected the expected increase of macrophage proportion with disease progression. Apart from this, MuSiC also found fibroblasts, B-, and T-lymphocytes to increase in proportion with disease progression, giving a consistent immune signature that is reproduced across mouse models. These findings are consistent with clinical and histological observations, indicating tissue inflammation is a consistent feature of kidney fibrosis. Such reproducible signatures were not found by other methods, which show much less agreement across mouse models. The weight ordered gene list for the three mouse models are provided in Table 12-13.

Finally, to illustrate MuSiCs cross-species applicability, we used the mouse kidney scRNA-seq reference from Park et al. (2018) to deconvolve the micro-dissected segment aggregated rat RNA-seq data from Lee et al. (2015), which contains 105 samples obtained from 14 segments spaced along the renal tubule. Cell type proportions are estimated with homologous genes between mice and rat. We mapped samples to their physical locations, and computed correlations between their cell type proportions (Figure 9). Reassuringly, cell types recovered by MuSiC for each segment agree with knowledge (Zhai et al., 2006) about

the dominant cell type at its mapped position, e.g. DCT cells come from the DCT segment. Correlation between samples is also high within anatomically distinct segments.

3.3.3. Evaluation of robustness for MuSiC

A good deconvolution method should be robust to the choice of single-cell reference. We conducted additional experiments to evaluate the robustness of MuSiC and other existing methods. First, we considered the case where cell type proportions in the single cell data are drastically different from those in the bulk data. Our results indicate that, under this scenario, MuSiC recovers the true cell type composition, improving upon the severely biased estimates produced by other existing approaches (Appendix A.2.3, Figure 21c). One limitation of scRNA-seq is that it may fail to recover some cell types, in particular, rare cell types may be missed. We next created considered the setting where the single-cell reference is incomplete, and found that MuSiC estimation is still accurate as long as the missing cell type is not the dominant cell type in bulk tissue (Appendix A.2.4, Figure 22, and Table 7). MuSiC is also tolerant of different scRNA-seq protocols. This has already been shown through the above analyses, where accurate deconvolution results were obtained using single cell reference generated using the Smart-seq2, inDrop, and 10x Chromium protocols. To probe this further, we directly investigated the impact of using biased values of relative abundance θ_{jg}^k in MuSiCs deconvolution step, and found that MuSiC estimated cell type proportions remain accurate, still improving upon existing methods, even though unbiased relative abundance values were provided to the existing methods as input (Appendix A.2.5, Figure 23c). Finally, we evaluated the impact of dropout in the single cell reference, by introducing dropout according to Jia et al. (2017) and varying the dropout rate in the benchmark experiment of Figure 8. MuSiC estimation is still accurate even when dropout rate is around 30% (Appendix A.2.6, Figure 23a-b).

3.4. Discussion

Knowledge of cell type composition in disease relevant tissues is an important step towards the identification of cellular targets in disease. Although most scRNA-seq data do not reflect true cell type proportions in intact tissues, they do provide valuable information on cell type-specific gene expression. Existing cell type deconvolution methods rely on pre-selected marker genes and ignore subject-to-subject variation and cross-cell consistency in gene expression. Through comprehensive benchmark evaluations and analysis of multiple real datasets, we show that both cross-subject and cross-cell consistency in gene expression need to be considered in deconvolution. By incorporating both types of consistency, MuSiC allows for scRNA-seq datasets to serve as effective references for independent bulk RNA-seq datasets involving different individuals. Harnessing multi-subject scRNA-seq reference data, MuSiC reliably estimates cell type proportions from bulk RNA-seq, therefore enabling the transfer of cell type-specific gene expression from one dataset to another. As bulk tissue data are more easily accessible than scRNA-seq, MuSiC allows the utilization of the vast amounts of disease relevant bulk tissue RNA-seq data for elucidating cell type contributions in disease.

Although this paper uses read counts as the measures of mRNA abundance, there are many other commonly used measures, such as RPKM and TPM. MuSiC can utilize RPKM if estimates of cell type specific total RNA abundance can be provided (e.g. estimated from another data set). However, cell type proportions cannot be estimated with TPM as the input. Detailed interconversion between read counts and other gene expression measures have been discussed in Section 3.2.

CHAPTER 4 : Cell type specific differential expression from bulk tissue with single cell reference

4.1. Introduction

In Chapter 3, we developed a deconvolution method, MuSiC, which estimates bulk tissue cell type proportions with single cell reference without requiring the pre-selection of marker gene. The real data examples in Section 3.3 confirmed the heterogeneity of cell type proportions across subjects and conditions. In the pancreas islet analysis, beta cell proportions associated negatively with HbA1c level, which is an indicator of diabetes. In this chapter, we examined how these estimated proportions can be used to more detailed characterize of gene expression changes across biological conditions.

Differential expression analysis is one of the common tools for discovering quantitative changes in expression levels between biological conditions, by testing if the difference of expression levels between biological conditions is greater than that due to natural random variations. Methods for differential expression analysis with bulk RNA-seq data have been developed. Currently, popular ones include DESeq2(Love et al., 2014), edgeR(Robinson and Oshlack, 2010), and more in Costa-Silva et al. (2017). However, differential expression analysis using bulk RNA-seq data is a product of the cell-type specific contribution and the cell type proportion changes, which is inadequate for cell type level study. With help of single cell reference, we are now able to quantify, using MuSiC, cell type proportions but the cell type specific differential expression are still unclear to us. Although, scRNA-seq make it possible to compare the cell type specific expression between conditions, the cost and labor usually limit the application of scRNA-seq. With cell type proportions, the large quantities to large cohorts of bulk sequencing data is still useful for cell-type specific differential expression analysis.

As stated in Chapter 3, bulk sequencing data and cell type proportions of the same tissue usually can not be observed at the same time. Hence, deconvolution is needed before

cell-type-specific differential expression analysis, for estimating the cell type proportions of the same tissue. Deconvolution methods such as CIBERSORT(Newman et al., 2015) and BSEQ-sc(Baron et al., 2016), as we mentioned in Chapter 3, pre-select marker genes. Deconvolution methods with marker genes assume that the expression of marker genes are consistent across conditions, which is not always true consider, for example, beta cell in the pancreas, whose main role is the production of insulin. *INS* is the gene responsible for insulin production and has always considered as a marker gene for beta cells. During the progression of type 2 diabetes, beta cells representation in the pancreas, decrease in number, but within the beta cells that are retained, there is also a functional decline (Porte and Kahn, 2001), where the expression of *INS* is lower in diseased subjects than healthy subjects. Although the expression of *INS* can still be used as marker genes for beta cell in diseased subjects, using only *INS* as the marker gene for deconvolution will attribute all of the differential expression to cell type proportions changes, while some of it is due to within cell type expression shift.

To avoid the biases of estimated proportions introduced by marker genes, we use MuSiC for deconvolution without marker genes. When incorporating both diseased and healthy single cell profile, we can automatically up-weigh genes that have consistent expression across different cohorts by MuSiC. However, using all genes for proportion estimation would confound the results of differential expression tests, where the estimated cell type specific gene expressions from both cohorts are driven towards the expression from single cell profile. Here we developed a method, MuSiC-DE, for finding cell type specific differential expressed (DE) gene from bulk data and single cell reference. We first estimate cell type proportions using a subsample of genes and then test differential expression on the rest of the genes, conditioning on the estimated cell type proportions. Repeating the subsampling-testing procedure multiple times, each gene in the bulk expression data is tested at least one time and gets a least one p-value. Since the p-values across repetitions are correlated, we apply the empirical Bayes framework to the p-values and determine the DE genes.

In this chapter, we focus on testing cell type specific differential expression with examples of human pancreas islet. This chapter is organized as follows: Section 4.2.1 introduced the null hypothesis and notations for DE test. In Section 4.2, we described the MuSiC-DE model, detailed estimation procedures and expected results. Null distribution validation using Fadista et al. (2014) data, benchmark examination with Xin et al. (2016), and the DE test results of Fadista et al. (2014) are shown in Section 4.3. We summarized methods and results and discussed future directions in Section 4.4.

4.2. Method

4.2.1. Model Set-up and Notations

We consider, for simplicity, the two-group scenario. Our method and analysis can be easily extended to multiple-group scenarios.

Suppose there are subjects $j = 1, \ldots, N$, gene $g = 1, \ldots, G$ and cell types $k = 1, \ldots, K$. For each subject sequenced by bulk RNA-seq, we have an indicator of whether the subject is healthy or diseased, t_j:

$$t_j = \begin{cases} 1 & \text{disease} \\ 0 & \text{normal} \end{cases}$$

For subject j and cell type k, the cell type proportions are denoted by p_{jk}.

For subject j and gene g, bulk expression can be view as weighted sum of cell type specific expression β_{kg}, where the weights are cell type proportions p_{jk}. There are some rare cell types not presenting in the single cell profile and the intercepts are required to incorporate the expression of those rare cell types. We assume that bulk expression, denoted as X_{jg} follows a linear additive model

$$X_{jg} = \beta_{0g} + \sum_{k=1} p_{jk}\beta_{kg} + \alpha_{0g}t_j + \sum_{k} p_{jk}t_j\alpha_{kg} + \epsilon_{jg}. \tag{4.1}$$

We called model (4.1) full model. There are two different interpretation of this model,

depending on the data what we input for X_{jg}. In the first case, X_{jg} is the normalized read counts, counts per million (CPM). Correspondingly, β_{0g} is the mean of missing cell types' expression and β_{kg} is the cell type specific expression for cell type k. The cell type specific differences between diseased and healthy subjects are modeled by α_{kg}. ϵ_{jg}, the noise term, does not follow normal distribution in general. In the second case, $X_{jg} = \log(\text{CPM} + 1)$ is log-transformed CPM and the interpretation above no longer holds. β_{kg} is the cell type specific parameters, instead of cell type specific gene expression. The cell type specific difference between diseased and healthy subjects are modeled by α_{kg}. In this case, it is more reasonable to assume that the noise term ϵ_{jg} normal homoscedasticity.

Now, under the context of this model (4.1), our goal is to detect cell type specific differential expression for each gene. The cell type proportions are not known in general, but can be estimated by deconvolution. In our derivation of test statistics, we will assume the scenario where proportions(p_{jk}) are known and then treat the case where it is unknown.

When the cell type proportions p_{jk} are known, the differential expression analysis can be reduced into two hypothesis testing problems. The first hypothesis is

$$H_{01} : \alpha_{kg} = 0 \quad \text{for all } k \in \{0, 1, \ldots, K\}. \tag{4.2}$$

Testing H_{01} is equivalent to testing that after accounting for proportions difference, which gene (genes) are still differential expressed across all of the cell types. For genes that reject H_{01}, we further specify in which cell type those genes differential expressed by testing the second hypothesis

$$H_{02}^{(k)} : \alpha_{kg} = 0 \quad \text{for a certain } k \in \{0, 1, \ldots, K\}. \tag{4.3}$$

Cell type proportions can be estimated by deconvolution methods with proper single cell reference. With estimated cell type proportions, $\hat{p}_{jk}$, we can run regression of equation (4.1) and estimate β_{0g}, β_{kg}, α_{0g} and α_{kg} at the same time. We call the regression that estimates

cell type specific expression with known cell type proportions as "Reverse regression". To test H_{01} and H_{02}, we compared equation (4.1) with regression under null hypothesis:

$$X_{jg} = \beta_{0g} + \sum_{k=1} p_{jk}\beta_{kg} + \epsilon_{jg}. \tag{4.4}$$

When treating X_{jg} as CPM and β_{kg} as cell type specific expression, we can not estimate cell type specific expression by ordinary least square due to the non-negative constraints on cell type specific expression, that is, $\beta_{kg} \geq 0$ and $\beta_{kg} + \alpha_{kg} \geq 0$ for all $k \in \{0, 1, \ldots, K\}$. Therefore, the null hypothesis H_{01} can not be tested by ANOVA test by comparing the residual of the full model (4.1) and cell type model (4.4). Through the literature, there are no test statistics for constrained linear regression. If we view X_{jg} as log-transformed CPM and β_{kg} as cell type specific parameters, full model (4.1) and cell type model (4.4) satisfy the requirement for ordinary least square estimation and H_{01} can be tested by ANOVA test. In this chapter, we examined both interpretations and found out the second interpretation, taking X_{jg} as log-transformed CPM, performs better in null distribution validation.

4.2.2. Hypothesis testing procedures

The reverse estimation and the hypothesis tests can be easily done when true cell type proportions of bulk tissues are available. However, this is usually not the case and we need to substituted with estimated cell type proportions from deconvolution methods. Using marker gene based deconvolution methods may bias cell type proportion estimates, which would lead to inaccurate cell type specific expression estimates from reverse regression. MuSiC can avoid bias from marker genes selection by assigning weights to all genes. The weights are calculated by the inverse of cross-subject variance and genes differential expressed across different conditions will receive low weights. However, there still exists a confounding problem in the reverse regression when a gene is tested and is used to estimate proportions at the same time. To tackle the confounding problem, we randomly partition common genes of bulk data and single cell data into two sets, deconvolution set and DE test set. Genes in DE test set are tested using proportions estimated with genes in the

deconvolution set. We repeat the partition-test steps many times so that all common genes get tested. Detailed procedures are described in Algorithm 4.2.2 and we call our method: MuSiC-DE. [H]

Matrix of bulk expression $\mathbf{X} = (X_{jg})$; Number of bulk subjects N; Number of common genes between bulk data and single cell data G; Number of cell types K; Repetition times R; size of deconvolution set m P-values $\mathbf{p} = (p^{(g)})$ for all repetitions and genes **Initialization:** $r = 1$; List $\mathbf{p} = (p^{(g)})$ of length G $r \leq R$ Randomly select m genes from G common genes Estimate cell type proportions by MuSiC with selected m genes, estimated proportions are $\hat{q}_{jk}$ Test H_{01} by comparing model (4.1) and model (4.4) with $\hat{q}_{jk}$ for each of the left-over $(G - m)$ genes Get p-values for each left-over genes: $p_r^{(g)}$ $r = r + 1$;

In a given repetition, a gene receives the p-value if it was not selected for deconvolution. Suppose there are R_g p-values for gene g, that is, R_g out of R repetitions do not select gene g for deconvolution. If the null hypothesis H_{01} is true, the p-values should be distributed uniformly on $[0, 1]$. However, the p-values across repetitions of a given are correlated. Next, we are going to model the correlation with z-values transformed from p-values under the null distribution.

4.2.3. Null distribution validation

Null distribution validation focuses on the healthy subjects in real bulk datasets where there are no true cell type specific differential expression for all genes. By randomly assigning pseudo-labels to the healthy subjects, we split them into two groups, whose gene expressions are of the same distribution. In this scenario, the null hypothesis H_{01} holds.

If the p-values derived from test statistics are accurate, we are expected to see the p-values follow the uniform distribution on $[0, 1]$.

In previous section, we introduced two models (equation (4.1) and (4.4)) and tested the null hypothesis by comparing them. Suppose there are L different random labelings and R

repetitions of partition-test procedure to get p-values and gene g receives $L \times R_g$ p-values. Let $p_{l,r}^{(g)}$ denote the p-value from lth label and rth repetition for g genes and $z_{l,r}^{(g)}$ denote the z-value transformed from p-value. Under the null hypothesis, the z-values should follow a standard normal distribution.

$$\begin{pmatrix} p_{1,1}^{(g)} & \cdots & p_{1,R_g}^{(g)} \\ \vdots & \ddots & \vdots \\ p_{L,1}^{(g)} & \cdots & p_{L,R_g}^{(g)} \end{pmatrix} \longrightarrow \begin{pmatrix} z_{1,1}^{(g)} & \cdots & z_{1,R_g}^{(g)} \\ \vdots & \ddots & \vdots \\ z_{L,1}^{(g)} & \cdots & z_{L,R_g}^{(g)} \end{pmatrix} \tag{4.5}$$

Under the null, the z-values for gene g and random labeling l follows a multivariate normal distribution. However, as we stated before, the z-values from each repetition are correlated with each other with correlation ρ_g. This is due to the correlated estimated cell type proportions across repetitions, which are generated by overlapped deconvolution genes across repetitions. The mean and variance of z-values are denoted as $\mu_g = 0$ and $\sigma_g^2 = 1$, respectively. The distribution for z-values $(z_{l,1}^{(g)}, \ldots, z_{l,R_g}^{(g)})^T$ are shown in equation (4.6) below.

$$\begin{pmatrix} z_{l,1}^{(g)} \\ \vdots \\ z_{l,R_g}^{(g)} \end{pmatrix} \sim N(\mu_g \mathbf{1}, \sigma_g^2 \begin{pmatrix} 1 & \cdots & \rho_g \\ & \ddots & \\ \rho_g & \cdots & 1 \end{pmatrix}). \tag{4.6}$$

Since we randomly assigning the labels for validation, the distribution of z-values across labelings are identical and independent.

For labeling l, The average across repetitions, denoted as $\bar{z}_l^{(g)}$, follows a normal distribution:

$$\bar{z}_l^{(g)} \sim N(\mu_g, \frac{1 + (R_g - 1)\rho_g}{R_g}\sigma_g^2), \quad \text{i.i.d.} \tag{4.7}$$

and the variance across labeling is

$$\text{Var.}l = \text{Var}[\bar{z}_l^{(g)}] = \frac{1 + (R_g - 1)\rho_g}{R_g}\sigma_g^2. \tag{4.8}$$

The empirical variance across repetition $z_{l,1}^{(g)}, \ldots, z_{l,R_g}^{(g)}$ is

$$\text{Var.}r = E[\frac{1}{R_g}\sum_{r=1}^{R_g}(z_{l,r}^{(g)} - \bar{z}_l^{(g)})^2] = \frac{R_g - 1}{R_g}(1 - \rho_g)\sigma_g^2. \tag{4.9}$$

We can identify ρ_g and σ_g^2 from cross labeling variance and cross repetition variance.

$$\hat{\sigma_g^2} = \text{Var.}r + \text{Var.}l \tag{4.10}$$

$$\hat{\rho}_g = 1 - \frac{R_g \cdot \text{Var.}r}{(R_g - 1)\hat{\sigma_g^2}} \tag{4.11}$$

The mean of z-values, μ_g, can be estimated from null distribution validation by the average across $\bar{z}_l^{(g)}$, denoted as $\bar{\bar{z}}^{(g)}$.

