Mathematical Modeling and Deconvolution for Molecular Characterization of Tissue Heterogeneity

Lulu Chen

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Computer Engineering

Yue Wang, Chair

Guoqiang Yu

William T. Baumann

T. Charles Clancy

Wenjing Lou

December 18th, 2019
Arlington, Virginia

Keywords: bioinformatics, deconvolution, unsupervised learning, convex analysis, feature selection, tissue heterogeneity, biomarkers

Mathematical Modeling and Deconvolution for Molecular Characterization of Tissue Heterogeneity

Lulu Chen

ABSTRACT (Academic)

Tissue heterogeneity, arising from intermingled cellular or tissue subtypes, significantly obscures the analyses of molecular expression data derived from complex tissues. Existing computational methods performing data deconvolution from mixed subtype signals almost exclusively rely on supervising information, requiring subtype-specific markers, the number of subtypes, or subtype compositions in individual samples. We develop a fully unsupervised deconvolution method to dissect complex tissues into molecularly distinctive tissue or cell subtypes directly from mixture expression profiles. We implement an R package, deconvolution by Convex Analysis of Mixtures (debCAM) that can automatically detect tissue or cell-specific markers, determine the number of constituent sub-types, calculate subtype proportions in individual samples, and estimate tissue/cell-specific expression profiles. We demonstrate the performance and biomedical utility of debCAM on gene expression, methylation, and proteomics data. With enhanced data preprocessing and prior knowledge incorporation, debCAM software tool will allow biologists to perform a deep and unbiased characterization of tissue remodeling in many biomedical contexts.

Purified expression profiles from physical experiments provide both ground truth and *a priori* information that can be used to validate unsupervised deconvolution results or improve supervision for various deconvolution methods. Detecting tissue or cell-specific expressed markers from purified expression profiles plays a critical role in molecularly characterizing and determining tissue or cell subtypes. Unfortunately, classic differential analysis assumes a convenient test statistic and associated null distribution that is inconsistent with the definition of markers and thus results in a high false positive rate or lower detection power. We describe a statistically-principled

marker detection method, One Versus Everyone Subtype Exclusively-expressed Genes (OVESEG) test, that estimates a mixture null distribution model by applying novel permutation schemes. Validated with realistic synthetic data sets on both type 1 error and detection power, OVESEG-test applied to benchmark gene expression data sets detects many known and *de novo* subtype-specific expressed markers. Subsequent supervised deconvolution results, obtained using markers detected by the OVESEG-test, showed superior performance when compared with popular peer methods.

While the current debCAM approach can dissect mixed signals from multiple samples into the ‘averaged’ expression profiles of subtypes, many subsequent molecular analyses of complex tissues require sample-specific deconvolution where each sample is a mixture of ‘individualized’ subtype expression profiles. The between-sample variation embedded in sample-specific subtype signals provides critical information for detecting subtype-specific molecular networks and uncovering hidden crosstalk. However, sample-specific deconvolution is an underdetermined and challenging problem because there are more variables than observations. We propose and develop debCAM2.0 to estimate sample-specific subtype signals by nuclear norm regularization, where the hyperparameter value is determined by random entry exclusion based cross-validation scheme. We also derive an efficient optimization approach based on ADMM to enable debCAM2.0 application in large-scale biological data analyses. Experimental results on realistic simulation data sets show that debCAM2.0 can successfully recover subtype-specific correlation networks that is unobtainable otherwise using existing deconvolution methods.

Mathematical Modeling and Deconvolution for Molecular Characterization of Tissue Heterogeneity

Lulu Chen

ABSTRACT (General Audience)

Tissue samples are essentially mixtures of tissue or cellular subtypes where the proportions of individual subtypes vary across different tissue samples. Data deconvolution aims to dissect tissue heterogeneity into biologically important subtypes, their proportions, and their marker genes. The physical solution to mitigate tissue heterogeneity is to isolate pure tissue components prior to molecular profiling. However, these experimental methods are time-consuming, expensive and may alter the expression values during isolation. Existing literature primarily focuses on supervised deconvolution methods which require *a priori* information. This approach has an inherent problem as it relies on the quality and accuracy of the *a priori* information. In this dissertation, we propose and develop a fully unsupervised deconvolution method - deconvolution by Convex Analysis of Mixtures (debCAM) that can estimate the mixing proportions and 'averaged' expression profiles of individual subtypes present in heterogeneous tissue samples. Furthermore, we also propose and develop debCAM2.0 that can estimate 'individualized' expression profiles of participating subtypes in complex tissue samples.

Subtype-specific expressed markers, or marker genes (MGs), serves as critical *a priori* information for supervised deconvolution. MGs are exclusively and consistently expressed in a particular tissue or cell subtype while detecting such unique MGs involving many subtypes constitutes a challenging task. We propose and develop a statistically-principled method - One Versus Everyone Subtype Exclusively-expressed Genes (OVESEG-test) for robust detection of MGs from purified profiles of many subtypes.

Acknowledgment

I would like to express my foremost thanks and deep gratitude to my advisor, Dr. Yue Wang. His constant support and encouragement throughout my Ph.D. work will always be remembered with great respect and admiration. I consider it a privilege to work closely with Dr. Wang in the Computation Bioinformatics & Bio-imaging Laboratory (CBIL). His sharp insights and profound knowledge helped me at various stages of my research. He is not only a great academic advisor but also a mentor and friend outside of the academic world. His caring attitude and profound words of wisdom have formed the finest part of my memories in graduate school, at Virginia Tech. He has always been patient and kind with me. I find his practical mindset very admirable. His full-of-life attitude and exuberance is something I wish to live my life by.

I would also like to extend my sincere gratitude to Dr. Guoqiang Yu, Dr. William T. Baumann, Dr. T. Charles Clancy and Dr. Wenjing Lou, who have generously served on my dissertation committee. Their insightful questions and helpful comments have helped a lot. I would also like to thank all of our collaborators at Wake Forest University, Cedars-Sinai Heart Institute, Georgetown University, Johns Hopkins Medical School and State University of New York Upstate Medical University for their generous help in the projects and publications.

My labmates at CBIL are more family than colleagues to me. We have shared our most pressing times, our greatest joys and very many adventures during my Ph.D. journey. Hiking, camping at the Shenandoah National Park and summer barbecues were some of the most fun times. Our regular 'Friday Group Meetings' were of invaluable support. Our discussions helped me get unstuck and advance in my research efforts. In particular, I want to thank Niya Wang, Chiung-Ting Wu, and Yingzhou Lu; their contributions to my dissertation are greatly appreciated.

There is a friend who sticks closer than a sister. I'm lucky to have found one such in Sneharaj Ramdaspalli. The mischief we had and the warm memories we share will be cherished for a lifetime. I look forward to realizing our beautiful future plans, with great excitement.

I would like to thank my dearest parents for their unconditional love and for always being there for me. Their immense patience and silent support have helped me scale the peaks I have never imagined.

Table of Contents

List of figures

Chapter 1

Introduction

1.1 Background and motivation

New high-throughput measurement technologies for biological molecules, *e.g.* RNA and proteins, have revolutionized biomedical research in recent decades. These technologies can simultaneously measure the expression abundance of thousands of molecules in tissues, casting light on differential diagnosis and monitoring of pharmacological efficacy [1, 2].

However, extensive yet complex heterogeneity has been revealed in tissues where multiple cell or tissue types are variably intermingled [1, 3]. Many diseases evolve with dynamic tissue remodeling, including changes in the subtype composition and changes in subtype-specific expressions, as well as interactions between subtypes [3, 4]. Global profiling can neither distinguish among the contributions of each subtype to the total measured signals nor identify differentially expressed molecules among different subtypes. Therefore, tissue heterogeneity becomes both a major confounding factor to obscure the studies of individual cellular subpopulations [1, 5, 6] and an underexploited information source in characterizing whole tissue systems [3, 4, 7]. Having analytic tools to define the molecular landscape of tissue heterogeneity and crosstalk, and to determine how subtypes are remodeled with phenotypic transitions, will be essential for the next step in systems biology research.

A physical solution to mitigate tissue heterogeneity is to isolate pure tissue components prior to molecular profiling, such as cell sorting [8] and tissue microdissection [9]. However, these experimental methods are expensive, time-consuming, and may alter the expression values during isolation [5, 6]. Many supervised methods have been reported to resolve tissue heterogeneity computationally, with *a priori* information on the constituent proportion [5, 10], purified expression profiles [11, 12] or markers of subtypes [4, 6, 7, 13]. Acquiring such prior information either from expensive physical experiments or from limited public datasets, these surprised methods rely on the quality of *a priori* information and also face the difficulty detecting subtypes that are subtle, condition-specific or previously unknown.

It is necessary to develop a blind source separation (BSS) algorithm which can deconvolute subtype-specific molecule expression profiles from mixed observations without any priors. Non-negative independent component analysis (nICA) [14] and non-negative matrix factorization (NMF) [15] are classical BSS algorithms. However, NICA requires that sources are mutually statistically independent, while most biological processes are highly correlated. NMF has no constraints on data distribution, but it easily gets stuck in local optima with various initial conditions and its deconvoluted signals are usually not biologically meaningful.

We have developed the first completely unsupervised deconvolution algorithm (deconvolution by convex analysis of mixtures – debCAM) that can estimate the mixing proportions by identifying vertices in scatter simplex of mixture observations and subsequently recover the underlying subtype-specific signals [16]. The molecules located at each vertex are exactly molecular markers whose expression patterns are most indicative of each identified subtype. debCAM requires no prior information on the number, composition, or profiling of any subtype present in mixture samples, and does not require the presence of pure subtypes among available samples.

1.2 Objectives and Statement of Problems

1.2.1 Unsupervised deconvolution

Most of the large datasets available for disease study are measured from complex tissue samples with unknown subtypes and where no supervising information is available. debCAM is capable of leveraging the mixing diversity across heterogeneous samples to distinguish between phenotypically similar subtypes, and thus provide new insight into old datasets. In practical data analysis, the difficulty of scatter simplex identification in debCAM could be increased by many factors, *e.g.* noise, outliers, high dimensionality, and batch effect. We need to enhance raw data preprocessing to ensure debCAM works well in real scenarios. Background knowledge or *a prior* information is necessary to explain the discoveries from debCAM and even can be integrated into debCAM framework to boost the performance.

Accordingly, for the debCAM applications to real datasets, we would like to seek the solutions to the following tasks:

1. Improve data preprocessing and optimal simplex identification to assure robust performance and the effective application of the entire debCAM algorithm pipeline;
2. Integrate supervising information, *e.g.*, known markers, to support semi-supervised learning for more plausible discovery;
3. Demonstrate real biomedical utilities of debCAM tool on molecular expression data, e.g. gene expression, proteomics, and methylation data;
4. Apply debCAM to real in-house datasets to help biologists uncover novel and context-specific markers and hidden molecularly-distinct subtypes in complex tissues.

