

A Likelihood-based Deconvolution of Bulk Gene Expression Data Using Single-cell References

*Justin Hong
Dan D. Erdmann-Pham
Jonathan Fischer
Yun S. Song*

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-21

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-21.html>

May 1, 2021



Copyright © 2021, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

**A Likelihood-based Deconvolution of Bulk Gene Expression Data Using
Single-cell References**

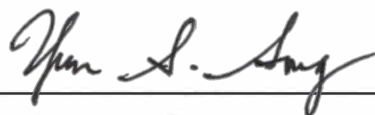
by Justin J. Hong

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley, in partial satisfaction of the requirements for the
degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:



Professor Yun S. Song
Research Advisor

May 27, 2020

(Date)



Professor Kannan Ramchandran
Second Reader/Co-Advisor

5/17/2020

(Date)

Abstract

A Likelihood-based Deconvolution of Bulk Gene Expression Data Using Single-cell References

by

Justin Hong

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Yun S. Song, Co-chair

Professor Kannan Ramchandran, Co-chair

A single bulk gene expression experiment estimates thousands of RNA transcript levels averaged over myriad cells. Unfortunately, direct comparison of different bulk expression profiles is complicated by the mixtures of distinct cell types in each sample, obscuring whether perceived differences are actually due to changes in expression levels themselves or simply cell type composition. Single-cell technology has made it possible to measure gene expression in individual cells, achieving higher resolution at the expense of increased noise. If carefully incorporated, such data can be used as references for the supervised deconvolution of bulk samples to yield accurate estimates of the true cell type proportions. These estimates permit us to disentangle the effects of differential expression and cell type mixtures, both of which are independently relevant to our understanding of aging and disease. We hence propose a generative model which uses asymptotic statistical theory and a robust estimation procedure to perform a supervised deconvolution of bulk RNA-seq samples to produce cell type proportion estimates. We demonstrate the effectiveness of our approach in several scenarios with real data and also discuss several novel extensions made uniquely possible by our paradigm.

To Umma and Appa

Contents

Contents	ii
List of Figures	iii
1 Introduction	1
1.1 Background	1
1.2 Problem Formulation	2
1.3 Related Work	2
1.4 Contributions	4
2 A Likelihood-based Deconvolution Method	5
2.1 Model	5
2.2 Algorithm	8
2.3 Results	12
3 Conclusion	20
3.1 Summary	20
3.2 Future Work	20
Bibliography	22

List of Figures

2.1	Log Likelihood Heatmaps on Real Dataset Pseudobulk	10
2.2	Comparison of Optimization Methods on Real Dataset Pseudobulk	11
2.3	Deconvolution results for liver pseudobulks	13
2.4	Deconvolution results for spleen pseudobulks	14
2.5	Deconvolution results for skin pseudobulks	15
2.6	Deconvolution results for large intestine pseudobulks	16
2.7	Deconvolution results for Fadista et al. bulks	18
2.8	Deconvolution results for <i>Tabula Muris Senis</i> bulks	19

Acknowledgments

First and foremost, I would like to thank Dan Erdmann-Pham, Jonathan Fischer, and Yun S. Song for their mentorship and for inviting me to work on this project with them. Thank you for giving me a proper introduction to the computational world of biology. I am very grateful to have worked with you during my time at Berkeley.

I would also like to thank Kannan Ramchandran for his support and advice through these past few years. Running EECS 126 and participating in research with you has been a fantastic experience.

Furthermore, I would like to thank Joy Xie for her love and support throughout my college career. Finally, I would like to give my utmost gratitude to my family for supporting my academic career in every way and motivating me to keep fighting through thick and thin.

Chapter 1

Introduction

1.1 Background

As intermediate molecules in protein production and in other cases non-coding, functional components of a cell, RNA transcripts can provide insight into the specialization and role of a cell that solely analyzing the genome cannot. With the development of Next-Generation Sequencing (NGS), quantifying RNA expression in a large population of cells has significantly improved in terms of throughput and coverage compared to RNA microarray methods through a method called RNA-seq [5]. However, unless the mixture proportion of the population of cells was known *a priori*, only conclusions about the bulk as a whole could be made.

Alternatively, single-cell sequencing (scRNA-seq) allows one to analyze transcriptome-wide RNA expression data for individual cells. The unparalleled level of resolution provided by this method allows one to identify cell types by clustering cells based on their gene expression, analyze rarer cell type populations, and understand changes in gene expression over time [21]. Single-cell sequencing has come into the spotlight in the field of computational biology, due to improvements in technology and sequencing methods making the data more accurate and accessible.

Despite RNA-seq having many successes, the full potential of this technology is inherently limited because each experiment measures the average gene expression among a large group of cells, the composition of which is unknown. Thus, despite the reduction in technical and biological variability they attain by averaging, bulk experiments are potentially confounded by cell type proportions when considering heterogeneous cell mixtures [14, 22]. Such confounding presents problems when comparing samples to identify differences with clinical importance and may result in the spurious or missed inference of biologically relevant genes. Moreover, cell type compositions are often independently informative of biological features including tissue function [3, 9, 12, 28] or development [9, 10]. For example, cell type infiltration has been found to be associated with disease progression, disease status, and complex processes such as aging. In this scenario, bulk RNA-seq data alone would typically be in-

sufficient to identify the presence of a certain cell type. By isolating the expression patterns of each measured cell type, single-cell expression data can provide a reference to infer the cell type composition of the bulk sample when the same cell types are captured by both experiments; this is known as deconvolution.

1.2 Problem Formulation

Statistical deconvolution of bulk RNA-seq data is attractive for several reasons. First is the aforementioned confounding potential of different cell type compositions present in bulk samples. Second is that single-cell experiments are more expensive than their bulk counterparts and more difficult to perform, often making the large-scale generation of single-cell expression data infeasible. Furthermore, certain protocols do not capture cell types in an unbiased fashion, so the empirical sample proportions don't always serve as reliable estimators of the true values. Finally, accurate deconvolution permits the re-visiting of old data sets to investigate questions which were previously unanswerable.

The deconvolution of expression data has recently attracted the interest of computational researchers. The problem is frequently framed mathematically as

$$\Omega\alpha = B, \tag{1.1}$$

where, for K cell types and G genes, $\Omega \in \mathbb{R}^{G \times K}$ is a matrix of single-cell gene expression averages, $\alpha \in \mathbb{R}^K$ a vector of mixing proportions, and $B \in \mathbb{R}^G$ is a vector of expression values in a bulk sample. Depending on which of Ω , α , and B are measured, different approaches are necessary. We focus on the case in which both Ω and B have been observed (noisily) and it remains to infer α .

1.3 Related Work

Deconvolution Problem Setups

The deconvolution problem was first stated in [25] long before abundant scRNA-seq data was widely available. Instead, the problem was applied to alternative settings such as controlled experiments with model organisms [15] or with cell lines that could be grown to produce pure samples for the reference [1].

Prior to the advent of sufficient single-cell sequencing data, it was unclear whether Ω or α would have to be estimated. While RNA microarrays could provide bulk expression measurements, B , either reference matrices, Ω , had to be estimated from controlled settings or α had to be estimated using alternative methods to infer Ω [2]. For example in [24], the cell type proportions of prostrate samples were estimated by visual identification, and then Ω was inferred via RNA microarray data. Another proposed method attempts to infer a (Ω, α) pair by structuring the problem as a non-negative matrix factorization problem (NMF) [17].

By inferring both at the same time, the number of cell types in question could be tuned accordingly, allowing for the identification of cell types.

In this work, we focus on the deconvolution problem setup which rose to prominence with scRNA-seq data. In this case, scRNA-seq datasets provide a noisy reference with a large coverage of the genome for a predetermined number of cell types. After constructing Ω , the mixture proportion, α , of RNA-seq data, B , over the same set of genes is to be inferred.

Single-Cell Reference Methods

The most straightforward method to approach the deconvolution problem is with a least squares approach which minimizes the sum of squared differences across each gene. Since the mixture proportion vector, α , should exist on the probability simplex, the typical approach is to solve the non-negative least squares (NNLS) problem then projecting the vector onto the probability simplex.

$$\hat{\alpha}_{\text{NNLS}} = \underset{\alpha \geq 0; \|\alpha\|_1=1}{\operatorname{argmin}} \|\Omega\alpha - B\|_2^2 \quad (1.2)$$

With additional higher-order information about Ω and B , one can find a weighted non-negative least squares (W-NNLS) solution, where the contribution of each gene varies based on a given vector of weights, $w \in \mathbb{R}^G$, for G genes:

$$\hat{\alpha}_{\text{W-NNLS}} = \underset{\alpha \geq 0; \|\alpha\|_1=1}{\operatorname{argmin}} \sum_{g=1}^G w_g \left(\sum_{k=1}^K \alpha_k \Omega_{g,k} - B_g \right)^2 \quad (1.3)$$

where the subscript g selects the respective row for the gene $g \in [G]$, and the subscript k selects the respective column for the cell type $k \in [K]$. The operator $\langle \cdot, \cdot \rangle$ is the Euclidean inner product.