4.2.4. Differential expressed genes selection

The analysis with real labels only provides R_g p-values for gene g. If gene g is not differential expression in all cell types, the distribution is the same as (4.6) with $\mu_g = 0$ and $\sigma_g = 1$. The test statistics are

$$\bar{z}^{(g)} = \frac{1}{R_g}\sum_{r=1}^{R_g} z_r^{(g)} \sim N(0, \frac{1 + (R_g - 1)\rho_g}{R_g}) \tag{4.12}$$

and

$$S_g = \frac{1}{R_g}\sum_{r=1}^{R_g}(z_r - \bar{z}^{(g)})^2, \quad E[S_g] = \frac{R_g - 1}{R_g}(1 - \rho_g). \tag{4.13}$$

For gene g, we test the differential expression in 2 steps. First, we take genes with $\bar{z}^{(g)}$ less than 0.05 quantile for standard normal distribution as DE genes. This is a conservative test because $\text{Var}[\bar{z}^{(g)}] = \frac{1+(R_g-1)\rho_g}{R_g} < 1$ when $\rho_g \in [0, 1]$. Then we focused on those genes that do not passed the conservative test in the first step. We adjusted their variance by estimating ρ_g via empirical Bayes. Since $0 \leq \rho_g \leq 1$, it is natural that we use a beta prior

with parameter a and b for ρ_g:

$$\begin{aligned} p(\rho|z_1^{(g)},\ldots,z_{R_g}^{(g)}) &\propto p(z_1^{(g)},\ldots,z_{R_g}^{(g)}|\rho_g)\cdot p(\rho_g|a,b) \\ &= \frac{\exp(-\mathbf{z}_g\Sigma_g^{-1}\mathbf{z}_g/2)}{\sqrt{(2\pi)^{R_g}|\Sigma_g|}}\rho_g^{a-1}(1-\rho_g)^{b-1}, \end{aligned} \tag{4.14}$$

where

$$\mathbf{z}_g = \begin{pmatrix} z_1^{(g)} \\ z_2^{(g)} \\ \vdots \\ z_{R_g}^{(g)} \end{pmatrix} \sim N(\mathbf{0},\Sigma_g), \quad \Sigma_g = \begin{pmatrix} 1 & \rho_g & \cdots & \rho_g \\ \rho_g & 1 & \cdots & \rho_g \\ \vdots & \vdots & \ddots & \vdots \\ \rho_g & \rho_g & \cdots & 1 \end{pmatrix}_{R_g\times R_g}.$$

The log-likelihood of ρ_g given $\mathbf{z}_g$ is

$$\begin{aligned} \log(p(\rho_g|z_1^{(g)},z_2^{(g)},\ldots,z_{R_g}^{(g)})) =& \text{constant} - \frac{1}{2}\Big[\frac{\sum_{r=1}^{R_g}(z_r^{(g)})^2}{1-\rho_g} - \frac{\rho_g R_g^2\bar{z}_g^2}{(1-\rho_g)(1+(R_g-1)\rho_g)}\Big] \\ &- \frac{R_g-1}{2}\log(1-\rho_g) - \frac{1}{2}\log(1+(R_g-1)\rho_g) \\ &+ (a-1)\log(\rho_g) + (b-1)\log(1-\rho_g) \end{aligned} \tag{4.15}$$

The summary statistics for log-likelihood (4.15) are $S_g = \frac{1}{R_g}\sum_{r=1}^{R_g}(z_r^{(g)})^2$ and $\bar{z}_g = \frac{1}{R_g}\sum_{r=1}^{R_g} z_r^{(g)}$. We estimate ρ_g with EM algorithm by iterating between Expectation and Maximization step. The initial values for ρ_g is

$$\rho_g^{(0)} = 1 - \frac{R_g S_g}{R_g - 1}. \tag{4.16}$$

From iteration t to $t+1$, we update $\hat{\rho}_g^{(t)}$ to $\hat{\rho}_g^{(t+1)}$ as well as the beta distribution parameters $(a^{(t)}, b^{(t)})$.

E-step

$$\hat{\rho}_g^{(t+1)} = E[\rho|\mathbf{z}_g, a^{(t)}, b^{(t)}] = \frac{\int_0^1 p(\rho|\mathbf{z}_g, a^{(t)}, b^{(t)})\cdot\rho d\rho}{\int_0^1 p(\rho|\mathbf{z}_g, a^{(t)}, b^{(t)})d\rho}; \tag{4.17}$$

M-step

$$(a^{(t+1)}, b^{(t+1)}) = \text{ArgMax}_{a,b}\{(a-1)\log(\hat{\rho}_g^{(t+1)}) + (b-1)\log(1-\hat{\rho}_g^{(t+1)})\}. \quad (4.18)$$

With estimated $\hat{\rho}_g$ and the corresponding variance $\hat{S}_g = \frac{1+(R_g-1)\hat{\rho}_g}{R_g}$, we further select DE genes that $\bar{z}_g/\sqrt{\hat{S}_g}$ less than 0.05 quantile of standard normal distribution.

4.3. Results

4.3.1. Constrained linear regression on counts data

In section 4.2, we described the expected behavior of the p-values and corresponding z-values by comparing full model (4.1) and cell type model (4.4), when there is not differential expression between healthy and diseased status. Traditional ANOVA test is designed for comparing two linear models with homoscedasticity normal noises and is based on the χ^2 distribution of the sum-of-squared residuals. However, ANOVA test is not applicable for model (4.1) and model (4.4) with normalized read counts as X_{jg} because of the non-negativity constraints on the model parameters, β_{kg} and $\alpha_{kg} + \beta_{kg}$. Furthermore, the noise term is not homoscedastic normal. Hence, the squared residuals can not be approximated by a χ^2 distribution.

Even though, we still checked the residuals from two models and calculated p-values from residual sum of squares (RSS) under the null validation setting. The bulk RNA-seq data of Fadista et al. (2014) includes $N_d = 11$ type 2 diabetes(T2D) subjects and $N_h = 66$ healthy subjects. Their cell type proportions are estimated by MuSiC(Wang et al., 2019) with single cell reference from Segerstolpe et al. (2016). Summary of single cell data and bulk data are in Table 1. The deconvolution of Fadista et al. (2014) provides the estimated cell type proportions of 6 major cell types: alpha, beta, delta, gamma, acinar and ductal with single cell data from Segerstolpe et al. (2016).

To examine p-values distribution under the null hypothesis H_{01}, we randomly assigned

pseudo-labels to 66 healthy subjects, half labeled as healthy and half as diseased. Hence, there are no DE genes between healthy and diseased groups. P-values from null distribution validation are generated with the procedures in Section 4.2 and should follow uniform distribution. First we deconvolve Fadista et al. (2014) with single cell reference Xin et al. (2016) with a subset of the common genes (15659 genes) and repeated $R = 100$ times with different subsets of genes. The estimated proportions for all 77 subjects are boxplotted in Figure 10 with 80% of common genes selected (12527 genes) for deconvolution. Subjects are ordered by increasing HbA1c levels from bottom to top, among which 66 subjects are healthy and 11 subjects are diabetic. The estimated cell type proportions vary across repetitions and this is due to different input genes for deconvolution.

One might ask why we use 80% of the common genes for deconvolution. Using different number of genes for deconvolution will lead different variation of estimated proportions across repetitions. In general, Using more genes for deconvolution leads to more consistent estimated proportions across repetition. The estimated proportions directly influence the results of " reverse regression" of model (4.1) and (4.4) and we would lose power of detecting differential expression genes using highly biased proportions estimated with a small subset of genes. However, we can not afford of using too many genes for deconvolution while leave only a few genes for DE tests because numerous repetitions are needed to guarantee that all genes have multiple p-values. After tried different number of genes for deconvolution, we choose 80% of the common genes.

To save computation time, we filtered out lowly expressed genes by average and maximum counts per million (CPM) across subjects ($\bar{X}_{jg} > 0.25$ and $\max X_{jg} > 0.5$). For the those candidate genes, we repeated the randomly labeling $L = 100$ times and calculated the p-values from a faked F-statistics constructed by residual squared sum (RSS),

$$F_{K+1,N_h-K-1} = \frac{(\mathrm{RSS}_{ct} - \mathrm{RSS}_{full})/(K+1)}{\mathrm{RSS}_{ct}/(N_h - K - 1)},$$

where RSS_{full} and RSS_{ct} are the RSSs from full model (4.1) and cell type model (4.4),

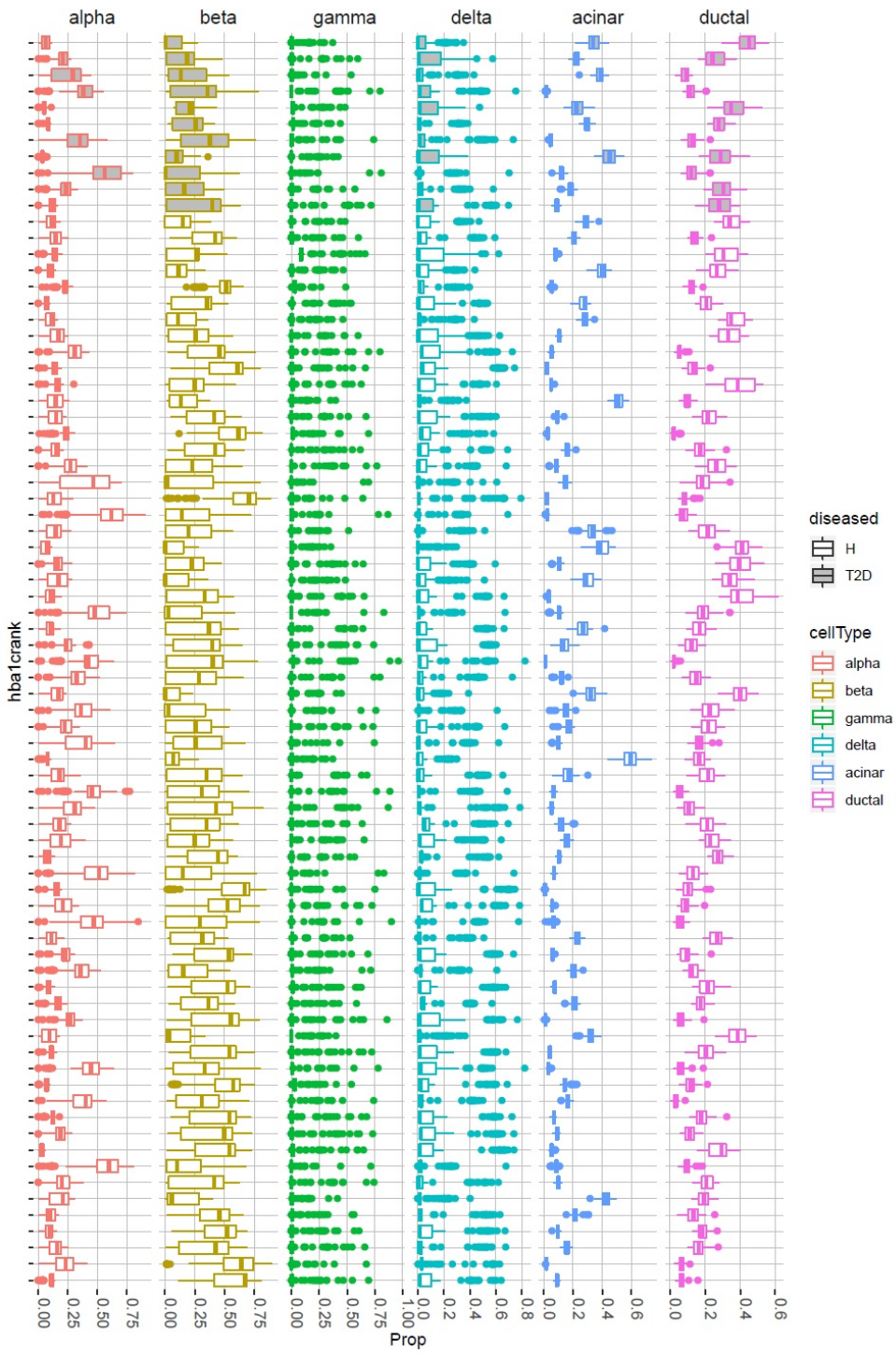


Figure 10: Boxplot of estimated cell type proportions for 100 repetitions of Fadista et al. (2014) dataset.
The bulk data is from Fadista et al. (2014) and the single cell reference for deconvolution is from Segerstolpe et al. (2016). Only 6 major cell types (color coded) are used for deconvolution and input with randomly selected 80% of common genes.

respectively and N_h is the number of healthy subjects. Even though F_{K+1,N_h-K-1} does not follow F distribution, the p-values from F-distribution can still guide for finding DE genes by constructing empirical distribution for p-values from the null distribution validation (see Section 4.2). Here, we used the test statistics from KolmogorovSmirnov test (KS test) (Stephens, 1974) to characterize the goodness-of-fit between observed average p-values and uniform distribution on $[0, 1]$. The KS test statistics is the maximum absolute difference between empirical distribution function and uniform distribution function, and small test statistics means the data is more likely come from uniform distribution. The test statistics calculated from the CPM of the healthy subjects from Fadista et al. (2014) are shown in Figure 11a. Moreover, the average p-values from worst cases and those p-values are concentrated near 1 (Figure 11b).

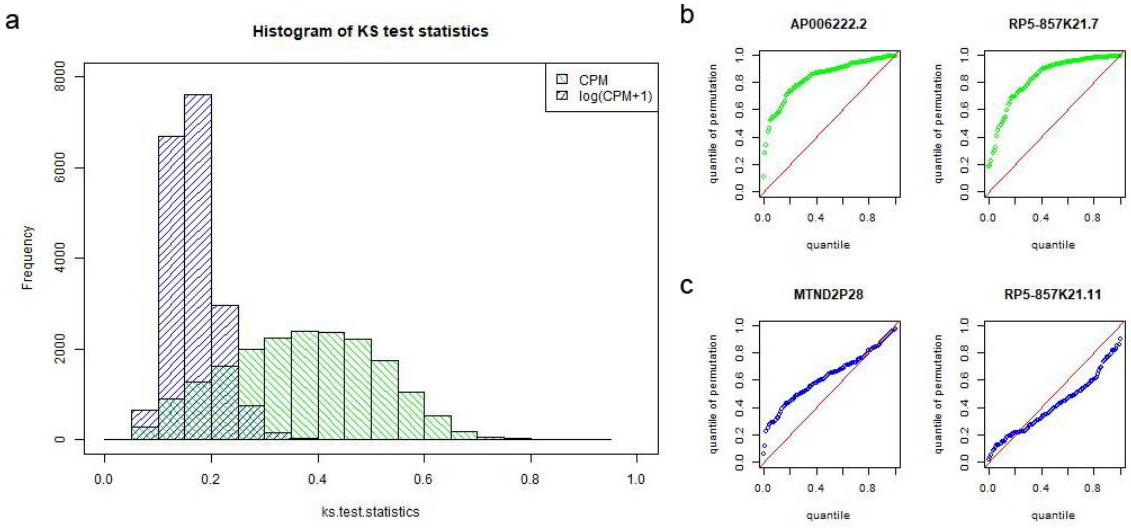


Figure 11: Null distribution validation with CPM and log-transformed CPM on Fadista et al. (2014) dataset.
a. Histograms of KS test statistics. The histogram of test statistics when bulk expressions are measured by CPM and log-transformed CPM are shown with green shadows and blue shadows, respectively. The KS test are performed on all genes. **b** and **c.** QQ plot of average-across-repetition p-values for all labels. **b.** Two genes with largest KS test statistics when bulk expressions are measured by CPM. **c.** Two genes with largest KS test statistics when bulk expression are measured by log-transformed CPM.

The faked F-statistics and the corresponding p-values are not reliable for testing cell type

specific differential expression. There are proposed reasons why the faked F-statistics is not reliable: (i) The noise term of model (4.1) and (4.4) do not follow normal distribution; (ii) Residuals from constrained linear regression does not follow normal distribution even when the noise term is from normal distribution; (iii) The degree of freedom for F-statistics is not $(K+1, N_h - K - 1)$ when there exists an β_{kg} or α_{kg} on boundary of the constraints.

4.3.2. Log-transformed CPM

Currently, there is no good test statistics for comparing constrained linear regressions. Furthermore, the noise is not normal homoscedastic. To eliminate the non-negative constraints on expressions, we transformed CPM to log scale. Let $X_{jg} = \log(\text{CPM} + 1)$ and the parameters in full model (4.1), β_{kg} and α_{kg}, no longer represent cell type specific expression and cell type specific differential expression, respectively, but represent cell type specific parameters and the cell type specific differences between two conditions. Although we loose part of the interpretability of our model, there are no constraints on parameters and it is reasonable to assume homoscedastic normal for the noise term.

When $X_{jg} = \log(\text{CPM} + 1)$, the regression of model (4.1) estimates the parameters for cell type proportions, the parameter for healthy/diseased conditions and the interaction between proportions and conditions. The null hypothesis with log-transformed CPM models is the same as H_{01} and tests the interaction between cell types and conditions. Now the null hypothesis focuses on the difference of cell type parameters between healthy and diseased conditions, rather than the cell type specific differential expression is zero or not. We test H_{01} and generate the p-values by the ANOVA test.

With the same analysis steps as we did in the previous section and the same datasets from Fadista et al. (2014), we checked the p-values under the null hypothesis, which are calculated when bulk expressions are measured with log-transformed CPM (Figure 11a and c). We first checked the goodness-of-it by comparing average-across-repetition p-values with uniform distribution on $[0, 1]$ by KS test, and the test statistics are plotted in Figure 11a.

Comparing with the test statistics from modeling CPM, modeling log-transformed CPM generates p-values that fit uniform distribution better. The worst case averaged-across-repetition p-values align better along the $x = y$ line in the QQ-plots (Figure 11c). This confirms that the p-values from log-transformed CPM models are convincing and proper for later analysis.

4.3.3. Benchmark validation

To examine the performance of MuSiC-DE, we construct a benchmark dataset with known cell type specific DE genes. The benchmark dataset is an artificial bulk dataset that is generated using single cell data from multiple subjects of both healthy and diseased status, similar to the procedure described in Appendix A.2.2. In this section, we use the single cell dataset from Xin et al. (2016) and there are 18 subjects (12 healthy and 6 T2D) with 4 cell types: alpha, beta, gamma and delta. We constructed the cell type specific artificial bulk data by summing up single cell read counts from cells of the same type for each subject. Then, for each cell type, cell-type specific DE genes are identified by comparing the artificial bulk data of that cell type between healthy and diseased conditions. We applied DESeq2 (Love et al., 2014) and selected genes with adjusted (by false discovery rate) p-values less than 0.05. This yielded the 27, 96, 102, 13 DE genes respectively for alpha, beta, gamma and delta cell types, which we will use as a benchmark for comparison.