1.2.2 Robust detection of subtype-specific markers

Subtype-specific markers are defined as being exclusively and consistently expressed in a particular tissue or cell subtype across varying conditions. Markers identified from tissue or cell-specific molecular expression profiles, as *a prior* information, can not only facilitate supervised deconvolution but also help determine whether novel markers deducted by debCAM from mixture tissues are associated with known subtypes or novel ones. Nowadays, biologists invest huge money on purified cell line profiling and even single-cell profiling. However, detecting MGs using purified profiles remains a challenging task. The most frequently used method is One-Versus-Rest Fold Change (OVR-FC), which has been demonstrated to be less effective than our earlier work, One-Versus-Everyone Fold Change (OVE-FC), in classifications [17]. Furthermore, there is no suitable statistical test method that explicitly matches the definition of markers and thus assesses the significance level of markers accurately. We proposed to extend our OVE strategy to a statistically-principled marker detection method, One Versus Everyone Subtype Exclusively-expressed Gene test (OVESEG-test), as a supplementary to OVE-FC in robust detection of subtype-specific markers.

For developing OVESEG-test method, we set the following research objectives:

1. Design OVESEG-test test statistic that mathematically matches the definition of subtype-specific markers;

2. Propose a new permutation scheme to estimate the corresponding distribution under the null hypothesis and evaluate the significance level of markers;
3. Validate the performance of OVESEG-test on extensive simulation data, in terms of type 1 error rate, False Discovery Rate (FDR), partial area under the receiver operating characteristic curve (pAUC), and in comparison with top peer methods;
4. Demonstrate the utility of OVESEG-test by applying it to benchmark public data, and assessing the performance by comparing with known markers;
5. Assess the performance of OVESEG-test by the accuracy of supervised deconvolution that uses the *de novo* markers detected by OVESEG-test.

1.2.3 Sample-specific deconvolution

While the current debCAM algorithm can dissect mixed signals of multiple samples into the ‘common’ expression profiles of subtypes, many subsequent molecular analyses of complex tissues require sample-specific signal deconvolution where each sample is a mixture of ‘individualized’ subtype-specific expression profiles. The ability to obtain between-sample variation carried by sample-specific signals, for detecting specific biological associations or networks in different subtypes, is both novel and timely [18, 19]. Such sample-specific Blind Source Separation (sBSS) problems have a much larger number of variables than the number of observations, with high variance and overfitting becoming a major concern. People usually adopt simple and highly regularized approaches according to practical application scenarios. We propose a new algorithm called debCAM2.0 as an extension of debCAM to solve sBSS problem for molecule expression data analyses.

For sample-specific deconvolution, we set the following research objectives:

1. Formulate the objective function for sample-specific deconvolution problem and design regularization terms that mathematically match the expected subtype-specific expression patterns;
2. Solve the problem with efficient optimization methods to make debCAM2.0 applicable in large-scale biological data;
3. Design a cross-validation strategy to decide the penalty parameter to avoid underfitting and overfitting;

4. Validate the performance of debCAM2.0 algorithm using simulations, by comparing the recovered sample-specific signals and recognized function modules to the ground truth.

1.3 Outline of the dissertation

The remainder of this dissertation is organized as follows. In chapter 2, we introduce the framework of unsupervised deconvolution algorithm, debCAM, and its supported principles, implementation details, validation in real benchmark datasets, as well as application in biological projects in which debCAM provides novel insights. In chapter 3, we extend OVE-FC to OVESEG-test, with a novel permutation scheme to evaluate the significance level of markers. We evaluate OVESEG-test via type 1 error control, FDR control, detection power and surpervised deconvolution performance. In chapter 4, we formulate the objective function for sample-specific deconvolution method, debCAM2.0, and introduce an efficient optimization method based on ADMM and a cross-validation strategy to determine the regularization parameter. Finally, the major contributions in this dissertation and the future work are summarized in chapter 5.

Chapter 2

Unsupervised deconvolution

2.1 Introduction

The existing machine learning methods for signal deconvolution are either supervised or unsupervised with certain assumptions. Many supervised methods have been applied to resolve tissue heterogeneity, with *a priori* information on the constituent proportion [5, 10], purified expression profiles [11, 12] or subtype markers [4, 6, 7, 13], based on linear latent variable model (Fig. 2.1). However, their performances rely on the quality of prior knowledge and are poor in finding subtypes that are condition-specific due to different microenvironments, or previously undetectable by physical methods. Other unsupervised approaches to solve blind source separation (BSS) problems have their own limits in biological applications. For instance, non-negative independent component analysis (nICA) [14] requires that original sources should be independent and non-negative matrix factorization (NMF) [15] decomposes mixtures into two matrices without plausible biological interpretations.

Motivated by the parallelism between the latent variable model and the theory of convex sets, we have developed a novel method to solve nonnegative BSS problems, called Convex Analysis of Mixtures (CAM) [20-22]. Within the CAM framework, the independent/uncorrelated assumption for underlying sources is no longer required. Instead, the CAM principle is derived from the assumption of well-grounded points among the sources [20, 21], which can be connected to the concept of molecular markers whose expressions are exclusively enriched in a specific subtype. Similar ideas also appear in works [23, 24] that analyze the convex set in sample space. CAM operates in molecular scatter space. Unsupervised deconvolution in sample space requires the presence of pure samples from each subtype, which is unlikely within complex tissues. On the contrary, only a limited number of subtype-specific markers are needed to guarantee the successful deconvolution by CAM.

We have also developed the first fully unsupervised deconvolution method, namely, deconvolution by Convex Analysis of Mixtures (debCAM), to dissect complex tissues into molecularly distinctive tissue or cell subtypes based on bulk expression profiles [22, 25, 26]. Importantly, debCAM requires no *a priori* information on the number, signatures, or compositions of the subtypes present in heterogeneous samples, and does not require pure subtype references. Supported by a well-grounded mathematical framework [22, 25], debCAM automatically detects tissue/cell-specific markers, determines the number of constituent subtypes, calculates subtype proportions in individual samples, and estimates tissue/cell-specific expression profiles.

To help biologists to conduct comprehensive analysis and characterization of tissue heterogeneity, we released the debCAM R package in Bioconductor that implemented and tested the latest functionalities of the debCAM algorithm pipeline in the literature. We have demonstrated real biomedical utilities of debCAM tool on gene expression [22], proteomics [27], imaging [28], and methylation data. These applications have led to novel findings and hypotheses. The debCAM R package also added several new features to our old CAM tool: (1) Enhanced data preprocessing, *e.g.,* dimension reduction and outlier filtering to assure robust performance and the effective application of the entire debCAM algorithm pipeline; (2) Improved vertex detection by minimizing reconstruction errors; (3) Provided a new 'speed-up' option using a greedy search method; (4) Visualization of high-dimensional simplex while preserving vertices; (5) Identification of all markers by one-versus-everyone fold change statistics with bootstrapping; (6) Support for supervised/semi-supervised deconvolution; (7) Enable methylation data deconvolution. This Chapter will introduce the mathematical foundation and technical details of the implemented debCAM algorithm, followed by multiple application examples using real data and validation with ground truth or cross-validation.

2.2 Method

2.2.1 Formulation and Notations

Most computational deconvolution methods are based on a linear latent variable model $\mathbf{X} = \mathbf{AS}$ (Fig. 2.1), where $\mathbf{X} \in \mathbb{R}^{M \times L}$ is the observation matrix containing M mixed signals with L dimensions, $\mathbf{A} \in \mathbb{R}^{M \times K}$ is the unknown mixing matrix, and $\mathbf{S} \in \mathbb{R}_{\geq 0}^{K \times N}$ is the unknown source matrix containing K non-negative source signals with L dimensions[5, 11, 21, 22]. The goal of blind source separation (BSS) is to deconvolute mixing matrix $\mathbf{A}$ and source matrix $\mathbf{S}$ from observation matrix $\mathbf{X}$ without any prior information. We can rewrite the model in vector-matrix notation as

$$\boldsymbol{x}(j) = \mathbf{A}\boldsymbol{s}(j) = \sum_{k=1}^{K} \boldsymbol{a}_k s_k(j)\,, \mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_K], j = 1,2, \dots L, \tag{2.1}$$

where $\boldsymbol{x}(j), \boldsymbol{a}_k, \boldsymbol{s}(j)$ are the column vectors of matrices $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{S}$, respectively, and $s_k(j)$ denotes the kth entry in $\boldsymbol{s}(j)$. When this model is applied for molecular expression data for tissue heterogeneity characterization, row vectors in $\mathbf{X}$ represent the measurements of M samples mixed by K underlying subtypes, and row vectors in $\mathbf{S}$ represent the pure subtype-specific expression profiles. A column vector in $\mathbf{X}$ or $\mathbf{S}$ is a data point corresponding to one expression unit (molecular feature).

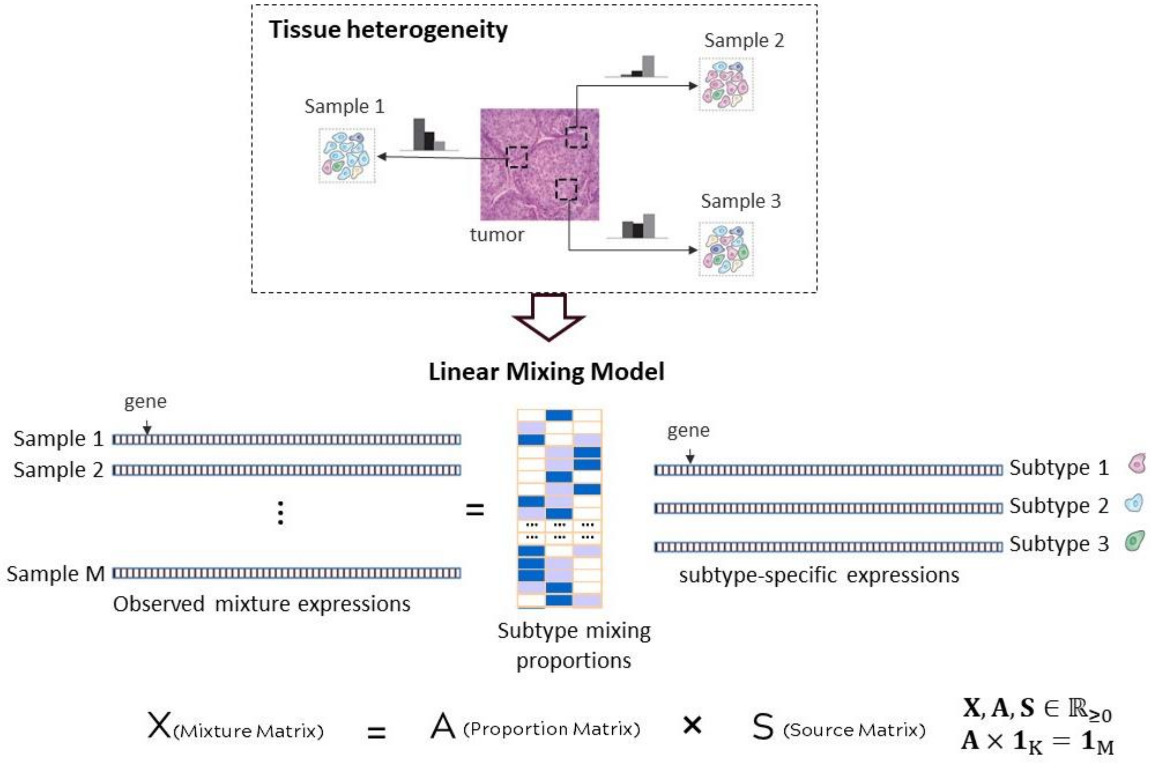

Tissue heterogeneity
Sample 1
Sample 2
Sample 3
tumor
Linear Mixing Model
gene
Sample 1
Sample 2
Sample M
Observed mixture expressions
=
Subtype mixing proportions
gene
Subtype 1
Subtype 2
Subtype 3
subtype-specific expressions
X (Mixture Matrix) = A (Proportion Matrix) × S (Source Matrix)
$\mathbf{X}, \mathbf{A}, \mathbf{S} \in \mathbb{R}_{\geq 0}$
$\mathbf{A} \times \mathbf{1}_\mathrm{K} = \mathbf{1}_\mathrm{M}$

Figure 2.1 Linear latent variable model for signal deconvolution to characterize tissue heterogeneity, *e.g.* tumor tissues are modeled as the mixtures of multiple cell types with varying mixing proportions and measured gene expressions are linearly combined by cell-type specific expressions.