Proposed methods such as MuSiC [26], DWLS [8], and dtangle[11] use some form of W-NNLS with different weights to infer the mixture proportion. Other approaches such as CPM [8] and CIBERSORTx [18] utilize support vector regression, which aims to find a hyperplane that separates classes by at least a constant margin. A recent work proposes Scaden [16], which infers the mixture proportion with deep learning.

However, all of these methods rely on discriminative models that fail to replicate the benefits of explicit generative modeling. Our proposed method estimates several latent parameters describing the underlying model of the RNA-seq counts. Beyond inferring a mixture proportion for the bulk, one can perform hypothesis tests to answer additional questions. For example, unlike the deconvolution problems described in Section 1.3, the reference requires a pre-determined set of cell types. Invasion of the bulk sample by cells from surrounding tissues can confound proportion estimates if not accounted for. With our proposed framework, we can detect the presence of such unidentified cell types by formulating hypothesis tests with respect to our algorithm’s estimates of the model parameters. While we

focus on the performance of our method benchmarked against these discriminative methods, we affirm that further value in our method lies in underlying parametric model.

1.4 Contributions

We propose a generative model and develop an associated optimization algorithm, RNA-Sieve, to estimate mixture proportions, α .

We structure this thesis as follows:

- In Section 2.1, we provide a detailed description of the generative model.
- in Section 2.2, we present the associated algorithm for estimating mixture proportions.
- In Section 2.3, we demonstrate the desirable performance of our method using simulated and real data. In addition, we apply our algorithm to deconvolve bulk samples generated as part of the *Tabula Muris Senis* project [19].

Chapter 2

A Likelihood-based Deconvolution Method

2.1 Model

Preliminaries

We assume that for each gene $g \in \{1, \dots, G\} = [G]$, and cell type $k \in \{1, \dots, K\} = [K]$, there exists a distribution $\nu_{g,k}$ describing the expression of gene g in cell type k . As tissues are comprised of multiple cell types, the expression of gene g in a cell drawn at random from a tissue is governed by the mixture distribution

$$\nu_g = \sum_{k=1}^K \alpha_k^* \nu_{g,k}, \quad (2.1)$$

where $\alpha^* = (\alpha_k^*)_{k \in [K]} \in \Delta^{K-1}$ contains the proportions of each cell type in the tissue of interest. Despite the infinite-dimensional setting, if $G > k$ and ν_g and $\{\nu_{g,k}\}_{k \in [K]}$ are fully known and sufficiently distinct, the convex combination of (2.1) immediately implies that α^* can be recovered as the unique solution of the finite-dimensional problem

$$\underbrace{\begin{bmatrix} f(\nu_{1,1}) & \dots & f(\nu_{1,K}) \\ \vdots & & \vdots \\ f(\nu_{G,1}) & \dots & f(\nu_{G,K}) \end{bmatrix}}_{\Omega} \cdot \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\alpha} = \underbrace{\begin{bmatrix} f(\nu_1) \\ \vdots \\ f(\nu_G) \end{bmatrix}}_B, \quad (2.2)$$

where $f \in \mathcal{M}(\mathbb{R})^*$ is any suitable linear functional on the space of measures on \mathbb{R} . Natural f to consider include the family of point evaluations $\{\delta_x : \nu \mapsto F_\nu(x)\}_{x \in \mathbb{R}}$ (where F_ν denotes the CDF of ν) or its k^{th} moments, both of which enjoy a wealth of statistical theory and proposed estimators. In experimental settings, exact expression distributions are not accessible and instead must be estimated, so utilizing easily and robustly inferable f becomes crucial. In

addition to not having direct access to $\{\nu_g\}_{g \in [G]}$, any analysis is further complicated by the fact that bulk sequencing only yields expression levels over whole samples and not for particular cells or cell types. That is, the output is effectively a random variable $X_g = \sum_{i=1}^n X_g^i$ where $X_g^i \stackrel{iid}{\sim} \nu_g$ gives the measured expression of gene g aggregated over the $n \in \mathbb{N}$ individual cells comprising the sample. Since $X_g \sim (\nu_g)^{\otimes n}$ is an n -fold convolution, it is expedient to choose an f in (2.2) that is not only linear on the space of measures, but also on the corresponding space of measurable functions. Let us define the k^{th} moments of ν as $\{\mu_k : \nu \mapsto \int x^k \nu(dx)\}_{k \in \mathbb{N}}$. Setting $f = \mu_1 = \mathbb{E}$ turns (2.2) into

$$\Omega\alpha = B/n, \quad \text{where} \quad \Omega_{g,k} = \mathbb{E}\nu_{g,k} \quad \text{and} \quad B_g = \mathbb{E}X_g. \quad (2.3)$$

Incorporating the fact that we only observe noisy bulk samples X_g instead of B directly results in

$$\hat{B}/n = (B + \varepsilon_B)/n = \Omega\alpha + \varepsilon_B/n, \quad (2.4)$$

where $(\varepsilon_B)_g \sim X_g - B_g = \sum_{i=1}^n (X_g^i - B_g/n) \sim \mathcal{N}(0, n v_g)$ for large n by the central limit theorem (with $v_g = \text{Var}(\nu_g)$). When ignoring the dependence of v_g on α , (2.4) lends itself to a simple (perhaps weighted) least squares scheme by first solving

$$\Omega\alpha^{LS} = \hat{B} \quad (2.5)$$

while possibly enforcing non-negativity constraints, though these are typically satisfied for most reasonable values of n and v_g . This yields a solution α^{LS} of roughly $\|\alpha^{LS}\|_1 \approx n$ that simply requires re-scaling to fit onto the simplex. This (plus a few data-driven modifications) is the approach recently developed in [26], where it is argued that (2.5) outperforms previous methods. It is thus natural to ask whether incorporating the dependence of v_g on α can improve accuracy even further.

Likelihood Formulation

To answer this question, we first make explicit this α -dependence by computing the bulk variance vector, $v \in \mathbb{R}^G$:

$$\begin{aligned} v_g(\alpha, \Omega, \sigma) &= \text{Var}(\nu_g) = \mu_2(\nu_g) - \mu_1(\nu_g)^2 \\ &= \left(\sum_{k=1}^K \alpha_k \mu_2(\nu_{g,k}) \right) - \left(\sum_{k=1}^K \alpha_k \Omega_{g,k} \right)^2 \\ &= \left(\sum_{k=1}^K \alpha_k \left[\sigma_{g,k}^2 + (\Omega_{g,k})^2 \right] \right) - \left(\sum_{k=1}^K \alpha_k \Omega_{g,k} \right)^2, \end{aligned} \quad (2.6)$$

where $\sigma_{g,k}^2 = \text{Var}(\nu_{g,k})$. The likelihood of observing data \hat{B} then follows straightforwardly from the central limit theorem:

$$\mathbb{P}_{\Omega,\sigma}^{\alpha,n} \left(\hat{B} \in db \right) = \prod_{g=1}^G \frac{1}{\sqrt{2\pi n v_g(\alpha, \Omega, \sigma)}} \exp \left\{ \frac{-(b_g - B_g(\alpha))^2}{2n v_g(\alpha, \Omega, \sigma)} \right\}, \quad (2.7)$$

where $B^g(\alpha) = n \sum_{k=1}^K \alpha_k \Omega_{g,k}$.