Even though this is an intuitively logical way of developing a benchmark, we do not expect complete overlap with the DE genes under our model, due to the fundamental difference in modeling. DESeq2 selects genes by comparing the cell-type specific bulk data between two conditions while MuSiC-DE selects genes from tissue-level bulk data by comparing two models. When cell-type specific bulk is not observed, there are many limitations as to what is estimable. In our results, we will explore these limitations.

Next, we tried to mimic the scenario where only tissue-level bulk sequencing is performed for each subject, that is, the tissue-level artificial bulk data constructed by summing up

single cell expression of the same subject from Xin et al. (2016) (see Appendix A.2.2). Deconvolution, the estimation of the cell type proportions, uses single cell reference data from Segerstolpe et al. (2016). In each partition, 80% of the common genes, in total 16101 genes are used for deconvolution. Comparing the estimated cell type proportions with true proportions (Figure 12), the estimated proportions for alpha cells and beta cells more accurate. For gamma cells and delta cells, which are relatively rare in the artificial bulk data, it is more difficult to get accurate estimations.

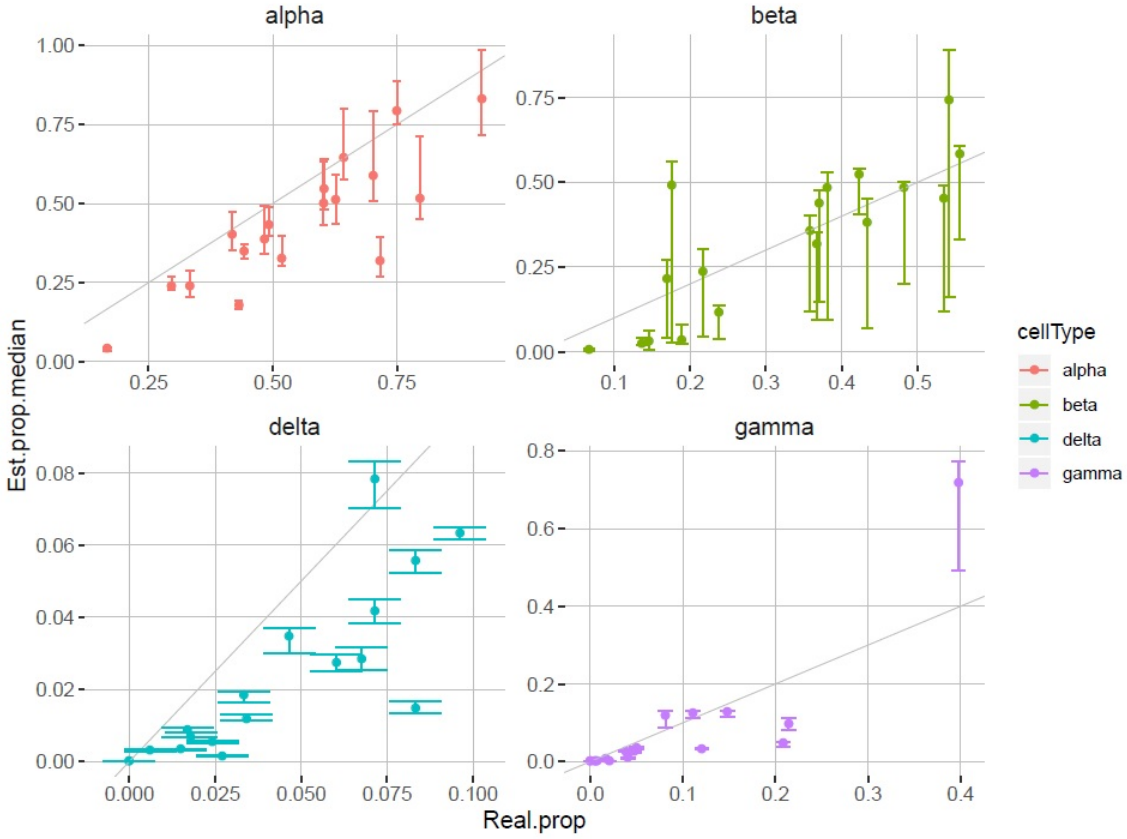


Figure 12: Scatter plot of true cell type proportions versus estimated cell type proportions from MuSiC over 100 repetitions.
The artificial bulk data is constructed from Xin et al. (2016) and the single cell reference is from Segerstolpe et al. (2016). For each repetition, we randomly selected 80% of the common genes between bulk data and single cell data. The estimate proportions for 4 cell types (color coded) are shown on the y-axis while the true proportions are on x-axis. The dot shows the median of estimated proportions and the errorbars indicate $0.3-0.7$ quantile of estimated proportions across 100 repetitions.

Within each repetition, we randomly selected 80% of the common genes for deconvolution and test DE for the remaining genes that have average CPM greater than 0.03. For the rest of the analysis, we transformed the p-values from the ANOVA test to z-values and investigated the z-values as derived in Section 4.2. First assume the z-values have mean $\mu_g = 0$ and variance $\sigma_g^2 = 1$ for all genes under the null, we reject H_{01} for genes with low z-values by the conservative test : $\bar{z}_g < \text{qnorm}(0.05)$. For genes that do not passed the conservative test, we adjusted their variance by estimating ρ_g from empirical Bayes using all of the remaining genes after first conservative test. We assigned a beta prior to $\rho_g \in [0, 1]$ and the log-likelihood of ρ_g given $\mathbf{z}_g$ is in equation (4.15). A total of 693 genes were selected from conservative test, and in the second round, a total of 181 genes were selected by the z-score computed using the adjusted variance. The null distribution validation of artificial bulk data are discussed in Appendix A.3.1.

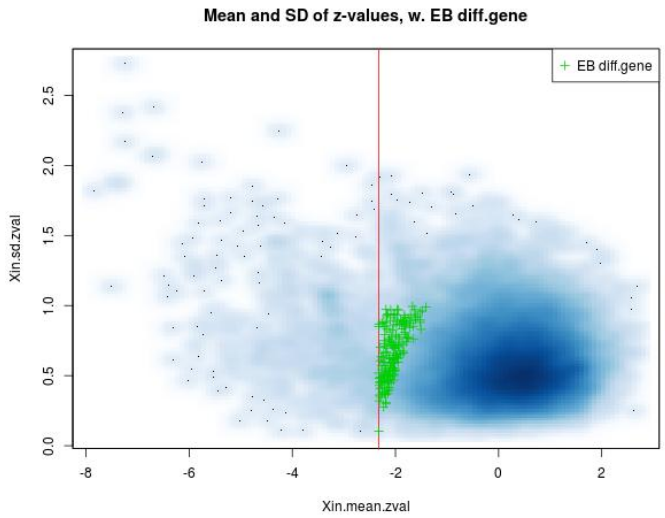


Figure 13: Smooth scatter plot of mean($\bar{z}$) versus variance($\sqrt{S}$) across genes. The red line is the cut-off for conservative test. The genes on the left of red line are the significant genes from conservative test. The genes that are significant after adjustment of their variance are denoted with green plus sign.

Table 3 shows the degree of overlap of MuSiC-DE with the benchmark DE genes favored with cell-type specific artificial bulk. However, MuSiC-DE can only select a small part of

the true cell type specific DE genes (Table 3), that is, MuSiC-DE tend to selects genes that DESeq2 does not select. Seyednasrollah et al. (2013) argues that DESeq2 is a conservative differential expression gene selection tool, which only selects part of the true DE genes. One possible explanation for MuSiC-DE selecting more significant genes than DESeq2 is the conservation of DESeq2. Another possible explanation for this phenomenon is that MuSiC-DE tests the difference between two models by residuals rather than tests the difference between expressions of healthy and diseased condition cell type by cell type. For example, if the expression of a gene is different between healthy and diseased conditions for all cell types, say the expression of healthy condition is higher than that of diseased condition. Suppose the differences are too little to be detected by DESeq2 with adjusted p-values less than 0.05. However, this gene has a high chance that will be selected by our method because our method accumulates the differences between healthy and diseased conditions from all cell types and tests the total difference by testing the difference of RSSs. To test differential expression of a specific cell type, we need hypothesis H_{02} for all cell types. Even though our

cell type	alpha	beta	delta	gamma	total(unique)
DESeq2 DE genes	27	96	102	13	235
MuSiC-DE & DESeq2	10	13	1	0	23
!MuSiC-DE& DESeq2	17	83	101	13	212
MuSiC-DE& !DESeq2	864	861	873	874	874
!MuSiC-DE & !DESeq2	19236	19170	19152	19240	19253

Table 3: Number of differential genes from cell type specific artificial bulk data.

method selects more genes than needed, we still can not detect true DE genes for delta and gamma cells. One might suspect that the inaccurate estimated proportions of delta and gamma cells lead to the lack of power in detecting true DE genes for those two cell types. We can investigate this, as for this artificial dataset we know the true proportions. Therefore, we use the true proportions in the full model 4.1 and cell type model 4.4 to compute a z-value for each gene. This z-value can be compared to the average z-value we computed across random partitions, where the true proportions are masked (Figure 12). From here, we learn two things: One, the z-values from estimated proportions do correlate with the

z-values from true proportions, but substantial error is introduced due to estimation of proportions. Second and most importantly, even with the true proportions, MuSiC-DE still missed most of the the DE genes detected DESeq2 for delta cells and gamma cells. The lack of power of detecting DE genes for rare cell type is due to reasons other than inaccurate estimated proportions.

Here we propose another explanation for the failure of detecting DE genes for rare cell types. In linear regression, if there is no variation for a covariate in the design matrix, we can not estimate the coefficient for that covariate because the covariate is collinear with the intercept. In model (4.1), the covariates for α_{kg} are $(p_{1k}t_1, \ldots, p_{Nk}t_N)^T$. For a rare cell type k, the covariates are close to 0 or equal to 0 and eliminating this covariates does not change the RSS of linear regression even if the estimated α_{kg} differs from 0. This limits the power of detecting the true DE genes for rare cell types.

Now, let's look at alpha and beta cells. Although MuSiC-DE detected some of DE genes DESeq2 suggests, there still exists a group of DE genes MuSiC-DE missed. The genes that DESeq2 selected for alpha and beta cell types are differential expressed and their p-values, computed DESeq2, are less 0.05 after adjusting for FDR. In general, among those DE genes, the gene are also chosen by our MuSiC-DE have lower p-values, provided by DESeq2, than those that are not selected (Figure 15). However, we still missed some significantly differential expressed gene with extremely low p-values from DESeq2. Some of the missed DE genes even have their average z-value greater than 0, which means the variance adjustment of z-values does not help correct the test statistics in this case. Explored the coefficients from the regression model (4.1) of the genes that are significant in DESeq2 but are missed by MuSiC-DE, we proposed an explanation for this phenomenon. This may due to the variation across subjects are largely explained by the rare cell type proportions. Adding extra covariates for differential expression does not make more contribution for explaining the variance of the bulk expression. In the space spanned by 4 cell type proportions, most of the bulk expression can be projected on the the direction of rare cell types. Additional

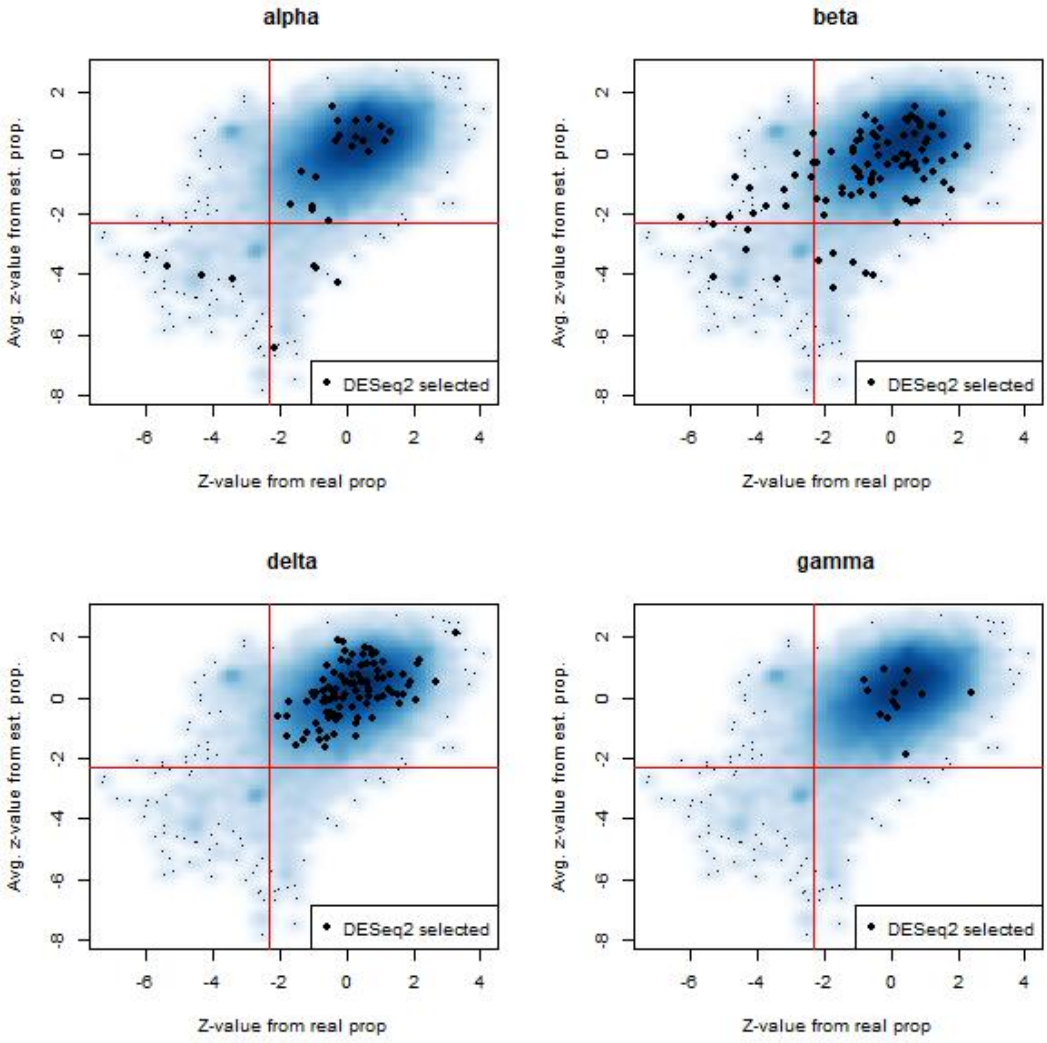


Figure 14: Smooth scatter plots of z-values from true proportions and average z-values from estimated proportions. The artificial bulk data constructed from Xin et al. (2016). The black line is the $x = y$ line and the red lines are the cut-off for DE. The small plots pointed out the DESeq2 selected genes for each cell type.

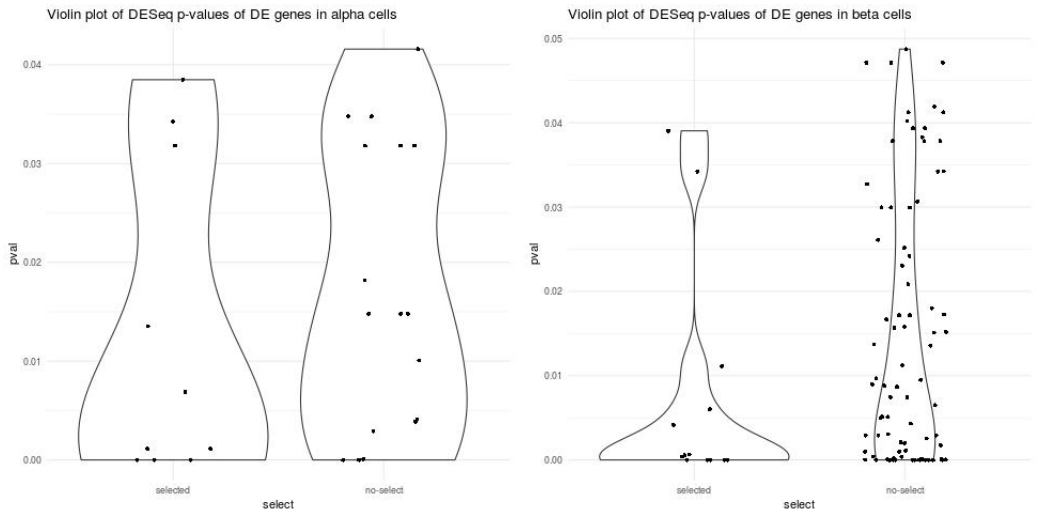


Figure 15: Violin plot of p-values from DESeq2 for alpha and beta cells

directions spanned by cell type specific differential expression can not more about the variation of bulk expression. In the linear regression context, the RSS with cell type model is very low and adding more covariates can not further reduce the RSS. Therefore, simply comparing the full model and cell type model is not enough to detect all true DE genes.

4.3.4. Results of Fadista et al. (2014) dataset

In this section, we analyze the dataset from Fadista et al. (2014) with MuSiC-DE and compared the significant genes selected from MuSiC-DE and the DE genes selected from bulk expression. We selected genes that have adjusted p-values less than 0.05 from DESeq2 as bulk DE genes. In total there are 713 genes selected.

The null distribution validation analysis using the expression of healthy subjects from Fadista et al. (2014) confirms that using log-transformed CPM as bulk expression fits the model assumption better and works better for the DE analysis (Section 4.3.2). We continued using log-transformed CPM to analyze the whole dataset, with both healthy and diseased subjects, from Fadista et al. (2014) and find DE genes between two conditions.

The cell type proportions were estimated with single cell dataset from Segerstolpe et al. (2016) (Figure 10) and the p-values are calculated by comparing the full model (4.1) and

the cell type model (4.4). We transformed p-values to z-values and used z-values to detect DE genes. After conservative test and variance adjustment by empirical Bayes with beta prior on the correlation between repetitions, we selected 4586 differential expressed genes (Figure 16). Compared with bulk DE genes selected by DESeq2 (Table 4), we focused on genes that only selected by MuSiC-DE or only bulk DE genes.