In this model, we have assumed linear mixing and source non-negativity, which can be widely satisfied in real-world problems including molecular expression data. Since the entries in one row of mixing matrix represent the proportions of constituent sources in one mixture, there is an implied limitation that each row in $\mathbf{A}$ must sum to one, i.e. $\sum_{j=1}^{K} a_{ij} = 1$, given that the mixtures have been well normalized. Besides, $\mathbf{A}$ is of full column rank, i.e. rank $(\mathbf{A})$=K, which is a fundamental requirement for separating K sources/subtypes in model identification problems.

2.2.2 Mathematical Foundation

The linear latent variable model (Eq. 2.1) can be linked to the theory of convex sets in two different views. One is observed in the sample space, where the scatter plot of samples in high-dimensional space are generated by using molecule features (*e.g.* gene expressions) as dimensions (Fig. 2.2a), and the other in scatter space, where the scatter plot of molecule features are generated by using samples as dimensions (Fig. 2.2b).

In the sample space, all samples can be demonstrated to be located within a convex hull (simplex), because each sample's expression vector ($\boldsymbol{x}_i, i = 1, \dots, M$) is a linear combination of subtypes' expression vectors ($\boldsymbol{s}_k, k = 1, \dots, K$) with non-negative and sum-to-one weights (Fig. 2.2a). The extreme points, i.e. vertices of a convex set, are exactly the expression vectors of pure subtypes. The samples located at the vertices are pure samples containing only one subtype. Thus, if the observed samples are enough for us to identify the convex hull, *i.e.* to locate the vertices, we can recover the subtypes' expression vectors. This idea has been applied to other BSS problems when there are sufficient samples among which some are close to pure subtypes [20, 29, 30]. However, most molecular expression datasets only measure a limited number of tissues that are mixture samples far away from pure subtypes. Therefore, it is impossible to recover the locations of pure subtypes in the sample space by an unsupervised approach.

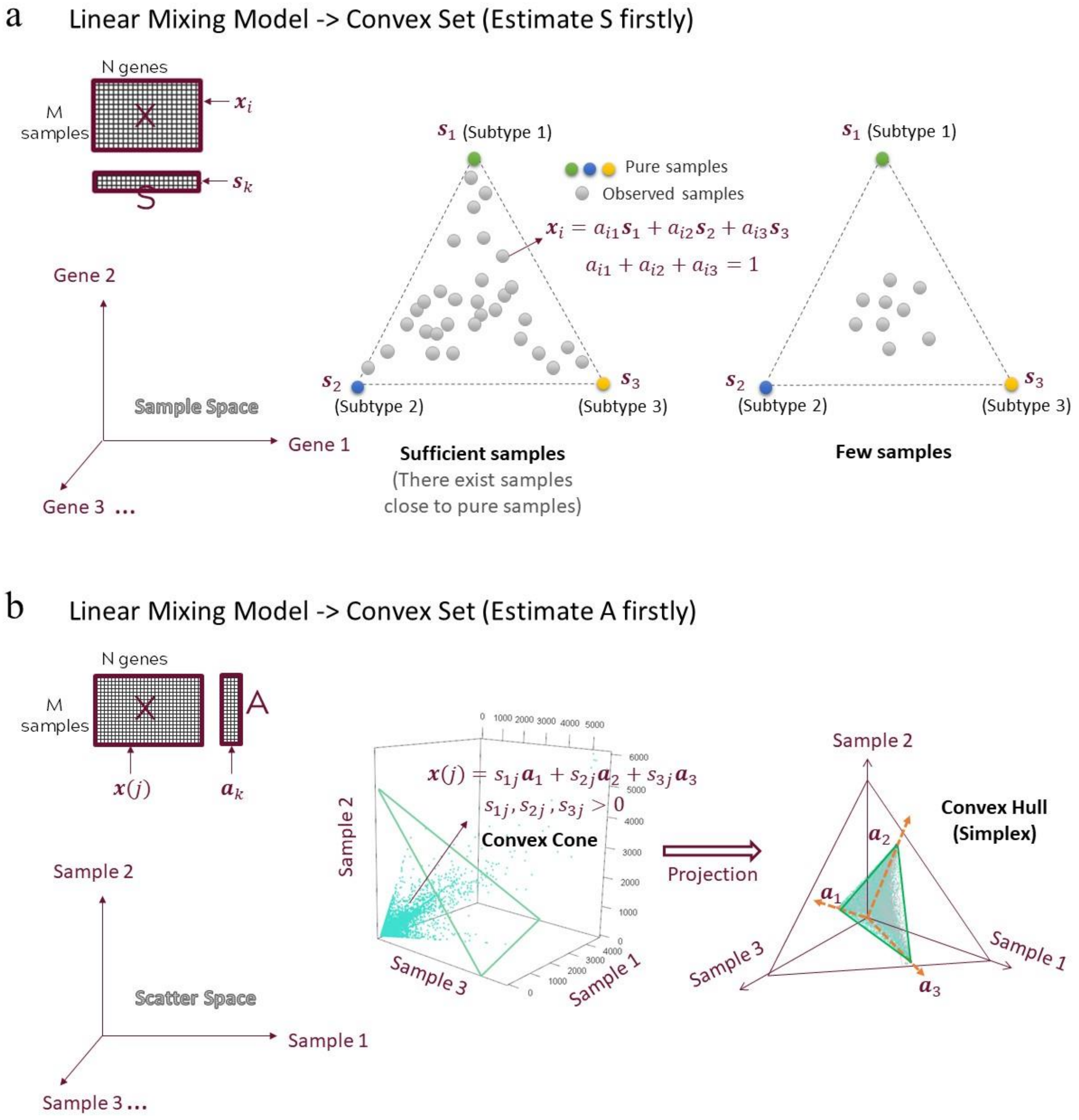


Figure 2.2 The parallelism between linear latent variable model and convex set theory in (a) sample space where samples are confined within a simplex, or (b) scatter space where genes are confined within a convex cone and projected to be within a simplex.

Our novel insight comes from the scatter space. All molecule features (*e.g.* gene expressions) can be proved to be located within a convex cone, because each molecule's expression vector ($\boldsymbol{x}(j), j = 1, \dots, N$) is a linear combination of $\mathbf{A}$ matrix's column vectors ($\boldsymbol{a}_k, k = 1, \dots, K$) with non-negative weights (Fig. 2.2b). After perspective projection, all molecule features are also located within a convex hull (or a simplex when using a certain projection hyperplane). The vertices of the convex hull (simplex) formed by $\boldsymbol{x}(j), j = 1, \dots, N$, coincide with $\boldsymbol{a}_k, k = 1, \dots, K$ with appropriate rescaling (**Theorem 1**) [22]. When $\boldsymbol{a}_k$'s are linearly independent and molecule features are enough to identify the projected convex hull (simplex), the directions of $\boldsymbol{a}_k$'s can be recovered as the locations of the vertices, with magnitudes yet unknown. Luckily, magnitudes of $\boldsymbol{a}_k$'s can be recovered based on the sum-to-one constraints for each row in **A**.

The scatter simplex formed by $\boldsymbol{x}(j), j = 1, \dots, N$, can also be regarded as a rotated and compressed version of the scatter simplex formed by $\boldsymbol{s}(j), j = 1, \dots, N$, where **A** works like a geometric transformation matrix in the mixing process (Fig. 2.3) (**Lemma 1**) [22]. Once **A** matrix is identified, **S** scatter simplex can be recovered from **X** scatter simplex. Compared to much fewer samples presented in sample space, there are usually a huge number of molecule features presented in scatter space to help the identification of simplex, which sheds light on the success of unsupervised deconvolution of heterogeneous tissues.

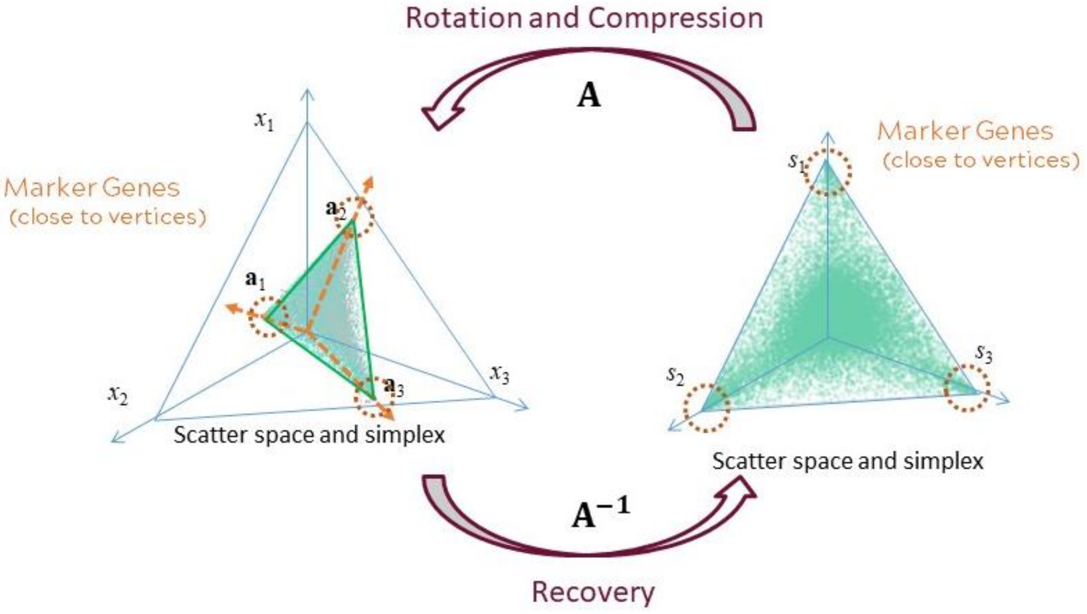


Figure 2.3 Geometry of the mixing operation that rotates and compresses the simplex in scatter space whose vertices coincide with mixing proportions and reside subtype-specific markers.