The above assumes exact knowledge of the individual distributions $\nu_{g,k}$ (or rather their expectations $\Omega_{g,k}$), which is implausible in experimental settings. Instead, Ω needs to be estimated from data through some estimator $\hat{\Omega}$, which we conveniently take to be the sample mean of expression across cells within each cell type, $\hat{\Omega}_{g,k} = \frac{1}{m_k} \sum_{i=1}^{m_k} S_{g,k}^i$, where $S_{g,k}^i \stackrel{iid}{\sim} \nu_{g,k}$ and m_k denotes the number of cells of type $k \in [K]$ used to estimate $\hat{\Omega}_{g,k} \forall g \in [G]$. With this additional correction, (2.4) becomes

$$\hat{B}/n = \hat{\Omega}\alpha + \varepsilon_B/n \quad \text{and} \quad \hat{\Omega} = \Omega + \varepsilon_\Omega \quad (2.8)$$

where ε_Ω is a matrix of entries $(\varepsilon_\Omega)_{g,k}$ independently following $\mathcal{N}(0, \sigma_{g,k}^2/m_k)$ distributions. The new likelihood thus becomes

$$\begin{aligned} \mathbb{P}_{\Omega,\sigma}^{\alpha,n,m} \left(\hat{B} \in db, \hat{\Omega} \in d\omega \right) &= \prod_{g=1}^G \frac{1}{\sqrt{2\pi n v_g(\alpha, \Omega, \sigma)}} \exp \left\{ \frac{-(b_g - B_g(\alpha))^2}{2n v_g(\alpha, \Omega, \sigma)} \right\} \\ &\times \prod_{g \in [G], k \in [K]} \frac{1}{\sqrt{2\pi \sigma_{g,k}^2/m_k}} \exp \left\{ \frac{-(\omega_{g,k} - \Omega_{g,k})^2}{2\sigma_{g,k}^2/m_k} \right\}. \end{aligned} \quad (2.9)$$

This is the likelihood with which we work in our model, and there are a handful of implicit assumptions made here which are worth examining. The first is that the large number of cells assayed in an experiment permits us to use asymptotic theory and apply the classical CLT. As a result, we can write down a likelihood for our observations using normal distributions as long as $\mu_2(\nu_g) < \infty$, which is trivially true since expression profiles are necessarily bounded. Secondly, we suppose that the errors arising from estimating \hat{B} and $\hat{\Omega}$ are independent. This seems reasonable as the bulk and single-cell experiments are performed separately. We additionally presume that expression levels in different genes are independent, as are those in different cells. It is unclear whether the latter is completely true in practice, though it is likely an accurate approximation of the truth. On the other hand, expression levels across genes within samples (either bulk or individual cells) are liable to be somewhat dependent due to expression co-regulation and the RNA sampling performed in RNA-seq. Given the large number of genes assayed, the latter co-dependence is likely to be fairly small. Meanwhile, co-expression estimation in single cells remains an open problem, and the independence assumption is currently required to ensure computational tractability.

Remark 1. (*Curse of universality.*) We note that the presence of the CLT in (2.4) seems to indicate that not much more power can be gained beyond second order methods.

Remark 2. (*Higher-Order models.*) It is reasonable to include the sample variances \hat{v} and $\hat{\sigma}^2$ into the formulation of (2.9). However, since \hat{v} and $\hat{\sigma}^2$ themselves satisfy CLTs that depend on the fourth moments of $\nu_{g,k}$, $\mu_4(\nu_{g,k})$ would need to be added as parameters to the model. In practice, those must be estimated by some $\hat{\mu}_4(\nu_{g,k})$ from data, which in turn prompts us to include $\hat{\mu}_4(\nu_{g,k})$ in (2.9). Of course, according to this reasoning, we would end up including infinitely many higher moments into our likelihoods, and it is interesting to consider at what point we reach the optimal balance between accuracy and computational tractability.

Remark 3. (*Joint deconvolution of multiple bulk samples.*) If it is known that multiple bulk expression vectors share the same constituent cell type expression profiles, we can gain statistical strength and decrease the computational burden by inferring their mixture proportions jointly rather than individually. Assuming statistical independence of the bulk sample observations, we must simply augment the likelihood in (2.9) by including the $D - 1$ additional mixtures in $\alpha = (\alpha^1, \dots, \alpha^D) \in \mathbb{R}^{K \times D}$, $\hat{B} = (\hat{B}^1, \dots, \hat{B}^D) \in \mathbb{R}^{G \times D}$, and the associated parameters in $n = (n^1, \dots, n^D) \in \mathbb{R}^D$:

$$\begin{aligned} \mathbb{P}_{\Omega, \sigma}^{\alpha, n, m} \left(\hat{B} \in db, \hat{\Omega} \in d\omega \right) &= \prod_{d=1}^D \prod_{g=1}^G \frac{1}{\sqrt{2\pi n^d v_g(\alpha^d, \Omega, \sigma)}} \exp \left\{ \frac{-(b_g^d - B_g^d(\alpha^d))^2}{2n^d v_g(\alpha^d, \Omega, \sigma)} \right\} \\ &\times \prod_{g \in [G], k \in [K]} \frac{1}{\sqrt{2\pi \sigma_{g,k}^2 / m_k}} \exp \left\{ \frac{-(\omega_{g,k} - \Omega_{g,k})^2}{2\sigma_{g,k}^2 / m_k} \right\}. \end{aligned} \quad (2.10)$$

2.2 Algorithm

Data Pre-Processing Procedure

Due to the well-known influence of technical variability in scRNA-seq data, we suggest that users of RNA-Sieve perform their own quality control filtering of cells and genes prior to running our software. Given the potential complexity of these patterns in general, we feel that manual cleaning is more reliable than automated procedures. Nonetheless, we implement a simple (largely optional) cell filtering scheme to ensure the accuracy of results when the user has not chosen to perform his or her own quality control. Our procedure attempts to do the following: 1) remove low-quality cells with anomalously low or high library sizes, 2) identify and remove cells which may be mislabeled or are simply extremely different from other cells with the same cell type label, and 3) identify and retain genes which are expressed sufficiently often in at least one cell type. We note that steps 1 and 2 are optional whereas 3 is necessary to remove lowly expressed genes whose presence results in poor optimization outcomes.

Algorithm 1: Find MLE of α

Data: Single cell expression vectors $\{v_k^i\}_{k \in [K], i \in m_k} \subset \mathbb{R}^G$, bulk expression vector $\hat{B} \in \mathbb{R}^G$

Result: Mixture proportions $\{\alpha_k^*\}_{k \in [K]}$ of cell types in the bulk, number of cells $n^* \in \mathbb{R}_+$ in bulk, mean expression $\Omega^* \in \mathbb{R}^{G \times K}$ of cell types

1 **begin**

2 $\hat{\Omega}_{g,k} \leftarrow \frac{1}{m_k} \sum_{i=1}^{m_k} (v_k^i)_g, \sigma_{g,k}^2 \leftarrow \frac{1}{m_k} \sum_{i=1}^{m_k} \left((v_k^i)_g - \hat{\Omega}_{g,k} \right)^2$

3 $\alpha^0 \leftarrow \arg \min_{\alpha \in \mathbb{R}_+^K} \|\hat{\Omega}\alpha - \hat{B}\|_2^2, n^0 \leftarrow \|\alpha^0\|_1, \alpha^0 \leftarrow \alpha^0 / \|\alpha^0\|_1, \Omega^0 \leftarrow \hat{\Omega}$

4 $j \leftarrow 0$

5 **while** $\mathbb{P}_{\Omega^{j+1}, \sigma}^{\alpha^{j+1}, n^{j+1}, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right) - \mathbb{P}_{\Omega^j, \sigma}^{\alpha^j, n^j, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right) > \delta$ **do**

6 $\Omega^{j+1} \leftarrow \arg \max_{\Omega} \mathbb{P}_{\Omega, \sigma | v(\alpha^j, \Omega^j, \sigma)}^{\alpha^j, n^j, m} \left(\Omega' \in d\hat{\Omega}, B \in d\hat{B} \right)$

7 $\alpha^{j+1} \leftarrow \arg \max_{\alpha \in \Delta^{K-1}} \mathbb{P}_{\Omega^{j+1}, \sigma | v(\alpha^j, \Omega^{j+1}, \sigma)}^{\alpha, n^j, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right)$

8 $n^{j+1} \leftarrow \arg \max_{n \in \mathbb{R}_+} \mathbb{P}_{\Omega^{j+1}, \sigma}^{\alpha^{j+1}, n, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right)$

9 $j \leftarrow j + 1$

10 **end**

11 $(\alpha^\ell, \Omega^\ell, n^\ell) \leftarrow$ Last $(\alpha^j, \Omega^j, n^j)$ iterate returned in line 7

12 **while** $\mathbb{P}_{\Omega^{\ell+1}, \sigma}^{\alpha^{\ell+1}, n^{\ell+1}, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right) - \mathbb{P}_{\Omega^\ell, \sigma}^{\alpha^\ell, n^\ell, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right) > \delta$ **do**

13 $\Omega^{\ell+1} \leftarrow \arg \max_{\Omega} \mathbb{P}_{\Omega, \sigma}^{\alpha^\ell, n^\ell, m} \left(\Omega' \in d\hat{\Omega}, B \in d\hat{B} \right)$