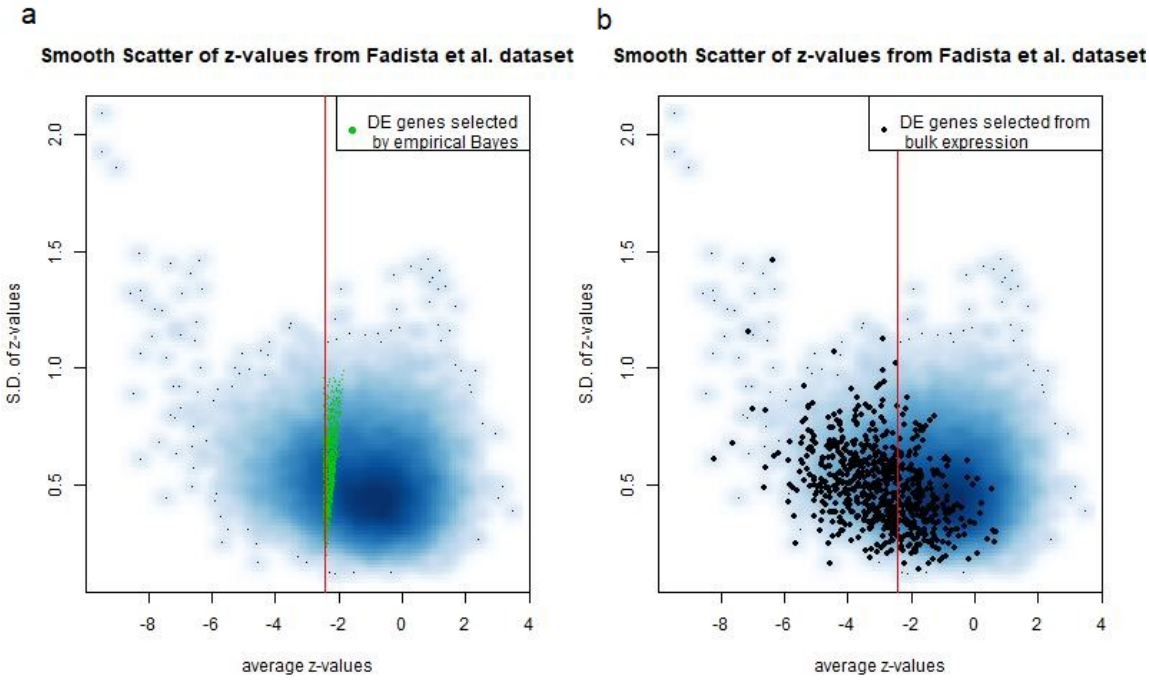


Figure 16: Smooth Scatter plot of z-values from the analysis of Fadista et al. (2014). The red line is the cut-off of conservative test.
a. The DE genes selected by MuSiC-DE includes the genes selected by conservative test, which are shown on the left of the red line, and the genes selected by empirical Bayes, which are shown with the green dots. **b.** The bulk DE genes are shown with black dots.

	Bulk DE	Bulk not DE	Total
MuSiC-DE selected	505	4081	4586
MuSiC-DE not selected	208	14094	14302
Total	713	18175	18888

Table 4: Number of genes that are selected as differentially expressed by MuSiC-DE of cell-type level and by DESeq2 of tissue level.

4.4. Discussion

Diseases progression involves the shift of cell type compositions as well as the change of cell type specific transcriptomic phenotype. The differential expression of bulk tissues between disease and healthy cohorts is a mixture of the two shifts. To investigate the cell type specific phenotype shifts from bulk tissues between cohorts, we need to separate the phenotype shifts from cell type proportions. Deconvolution methods provide a way to estimate cell type proportions. Some deconvolution methods require using marker genes, that is, the genes only expressed in a specific cell type. However, cell type specific expression changes can also be observed in marker genes and the deconvolution methods with markers may lead to biased proportions that treat the cell type specific expression changes as cell type proportion changes.

In this chapter, we proposed a method to study the cell type specific differential expression using bulk sequencing data on samples between healthy and diseased condition using estimated cell type proportions by reverse regression. Differential expression of each gene is tested multiple times with cell type proportions estimated by MuSiC, a deconvolution method without marker gene selection, excluding this gene from input. We compared two linear models, the full model and the cell type model, to calculated p-values. Compared to cell type model, the full model include the interaction term between diseased indicator and cell type proportions. We modeled the p-values to select cell type specific DE genes and called this method as MuSiC-DE. The p-values generated with null distribution using dataset from Fadista et al. (2014) suggests that we should use log-transformed CPM to measure bulk expression for detecting cell type specific DE genes. We also examined our method on benchmark dataset from Xin et al. (2016) by comparing the cell type specific DE genes selected from cell type specific artificial bulk data by DESeq2 and the genes selected by MuSiC-DE. The fundamental difference between two methods yields that the DE genes selected by our method will not full agree with the DE genes selected by DESeq2 and we investigated and explained the discrepancies between DE genes selected by two methods.

We also studied the performance of MuSiC-DE on datasets from Fadista et al. (2014).

The analysis of MuSiC-DE is not completed. In future study, we want to benchmark MuSiC-DE in a better way, rather than comparing MuSiC-DE selected genes with genes selected by DESeq2, which is not a golden standard of differential expression. We also would like to extend our method such that we are able to match DE genes to certain cell types. More datasets should be involved for the examination of MuSiC-DE. We will improve MuSiC-DE in our future research.

APPENDIX

A.1. Supplementary Information for Allele specific Mendelian randomization

A.1.1. Simulation results with extreme instrumental variable strength

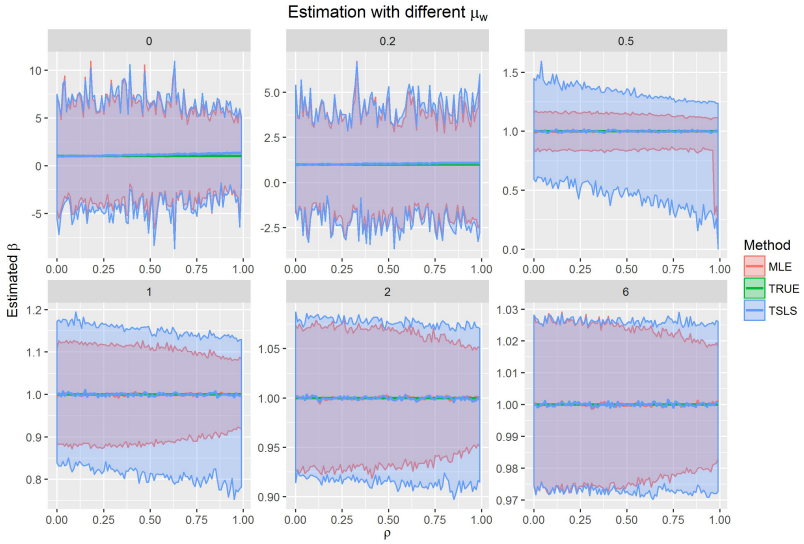


Figure 17: Simulation results when changing instrument strength by changing the mean of W, μ_w.

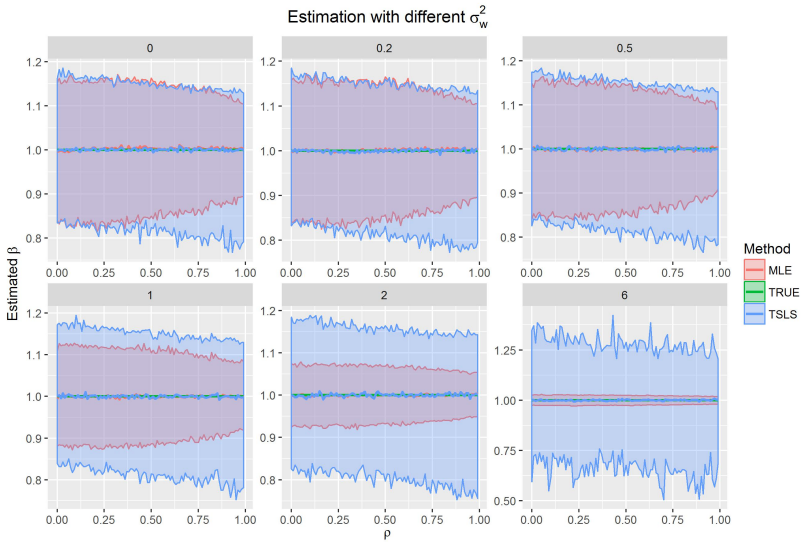


Figure 18: Simulation results when changing instrument strength by changing the variance of W, σ_w.

A.2. Supplementary Information for bulk tissue cell type proportion deconvolution with single cell reference

A.2.1. Linear regression to examine the relationship between estimated cell type proportion and HbA1c level

The linear regression results, mentioned in section 3.3, that examine the relationship between estimated cell type proportion and HbA1c level are shown Table 5 and 6. Table 5 contains the results with single cell data from Segerstolpe et al. (2016) as reference while Table 6 uses Baron et al. (2016) as reference. Full deconvolution results with Baron et al. (2016) are shown in Figure 19.

A.2.2. Construction of artificial bulk tissue RNA-seq data

We construct artificial bulk tissue RNA-seq data by summing up read counts across all cells from the same subject in the single-cell RNA-seq data. By way of construction, the cell type proportions of the artificial bulk data are equal to the observed cell type proportions in the single-cell data, and this allows us to compare estimated cell type proportions from various methods with the true proportions. Figure 20b shows that the artificial bulk tissue RNA-seq data have similar gene expression as the real bulk RNA-seq data generated from the same subjects.

A.2.3. Impact of vary cell type proportions of artificial bulk data in deconvolution

Figure 8b in the main text shows the deconvolution results from MuSiC, NNLS, BSEQ-sc and CIBERSORT, and these results indicate that the alpha cell proportion is over-estimated by all methods except for MuSiC. To evaluate the impact of different cell type proportions in the bulk data on deconvolution estimates, we generated additional artificial bulk data to show that MuSiC can still reliably estimate cell type proportions even when the true cell type proportions in the bulk data are very different from the cell type proportions in the single-cell reference. In this newly constructed benchmark data, the single-cell reference stays the

Cell type		MuSiC			BSEQ-sc		
		Estimate	Std.Error	P value	Estimates	Std.Error	P value
alpha	(Intercept)	0.380382	0.207754	0.07125	1.351464	0.240052	3.26E-07
	HbA1c	-0.00203	0.027737	0.941834	-0.07377	0.032049	0.024249
	Age	-0.00097	0.001935	0.617836	0.002753	0.002236	0.222198
	BMI	-0.00167	0.007945	0.834127	-0.01711	0.00918	0.066449
	Gender	0.033135	0.042881	0.442221	-0.00638	0.049548	0.897869
beta	(Intercept)	0.877022	0.190276	1.71E-05	0.065847	0.046433	0.16047
	HbA1c	-0.0614	0.025403	0.01819	-0.00295	0.006199	0.635957
	Age	0.002639	0.001772	0.140873	0.000576	0.000433	0.187339
	BMI	-0.01362	0.007276	0.065293	-0.00162	0.001776	0.365258
	Gender	-0.07987	0.039274	0.04566	-0.00541	0.009584	0.574159
gamma	(Intercept)	0.008556	0.010504	0.417988	0.102201	0.024366	7.69E-05
	HbA1c	0.001047	0.001402	0.457785	-0.00278	0.003253	0.396334
	Age	9.21E-05	9.78E-05	0.349431	-0.00013	0.000227	0.570225
	BMI	-0.00057	0.000402	0.160731	-0.00207	0.000932	0.029738
	Gender	-0.00165	0.002168	0.450416	-0.00092	0.005029	0.855252
delta	(Intercept)	0.057678	0.010592	6.81E-07	0.015539	0.018715	0.409122
	HbA1c	-0.00106	0.001414	0.455427	0.002017	0.002499	0.422131
	Age	-0.00016	9.87E-05	0.12039	9.99E-05	0.000174	0.568316
	BMI	-0.0011	0.000405	0.008142	-0.00103	0.000716	0.154263
	Gender	0.000424	0.002186	0.846817	-0.00254	0.003863	0.512616
acinar	(Intercept)	-0.10619	0.131102	0.420638	-0.14553	0.052092	0.006672
	HbA1c	0.034967	0.017503	0.049519	0.019075	0.006955	0.007684
	Age	-0.00247	0.001221	0.046841	0.00066	0.000485	0.178153
	BMI	0.00662	0.005013	0.190883	0.002008	0.001992	0.316847
	Gender	0.05332	0.02706	0.052632	-0.02338	0.010752	0.032985
ductal	(Intercept)	-0.21745	0.141008	0.127428	-0.38952	0.232841	0.098686
	HbA1c	0.028474	0.018826	0.134781	0.058397	0.031086	0.064353
	Age	0.000863	0.001313	0.513005	-0.00396	0.002169	0.072066
	BMI	0.010341	0.005392	0.059097	0.019814	0.008904	0.029191
	Gender	-0.00536	0.029105	0.854406	0.038631	0.048059	0.424144

Table 5: Linear regression to examine the relationship between estimated cell type proportions (Segerstolpe et al. (2016) as reference) and HbA1c levels. The fitted linear model is estimated cell type proportion $\sim$ HbA1c + Age + BMI + Gender. Significant results (p value < 0.05) are highlighted.

Cell type		MuSiC			BSEQ-sc		
		Estimate	Std.Error	P value	Estimates	Std.Error	P value
alpha	(Intercept)	1.000504	0.275906	0.000533	1.220529	0.187349	8.56E-09
	HbA1c	-0.0259	0.036835	0.48424	-0.06398	0.025012	0.012632
	Age	0.000234	0.00257	0.927855	0.001921	0.001745	0.274661
	BMI	-0.01137	0.010551	0.28475	-0.00681	0.007164	0.345275
	Gender	0.038364	0.056948	0.502676	-0.02104	0.038669	0.588048
beta	(Intercept)	0.315176	0.09427	0.001316	0.011001	0.016796	0.51455
	HbA1c	-0.02843	0.012586	0.026936	-3.70E-05	0.002242	0.986889
	Age	-0.00081	0.000878	0.361952	0.000142	0.000156	0.366396
	BMI	-0.00158	0.003605	0.661813	-0.00044	0.000642	0.498345
	Gender	-0.00927	0.019458	0.635249	-0.00079	0.003467	0.819685
gamma	(Intercept)	-0.0172	0.055935	0.759333	0.040372	0.011566	0.000827
	HbA1c	0.001227	0.007468	0.869925	7.31E-05	0.001544	0.962362
	Age	0.00085	0.000521	0.107295	-8.47E-05	0.000108	0.434521
	BMI	-0.00042	0.002139	0.843112	-0.0011	0.000442	0.015394
	Gender	-0.00998	0.011545	0.390355	-0.00048	0.002387	0.842519
delta	(Intercept)	0.043785	0.009622	2.12E-05	0.012347	0.016882	0.466922
	HbA1c	-0.00121	0.001285	0.349663	0.002763	0.002254	0.224153
	Age	-8.79E-05	8.96E-05	0.330262	5.00E-05	0.000157	0.751577
	BMI	-0.00093	0.000368	0.013618	-0.00101	0.000646	0.1226
	Gender	-0.00063	0.001986	0.753674	-0.00098	0.003484	0.780352
acinar	(Intercept)	0.002232	0.042169	0.957925	-0.23299	0.083467	0.006714
	HbA1c	0.013032	0.00563	0.023475	0.034902	0.011143	0.00251
	Age	-0.00062	0.000393	0.119068	-0.00015	0.000777	0.848086
	BMI	-0.0008	0.001613	0.621478	0.006564	0.003192	0.043362
	Gender	0.013342	0.008704	0.129687	-0.01866	0.017228	0.282488
ductal	(Intercept)	-0.3445	0.218745	0.119669	-0.05126	0.14	0.715354
	HbA1c	0.041276	0.029204	0.161852	0.026281	0.018691	0.164004
	Age	0.00043	0.002038	0.833485	-0.00188	0.001304	0.153951
	BMI	0.015109	0.008365	0.075051	0.002786	0.005354	0.604398
	Gender	-0.03183	0.04515	0.483036	0.04194	0.028896	0.151016

Table 6: Linear regression to examine the relationship between estimated cell type proportions (Baron et al. (2016) as reference) and HbA1c levels. The fitted linear model is estimated cell type proportion $\sim$ HbA1c + Age + BMI + Gender. Significant results (p value < 0.05) are highlighted.