The molecule features located at the kth vertice of the scatter simplex of **X** or **S** fulfill $s_k(j) > 0$ while $s_i(j) = 0, \forall i \neq k$, such that $\boldsymbol{s}(j) = \lambda \boldsymbol{e}_k, \lambda > 0$, where $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_K\}$ is the standard basis of K-dimensional space (Fig. 2.3). These special molecule features are called well-ground points (WGPs) in general BSS problems. The sources are well-grounded when there exists at least one data point in each source with significantly high values relative to the remaining sources. Molecular expression data are usually well-grounded with many subtype-specific markers which have high expression value in one subtype, but low in the other ones [5]. These markers, serving as well ground points, are usually called molecular markers, molecular signatures or tissue/cell-specific marker genes/proteins according to different biological backgrounds. The existence of markers guarantees the identification of simplex in the scatter space and motivates us to propose the first fully unsupervised deconvolution algorithm - deconvolution by Convex Analysis of Mixtures (debCAM). We will conduct more discussions on subtype-specific marker detection in supervised methods in Chapter 3.

2.2.3 debCAM

Under the aforementioned assumptions: (1) linear mixing, (2) non-negativity of entries in **S**, (3) sum-to-one constraints in each row of **A**, (4) full rank of **A**, and (5) the existence of well-grounded points in each source, we proposed the mathematical principle of debCAM in ideal noise-free scenarios: well-grounded points can be blindly detected by identifying the vertices of multifaceted simplex that most tightly encloses the scatter plot of **X**; then the expression vectors of well-grounded points can be used to estimate **A** matrix and further recover **S** matrix; the number of vertices is exactly the number of underlying subtypes in the mixtures.

Under real scenarios of biological data, to identify vertices of a noisy simplex in the high-dimensional scatter space is yet a challenging task. We have designed the following major analytics pipelines in debCAM (Fig. 2.4) to assure robust performance and effective applications:

(1) **Data preprocessing**. Molecule features whose expression levels are lower or higher than a pre-fixed threshold are removed as noise or outliers. The number of mixture profiles is reduced by principal component analysis (PCA). On the scatter simplex constructed by perspective projection and local outlier factoring (LOF [31]), molecule features are aggregated into representative cluster centers using K-means or Affinity Propagation

Clustering (APC [32]). The interior cluster centers are removed by QuickHull [33] and a greedy/floating research algorithm [34].

(2) **Simplex and Marker detection**. An exhaustive combinatorial search is conducted to identify the optimal scatter simplex for a given number of vertices among the peripheral cluster centers, guided by a convex-hull-to-data fitting criterion. The members of the identified simplex vertex clusters are considered subtype-specific markers.

(3) **Deconvolution of mixed expression profiles**. The mixing proportions are first estimated by the collective expression levels of subtype-specific markers, followed by non-negative least squares estimation of subtype-specific expression profiles.

(4) **Model selection**. The optimal number of subtypes present in samples is determined by the MDL information criterion among the competing models.

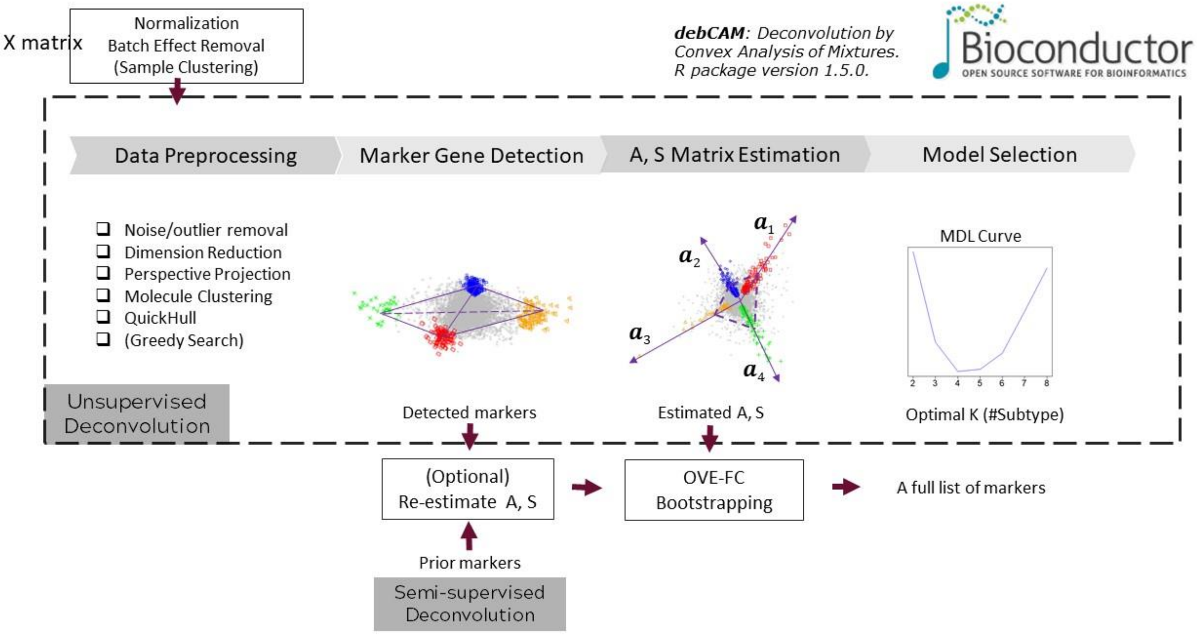


Figure 2.4 The unsupervised deconvolution workflow in debCAM with four major functional modules: data preprocessing, simplex and marker detection, deconvolution of mixed expression profiles, and model selection. Combining blindly detected markers and available prior markers can achieve supervised/semi-supervised deconvolution workflow.

Before debCAM, we also need to normalize data and remove batch effects to reduce non-biological variations. Sample clustering is an optional method to reduce the input mixtures and possibly make input mixtures more balanced. The final input data must be in non-log linear space with non-negative entries. Beside the unsupervised deconvolution pipeline, replacing blindly detected markers with available *a priori* markers, or combining both types of markers can achieve supervised/semi-supervised deconvolution. The complete description of each procedure implementation will be described in Section 2.3.

2.2.4 UNDO2.0

UNsupervised DecOnvolution version 2 (UNDO2.0) software tool [35] is the spin-off from the debCAM framework for deconvolving mixtures of two subtypes in heterogeneous tissues [36]. UNDO2.0 can automatically detect cell-specific markers located on the scatter radii (convex cone without perspective projection) of mixed gene expressions, estimate mixing proportions in each sample, and deconvolve mixed expressions into subtype-specific expression profiles. Because the deconvolution is limited to only two subtypes, UNDO2.0 produces the two most differential subpopulations, while not necessarily the most dominant ones. Our experimental results show that data preprocessing may have a significant impact on deconvolution. For example, when sample clustering is performed prior UNDO2.0, the input mixtures would represent dominant subpopulations; while extracted mixtures by e.g. principal component analysis may favor the most differential subpopulations. Furthermore, feature selection (e.g., genes) based on *a priori* would produce the subtypes that may be focused on certain biological questions.

2.3 Implementation

The debCAM package implemented the pipeline in R language and is freely available at the Bioconductor repository. The package provides core functions (Fig. 2.5) to support each procedure in the unsupervised deconvolution workflow (Fig. 2.4) and auxiliary functions, e.g. simplexplot(), to help understand results. The package also implements functions to perform supervised deconvolution based on prior knowledge of molecular markers, **A** matrix or **S** matrix. Semi-supervised deconvolution can be achieved by combining molecular markers from debCAM with prior knowledge to analyze mixture expressions. The following subsections give more details about the requirements of input data and the techniques used in each procedure. The package and code is available in Bioconductor repository:

http://bioconductor.org/packages/debCAM

User manual:

https://bioconductor.org/packages/release/bioc/manuals/debCAM/man/debCAM.pdf

Vignette and examples:

https://bioconductor.org/packages/release/bioc/vignettes/debCAM/inst/doc/debcam.html

2.3.1 Data Input

Normalization and batch effect removal are conventional procedures to reduce non-biological variations in molecular expression profiles prior to any further analysis. Besides, well-designed normalization will support the assumption that each row in **A** sums to one; removing batch effects is critical to identify the valid scatter simplex for unsupervised deconvolution.

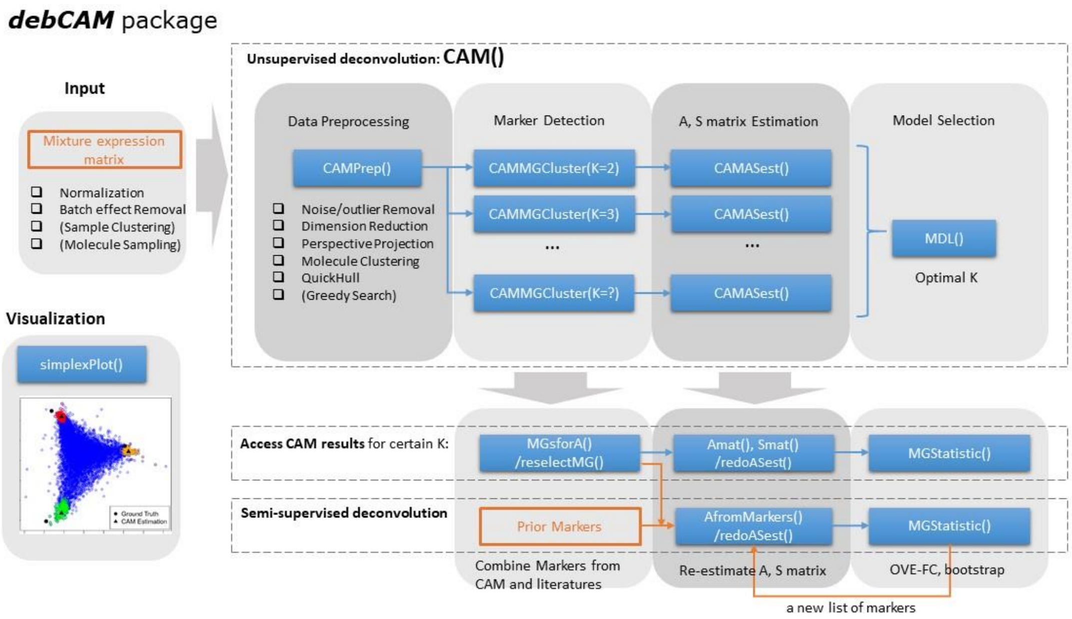


Figure 2.5 The structure and functions implemented in the debCAM package, including necessary or optional procedures for unsupervised/semi-unsupervised deconvolution and visualization.

Normalization

Many normalization pipelines or packages have been developed to reduce technical variations in a specific molecular measurement platform, *e.g.*, RMA and PLIER for Affymetrix Microarray, SWAN and BMIQ for Infinium HumanMethylation450 array. Most of the public dataset repositories provide normalized data and the description of normalization workflow.