14 $\alpha^{\ell+1} \leftarrow \arg \max_{\alpha \in \Delta^{K-1}} \mathbb{P}_{\Omega^{\ell+1}, \sigma | v(\alpha^\ell, \Omega^{\ell+1}, \sigma)}^{\alpha, n^\ell, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right)$

15 $n^{\ell+1} \leftarrow \arg \max_{n \in \mathbb{R}_+} \mathbb{P}_{\Omega^{\ell+1}, \sigma}^{\alpha^{\ell+1}, n, m} \left(\Omega \in d\hat{\Omega}, B \in d\hat{B} \right)$

16 $\ell \leftarrow \ell + 1$

17 **end**

18 $(\alpha_*, \Omega^*, n^*) \leftarrow$ Last iterate returned in line 13

19 $\alpha^* \leftarrow \arg \min_{\alpha \in \Delta^{K-1}} \|\Omega^* \alpha - \hat{B}\|_{v(\alpha_*, \Omega^*)}^2$

20 Return $(\alpha^*, \Omega^*, n^*)$.

21 **end**

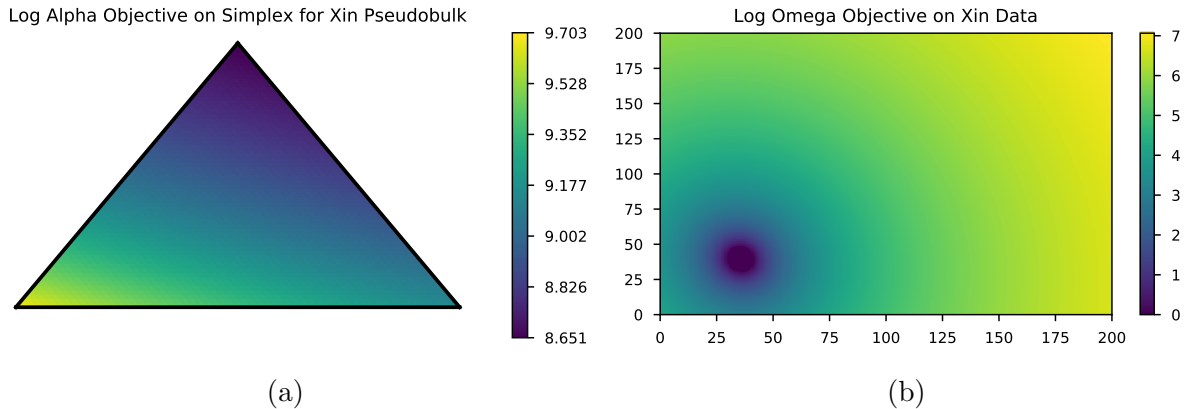


Figure 2.1: **Log Likelihood Heatmaps on Real Dataset Pseudobulk.** For a select number of cell types, we visualize the objective function for both the α and Ω optimization steps. (a) A heatmap of negative log likelihood values (ignoring constant scaling factors) over α values in the 2-simplex. (b) A heatmap of negative log likelihood values (ignoring constant scaling factors) over $\{\Omega_{g,k}\}_{k \in [K]}$ values in the non-negative orthant for an example g .

Algorithm Outline

We estimate α , the cell type proportions for a given bulk sample, using the MLE which arises from maximizing (2.9). Given the number of free parameters ($GK + K + 1$) and structure of the likelihood, this is non-trivial, with standard optimization schemes commonly failing or returning sub-optimal solutions. We thus propose an alternating maximization scheme which iteratively estimates and updates α , Ω , and n (and consequently v) via a combination of quadratic programming and gradient descent. Despite the increased computational burden relative to W-NNLS or similar techniques, we find that convergence remains fairly fast, requiring no more than two minutes on typical data sets of 10,000+ genes and 6 cell types using a modern laptop computer. We sketch an overview of our optimization procedure below (where $\mathbb{P}_{\Omega, \sigma | v}^{\alpha, n, m}$ refers to (2.9) with v kept fixed at $v(\alpha', \Omega', \sigma)$).

Implementations of Algorithm 1 are currently available in Python and Mathematica.

Optimization

In implementing the algorithm, we made the following design choices:

1. Instead of maximizing $\mathbb{P}_{\Omega, \sigma}^{\alpha, n, m}$, we minimize $-\log \mathbb{P}_{\Omega, \sigma}^{\alpha, n, m}$, rendering lines 6 and 7 as quadratic programs, which can be solved efficiently.
2. Lines 8 and 15 can be solved explicitly by differentiating (2.9) and finding the zeros of

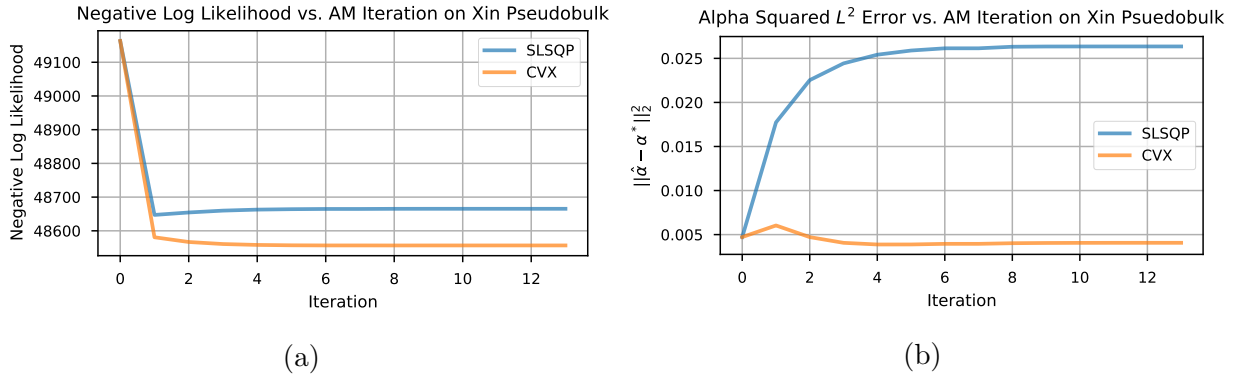


Figure 2.2: **Comparison of Optimization Methods on Real Dataset Pseudobulk.** We compare the performance of SLSQP vs. iterative CVX (OSQP) in the α minimization step on the deconvolution of a pseudobulk randomly generated from the Xin et al. dataset. For the first five iterations, we use iterative CVX steps for Ω , and then we switch to SLSQP for the final five iterations. (a) A plot of the negative log likelihood values (ignoring constant scaling factors) vs. alternating minimization iterations. (b) A plot of the squared L^2 error of the mixture proportion estimate, $\hat{\alpha}$, vs. alternating minimization iterations.

the resulting algebraic fractions in n :

$$0 = \left(\sum_{g=1}^G \frac{(\sum_{k=1}^K \alpha_k \Omega_{g,k})^2}{v_g(\alpha, \Omega, \sigma)} \right) n^2 + Gn - \sum_{g=1}^G \frac{(\hat{B}_g)^2}{v_g(\alpha, \Omega, \sigma)} \quad (2.11)$$

Thus, these steps do not require any explicit optimization scheme.

3. The optimization in line 13 proceed via gradient descent (or a variation thereof), and so could possibly require long run-times. However, the coarser maximization (minimization, cf. item 1) in lines 6 – 7 and 14 typically improves the objective function to such an extent that only two or three more iterations are required. Moreover, both sets of optimizations are amenable to parallelization.
4. Algorithm 1 straightforwardly generalizes to the setting of jointly inferring mixture proportions in an arbitrary number D of bulk samples (cf. Remark 3). Both our implementations support this generalized deconvolution.

Lastly, we note that although the alternating optimization in lines 5 – 7 is not guaranteed to converge; however, we observe empirically that the coarser maximization steps perform better with regards to converging to the true α . We observe in Figure 2.1 the objectives in each step have a smooth landscape with a single local minimum within the constraint set despite the objective being generally non-convex. However, we find that using local minimization programs such as Sequential Least Squares Quadratic Programming (SLSQP),

a constrained local minimization package [13], the mixture proportion estimate diverges from the true α . In Figure 2.2, we compare the use of SLSQP to find a local minimum in each α minimization step without fixing the mixture variances and iteratively using CVX (applying the OSQP solver [23]) to find a fixed point for the quadratic program where we fix the mixture variances. We find the local constrained optimization worsens our estimate both in the objective value as well as in squared L^2 error of the mixture proportion estimate, $\hat{\alpha}$. We conjecture that the poor performance of the local constrained minimization program can be attributed to termination at poor local minima in cases not shown in Figure 2.1. Otherwise, since σ is fixed throughout our algorithm, the global minimum for this objective will generally be mis-specified which may result in poor estimates for exact optimization programs. As a result, we choose to apply the coarser minimization steps which finds stronger empirical performance than local constrained optimization steps.