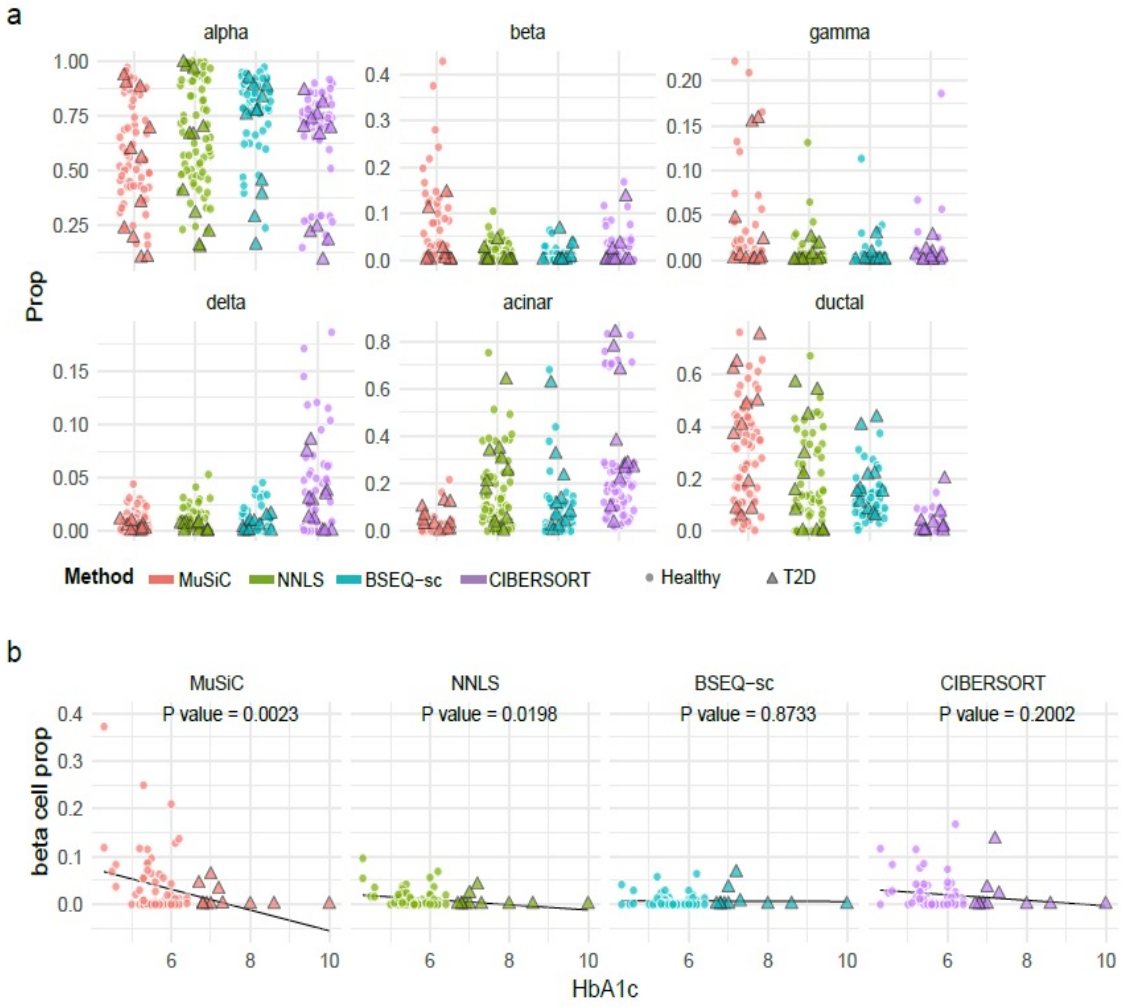
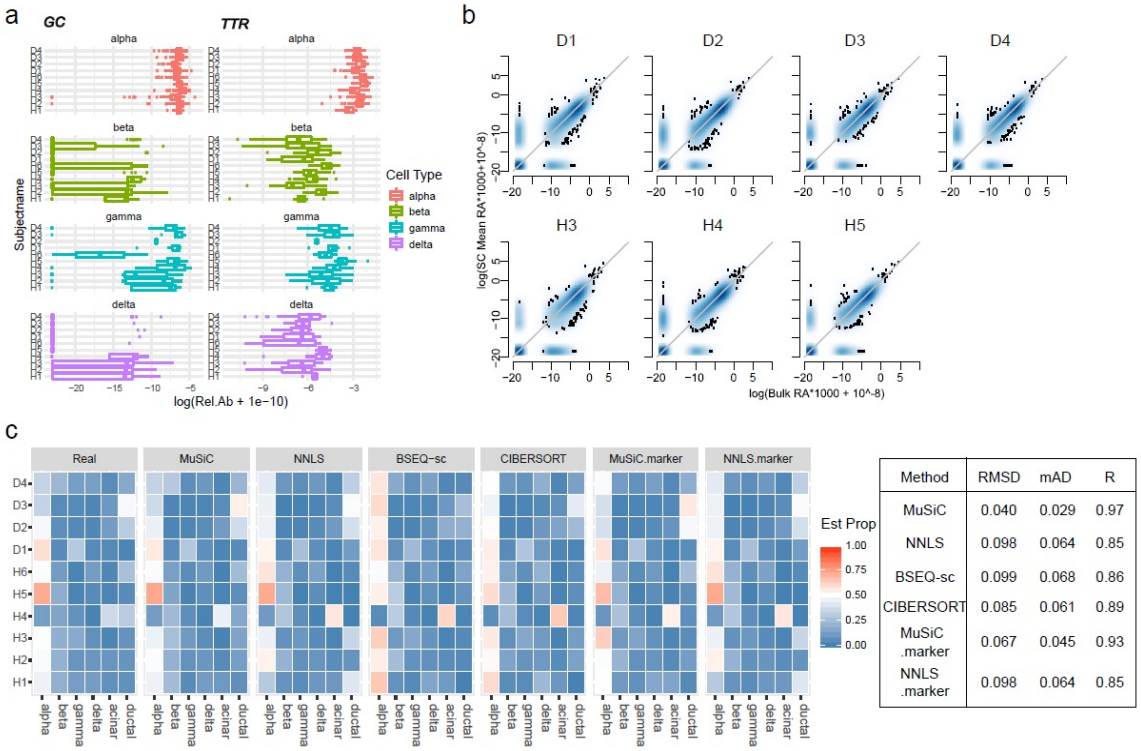


Figure 19: Estimated cell type proportions of the pancreatic islet bulk RNA-seq data in Fadista et al. (2014) with single cell reference from Baron et al. (2016). The analysis is similar to Figure 8c-d in the main text except that the single-cell reference are based on the three healthy subjects from Baron et al. (2016) and the MuSiC estimation was adjusted for protocol bias as described in section 3.2. **a.** Jitter plot of the estimated cell type proportions for Fadista et al. (2014) subjects, color-coded by deconvolution methods. 77 out of the 89 subjects from Fadista et al. (2014) that have recorded HbA1c levels are plotted. T2D subjects are denoted as triangles. **b.** HbA1c levels vs beta cell type proportions estimated by each of the four methods. The reported p-values are from single variable regression beta cell proportions HbA1c. Multivairiable regression results adjusting for age, BMI and gender are reported in Table 6.



Method	RMSD	mAD	R
MuSiC	0.040	0.029	0.97
NNLS	0.098	0.064	0.85
BSEQ-sc	0.099	0.068	0.86
CIBERSORT	0.085	0.061	0.89
MuSiC .marker	0.067	0.045	0.93
NNLS .marker	0.098	0.064	0.85

Figure 20: Exploratory analysis of single-cell RNA-seq data from Segerstolpe et al. (2016). **a.** Example of cross-subject and cross-cell variation in cell type specific gene expression. The boxplot contains 4 cell types: alpha, beta, gamma, and delta cells from Segerstolpe et al. (2016) single-cell RNA-seq data. The x-axis is the log transformed average relative abundance across cells from the same cell type, and the y-axis is the subject label. The relative abundance of gene *GC* is widely spread across the x-axis while the relative abundance of gene *TTR* is more concentrated across subjects. We consider gene *GC* as non-informative and *TTR* as informative. **b.** Comparison of log transformed relative abundance levels between real bulk tissue RNA-seq data and artificially constructed bulk RNA-seq data for the same subject. Single-cell and bulk tissue RNA-seq data are both from Segerstolpe et al. (2016). Each dot represents a gene and the gray line is $x = y$. **c.** Heatmap of true and estimated cell type proportions. In addition to the four methods described in the main text, we also evaluated the estimates given by MuSiC and NNLS when using only the marker genes used in BSEQ-sc.

same while we construct the artificial bulk data from Xin et al. (2016) by combining cells from 2 subjects with 75% alpha cells dropped. In this way, beta cells become the dominant cell type in the artificial bulk data, as expected for real bulk tissue. Figure 21c shows that only MuSiC recovers the true cell type composition, revealing that beta cells are the major cell type in the artificial bulk data, whereas the other methods overestimate the alpha cell proportion, indicating that these methods are more likely to be influenced by the cell type proportions in the single-cell reference. This analysis also gives the likely explanation for why, in the Fadista et al. (2014) data, all methods that rely on CIBERSORT marker genes grossly overestimate alpha cell proportion.

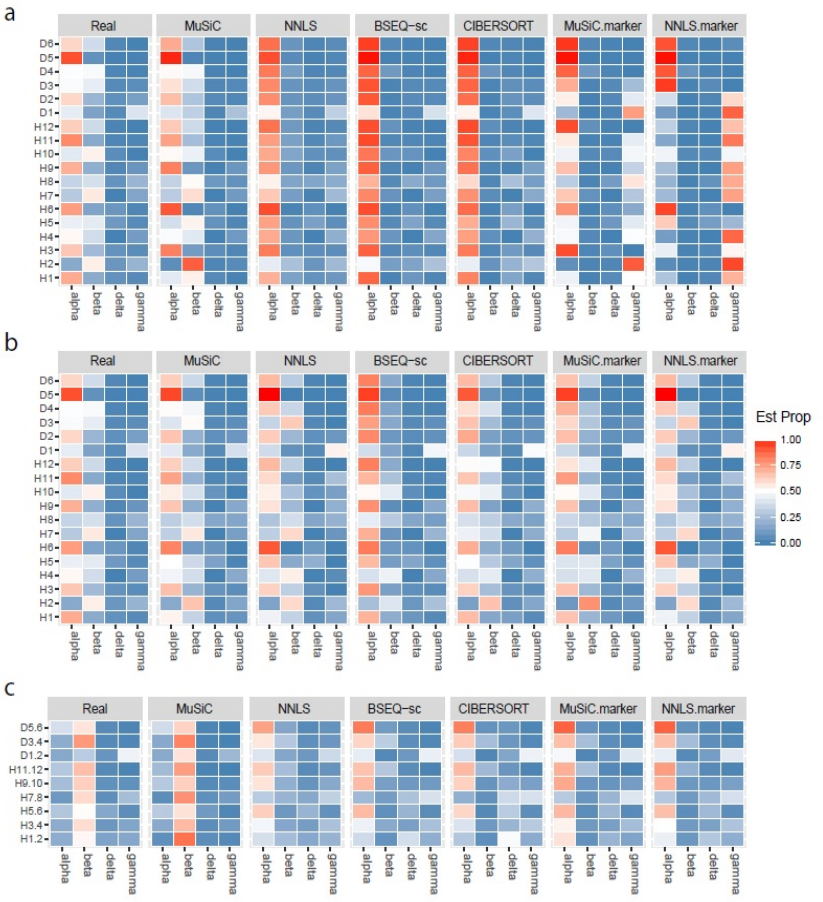
A.2.4. Impact of missing cell types in single-cell reference on deconvolution

One of the limitations of single-cell RNA-seq is cell loss during cell dissociation. This not only biases cell type proportions, but also leads to failure of detecting certain cell types, especially those rare cell types. In practice, the single-cell reference dataset might be incomplete, and not every cell type present in the bulk data is included in the single-cell reference. Since the deconvolution methods rely on observed cell types in the single-cell reference, it is important to evaluate whether cell type proportions can be reliably estimated when some cell types are missing in the single-cell reference.

We evaluate MuSiC, NNLS, BSEQ-sc and CIBERSORT with missing cell types (Figure 22, Table 7). The artificial bulk data consist of 6 cell types while the single-cell reference only consists of 5 cell types. The evaluation shows that when major cell types are missed, none of the methods can give accurate estimates. However, the cell type proportions are estimated accurately by MuSiC when the missing cell type is not the dominant cell type in the bulk tissue.

A.2.5. Tolerance of bias in single-cell relative abundance on deconvolution

The protocol discrepancies between bulk and single-cell datasets may lead to estimation bias. To evaluate the degree of bias tolerance relative to the biological signal, we manually



Segerstolpe et al. as reference

Method	RMSD	mAD	R
MuSiC	0.10	0.06	0.94
NNLS	0.17	0.12	0.81
BSEQ-sc	0.22	0.15	0.79
CIBERSORT	0.21	0.15	0.76
MuSiC .marker	0.25	0.18	0.60
NNLS .marker	0.34	0.26	0.30

Xin et al. as reference

Method	RMSD	mAD	R
MuSiC	0.05	0.033	0.98
NNLS	0.10	0.072	0.92
BSEQ-sc	0.13	0.088	0.87
CIBERSORT	0.07	0.053	0.95
MuSiC .marker	0.07	0.048	0.96
NNLS .marker	0.10	0.072	0.92

Method	RMSD	mAD	R
MuSiC	0.10	0.07	0.94
NNLS	0.25	0.19	0.23
BSEQ-sc	0.28	0.21	0.16
CIBERSORT	0.30	0.24	0.03
MuSiC .marker	0.31	0.23	0.10
NNLS .marker	0.31	0.23	0.08

Figure 21: Heatmaps of true and estimated cell type proportions of artificial bulk data constructed using single-cell RNA-seq data from Xin et al. (2016). **a.** Deconvolution results when the single-cell reference is from the 6 healthy subjects of Segerstolpe et al. (2016) with leave-one-out, i.e., for each subject under deconvolution, only single-cell data from the remaining 5 subjects were used as single-cell reference. **b.** Deconvolution results when the single-cell reference is from the 12 healthy subjects of Xin et al. (2016) with leave-one-out, i.e., for each subject under deconvolution, only single-cell data from the remaining 11 subjects were used as single-cell reference. **c.**The cell type proportions for the artificial bulk data are manually adjusted so that beta cells are the dominant cell type, as expected in real bulk tissue. Alpha cells dominate in the scRNA-seq data due to dissociation and capture bias. Thus, this analysis mirrors the real data analysis scenario where cell type proportions differ substantially between scRNA-seq reference and bulk tissue. In more detail, we combined cells from two subjects as one artificial bulk tissue RNA-seq dataset, for example, H1.2 combined cells from subject H1 and H2. Then we dropped 75% of the alpha cells at random. The single-cell reference is from the 6 healthy subjects of Segerstolpe et al. (2016). Here, all methods that rely on pre-selected marker genes from CIBERSORT are heavily biased by the cell type proportions in the single cell reference, and miss the true cell type proportions in the bulk tissue data. In comparison, MuSiC is able to adjust to the difference between scRNA-seq reference and bulk data.

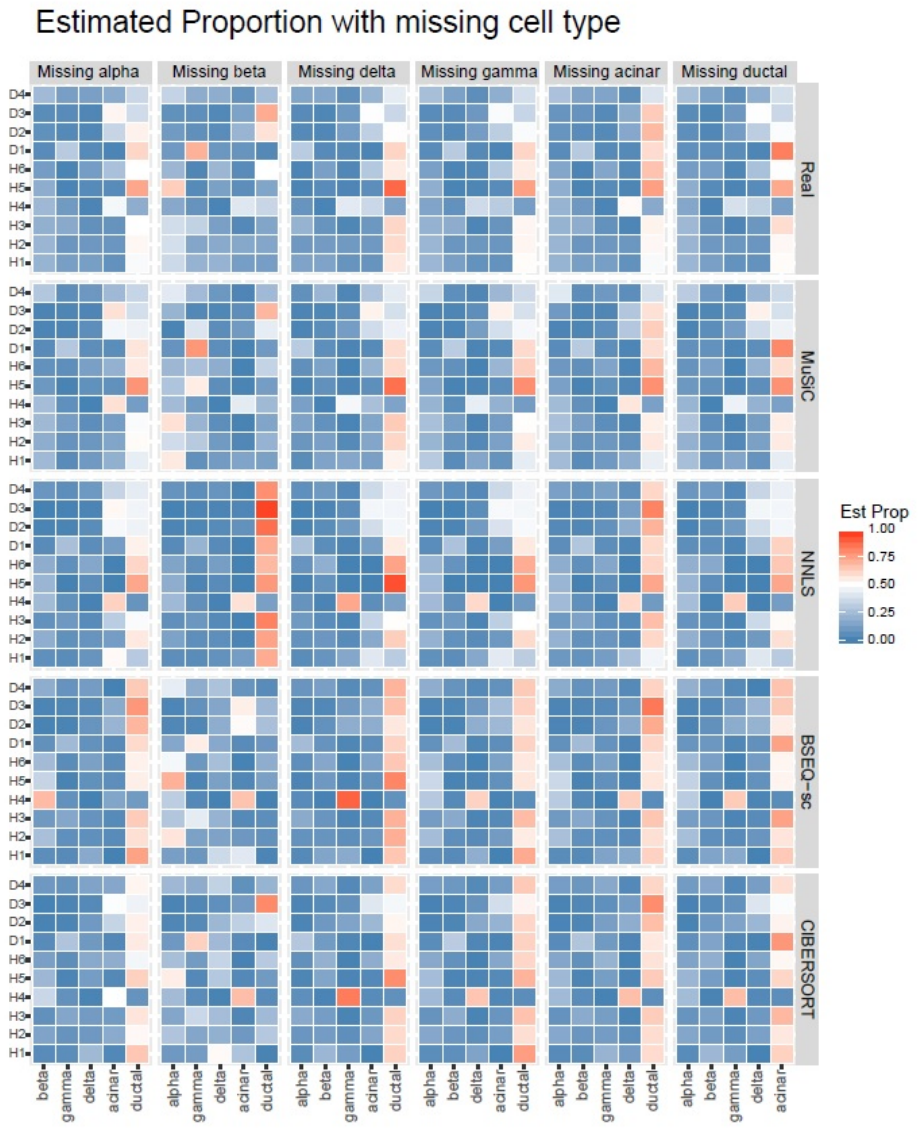


Figure 22: Heatmaps of true and estimated cell type proportions with missing cell types in single-cell reference.

The artificial bulk data and the single-cell reference are both from Segerstolpe et al. (2016). We constrained our analysis to the 6 major cell types: alpha, beta, delta, gamma, acinar and ductal cells. The artificial bulk data is constructed by summing read counts from the 6 major cell types while the single-cell reference contains only 5 cell types (the column header shows the cell type that is missing in the single-cell reference). The x-axis labels cell types used in the single-cell reference and the y-axis shows the subject label. The top panel shows the true composition, while panels below it show the results from each method. See Table 7 for detailed evaluation results.

alpha(0.447)	**RMSD**	**mAD**	**R**	**beta**(0.137)	**RMSD**	**mAD**	**R**
MuSiC	0.13	0.09	0.72	MuSiC	0.04	0.03	0.98
NNLS	0.27	0.18	0.42	NNLS	0.12	0.08	0.86
BSEQ-sc	0.17	0.12	0.58	BSEQ-sc	0.12	0.08	0.87
CIBERSORT	0.12	0.09	0.77	CIBERSORT	0.09	0.06	0.91
delta(0.092)	**RMSD**	**mAD**	**R**	**gamma**(0.062)	**RMSD**	**mAD**	**R**
MuSiC	0.04	0.03	0.98	MuSiC	0.05	0.038	0.97
NNLS	0.12	0.08	0.82	NNLS	0.12	0.081	0.84
BSEQ-sc	0.12	0.08	0.85	BSEQ-sc	0.12	0.083	0.86
CIBERSORT	0.10	0.07	0.90	CIBERSORT	0.10	0.070	0.90
acinar(0.092)	**RMSD**	**mAD**	**R**	**ductal**(0.062)	**RMSD**	**mAD**	**R**
MuSiC	0.05	0.04	0.97	MuSiC	0.050	0.037	0.97
NNLS	0.11	0.07	0.85	NNLS	0.067	0.046	0.96
BSEQ-sc	0.14	0.10	0.79	BSEQ-sc	0.084	0.064	0.93
CIBERSORT	0.07	0.05	0.93	CIBERSORT	0.076	0.058	0.94

Table 7: Evaluation of deconvolution methods when there are missing cell types in the single-cell reference. The missing cell type is shown in bold and the proportions in the bulk tissue data are shown in parentheses.

introduce noise to cross-subject average of the single-cell obtained relative abundance θ_g^k. Because of the constraint that $\sum_{g=1}^{G} \theta_g^k = 1$, we generate biased relative abundance by Dirichlet distribution, denoted by $\theta_g^{k\prime}$. Consider one cell type only. For simplicity, we drop the superscript k for cell type. We assume the relative abundances of G genes follow a Dirichlet distribution,

$$(\theta_1', \ldots, \theta_G') \sim \text{Dirichlet}(t \times (\theta_1, \ldots, \theta_G)), \tag{A.1}$$

where t is a scaling factor. The mean and variance of θ_g' are θ_g and $\frac{\theta_g(1-\theta_g)}{t+1}$, respectively. By setting $t = 999, 1332, 1999$ and 3999, corresponding to $\dfrac{\text{Var}[\theta_g']}{E^2[\theta_g']} \approx (\theta_g(1+t))^{-1} \geq 2, 1.5, 1$ and 0.5, we simulated 100 times the cross-subject average of relative abundance of 6 major cell types from Segerstolpe et al. (2016). We deconvolved the artificial bulk data constructed by Xin et al. (2016) (Figure 23) and MuSiC provides accurate cell type proportions even with biased relative abundance as input.