Batch Effect Removal

Batch effect is the systematic error introduced by the time and place-dependent experimental variations. ComBat [37], a popular method to remove batch effects, models the expressions of gene j with the additive and multiplicative batch effects:

$$\boldsymbol{x}_i(j) = \mathbf{A}_i\boldsymbol{s}(j) + \gamma_i(j)\mathbf{1}_{M_i} + \delta_i(j)\boldsymbol{\varepsilon}_{M_i}, \tag{2.2}$$

where i is the batch index, $\gamma_i(j)$ and $\delta_i(j)$ are additive and multiplicative batch effects of batch i. M_i is the number of samples/tissues in batch i. $\mathbf{1}_{M_i}$ is a vector of M_i ones and $\boldsymbol{\varepsilon}_{M_i}$ is a vector of M_i measurement errors. Each column in matrix $\mathbf{A}_i$, i.e. proportions of underlying subtypes, are variables of interest that should be preserved while removing batch effects. Combining b batches, we get

$$\boldsymbol{x}(j) = \mathbf{A}\boldsymbol{s}(j) + \mathbf{B}\boldsymbol{\gamma}(j) + \mathbf{E}\boldsymbol{\delta}(j), \tag{2.3}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{1}_{M_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{M_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{M_b} \end{bmatrix}, \boldsymbol{\gamma}(j) = \begin{bmatrix} \gamma_1(j) \\ \gamma_2(j) \\ \vdots \\ \gamma_b(j) \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} \boldsymbol{\varepsilon}_{M_1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\varepsilon}_{M_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\varepsilon}_{M_b} \end{bmatrix}, \boldsymbol{\delta}(j) = \begin{bmatrix} \delta_1(j) \\ \delta_2(j) \\ \vdots \\ \delta_b(j) \end{bmatrix}$$

Here, $\gamma_i(j)$ and $\delta_i(j)$ could be positive and negative. Let $sgn\{\boldsymbol{v}\}$ denote the vector of signs of entries in $\boldsymbol{v}$ and $abs\{\boldsymbol{v}\}$ the vector of absolute values of entries in $\boldsymbol{v}$. Extracting signs of $\gamma_i(j)$ and $\delta_i(j)$, we can get

$$\begin{aligned}\boldsymbol{x}(j) &= \mathbf{A}\boldsymbol{s}(j) + \mathbf{B}sgn\{\boldsymbol{\gamma}(j)\}abs\{\boldsymbol{\gamma}(j)\} + \mathbf{E}sgn\{\boldsymbol{\delta}(j)\}abs\{\boldsymbol{\beta}(j)\} \\ &= [\mathbf{A}, \mathbf{B}sgn\{\boldsymbol{\gamma}(j)\}, \mathbf{E}sgn\{\boldsymbol{\delta}(j)\}]\begin{bmatrix} \boldsymbol{s}(j)) \\ abs\{\boldsymbol{\gamma}(j)\} \\ abs\{\boldsymbol{\beta}(j)\} \end{bmatrix}\end{aligned} \tag{2.4}$$

which implies that every observed mixture expression data point $\boldsymbol{x}(j)$ is confined within the simplex (convex hull) defined by column vectors in $[\mathbf{A}, \mathbf{B}sgn\{\boldsymbol{\gamma}(j)\}, \mathbf{E}sgn\{\boldsymbol{\delta}(j)\}]$, *i.e.,* $\boldsymbol{a}_k$ and batch-associated vectors in either positive or negative directions (**Lemma 1**) [22]. Thus, without batch effect removal, detected vertices from the simplex might coincide with column vectors of **A**, hosting those marker genes barely affected by batch effects, or coincide with a linear combination of column vectors in **B** and **E**, hosting those genes significantly up/down-regulated in the certain batch(es).

ComBat can remove batch effects orthogonal to known variables of interest. However, subtype-specific proportions are yet unknown. Directly applying ComBat to the data before deconvolution may remove meaningful variations unless the batch effects are not correlated with proportion changes. For example, experiments were well-designed to generate similar constituent distributions in each batch. We can reorganize the dataset by manually picking suitable tissue samples or batch(es) for data deconvolution. If well-designed datasets are not available, we suggest the following steps:

- Step 1: Perform debCAM on data (no batch effect removal) to detect all vertices in simplex;
- Step 2: Identify those vertices highly correlated with batch effects (represented by the design matrix of categorical variables) and remove them from vertex sets;
- Step 3: Treat the remaining vertices as vectors associated with proportion dynamics and use them as variables of interest in batch effect removal by ComBat;
- Step 4: Perform debCAM on data after batch effect removal.

Sample Clustering (optional)

Performing affinity propagation (AP) clustering [32] among samples and using resulted cluster exemplars/means as the input can reduce data dimensionality. The new data after sample clustering should have more balanced proportions among mixtures, which may decrease the condition number of the proportion matrix and lead to a more robust deconvolution. However, sample clustering also increases the risk that subtle subtypes could be omitted.

Molecule Sampling (used in methylation data)

Molecule sampling is used to select a portion of expression units randomly as the input, as long as debCAM can detect vertices successfully from the downsampled scatter simplex. Molecule sampling is useful when limited computer memory cannot load all data points or the run-time is too long to operate all data points, *e.g.*, a very large number of CpG sites in Methylation 450K and Methylation EPIC BeadChip.

2.3.2 Unsupervised Deconvolution Workflow

Noise/Outlier Removal

We use $\|\boldsymbol{x}(j)\|_2$, L2 norm of expression vectors, to quantify the signal intensity of molecule j. Molecules with low/high signal intensity could be noise/outliers and disrupt the scatter simplex. debCAM filters these molecules with thresholds that depend on the data platform. Typically, 30% ~ 50% low-expressed genes can be removed from gene expression data. Fewer low-expressed proteins are removed, *e.g.*, 0% ~ 10%, due to the limited number of proteins in proteomics data. Removal of high-expressed molecules has much less impact on the results and is usually set at 0% ~ 10%.

Dimension Reduction

Data points, $\boldsymbol{x}(j)$, have M dimensions where M is the number of tissue samples. High dimensionality will slow subsequent analyses, *e.g.*, vertex detection. Sample clustering prior to debCAM is one choice to reduce dimension. We also implemented PCA within the debCAM pipeline to reduce dimension and reduce noise. When PCA is enabled, debCAM tends to catch the subtypes with significantly varied proportions.

Perspective Projection

Without PCA to reduce dimension, perspective projection is to project all data points into the hyperplane $\boldsymbol{H} = \{\boldsymbol{x} \in \mathbb{R}^n | \mathbf{1}^\mathrm{T}\boldsymbol{x} = 1\}$ to form a simplex, which is equivalent to scaling each vector $\boldsymbol{x}(j)$ by its sum. If PCA has been applied, data points are rotated and thus no longer confined within the hyperspace $\mathbb{R}_{+}^{n}$. Thus, we use the first principal component as our projection direction instead of using $\mathbf{1}$.

Molecule Clustering and Outlier Filtering

Aggregation of data points into representative clusters can further reduce the impact of noise/outliers and the time needed for subsequent vertex detection. AP clustering is initialization-free and can obtain a near-global-optimum [32]. However, convergence is slow when clustering more than a few thousand data points. K-means clustering is faster but sensitive to the selection of the initial clusters. We recommend using AP clustering for proteomics data and K-means with multiple runs for gene expression data and methylation data. The final cluster centers are located at the space median of the cluster members, estimated by the "l1median" function in R package "pcaPP"[38].

Filtering outliers far away from other data points in the scatter simplex can avoid them being detected mistakenly as vertices. Before molecule clustering, we apply LOF [31] to filter density-based local outliers. After molecule clustering, those clusters with only a few data points are also filtered as outliers.

QuickHull

Since cluster centers within the simplex are not likely to be vertices, we can apply the QuickHull [33] algorithm to exclude cluster centers located within the convex hull that encloses all cluster centers. The remaining cluster centers are vertex candidates to be searched to find the most probable vertices.

Greedy/floating Search (optional)

If the number of vertex candidates remaining after QuickHull is still large, *e.g.*, >40, we suggest using a greedy search algorithm, Sequential Floating Forward Selection (SFFS) [34], prior to vertex detection to narrow the search space. SFFS was designed to select feature subsets effectively. We modified SFFS to select a subset of vertex candidates by the defining cost function as the sum of margin-of-errors between the convex hull and the remaining "exterior" cluster centers:

$$\Delta_j^{(1)} = \min_{\theta_1,\ldots,\theta_K \in R_+} \left\| \boldsymbol{\mu}_j - \sum_{k=1}^{K} \theta_k \boldsymbol{\mu}_{C_k} \right\|_2, \sum_{k=1}^{K} \theta_k = 1, j = 1,2,\ldots,J, \tag{2.5}$$

where $\boldsymbol{\mu}_j, j = 1, \ldots, J$ is all the cluster centers, $\{\boldsymbol{\mu}_{C_1}, \boldsymbol{\mu}_{C_2}, \ldots, \boldsymbol{\mu}_{C_K}\}$ is a subset of vertex candidates after QuickHull .

The subset of vertices is initialized as the empty set. At each step, a new vertex is added yielding the minimum cost. Subsequently, the algorithm searches for vertices that can be removed from the optimal subset until the cost does not decrease. This greedy search process is fast but cannot guarantee the global optimum. Thus, we use SFFS to reduce vertex candidates to a limited number yet larger than the number assumed to be present, *e.g.,* 20 candidates after SFFS while 10 is the target. The tradeoff is between the time complexity and the risk of excluding true vertices.

Marker Detection

After the data preprocessing procedures above, we assumed K true vertices and conducted an exhaustive combinatorial search among the candidates based on convex-hull-to-data fitting criterion. Minimizing the sum of margin-of-errors (Eq. 2.5) or minimizing the sum of reconstruction-errors:

$$\Delta_j^{(2)} = \min_{\theta_1,\ldots,\theta_K \in R_+} \left\| \boldsymbol{x}(j) - \sum_{k=1}^{K} \theta_k \boldsymbol{\mu}_{C_k} \right\|_2, j \in D, \tag{2.6}$$

where D is the set of indexes of data points left after outlier filtering, we identify the most probable K clusters whose centers are the locations of vertices and whose members are detected markers:

$$(C_1^*, C_2^*, \ldots, C_{\mathrm{K}}^*) = \underset{(C_1, C_2 \ldots, C_K)}{\operatorname{argmin}} \sum_{j=1}^{J} \Delta_j^{(1)} \; or \; \underset{(C_1, C_2 \ldots, C_K)}{\operatorname{argmin}} \sum_{j \in D} \Delta_j^{(2)} \; . \tag{2.7}$$

The former criterion only depends on cluster centers and is insensitive to outliers. The latter criterion makes a decision based on all data points and so is less affected by unstable clustering. This latter criterion considers data point distribution in scatter space before projection and thus favors subtypes abundant in the mixture.