2.3 Results

Method Performance

To validate the ability of our method to accurately deconvolve heterogeneous mixtures, we applied RNA-Sieve to data falling into a wide range of different scenarios. While these mostly consist of the deconvolution of pseudobulk samples, we also demonstrate our model’s performance on several bulk data sets with strong biological priors for the cell type proportions.

We benchmark our approach on scRNA-seq data from several different tissues in the *Tabula Muris Senis* experiment [19]. This allows us to show that our performance is strong across a number of different biological settings, including different numbers of cell types and different degrees of similarity between cell types. In the liver, for example, we performed 15 different deconvolution of pseudobulk samples with six different cell type shown in Figure 2.3. Averaged over these trials, RNA-Sieve produced an mean absolute error of 0.00618 while MuSiC produced an mean absolute error of 0.00883.

We further highlight our performance by showing RNA-Sieve’s robustness in a number of different contrived situations which could plausibly occur in real settings. We briefly summarize our takeaways below, and full results are presented in the supplement.

Analysis of Beta Pancreatic Islet Cell Type Proportion in Healthy and T2D Human Samples

Glycated hemoglobin, HbA1c, levels rise in the presence of excess sugar in the bloodstream. HbA1c is typically used a diagnostic test for type 2 diabetes [7], a condition triggered by the failure and loss of beta islet cells in the pancreas [4]. With HbA1c concentration as a covariate, we estimate the beta islet cell proportion from bulk RNA-seq samples from subjects of varying diagnosis. We validate our method’s performance by showing our estimates are

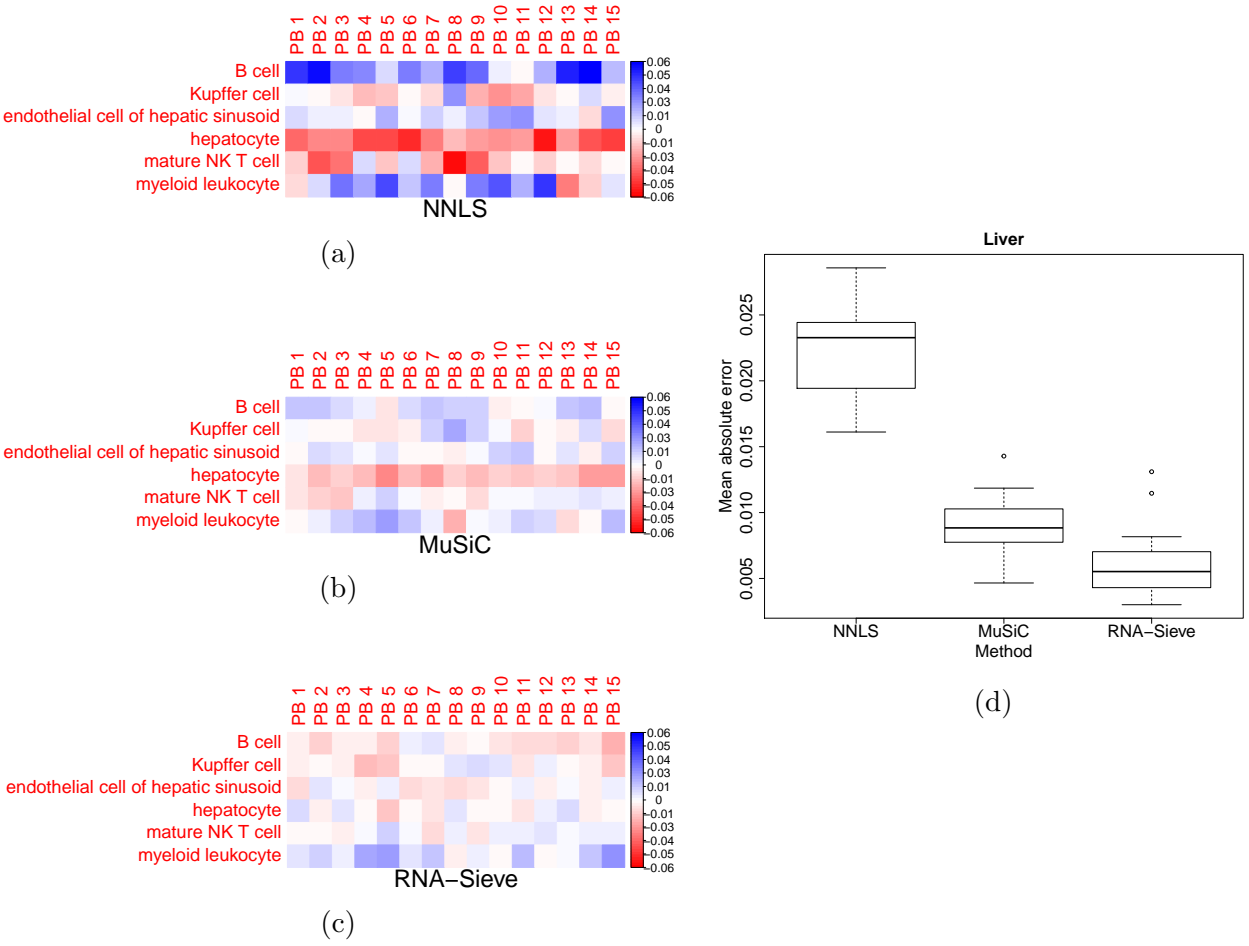


Figure 2.3: **Deconvolution results for liver pseudobulks.** Five references and 15 pseudobulks were randomly constructed by sampling cells from the liver; each reference was used for three pseudobulks. The matrices display the differences between estimated and true proportions while the boxplot shows the distribution of the mean absolute error across pseudobulk samples.