A.2.6. Robustness to single-cell dropout noise on deconvolution

Single-cell RNA-seq data are prone to gene dropout and the dropout rates can differ across datasets. To evaluate the robustness of MuSiC, NNLS, BSEQ-sc and CIBERSORT to dropout in single-cell data, we constructed artificial bulk data from the original scRNA-seq data and deconvolve it by single-cell data with additional dropout noise. Following Jia et al. (2017), the dropout rate π_{jgc} is generated by

$$\pi_{jgc} = \frac{1}{1 + k \exp\left(k \ln X_{jgc}\right)}, \tag{A.2}$$

where X_{jgc} is the observed read counts, k is the dropout rate parameters. The simulated read count $X_{jgc}{}'$ follows distribution such that

$$P(X_{jgc}{}' = X_{jgc}) = \pi_{jgc}, P(X_{jgc}{}' = 0) = 1 - \pi_{jgc}. \tag{A.3}$$

We evaluated four different dropout rates with $k = 1, 0.5, 0.2$ and 0.1 (Figure 23). In general, adding more dropout noise leads to lower MuSiC estimation accuracy. Compared with other methods, MuSiC consistently performs better in the presence of dropout noise.

A.2.7. Convergence of MuSiC with different starting points

MuSiC estimates cell type proportions by weighted non-negative least square (W-NNLS), which might be sensitive to choice of starting values. To examine the convergence property of MuSiC, we re-analyzed the data in Figure 8 to show convergence with different starting points. The artificial bulk data is constructed by Xin et al. (2016) while the single-cell reference consists of 6 healthy subjects from Segerstolpe et al. (2016). The cell type proportions of four cell types: alpha, beta, delta and gamma are estimated using MuSiC with different starting points are shown in Table 8. W-NNLS converges to the same value regardless of the starting points (Figure 24).

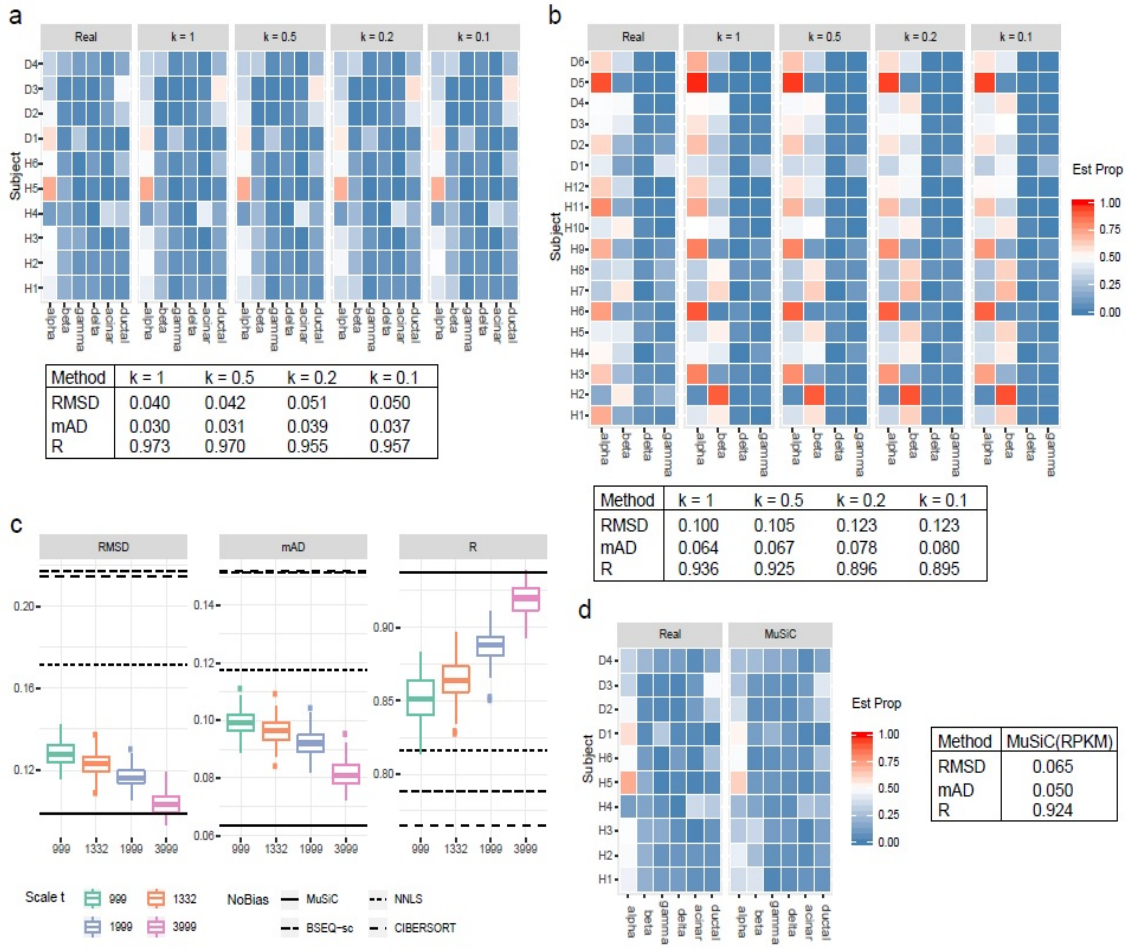


Figure 23: Benchmark evaluation of robustness of MuSiC.
a. and **b.** evaluate the impact of different dropout rate in scRNA-seq (Section A.2.6). **a.** and **b.** show heatmaps of MuSiC estimated cell type proportions. The single-cell reference is based on six healthy subjects from Segerstolpe et al. (2016) with different dropout rates. The artificial bulk data of **a.** is constructed by Segerstolpe et al. (2016) while **b.** is constructed by Xin et al. (2016) **c.** Evaluation of the impact of biased relative abundance 84 in the single-cell reference (Section A.2.5). Boxplot shows three evaluation metrics from 100 simulations of MuSiC estimated cell type proportions with biased relative abundance, color-coded by scale parameter of Dirichlet distribution. The horizontal lines show the evaluation metrics of four methods without bias in the single-cell reference. **d.** Heatmap of MuSiC estimated cell type proportions with RPKM as the input. The artificial bulk data and single-cell reference are both from Segerstolpe et al. The estimation follows leave-out-one rule. We utilized the average library size ratio of the six healthy subjects from Segerstolpe et al. (2016) as the ratio of cell size.

Cell type	EQ	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8
alpha	0.25	0.4	0.2	0.2	0.2	0.7	0.1	0.1	0.1
beta	0.25	0.2	0.4	0.2	0.2	0.1	0.7	0.1	0.1
delta	0.25	0.2	0.2	0.4	0.2	0.1	0.1	0.7	0.1
gamma	0.25	0.2	0.2	0.2	0.4	0.1	0.1	0.1	0.7

Table 8: Starting points for convergence analysis

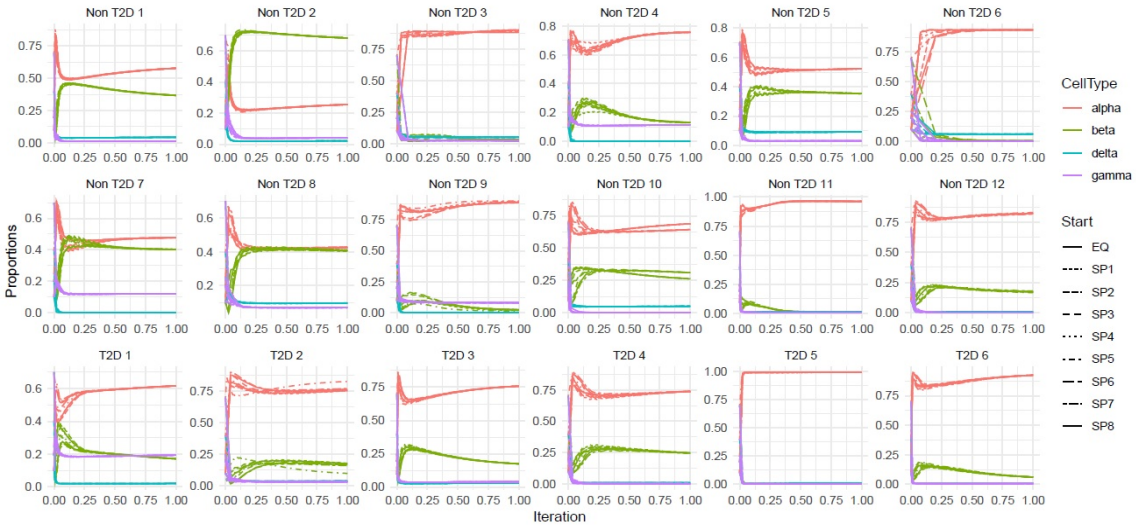


Figure 24: Convergence of MuSiC with different starting points.
The evaluation is performed on artificial bulk data, constructed by single-cell data from Xin et al. (2016) while the single -cell reference is from Segerstolpe et al. (2016). We evaluate the convergence of MuSiC with nine different starting points of four cell types in Table 8. The iteration numbers are normalized between 0 and 1 for comparison. We plotted the normalized iteration against estimated proportions for each subjects in Xin et al. (2016) colored by cell types. From different starting points, estimated cell types converged to the same proportions.

A.2.8. Additional information of mouse kidney data

The benchmark evaluation procedure is the same as described in section3.3.

Method	MuSiC	NNLS	BSEQ-sc	CIBERSORT
RMSD	0.030	0.115	0.033	0.031
mAD	0.018	0.061	0.018	0.019
R	0.981	0.711	0.977	0.980

Figure 25: Benchmark evaluation using mouse kidney single-cell RNA-seq data from Park et al. (2018). The artificial bulk RNA-seq data is constructed by summing read counts across cells in all 16 cell types while the single-cell reference only consists of 13 cell types. The other 3 cell types were discarded in the single-cell reference because they are too rare. **a.** Heatmap of estimated cell type proportions and evaluation results. **b.** Scatter plot of real cell type proportions versus estimated cell type proportions.

Estimated cell type proportions and correlation of the estimated cell type proportions across samples for bulk RNA-seq data o rat renal tubule segments from Lee et al. (2015). Results of MuSiC are shown in Figure 9e.

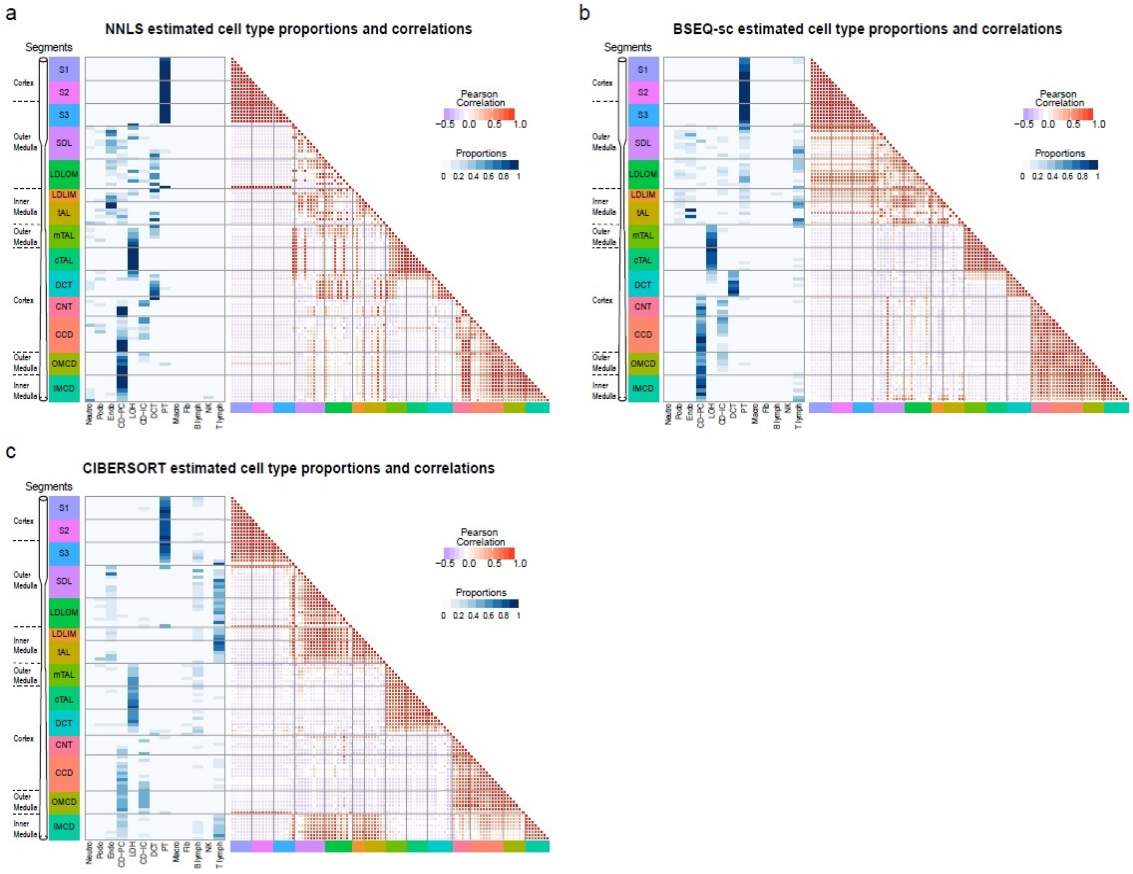


Figure 26: Estimated cell type proportions and correlation of the estimated cell type proportions across samples for bulk RNA-seq data of rat renal tubule segments (Lee et al., 2015). Park et al. (2018) mouse single-cell RNA-seq data are used as reference. **a.** NNLS. **b.** BSEQ-sc. c. CIBERSORT.

The full results with 13 cell types from bulk data deconvolution. Summary of cell types of

Figure 27: Estimated cell type proportions of the 13 cell types in three real mouse bulk RNA-seq datasets. **a.** Boxplot of estimated cell type proportions of 10 mice (4 APOL1 disease mice and 6 control mice) from Beckerman et al. (2017). **b.** Line plot of cell type proportion changes after FA induction (Craciun et al., 2016) at 6 time points. There are 3 replicates at each time point and the average proportions are plotted. N: normal. c. Line plot of cell type proportions of control (Sham operated mice), 2 days and 8 days after UUO (Arvaniti et al., 2016).

Park et al. (2018) single cell dataset.

Cell Type	**Abbr.**	**# Cell**	**% Cell**	**Cell Type**	**Abbr.**	**# Cell**	**% Cell**
Endothelial	Endo	1,001	2.29	Fibroblast	Fib	549	1.26
Podocyte	Podo	78	0.18	Macrophage	Macro	228	0.52
Proximal tubule	PT	26,482	60.54	Neutrophil	Neutro	74	0.17
Loop of Henle	LOH	1,581	3.61	B lymphocyte	B lymph	235	0.54
Distal convoluted tubule	DCT	8,544	19.53	T lymphocyte	T lymph	1,308	2.99
Collecting duct principal cell	CD-PC	870	1.99	Natural killer cell	NK	313	0.72
Collecting duct intercalated cell	CD-IC	1729	3.95	Novel cell type 1	Novel 1	601	1.37
Collecting duct transitional cell	CD-Trans	110	0.25	Novel cell type 2	Novel 2	42	0.10

Table 9: Summary of cell types of Park et al. (2018) single-cell dataset. Park et al. (2018) sequenced 57,979 cells from healthy mouse kidneys and identified 16 cell types. As suggested in Park et al. (2018), we limited our consideration to the 13 confidently characterized cell types and eliminated CD-Trans and 2 novel cell types in our deconvolution analyses.

Abbrv.	Full Segment Name	Abbrv.	Full Segment Name
S1	S1 proximal tubule	mTAL	medullary thick ascending limb
S2	S2 proximal tubule	cTAL	cortical thick ascending limb
S3	S3 proximal tubule	DCT	distal convoluted tubule
SDL	short descending limb	LDLIM	long descending limb, inner medulla
LDLOM	long descending limb	CCD	cortical collecting duct
CNT	connecting tubule	OMCD	outer medullary collecting duct
tAL	thin ascending limb	IMCD	inner medullary collecting duct

Table 10: Renal tubule segment names. Abbreviations and full names.