A, S Matrix Estimation

A matrix is estimated as $\left[\lambda_1\boldsymbol{\mu}_{C_1^*}, \lambda_2\boldsymbol{\mu}_{C_2^*}, \dots, \lambda_K\boldsymbol{\mu}_{C_K^*}\right]$, each column vector coinciding with the detected simplex vertices. If PCA has been applied, it needs an extra transformation back to the original space. The scales $\lambda_k, k = 1, \dots, K$, are recovered by row sum-to-one constraints:

$$\min_{\lambda_1,\dots,\lambda_K \in R_+} \left\| \mathbf{P}^T \mathbf{1}_M - \sum_{k=1}^{K} \lambda_k \boldsymbol{\mu}_{C_k} \right\|_2^2 \tag{2.8}$$

where **P** is the transformation matrix if PCA has been applied, otherwise it is the identity matrix. The resulting **A** matrix is then used to deconvolute the **X** matrix into **S** matrix by Non-negative Least-Square regression.

Re-estimation

Considering data preprocessing may have filtered some biologically-meaningful markers, it is necessary to check each molecule again to identify all the possible markers. Two statistics based on the estimated subtype-specific expressions in **S** matrix are used to select markers with certain thresholds. The first is OVE-FC (one versus everyone - fold change) [17]. The second is the lower confidence bound of bootstrapped OVE-FC at α level.

It is optional to refine **A** and **S** matrix by Alternating Least Squares (ALS) method to further reduce the mean squared error. The constraint for methylation data, $\mathbf{S} \in [0,1]$, will be imposed during this re-estimation process. Note that allowing too many iterations of ALS may cause a significant deviation from initial values.

Model Selection

debCAM exploits Minimum Description Length (MDL) [39], a widely-adopted and consistent information theoretic criterion, to guide model selection. MDL aims to find the optimal model that assigns high probabilities to the observed data and keeps the model simple. The underlying subtype number will be decided by minimizing the total description code length:

$$MDL(K) = -\,log\left(\mathcal{L}\left(\mathbf{X}\middle|\boldsymbol{\Theta}(K)\right)\right) + \frac{(K-1)M}{2} log(N) + \frac{KN}{2} log(M)\,, \tag{2.9}$$

where $\mathcal{L}(\cdot)$ denotes the joint likelihood function of observations, i.e. **X**, with estimated parameters Θ, i.e. **A** and **S**, conditioned on the given model, i.e. K. The first term (data term) represents the probability of those observations under the given model and corresponds to description length of the model fitting error. The first item is estimated by

$$-\log\left(\mathcal{L}\left(\mathbf{X}\middle|\boldsymbol{\Theta}(K)\right)\right) \propto \frac{NM}{2}\log\left(\hat{\sigma}^2(K)\right) = \frac{NM}{2}\log\left(\frac{1}{NM}\sum_{j=1}^{N}\|\boldsymbol{x}(j) - \mathbf{A}\boldsymbol{s}(j)\|_2^2\right). \qquad (2.10)$$

The second item corresponds to the description length of **A** matrix with $(K-1)M$ free variables decided by N data points. The third term corresponds to the description length of **S** matrix with KN free variables decided by M mixtures. The second and third terms represent the penalty on the model complexity.

2.3.3 Supervised/Semi-supervised Deconvolution

The debCAM algorithm decomposes an observation matrix based on blindly detected markers. Its framework can also achieve supervised deconvolution either by replacing blindly detected markers with *a priori* markers, or by semi-supervised deconvolution that combines blindly detected markers and *a priori* markers. The only step that needs modification is that each column vector in **A** matrix will locate at the space median of a new set of markers in the scatter simplex:

$$\tilde{\boldsymbol{a}}_k = \frac{\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|_1} = SpaceMedian\left\{\frac{\boldsymbol{x}(j_{Marker\text{-}k})}{\|\boldsymbol{x}(j_{Marker\text{-}k})\|_1}\right\}, \qquad (2.11)$$

where $\tilde{\boldsymbol{a}}_k$ is scaled $\boldsymbol{a}_k$ and $j_{Marker\text{-}k}$ is the set of marker indexes of subtype k. The scale, $\|\boldsymbol{a}_k\|_1$, can be estimated based on row sum-to-one constraint (Eq. 2.8). The conventional supervised deconvolution based on prior **A** matrix or **S** matrix is also supported in the package by applying Alternating Least Squares on markers.

2.3.4 Notes on methylation data deconvolution

Deconvolution on methylation beta values needs extra manipulations, some of which have been mentioned above.

(1) Perform ALS to refine **A** and **S** matrix with truncating **S** entries in the interval [0,1].

(2) Unmethylation values (1-beta) have similar characters as methylation values (beta) and also follow a linear mixing model:

$$\mathbf{X}^{(1-\beta)} = \mathbf{1}_{\mathrm{M}\times\mathrm{N}} - \mathbf{X}^{(\beta)} = \mathbf{A}\left(\mathbf{1}_{\mathrm{K}\times\mathrm{N}} - \mathbf{S}^{(\beta)}\right) = \mathbf{A}\mathbf{S}^{(1-\beta)}. \tag{2.12}$$

Therefore, we can also apply debCAM to (1-beta) matrix. Estimated proportions should be the same as those estimated from beta matrix. Estimated subtype-specific expressions are unmethylation levels. Detected unmethylation markers have exclusively low methylation levels in a particular subtype. When the quality and/or quantity of unmethylation markers is better than methylation markers, we suggest using (1-beta) values as the input data of debCAM.

(3) It is necessary to downsample CpG sites and enable SFFS for quick vertex selection when the input is Methylation 450K or Methylation EPIC data. After **A** matrix is recovered, subtype-specific expressions for all CpG sites can be estimated by least squares.

2.3.5 Simplex Visualization

Visualizing the simplex of the high-dimensional scatter space can help us understand the data better and interpret results more easily. However, human visual system is considered as "2.5D" and data visualization is typically restricted to two dimensions. The common methods to visualize

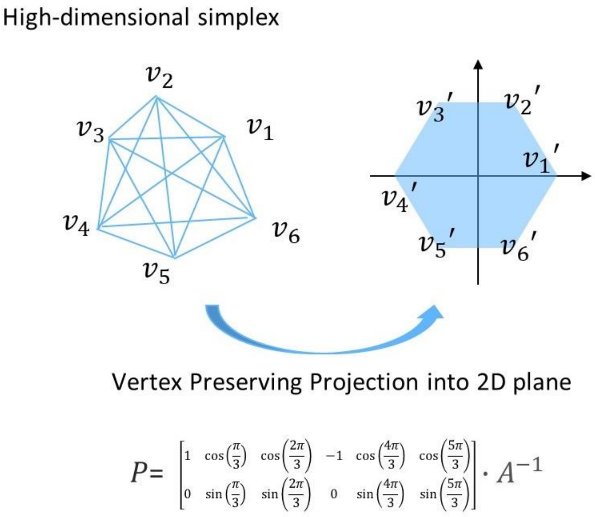

High-dimensional simplex
v_2
v_3
v_1
v_4
v_6
v_5
v_3'
v_2'
v_1'
v_4'
v_5'
v_6'
Vertex Preserving Projection into 2D plane
$P= \begin{bmatrix} 1 & \cos\left(\frac{\pi}{3}\right) & \cos\left(\frac{2\pi}{3}\right) & -1 & \cos\left(\frac{4\pi}{3}\right) & \cos\left(\frac{5\pi}{3}\right) \\ 0 & \sin\left(\frac{\pi}{3}\right) & \sin\left(\frac{2\pi}{3}\right) & 0 & \sin\left(\frac{4\pi}{3}\right) & \sin\left(\frac{5\pi}{3}\right) \end{bmatrix} \cdot A^{-1}$

Figure 2.6 Visualization of high-dimensional simplex in a 2D plane by vertex preserving projection.

high-dimensional data in a 2D plane include PCA, LDA, MDS and t-SNE. However, none of these methods can assure the preservation of simplex vertices after projection. We designed a transformation matrix

$$\mathbf{P} = \begin{bmatrix} 1 & \cos\left(\frac{2\pi}{K}\right) & \cos\left(\frac{2\pi}{K} * 2\right) & \dots & \cos\left(\frac{2\pi}{K} * (K-1)\right) \\ 0 & \sin\left(\frac{2\pi}{K}\right) & \sin\left(\frac{2\pi}{K} * 2\right) & \dots & \sin\left(\frac{2\pi}{K} * (K-1)\right) \end{bmatrix} * \widehat{\mathbf{A}}^{-1}$$

which projects the simplex in $\mathbb{R}^M$ in to a regular K-sided polygon in $\mathbb{R}^2$ whose K vertices are located at the coordinates given by the left matrix of **P**, and $\widehat{\mathbf{A}}$ is the estimated proportion matrix. To make the shape of simplex projected into a 2D plane close to that in high-dimensional space, two row-vectors in **P** are orthogonalized and normalized by Gram–Schmidt process. We call this dimension reduction method as Vertex Preserving Projection (VPP). Through this linear transformation method, the vertices of the high-dimensional simplex will still locate at extreme points of the new simplex in 2D plane and all interior points within the simplex in original scatter space will still be confined within the new simplex.

As BSS problems have inherent permutation ambiguity, the order of column vectors of estimated $\widehat{\mathbf{A}}$ can be changed to generate a different 2D simplex plot. It is better to choose an order which can maximize the distance between vertices and thus show a clear simplex.

2.4 Results

To demonstrate the wide application and biomedical utility of debCAM, we use datasets of three molecular omics types (mRNA, methylation, protein) and three validation schemes (gold standard, ground truth, and cross-validation) to assess the performance of debCAM R package.

2.4.1 Validation on RNA mixtures of immune and cancer cell lines

We first tested debCAM on biologically mixed gene expression profiles (GSE64385). The 12 samples represent the mixtures of five immune cell subtypes and one cancer cell line in various known proportions. Immune cell populations were sorted from healthy donors' peripheral blood and cancer cell population was from HCT116 colon cancer cell line [40].