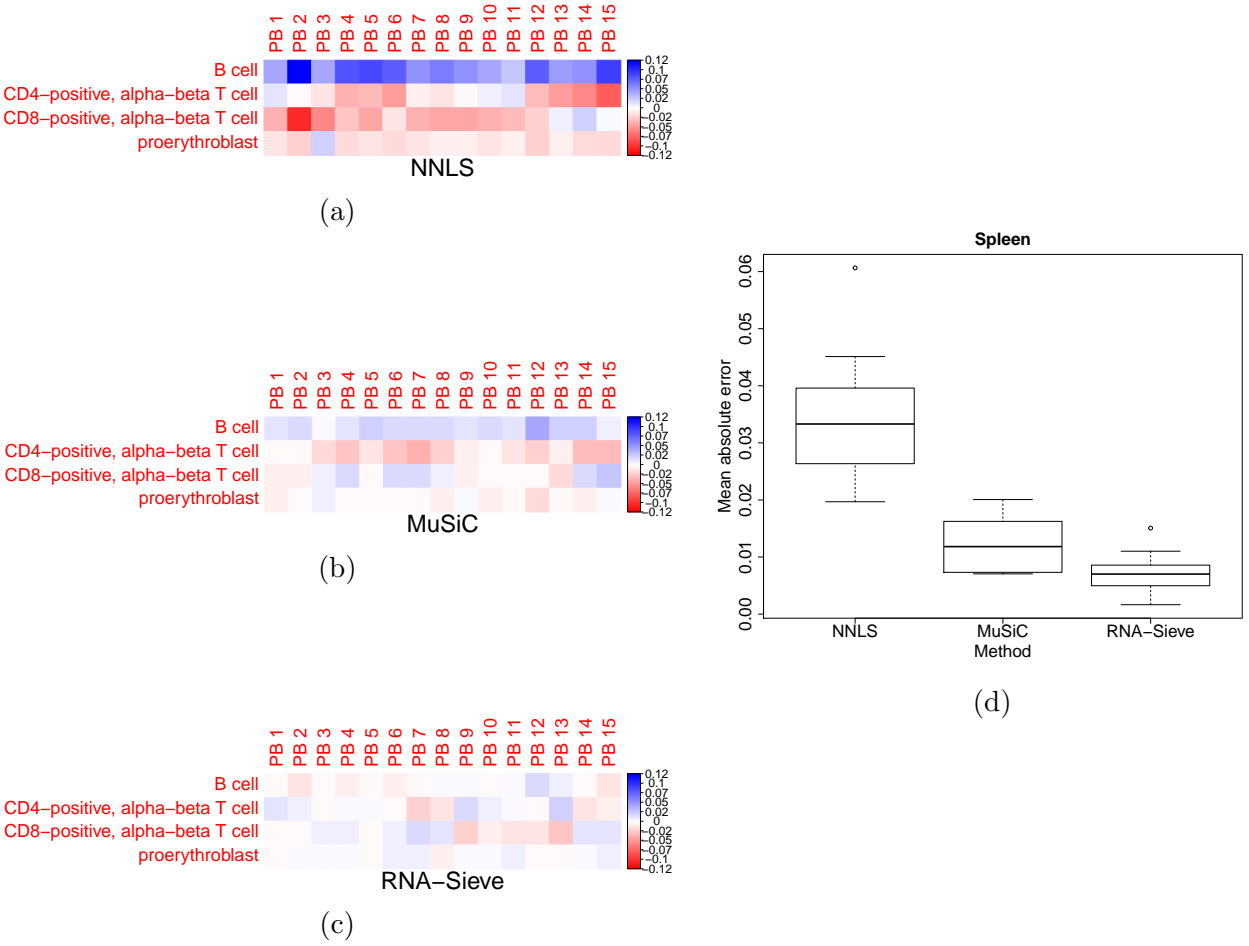


Figure 2.4: **Deconvolution results for spleen pseudobulks.** Five references and 15 pseudobulks were randomly constructed by sampling cells from the spleen; each reference was used for three pseudobulks. The matrices display the differences between estimated and true proportions while the boxplot shows the distribution of the mean absolute error across pseudobulk samples.

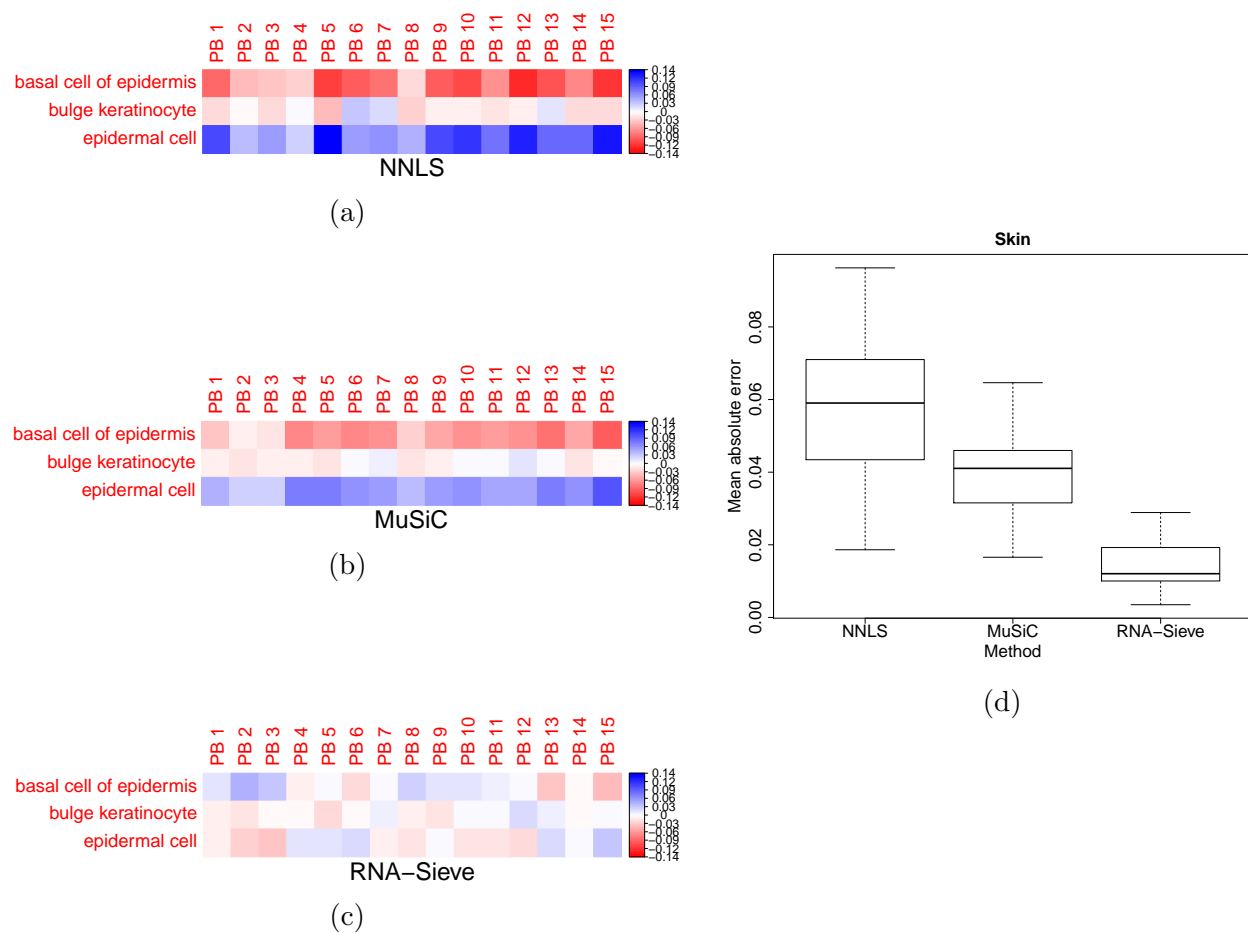


Figure 2.5: **Deconvolution results for skin pseudobulks.** Five references and 15 pseudobulks were randomly constructed by sampling cells from the spleen; each reference was used for three pseudobulks. The matrices display the differences between estimated and true proportions while the boxplot shows the distribution of the mean absolute error across pseudobulk samples.

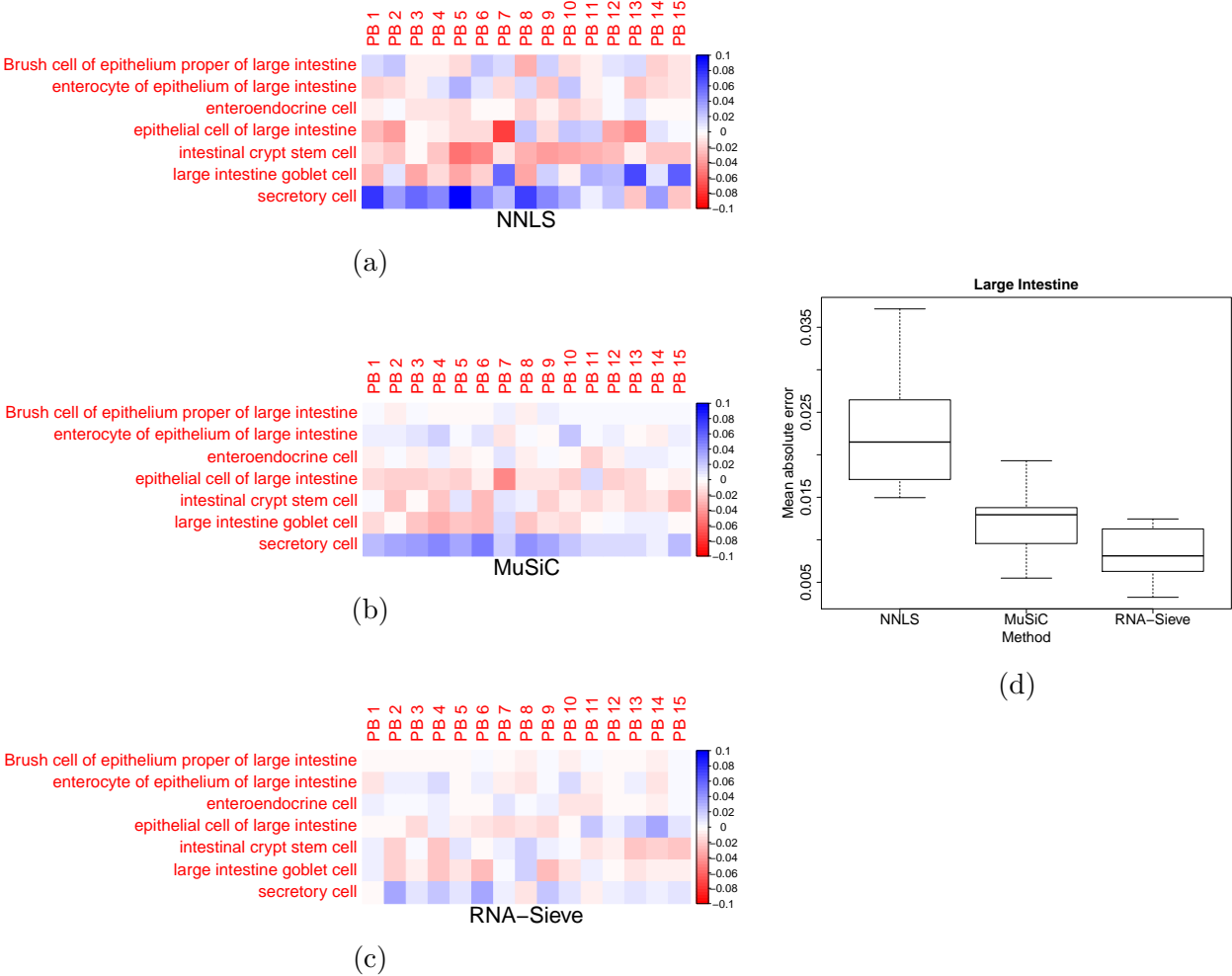


Figure 2.6: **Deconvolution results for large intestine pseudobulks.** Five references and 15 pseudobulks were randomly constructed by sampling cells from the spleen; each reference was used for three pseudobulks. The matrices display the differences between estimated and true proportions while the boxplot shows the distribution of the mean absolute error across pseudobulk samples.

inversely proportional to the recorded HbA1c concentrations, according to the underlying biological pattern.