Rank	Segerstolpe	Xin	Fadista	Rank	Segerstolpe	Xin	Fadista
1	GCG	GCG	MALAT1	51	ITM2B	EIF4A2	RPS3A
2	TTR	MALAT1	EEF1A1	52	ENPP2	CTSD	RPL9
3	MALAT1	INS	TTR	53	ATP1A1	RBP4	SOD2
4	SERPINA1	TTR	FTH1	54	ANXA4	HNRNPH1	EIF4B
5	SPP1	FTL	GCG	55	HNRNPH1	BSG	HSPA8
6	B2M	PPP1CB	CPE	56	ALDOB	EEF2	PKM
7	FTH1	PCSK1N	GNAS	57	CD164	RPS3	SCG2
8	CHGA	CHGB	RPL4	58	HLA-A	PDK4	RPS24
9	PIGR	PSAP	APP	59	RIN2	SSR1	CD74
10	IAPP	CHGA	CTSD	60	ASAH1	SCD	SQSTM1
11	SST	EGR1	HSP90AA1	61	TMSB10	DNAJC3	TMBIM6
12	FTL	SRSF6	RPLP0	62	BSG	SAR1A	TXNRD1
13	CALM2	FTH1	RPL7A	63	CLDN4	GPX4	LCN2
14	CHGB	HSP90AB1	HSP90AB1	64	TMEM59	PLD3	RPL14
15	SERPINA3	SPINT2	HSP90B1	65	PPY	ATP6AP1	PDIA3
16	ACTG1	MAP1B	UBC	66	C10orf10	ANP32E	HDLBP
17	SCG5	RIN2	CANX	67	HSPA8	TBL1XR1	HNRNPK
18	ALDH1A1	GNAS	PAM	68	REG1B	GNB2L1	SCARB2
19	TM4SF4	SCG5	RPS6	69	P4HB	SLC22A17	RPL13A
20	REG3A	CSNK1A1	SERPINA3	70	LCN2	PAFAH1B2	LINC00657
21	GAPDH	PTEN	EIF4G2	71	PKM	RTN4	DSP
22	PPP1CB	TSPYL1	RPS4X	72	ATP6V0B	TMED4	SPINT2
23	ACTB	C6orf62	HSPA5	73	PSAP	CST3	REG1B
24	PRSS1	RPL3	ITGB1	74	LRRC75A-AS1	CD63	HNRNPC
25	RBP4	DPYSL2	IAPP	75	S100A11	TOB1	RPL15
26	GDF15	UBC	TPT1	76	MUC13	HLA-A	ENO1
27	COX8A	SCG2	RPL5	77	MAP1B	CLU	RPS11
28	ALDOA	ALDH1A1	SLC7A2	78	CD59	TTC3	GANAB
29	PDK4	PFKFB2	HNRNPA1	79	SLC30A8	RPS11	CDH1
30	RPL8	CPE	ANXA2	80	CPE	G6PC2	PEG10
31	H3F3B	C10orf10	RPL7	81	CLPS	GRN	CLDN4
32	IGFBP7	TMBIM6	RPS18	82	CTSD	SERPINA1	GSTP1
33	S100A6	CRYBA2	PCSK1	83	ATP1B1	SSR4	TUBA1A
34	EEF2	FTX	ATP1A1	84	OLFM4	RPS6	RPS27A
35	TIMP1	HSPA8	IDS	85	TAGLN2	OAZ1	PRPF8
36	CFL1	HSP90AA1	GDF15	86	SCGN	MARCKS	HSPB1
37	GRN	H3F3B	RPS3	87	SERPING1	RPL15	RPS8
38	SPINT2	SLC30A8	RPSA	88	WFS1	SQSTM1	RPS12
39	SQSTM1	TLK1	CSDE1	89	LAPTM4A	RASD1	ACLY
40	KRT19	ETNK1	CLTC	90	TAAR5	DSP	MSN
41	CD63	B2M	RPL10	91	SLC22A17	COX8A	HNRNPA2B1
42	SLC40A1	DDX5	YWHAZ	92	RPL3	TIMP1	CTNNB1
43	G6PC2	FOS	RPL3	93	HERPUD1	ATP1B1	MORF4L1
44	REG1A	MAFB	SLC30A8	94	CD24	WFS1	SERINC1
45	DDX5	CD59	RPL6	95	CALR	PRDX3	KRT19
46	PCBP1	TM4SF4	TMSB10	96	CLDN7	CHP1	NCL
47	C6orf62	TMEM33	CD44	97	LAMP2	YWHAE	GPX4
48	CRYBA2	CAPZA1	NPM1	98	CST3	FAM46A	GNB1
49	CD74	CALM2	B2M	99	TMBIM6	RUFY3	RPS7
50	HLA-E	GPX3	PABPC1	100	CTSB	C4orf48	SEP2
	alpha	beta	delta		gamma	acinar	ductal

Table 11: Top 100 Genes with highest weights in the pancreatic islet analysis. The bulk/artificial bulk data are from Segerstolpe et al. (2016), Xin et al. (2016) and Fadista et al. (2014) and the single cell reference is obtained from 6 healthy subjects from Segerstolpe et al. (2016). This table is color-coded by well-known marker genes.

Rank	Beckerman	Craciun	Arvaniti	Rank	Beckerman	Craciun	Arvaniti
1	Kap	Malat1	Malat1	51	Cycs	Dbi	Rps14
2	mt-Atp6	Kap	Kap	52	Rplp1	Rps18	Cox4i1
3	Gpx3	mt-Atp6	Gpx3	53	Rpl23	Rps14	Rpl26
4	mt-Co1	Gpx3	S100g	54	Gatm	Cycs	Cox5a
5	mt-Cytb	mt-Co1	Ftl1	55	Rpl32	Cox4i1	Rps19
6	S100g	mt-Cytb	Fth1	56	Cyb5a	Uqcrb	Rpl10
7	mt-Co3	S100g	Rps29	57	Acsm2	Ndrg1	Ttc36
8	mt-Co2	mt-Co3	Xist	58	Guca2b	Rpl10	Rpl35
9	mt-Nd4	mt-Co2	Rpl37a	59	Uqcrb	Rpl26	Gm8730
10	mt-Nd1	mt-Nd4	Rpl41	60	Rps14	Rps19	Dnase1
11	Ftl1	mt-Nd1	Fxyd2	61	Cox4i1	Acsm2	Itm2b
12	Fth1	Ftl1	Rpl38	62	Rpl26	Rpl35	Rpl35a
13	Rps29	Fth1	Rpl37	63	Cox5a	Cyb5a	Rps24
14	mt-Nd2	Rps29	Miox	64	Rps19	Miox	Gm10260
15	mt-Nd3	mt-Nd2	Eef1a1	65	Ttc36	Itm2b	Atp5l
16	Rpl37a	mt-Nd4l	Rpl39	66	Rpl10	Rpl35a	Slc34a1
17	Rpl41	mt-Nd3	Cox6c	67	Dnase1	Atp5l	Aldob
18	Fxyd2	Rpl37a	Rps28	68	Rpl35	Gm8730	Cela1
19	Rpl38	Rpl41	Rps27	69	Rpl35a	Akr1c21	Ass1
20	Rpl37	Xist	Cndp2	70	Atp5l	Rpl28	Prdx1
21	Miox	Fxyd2	Cyp4b1	71	Rps24	Slc34a1	Rpl28
22	Eef1a1	Rpl37	Ndufa4	72	Slc34a1	Prdx1	Rpl23a
23	Rpl39	Rpl38	Akr1c21	73	Gm8730	Aldob	Rpl6
24	Cox6c	Eef1a1	Atp1a1	74	Itm2b	Rps27a	Pck1
25	Rps28	Spink1	Acy3	75	Aldob	Cox6a1	Gm10709
26	mt-Nd5	Rpl39	Atp5k	76	Cela1	Rps24	2010107E04Rik
27	Rps27	Rps28	Cox7c	77	Ass1	Rpl23a	Cox6a1
28	Cndp2	Cox6c	Klk1	78	Prdx1	Rps4x	Slc25a5
29	Cyp4b1	Rps27	Ubb	79	Rpl28	Gm10709	Rps4x
30	Ndufa4	mt-Nd5	Atp5e	80	Rpl6	Slc25a5	Rps27a
31	Akr1c21	mt-Atp8	Rps2	81	Rpl23a	Ppia	Ldhb
32	Atp1a1	Atp1a1	Ndrg1	82	Pck1	Cox5a	Cox6b1
33	Acy3	Cox7c	Rps23	83	2010107E04Rik	Rpl13	Rpl18a
34	Atp5k	Ubb	Gm10076	84	Cox6a1	Cox6b1	Calb1
35	Cox7c	Atp5e	Prdx5	85	Gm10709	Cox7a2	Rpl13
36	Klk1	Atp5k	Rps1	8 86	Slc25a5	Gatm	Atp5b
37	Atp5e	Ndufa4	Tpt1	87	Rps27a	Ass1	Rpl13a
38	Ubb	Rps2	Chchd10	88	Rps4x	Ndufa3	Cox7a2
39	Rps2	Rps23	Rplp0	89	Ldhb	Rpl18a	Ndufa3
40	Ndrg1	Gm10076	Dbi	90	Cox6b1	Cyp4b1	Slc27a2
41	Rps23	Klk1	Rpl29	91	Calb1	Atp5j	Actb
42	Gm10076	Rps21	Rps21	92	Atp5b	Cox8a	Ppia
43	Prdx5	Rpl29	Rplp1	93	Cox7a2	Acy3	Rpl36a
44	Chchd10	Prdx5	Cycs	94	Rpl18a	Rpl36a	Atp5j
45	Tpt1	Rplp1	Rpl23	95	Ndufa3	Actb	Chpt1
46	Rps18	Tpt1	Rpl32	96	Slc27a2	Ndufa13	Rps15a
47	Dbi	Rpl23	Gatm	97	Rpl13	Rpl13a	Hrsp12
48	Rps21	Rpl32	Acsm2	98	Rpl36a	Ttc36	Ndufa13
49	Rplp0	Chchd10	Guca2b	99	Ppia	2010107E04Rik	Cox8a
50	Rpl29	Rplp0	Uqcrb	100	Atp5j	Gm10260	Ugt2b38
	PT	DCT	CD-IC		Podo	T lymph	

Table 12: Top 100 genes with highest weights in the mouse kidney analysis in Step 1 of the tree-guided deconvolution procedure. The bulk/artificial bulk data are from Beckerman et al. (2017), Craciun et al. (2016) and Arvaniti et al. (2016) and the single cell reference is obtained by 7 healthy mice from Park et al. (2018). This table is color-coded by marker genes.

Immune	Rank	Beckerman	Craciun	Arvaniti	Rank	Beckerman	Craciun	Arvaniti
	1	Cd74	Apoe	Cd74	26	C1qb	Npc2	C1qb
	2	Lyz2	S100a6	Lyz2	27	Nkg7	Gzma	Nkg7
	3	Ccl5	S100a4	Ccl5	28	Ccl4	Capza2	Vim
	4	H2-Aa	Psap	H2-Aa	29	Vim	Ly6e	Ccl4
	5	H2-Ab1	Nkg7	H2-Ab1	30	Ly6c2	Ly6c2	Ly6c2
	6	Tmsb10	Crip1	Tmsb10	31	Ms4a4b	Serinc3	Ms4a4b
	7	Gzma	Cd3g	Gzma	32	Sat1	Fos	Sat1
	8	H2-Eb1	Ccl3	H2-Eb1	33	C1qc	Pou2f2	C1qc
	9	Plac8	Ccnd2	Plac8	34	S100a10	Ctsz	S100a10
	10	Cst3	Slpi	Cst3	35	H3f3a	Cd74	H3f3a
	11	Ifi27l2a	Gm2a	Ifi27l2a	36	Ctss	Il7r	Ctss
	12	Slpi	Ssr4	Slpi	37	Gngt2	H2afy	Gngt2
	13	Ifitm3	Lck	Ifitm3	38	S100a6	Ctsb	S100a6
	14	Apoe	Spi1	Apoe	39	S100a4	Ifngr1	S100a4
	15	Tyrobp	Fxyd5	Tyrobp	40	Lst1	Tgfb1	Lst1
	16	Actg1	Ccl4	Actg1	41	Klf2	Sub1	Klf2
	17	Crip1	Gzmb	Crip1	42	Msrb1	Socs2	Msrb1
	18	Fcer1g	Cnn2	Fcer1g	43	H2afz	Ifitm3	H2afz
	19	Cebpb	Id2	Cebpb	44	Wfdc17	Itgb7	Wfdc17
	20	C1qa	Cybb	C1qa	45	Arpc1b	Cd79a	Arpc1b
	21	AW112010	Sep1	AW112010	46	Ifitm2	Ltb	Ltb
	22	Ly6e	Hsp90b1	Ly6e	47	Ltb	Fyb	Ifitm2
	23	Id2	Itgb2	Id2	48	S100a11	Tspan32	S100a11
	24	Psap	Ccl6	Psap	49	Lgals3	Sat1	Mzb1
	25	Lgals1	Lsp1	Lgals1	50	Mzb1	Xbp1	Lgals3
Epithelial	Rank	Beckerman	Craciun	Arvaniti	Rank	Beckerman	Craciun	Arvaniti
	1	Hbb-bs	Hbb-bs	Hbb-bs	26	Gm5424	Slc12a3	Slc22a28
	2	Hba-a1	Hba-a1	Hba-a1	27	Slc12a3	Slc22a28	Slc22a29
	3	Umod	Slco1a1	Slco1a1	28	Nrp1	Slc22a29	Emcn
	4	Slco1a1	Slc22a6	Slc22a6	29	Igfbp5	Ly6c1	Car12
	5	Slc22a6	Pvalb	Nat8	30	Ehd3	Car12	Aspdh
	6	Pvalb	Nat8	Pvalb	31	Slc22a28	Aspdh	Akr1c14
	7	Nat8	Umod	Mep1a	32	Slc12a1	Igfbp5	Ly6c1
	8	Mep1a	Mep1a	Umod	33	Slc22a29	Akr1c14	Hexb
	9	Egf	Slco1a6	Slco1a6	34	Car12	Atp6v1g3	BC035947
	10	Slco1a6	Ces1f	Ces1f	35	Aspdh	Ehd3	Igfbp5
	11	Ces1f	Hbb-bt	Hbb-bt	36	Akr1c14	Hexb	Atp6v1g3
	12	Hbb-bt	Egf	Snhg11	37	Kdr	Slc12a1	Nrp1
	13	Snhg11	Snhg11	Tmigd1	38	Atp6v1g3	BC035947	Slc13a1
	14	Tmigd1	Tmigd1	Egf	39	Hsd11b2	Slc13a1	Slc12a1
	15	Acsm3	Acsm3	Acsm3	40	Hexb	Col6a6	Col6a6
	16	Slc22a30	Slc22a30	Slc22a30	41	Eng	Gm4450	Gm4450
	17	Gm11128	Cyp2a4	Gm11128	42	BC035947	Kdr	Adamts15
	18	Aqp2	Hba-a2	Cyp2a4	43	Pi16	Adamts15	Ehd3
	19	Cyp2a4	Aqp2	Hba-a2	44	Slc13a1	Hsd11b2	Aspa
	20	Fxyd4	Aqp1	Gm5424	45	Col6a6	Aspa	Mogat1
	21	Emcn	Gm5424	Slc17a1	46	Gm4450	Apela	D630029K05Rik
	22	Aqp1	Slc17a1	Aqp1	47	Egfl7	Mogat1	Gm15638
	23	Hba-a2	Plpp1	Aqp2	48	Adamts15	D630029K05Rik	Hsd11b2
	24	Ly6c1	Fxyd4	Slc12a3	49	Meis2	Eng	Akr1c18
	25	Slc17a1	Emcn	Fxyd4	50	Aspa	Gm15638	Smlr1
		PT	DCT	CD-IC	LOH	CD-PC	Endo	Podo
		Neutro	T lymph	Macro	Fib	B lymph	NK	

Table 13: Top 100 genes with highest weights in the mouse kidney analysis in Step 2 of the tree-guided deconvolution procedure (separated by epithelial and immune cells). The bulk data are from Beckerman et al. (2017), Craciun et al. (2016). and Arvaniti et al. (2016) and the single cell reference is obtained by 7 healthy mice from Park et al. (2018). This table is color-coded by marker genes.

A.2.9. Tables of high-weighted genes from pancreatic islet analysis and mouse kidney analysis

A.3. Supplementary Information for cell type specific differential expression via deconvolution

A.3.1. Null distribution validation on Xin et al. (2016) artificial dataset

The validation part utilized only 12 healthy subjects from Xin et al. (2016). We randomly assign diseased and healthy labels 100 times ($L = 100$) for 12 subjects (6 healthy and 6 diseased). To get the p-values from null distribution, We utilized the same $R = 100$ repetitions of gene subsets as in 4.3. There are 141 out of 20127 genes shows zero expression for the 12 healthy subjects and those zero-expressed genes are deleted in the null validation analysis. Before deleting, we matched those 141 genes to DE genes chosen by our method. Most of those genes (101 out of 141) overlapped with genes chosen by our method before the adjustment of variances.

For the 19986 genes tested in null validation, we first draw the QQ-plot for average p-values across repetitions (Figure 28). The best matched genes and worst match genes are shown in Figure 28. They actually lines pretty well as a line. However, not exactly on the $x = y$ line.

Those are not the genes that we are most interested in. We are more interested in genes that are actually cell type specific expressed but not selected by our method. For each cell type, we showed QQ-plot for 10 genes that are not selected from our method.

From those plots, there is a motivation to relax the assumption of μ_g and σ_g and include them as parameters in our model.