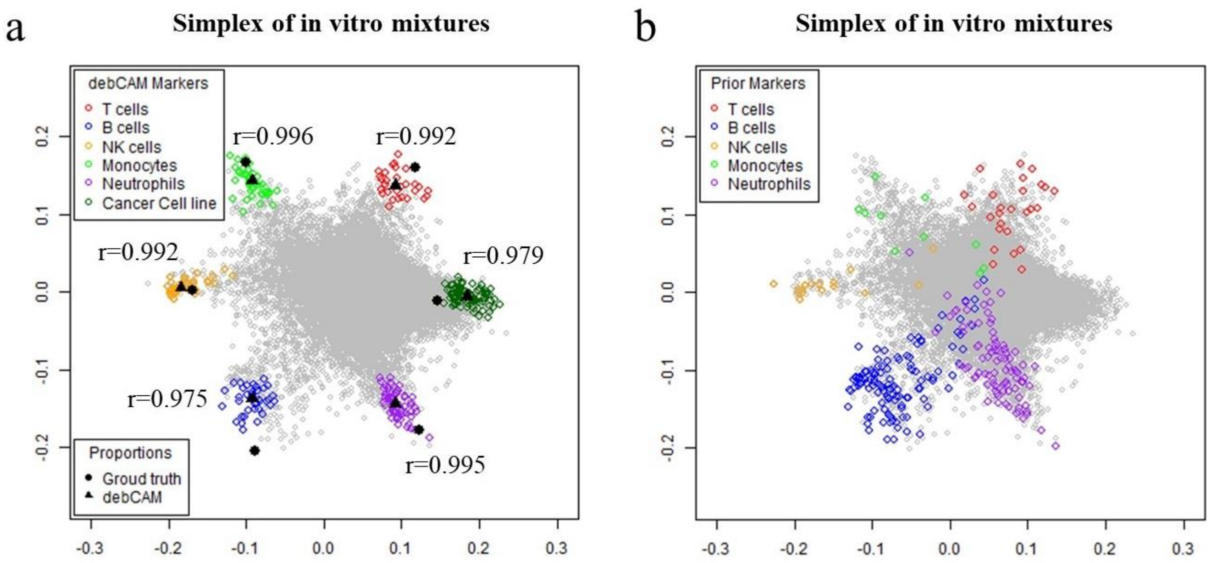


Figure 2.7 (a) Scatter simplex of 12 in vitro mixtures shows six vertices, hosting the subtype-specific markers (detected by debCAM). Column vectors of ground-truth/estimated proportion matrix reside around the vertices. (b) *A priori* known markers (five immune cell types) superimposed onto the scatter simplex are closely resided around the corresponding vertices.

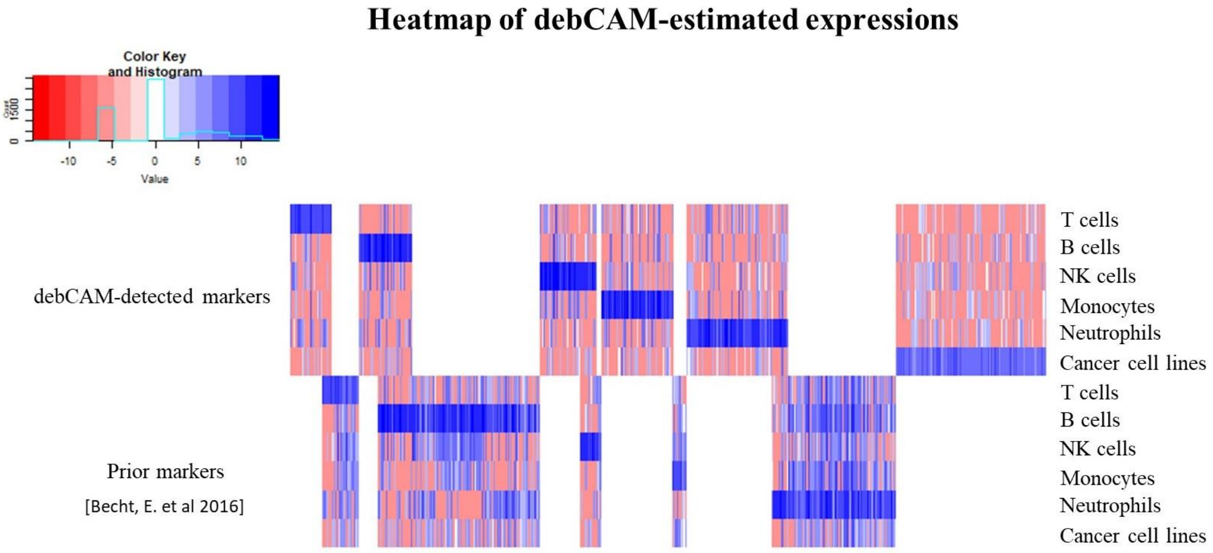


Figure 2.8 Heatmap of debCAM-estimated expressions over debCAM-detected markers of 6 underlying cell types or prior markers of 5 immune cell types. debCAM-detected markers and most of prior markers show their exclusively enriched expressions in the corresponding cell type.

The optimal scatter simplex blindly identified by debCAM is given in Fig. 2.7a, showing six distinctive vertices. To validate the accuracy of subtype-specific markers blindly detected by debCAM, we superimpose the color-coded known markers in the scatter simplex showing close proximity to the corresponding vertices (Fig. 2.7b, the immune cells' prior markers were derived from purified expression profiles in other experiments [40]; the cancer cell line-specific markers are unavailable for comparison). More convincingly, the sample-wise subtype proportions estimated by debCAM match the ground truth almost perfectly, with correlation coefficients of 0.975~0.996.

The debCAM-detected markers and most prior markers have exclusively enriched expression patterns in the corresponding cell type, based on debCAM-estimated cell-specific expressions (Fig. 2.8). Since GSE41826 did not measure expressions of each cell line, we cannot further evaluate the accuracy of debCAM-estimated cell-specific expressions.

2.4.2 Validation on mRNA expression datasets of cortex tissues

We cross-validated debCAM results in two benchmark human brain data sets, Human Brain Transcriptome (HBT) (GSE25219, Illumina Human 49K Oligo array, 923 samples with 17,565 probes) [41] and Braincloud (GSE30272, Affymetrix Human Exon 1.0 ST Array, 269 samples with 30,176 probes) [42]. With the time-courses of mixed gene expressions in human brain cortices sampled at various developmental/life periods, debCAM detected *de novo* subtype-specific markers whose intersection with *a priori* markers implied their associated cell types (Fig. 2.9; Table 2.1, 2.2; *a priori* markers are from the literature [43]). Two glia cell types and progenitor cell type were discovered in both data sets with concordant markers. The other distinctive subtype/states are novel ones, which were undetectable by either global profiling or supervised deconvolution. Their functions in human brain development deserve further investigation. The relative proportions of these subtypes estimated by debCAM, plotted as a function of time (Fig. 2.10), match well with the previously-validated repopulation dynamics during life span [44]. We also applied debCAM to HBT tissue specimens in each of four cortices independently. The obtained region-specific repopulation dynamics are quite similar to those estimated from the whole data set (Fig. 2.11), reflecting that spatial diversity of cell types within the brain cortex is changeless compared to temporal diversity.

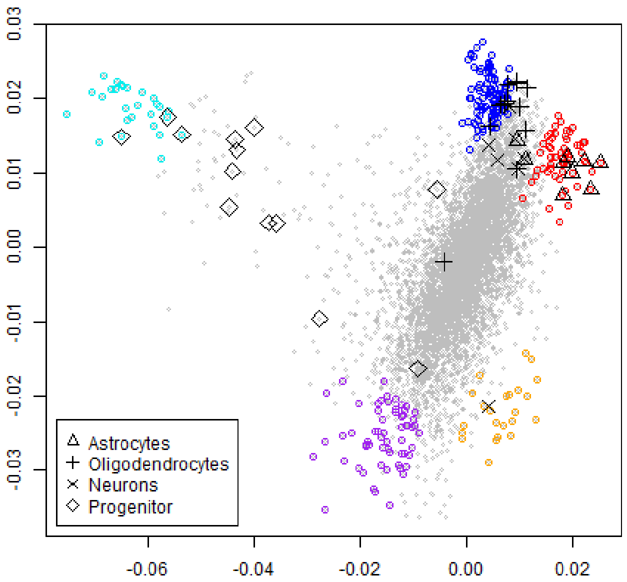


Figure 2.9 Scatter simplex of HBT cortex samples shows five distinctive vertices, three of which host a part of *a priori* markers (labeled) overlapping with debCAM-detected markers (colored).

Table 2.1 Counts of *a priori* markers enriched in debCAM-identified cell types (HBT)

Brain span	S1 red	S2 blue	S3 cyan	S4 purple	S5 orange
Astrocyte(18*)	15	0	0	0	0
Mature oligodendrocyte(18*)	0	10	0	0	0
Neuron(13*)	0	6	0	0	2
Progenitor(25*)	0	0	22	1	0

*count of probes measured in GSE25219 and linked to *a priori* marker genes

Table 2.2 Counts of *a priori* markers enriched in debCAM-identified cell types (Braincloud)

Braincloud	S1 red	S2 blue	S3 cyan	S4 green
Astrocyte(33*)	22	0	0	0
Mature oligodendrocyte(23*)	1	13	0	0
Neuron(11*)	0	3	1	3
Progenitor(38*)	3	0	15	0

*count of probes measured in GSE30272 and linked to *a priori* marker genes

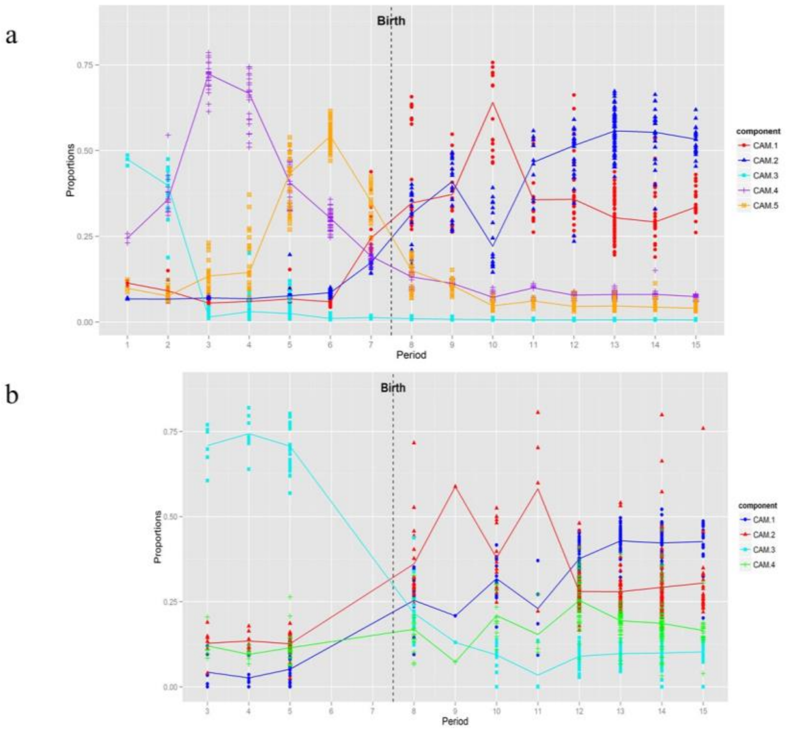


Figure 2.10 Repopulation dynamics of distinctive subtypes/states during life span estimated by debCAM applied to gene expression data from (a) HBT or (b) Braincloud cortex tissues.