As in [26], we applied our algorithm to the bulk RNA-seq data from Fadista et al [6]. Using single-cell data from Segerstolpe et al. [20] and Xin et al. [27] from both healthy and type-2 diabetes subjects as reference, we estimated the cell type decomposition of the bulk data from Fadista et al. [6].

As anticipated, we observe a negative correlation in Figure 2.7 between the HbA1c concentration and the estimated proportion of beta pancreatic islet cells in the bulk sample.

Analysis of the *Tabula Muris Senis*

Given the utility of our algorithm, we applied it to analyze the gene expression data of the *Tabula Muris Senis* experiment [19]. This unique data set contains bulk RNA-seq samples from many different tissues sampled at different ages of mouse and also includes single-cell measurements (both SmartSeq2 and Drop-seq) at a subset of these time points.

Using all of the single-cell measurements available for a given tissue as reference, we applied our algorithm to the bulk RNA-seq samples from the tissue. Since there is no ground truth for the cell type distribution of the bulk samples available for this data set, we judged the accuracy of our algorithm's estimates by the relative consistency of the proportion estimates across subjects of the same age group and the averages across age groups.

We highlight our algorithm's effectiveness in uncovering trends in cell type proportions across age groups. In Figure 2.8a, we observe a noticeable increase in skeletal muscle satellite cells and a significant decrease in the mesenchymal stem cell proportion in the muscle tissue of older mice. In Figure 2.8b, we observe a decrease in endothelial cell proportion in older mice in the mesenteric fat tissue. In both tissues, we observe consistency in proportion estimates both within age group and across age groups suggesting our estimates capture a significant signal found in common between the samples.

HbA1c levels vs. Beta cell proportion estimated by RNA-Sieve

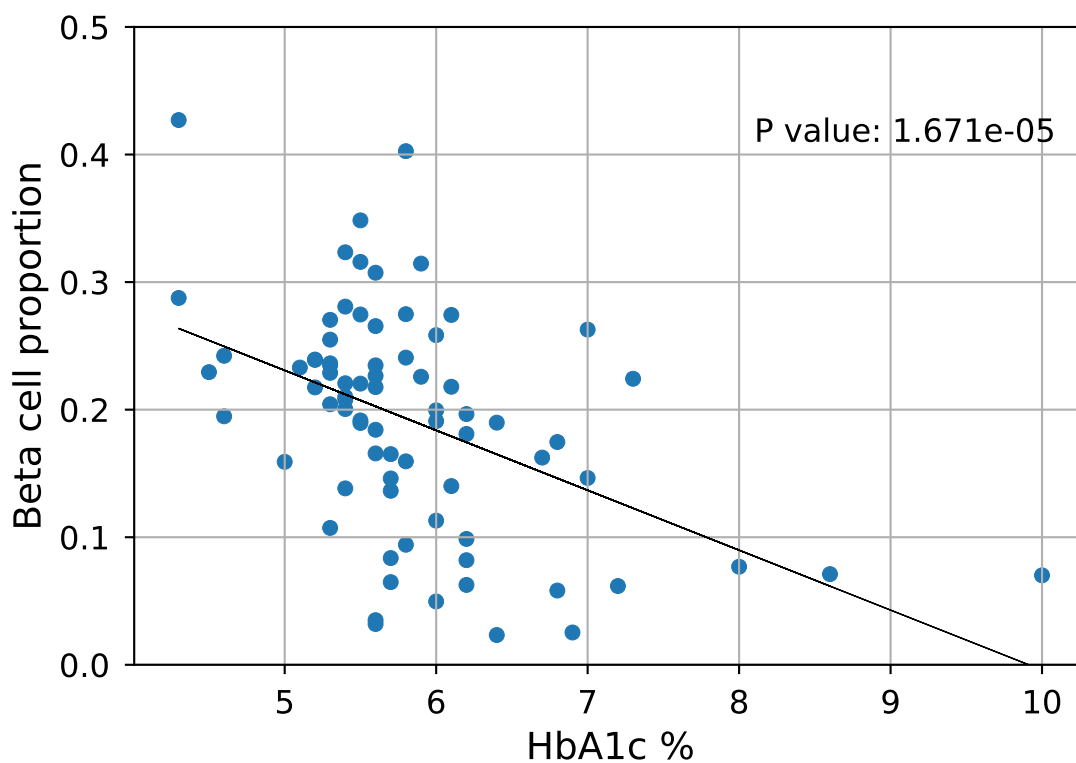


Figure 2.7: **Deconvolution results for Fadista et al. bulks.** Single-cell data was used as reference from Xin et al. and Segerstolpe et al. datasets. Each point represents the estimated beta pancreatic islet cell proportion one of 77 bulks with recorded HbA1c levels from Fadista et al. The p -value is for a single-variate regression on the estimated proportions.