We estimated the $\hat{\rho}_g$, $\hat{\sigma_g^2}$ and $\hat{\mu_g}$ from equations (4.10-4.10) and their distribution are shown in Figure 33. From the histogram, we can model ρ_g with beta distribution, σ_g^2 with gamma

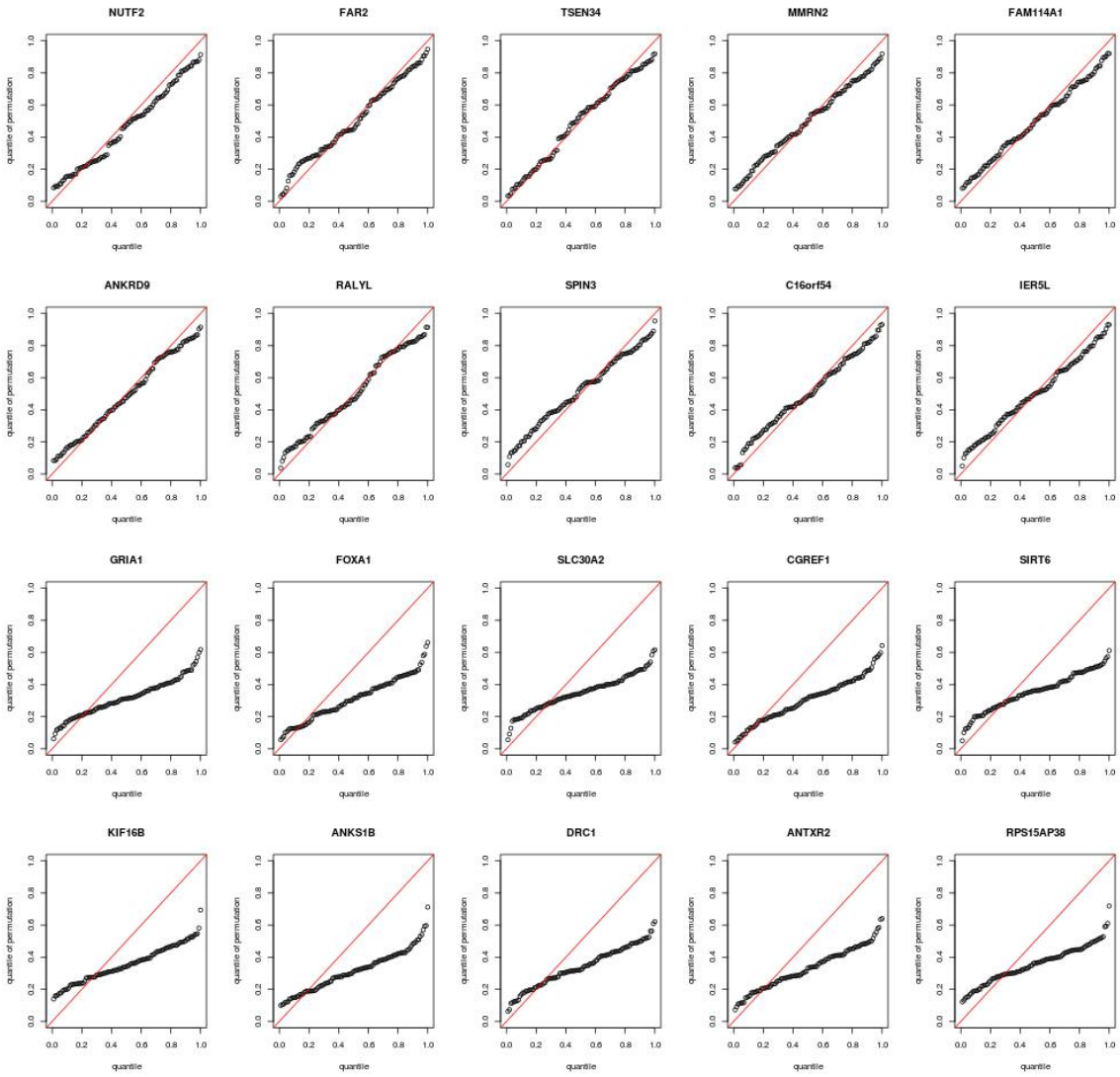


Figure 28: QQ plots of genes with best fit of uniform distribution or worst fit of uniform distribution.

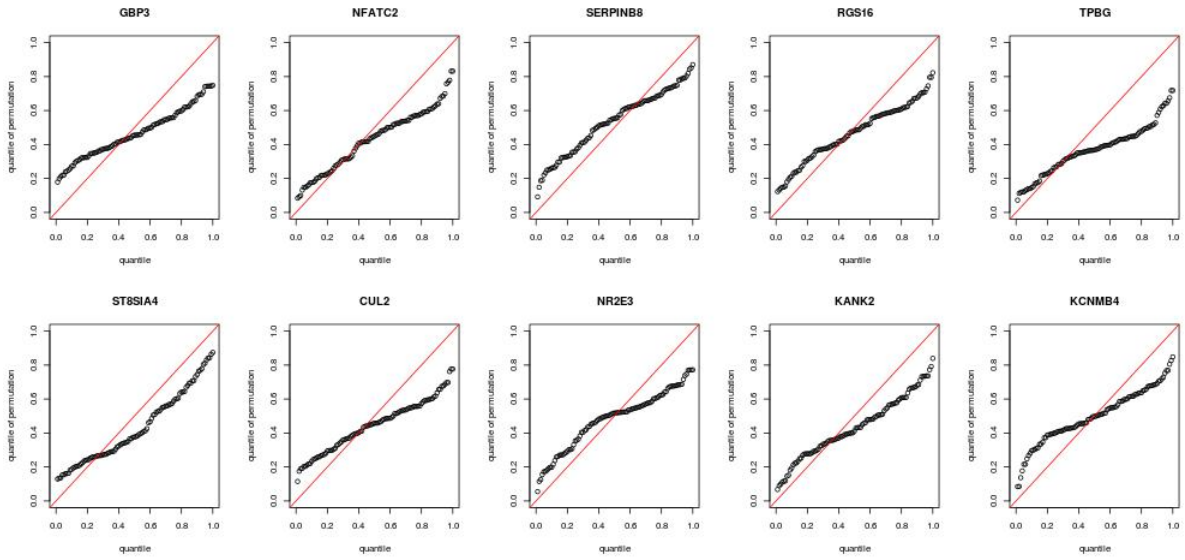


Figure 29: QQ plots of DE genes in alpha cells selected by DESeq 2, but not selected by MuSiC-DE.

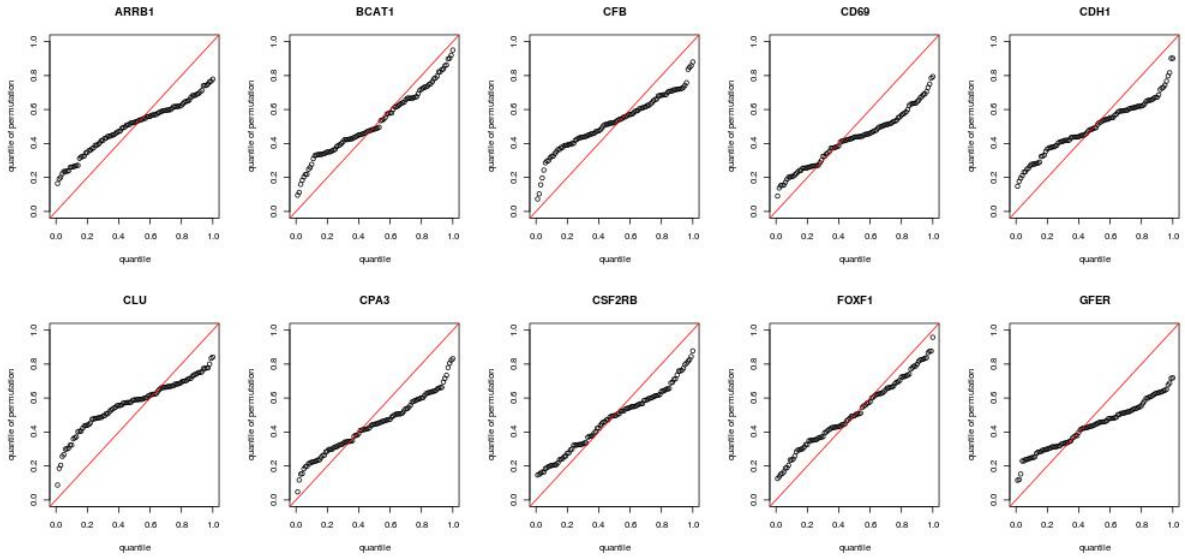


Figure 30: QQ plots of DE genes in beta cells selected by DESeq 2, but not selected by MuSiC-DE.

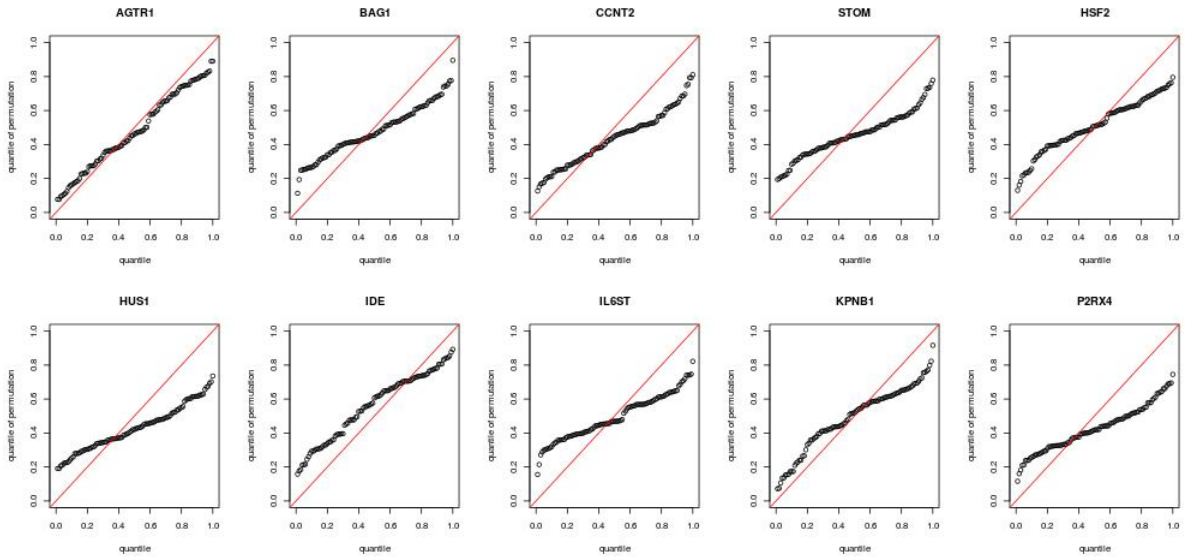


Figure 31: QQ plots of DE genes in delta cells selected by DESeq 2, but not selected by MuSiC-DE.

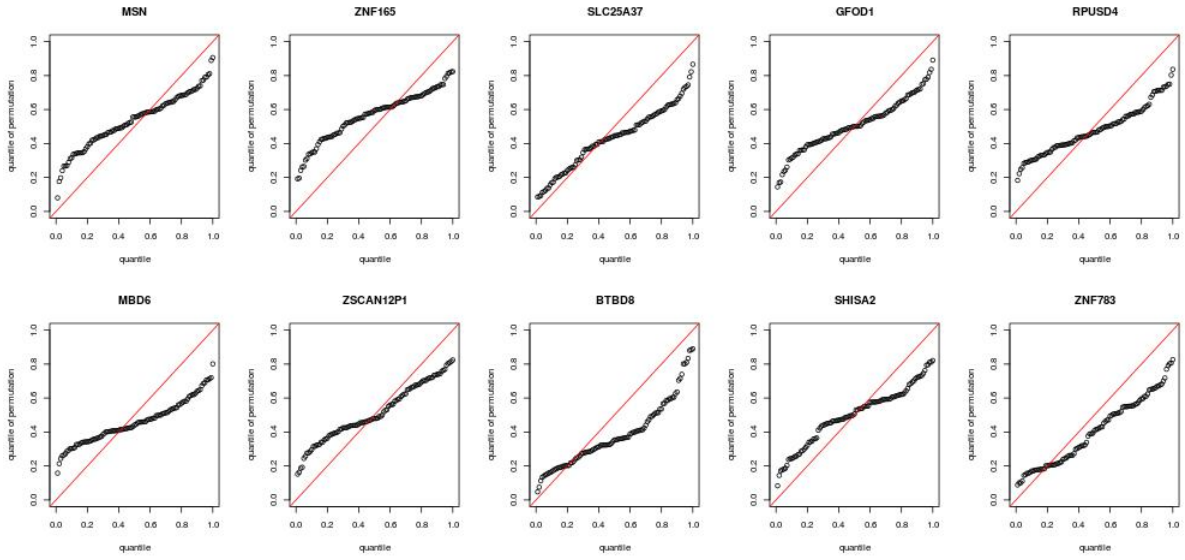


Figure 32: QQ plots of DE genes in gamma cells selected by DESeq 2, but not selected by MuSiC-DE.

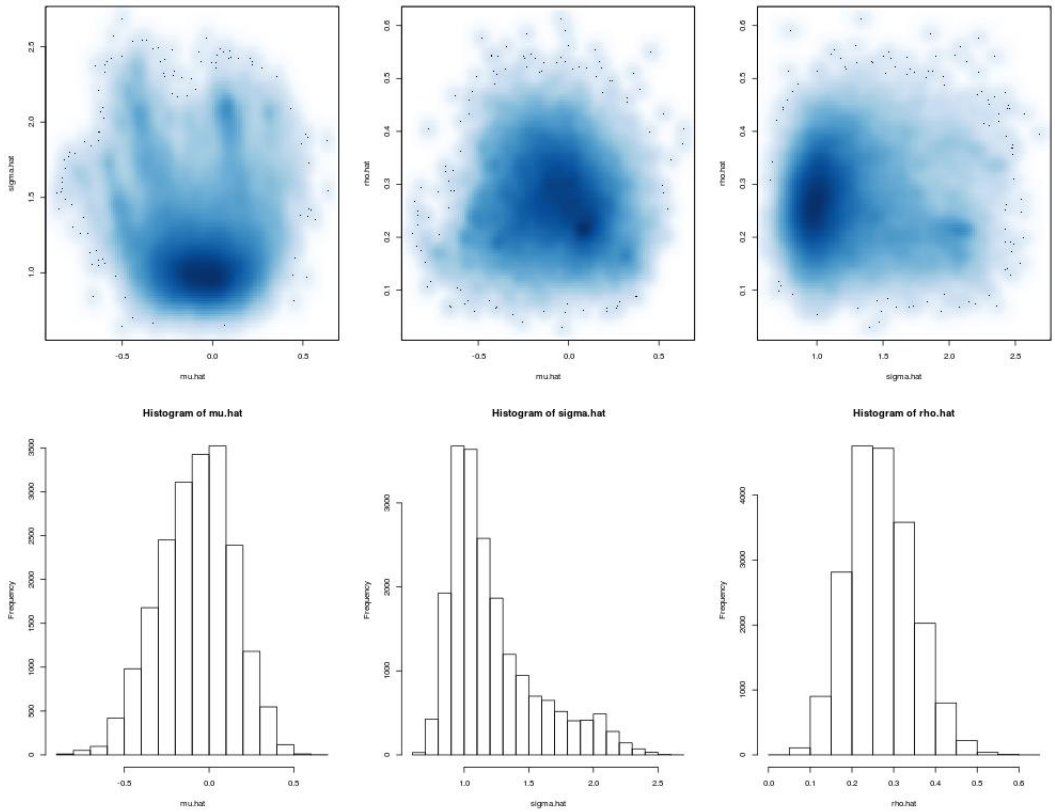


Figure 33: Estimated parameters from null distribution.

distribution and μ_g with normal distribution with mean zero. After that, we can use empirical Bayes estimated ρ_g, σ_g^2 and μ_g to test for DE.

BIBLIOGRAPHY

E. Arvaniti, P. Moulos, A. Vakrakou, C. Chatziantoniou, C. Chadjichristos, P. Kavvadas, A. Charonis, and P. K. Politis. Whole-transcriptome analysis of uuo mouse model of renal fibrosis reveals new molecular players in kidney diseases. *Scientific reports*, 6:26235, 2016.

F. Avila Cobos, J. Vandesompele, P. Mestdagh, and K. De Preter. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11): 1969–1979, 2018.

M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4): 346–360, 2016.

P. Beckerman, J. Bi-Karchin, A. S. D. Park, C. Qiu, P. D. Dummer, I. Soomro, C. M. Boustany-Kari, S. S. Pullen, J. H. Miner, C.-A. A. Hu, et al. Transgenic expression of human apol1 risk variants in podocytes induces kidney disease in mice. *Nature medicine*, 23(4):429, 2017.

S. Burgess and S. G. Thompson. *Mendelian randomization: methods for using genetic variants in causal estimation*. Chapman and Hall/CRC, 2015.

S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

O. Cabrera, D. M. Berman, N. S. Kenyon, C. Ricordi, P.-O. Berggren, and A. Caicedo. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proceedings of the National Academy of Sciences*, 103(7):2334–2339, 2006.

L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and biobanking*, 13 (5):311–319, 2015.

. G. P. Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571): 68, 2015.

J. Costa-Silva, D. Domingues, and F. M. Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PloS one*, 12(12):e0190152, 2017.

F. L. Craciun, V. Bijol, A. K. Ajay, P. Rao, R. K. Kumar, J. Hutchinson, O. Hofmann, N. Joshi, J. P. Luyendyk, U. Kusebauch, et al. Rna sequencing identifies novel translational biomarkers of kidney fibrosis. *Journal of the American Society of Nephrology*, 27 (6):1702–1713, 2016.

A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

J. Fadista, P. Vikman, E. O. Laakso, I. G. Mollet, J. L. Esguerra, J. Taneera, P. Storm, P. Osmark, C. Ladenvall, R. B. Prasad, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences*, 111(38):13924–13929, 2014.

Z. Ji and H. Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

C. Jia, Y. Hu, D. Kelly, J. Kim, M. Li, and N. R. Zhang. Accounting for technical noise in differential expression analysis of single-cell rna sequencing data. *Nucleic acids research*, 45(19):10978–10988, 2017.

M. B. Katan. Apolipoprotein e isoforms, serum cholesterol, and cancer. *International journal of epidemiology*, 33(1):9–9, 2004.

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

T. Lappalainen, M. Sammeth, M. R. Friedländer, P. ACt Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.

J. W. Lee, C.-L. Chou, and M. A. Knepper. Deep sequencing in microdissected renal tubules identifies nephron segment–specific transcriptomes. *Journal of the American Society of Nephrology*, 26(11):2669–2677, 2015.

B. Li, E. Severson, J.-C. Pignon, H. Zhao, T. Li, J. Novak, P. Jiang, H. Shen, J. C. Aster, S. Rodig, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome biology*, 17(1):174, 2016.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

J. Mattick. The state of long non-coding rna biology. *Non-coding RNA*, 4(3):17, 2018.

I. McDowell, A. Pai, C. Guo, C. M. Vockley, C. D. Brown, T. E. Reddy, and B. E. Engelhardt. Many long intergenic non-coding rnas distally regulate mrna gene expression levels. *BioRxiv*, page 044719, 2016.

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

K. Ozaki, Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, Y. Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650, 2002.

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

D. Porte and S. E. Kahn. beta-cell dysfunction and failure in type 2 diabetes: potential mechanisms. *Diabetes*, 50(suppl 1):S160, 2001.

J. L. Rinn and H. Y. Chang. Genome regulation by long noncoding rnas. *Annual review of biochemistry*, 81:145–166, 2012.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

T. J. Rothenberg et al. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4):593–607, 2016.

F. Seyednasrollah, A. Laiho, and L. L. Elo. Comparison of software packages for detecting differential expression in rna-seq studies. *Briefings in bioinformatics*, 16(1):59–70, 2013.

G. D. Smith and S. Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology*, 33(1):30–42, 2004.

M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737, 1974.

J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20 (4):518–529, 2002.

P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li. Bulk tissue cell type deconvolution

with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.

Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell metabolism*, 24(4):608–615, 2016.

X.-Y. Zhai, J. S. Thomsen, H. Birn, I. B. Kristoffersen, A. Andreasen, and E. I. Christensen. Three-dimensional reconstruction of the mouse nephron. *Journal of the American Society of Nephrology*, 17(1):77–88, 2006.