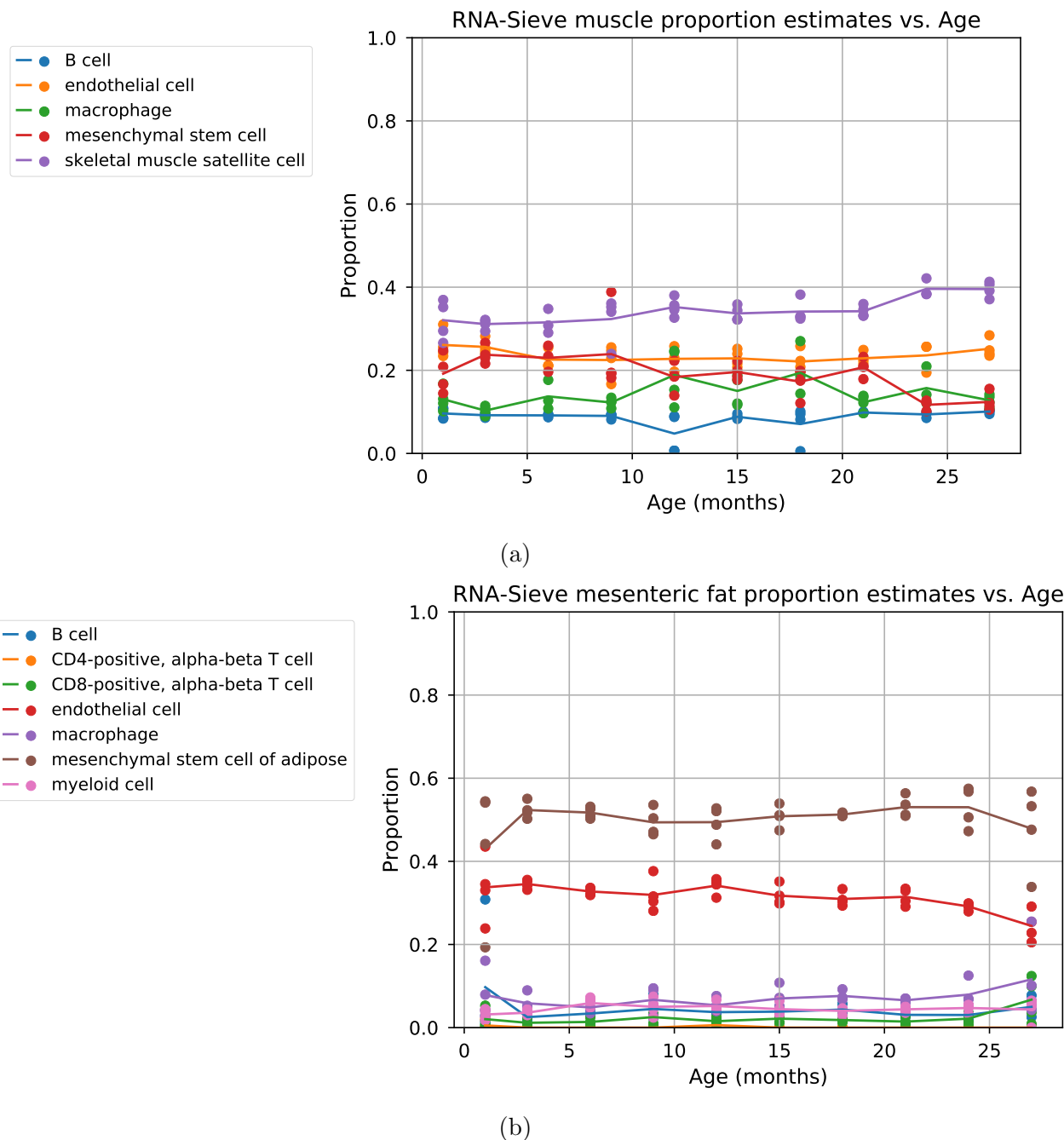


Figure 2.8: **Deconvolution results for *Tabula Muris Senis* bulks.** For each tissue, all the single-cell data available were used as a reference to estimate proportions in bulk RNA-seq samples across different age groups. Each point represents the proportion estimate for a given cell type and age group. The plotted line displays the trend in the average across individuals of an age group for a given cell type.

Chapter 3

Conclusion

3.1 Summary

In this paper, we introduce RNA-Sieve, a method for the deconvolution of bulk RNA-seq samples based on the likelihood of the data in a proposed generative model. We demonstrate the efficacy of our algorithm in both simulated and real settings compared against both leading and naive methods. Preliminary results on RNA-seq samples show our method can reveal biological trends in mixture proportion against age or other covariates. In addition to providing accurate mixture proportion estimates, RNA-Sieve can output intelligible parameter estimates for the proposed generative model.

3.2 Future Work

In this work, our empirical results mainly focused on the accuracy of the mixture proportion estimates. Unlike other methods which rely on variants of regression or the application of black-box machine learning algorithms, we place the deconvolution problem into a generative probabilistic framework by relying on distributional results from asymptotic theory. Our likelihood-based approach opens avenues for extensions which are intractable or difficult to pursue using current algorithms. Perhaps the most notable is a hypothesis test to determine whether the reference panel is missing cell types present in the bulk sample. While other methods have assessed their robustness to such misspecification, it would be beneficial to know whether the deconvolution performed was actually valid using a principled approach. Consequently, we have begun developing this test and are also undertaking a comprehensive evaluation of its performance while attempting to characterize the theoretical limits of the problem. In addition, it is possible to construct confidence regions by relying on the Fisher information matrix which can be computed from our likelihood. These confidence regions provide a more mathematically complete description of the certainty of estimates than marginal estimates of variance. We also note that the initialization of the cell type means and variances is separate from the rest of the algorithm, so it is simple to plug in

other estimation techniques which may be better suited for high-dimensional data and/or robust to outliers.

Additionally, modifications to the current method can be made to improve robustness to numerical instability or to account for additional variation. Our method utilizes optimization steps that do not have certain convergence guarantees, which could be improved by exploring other methods such as interior point methods. Furthermore, we fix the sample variances in our algorithm and aggregate our reference data without accounting for differences between individuals. Barring the potential infeasibility of addressing these simplifications, incorporating these additional details may improve performance.

Bibliography

- [1] Alexander R Abbas et al. “Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus”. In: *PloS one* 4.7 (2009).
- [2] Francisco Avila Cobos et al. “Computational deconvolution of transcriptomics data from mixed cell populations”. In: *Bioinformatics* 34.11 (2018), pp. 1969–1979.
- [3] Over Cabrera et al. “The unique cytoarchitecture of human pancreatic islets has implications for islet cell function”. In: *Proceedings of the National Academy of Sciences* 103.7 (2006), pp. 2334–2339.
- [4] Jae-Hyoung Cho et al. “ β -cell mass in people with type 2 diabetes”. In: *Journal of diabetes investigation* 2.1 (2011), pp. 6–17.
- [5] Yongjun Chu and David R Corey. “RNA sequencing: platform selection, experimental design, and data interpretation”. In: *Nucleic acid therapeutics* 22.4 (2012), pp. 271–274.
- [6] João Fadista et al. “Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism”. In: *Proceedings of the National Academy of Sciences* 111.38 (2014), pp. 13924–13929.
- [7] Chris Florkowski. “HbA1c as a diagnostic test for diabetes mellitus—reviewing the evidence”. In: *The Clinical Biochemist Reviews* 34.2 (2013), p. 75.
- [8] Amit Frishberg et al. “Cell composition analysis of bulk genomics using single-cell data”. In: *Nature methods* 16.4 (2019), pp. 327–332.
- [9] Megan Hastings Hagenauer et al. “Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis”. In: *PloS one* 13.7 (2018), e0200003.
- [10] Peng Hu et al. “Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq”. In: *Molecular cell* 68.5 (2017), pp. 1006–1015.
- [11] Gregory J Hunt et al. “dtangle: accurate and robust cell type deconvolution”. In: *Bioinformatics* 35.12 (2019), pp. 2093–2099.

- [12] Tomer Kalisky et al. “Analysis of Human Colon Tissue Cell Composition Using Single-Cell Gene-Expression PCR”. In: *Journal of biomolecular techniques: JBT* 24.Suppl (2013), S11.
- [13] D. Kraft. *A Software Package for Sequential Quadratic Programming*. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR, 1988. URL: <https://books.google.com/books?id=4rKaGwAACAAJ>.
- [14] Robert Lowe and Vardhman K Rakyan. “Correcting for cell-type composition bias in epigenome-wide association studies”. In: *Genome medicine* 6.3 (2014), p. 23.
- [15] Peng Lu, Aleksey Nakorchevskiy, and Edward Marcotte. “Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100 (Oct. 2003), pp. 10370–5. DOI: 10.1073/pnas.1832361100.
- [16] Kevin Menden et al. “Deep-learning-based cell composition analysis from tissue expression profiles.” In: *bioRxiv* (2019), p. 659227.
- [17] Richard A Moffitt et al. “Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma”. In: *Nature genetics* 47.10 (2015), p. 1168.
- [18] Aaron M Newman et al. “Determining cell type abundance and expression from bulk tissues with digital cytometry”. In: *Nature biotechnology* 37.7 (2019), pp. 773–782.
- [19] Angela Oliveira Pisco et al. “A Single cell transcriptomic atlas characterizes aging tissues in the mouse”. In: *bioRxiv* (2019), p. 661728.
- [20] Åsa Segerstolpe et al. “Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes”. In: *Cell metabolism* 24.4 (2016), pp. 593–607.
- [21] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. “Single-cell sequencing-based technologies will revolutionize whole-organism science”. In: *Nature Reviews Genetics* 14.9 (2013), pp. 618–630.
- [22] Yuh Shiwa et al. “Adjustment of cell-type composition minimizes systematic bias in blood DNA methylation profiles derived by DNA collection protocols”. In: *PloS one* 11.1 (2016), e0147519.
- [23] Bartolomeo Stellato et al. “OSQP: An Operator Splitting Solver for Quadratic Programs”. In: *Mathematical Programming Computation* (2020). DOI: 10.1007/s12532-020-00179-2. URL: <https://doi.org/10.1007/s12532-020-00179-2>.
- [24] Robert O Stuart et al. “In silico dissection of cell-type-associated patterns of gene expression in prostate cancer”. In: *Proceedings of the National Academy of Sciences* 101.2 (2004), pp. 615–620.
- [25] David Venet et al. “Separation of samples into their constituents using gene expression data”. In: *Bioinformatics* 17.suppl_1 (2001), S279–S287.

- [26] Xuran Wang et al. “Bulk tissue cell type deconvolution with multi-subject single-cell expression reference”. In: *Nature communications* 10.1 (2019), p. 380.
- [27] Yurong Xin et al. “RNA sequencing of single human islet cells reveals type 2 diabetes genes”. In: *Cell metabolism* 24.4 (2016), pp. 608–615.
- [28] Qianhui Yu and Zhisong He. “Comprehensive investigation of temporal and autism-associated cell type composition-dependent and independent gene expression changes in human brains”. In: *Scientific reports* 7.1 (2017), p. 4121.