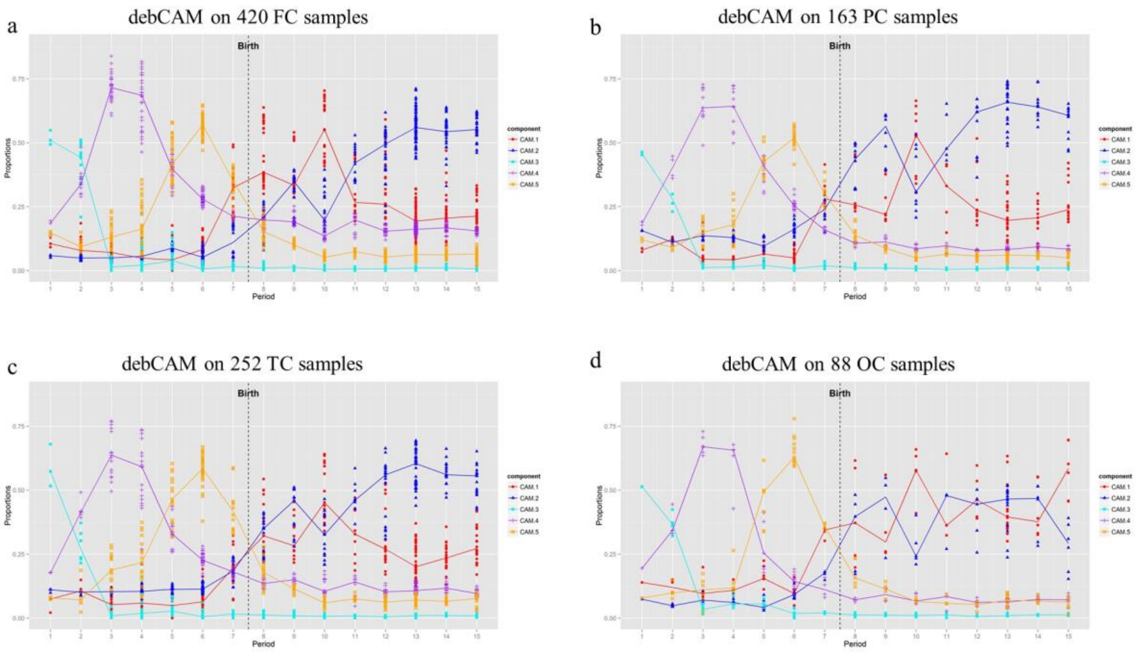


Figure 2.11 Repopulation dynamics of distinctive subtypes/states during life span in each of four cortices estimated by debCAM applied to gene expression data from HBT cortex tissues.

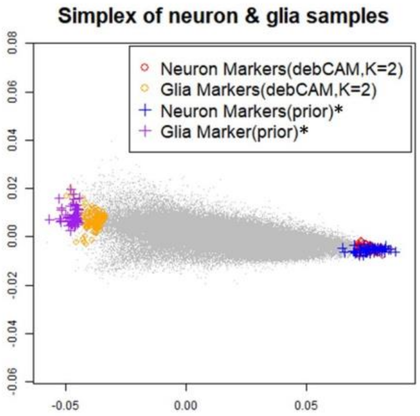
2.4.3 Validation on DNA methylation data of frontal cortex tissues

We further assessed debCAM on two methylation datasets, GSE41826 and GSE74193, measured from frontal cortex tissues. The former contains purified (flow-sorted) neuron and glia samples, providing *a priori* information on neuron/glia cell types. The later was sampled at various developmental/life periods, from which repopulation dynamics of detected subtypes can be estimated and compared to the dynamics estimated from gene expression data.

GSE41826, purified/bulk samples

GSE41826 consists of Illumina HumanMethylation450 DNA profiling for neuron (NeuN+) and glia (NeuN-) cellular populations purified post mortem by FACS from the frontal cortex of 58 individuals [45]. Some “bulk” tissues were also profiled as real mixtures. The deposited data have been normalized. We removed batch effects caused by position, slide, and plate successively using ComBat [46]. Then we applied debCAM to 116 purified samples, including both neuron and glia populations. As expected, the scatter simplex presented two dominant vertices that matched well with neuron/glia markers detected from differential test between neuron and glia samples (Fig. 2.12). However, we also observed two less significant vertices around the one corresponding to glia cell type (Fig. 2.13), which were detectable by debCAM with $K = 3$. We assumed that these vertices represent two glia subtypes, astrocytes and mature oligodendrocytes, coexisting with neurons.

To validate this deconvolution result, we applied debCAM to “bulk” samples independently. Though we had already performed batch effect removal, we only used 11 “bulk” samples measured in the same slide and the same plate, reducing the batch effect to the least possible. Note that all the 116 purified samples were involved in debCAM analysis because the balanced neuron and glia sample distribution in each slide and each plate made slide and plate effects orthogonal to proportion changes. However, the total 20 “bulk” samples in the dataset were not designed so well. debCAM on 11 “bulk” samples identified three cell types, each of which matched with one cell type detected by debCAM on purified samples according to the overlap of debCAM-detected markers from two datasets (Table 2.3). Their simplexes are like a rotated and compressed version of each other (Fig. 2.13). The consistency between cell-type-specific expression profiles deconvoluted from two datasets was demonstrated by almost perfect correlation coefficients of



*Prior markers are from t-test on neuron and glia samples

Figure 2.12 Scatter simplex of purified neuron and glia samples. Two dominated extremes indicate two underlying major cell types: neuron and glia, each of which matches the locations of neuron or glia markers from debCAM ($K = 2$) and from prior information. *A priori* markers were obtained from t-test on neuron and glia samples (p-value ranking top 50 among those with fc > 4).

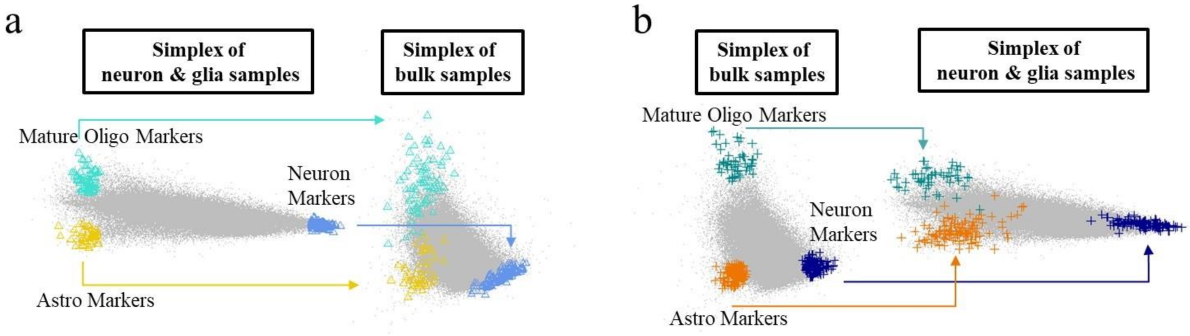


Figure 2.13 Scatter simplex of purified neuron and glia subtypes (astrocytes and mature oligodendrocytes) shows clearly three distinctive vertices. As expectedly, with varying mixing proportions, the scatter simplex of bulk tissue samples shows a rotated and compressed version of the original simplex.

0.982~0.992 over all CpG sites and correlation coefficient of 0.884~0.962 over CpG marker sites (Fig. 2.14). Mean expression values from purified neuron samples had been treated as neuron-specific expressions by many supervised deconvolution methods. We found that these samples were highly correlated with debCAM-estimated neuron-specific expressions from "bulk" samples but deviated a little from the diagonal line over markers (Fig. 2.15). The possible reason could be that physically sorted neuron populations were not pure or that neuron markers had low but non-negligible expression values in other cell types.

Table 2.3 Counts of methylation markers detected from purified and “bulk” samples

Bulk Purified	Neuron Markers (657)	Astro Markers (510)	MO Markers (137)
Neuron Markers (328)	218	0	0
Astro Markers (135)	0	96	0
MO Markers (144)	0	0	61

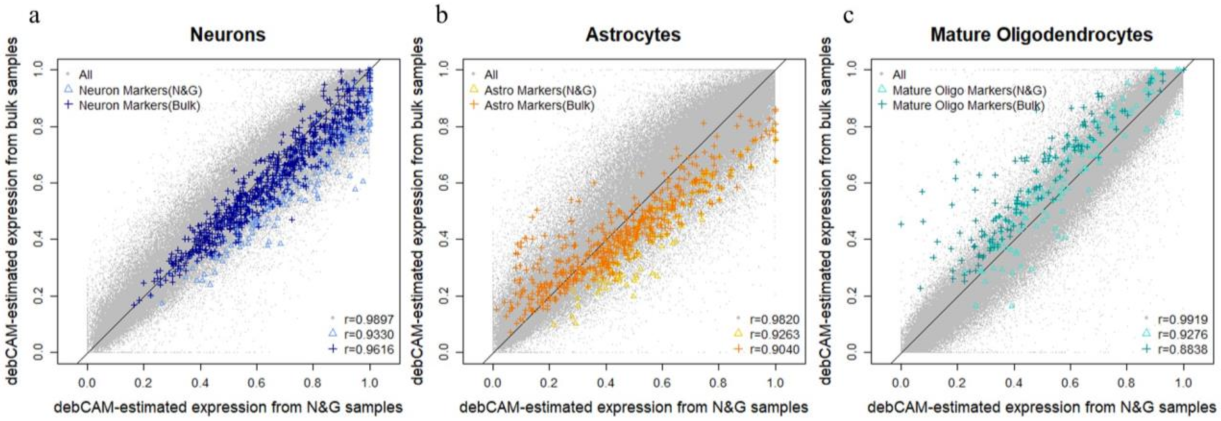


Figure 2.14 Subtype-specific expression profiles obtained by debCAM on bulk samples and on purified tissue samples are highly correlated over all CpG and marker sites.

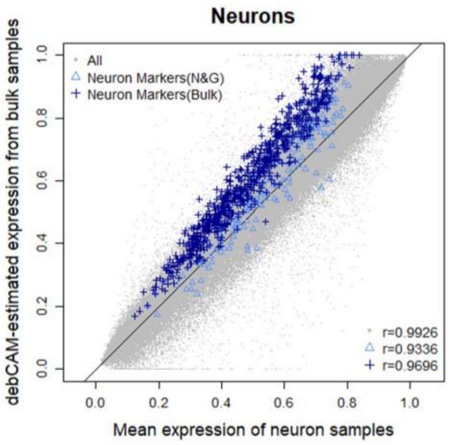


Figure 2.15 Neuron-specific expression profiles obtained by debCAM on bulk samples and by averaging purified neuron sample expressions are highly correlated over all CpG and marker sites.

Identification of two glia subtypes by GSE74193

While debCAM detected two glia subtypes blindly in GSE41826 datasets, no CpG site markers from the literature or any other prior information could tell us which one was astrocyte and which one was mature oligodendrocyte. As the two glia subtypes have different temporal dynamics across human cortex development, we took advantage of this feature to identify two subtypes. Briefly, we downloaded a set of prefrontal cortex samples in another dataset, GSE74193, ranging from fetal (negative ages) through aging (96 years). Two glia subtype markers detected in GSE41826 (in both purified and "bulk" samples) were used as prior knowledge for supervised deconvolution (Eq. 2.11). Considering that GSE74193 includes samples around birth date and thus detectable progenitor cells while GSE41826 does not have them, some CpG sites in GSE41826's marker list may also have non-negligible methylation level in progenitor cells. These CpG sites do not meet the definition of markers any more in GSE74193. Since they show a decreasing trend during brain development due to the infiltration of progenitor cells, we rank GSE41826 markers

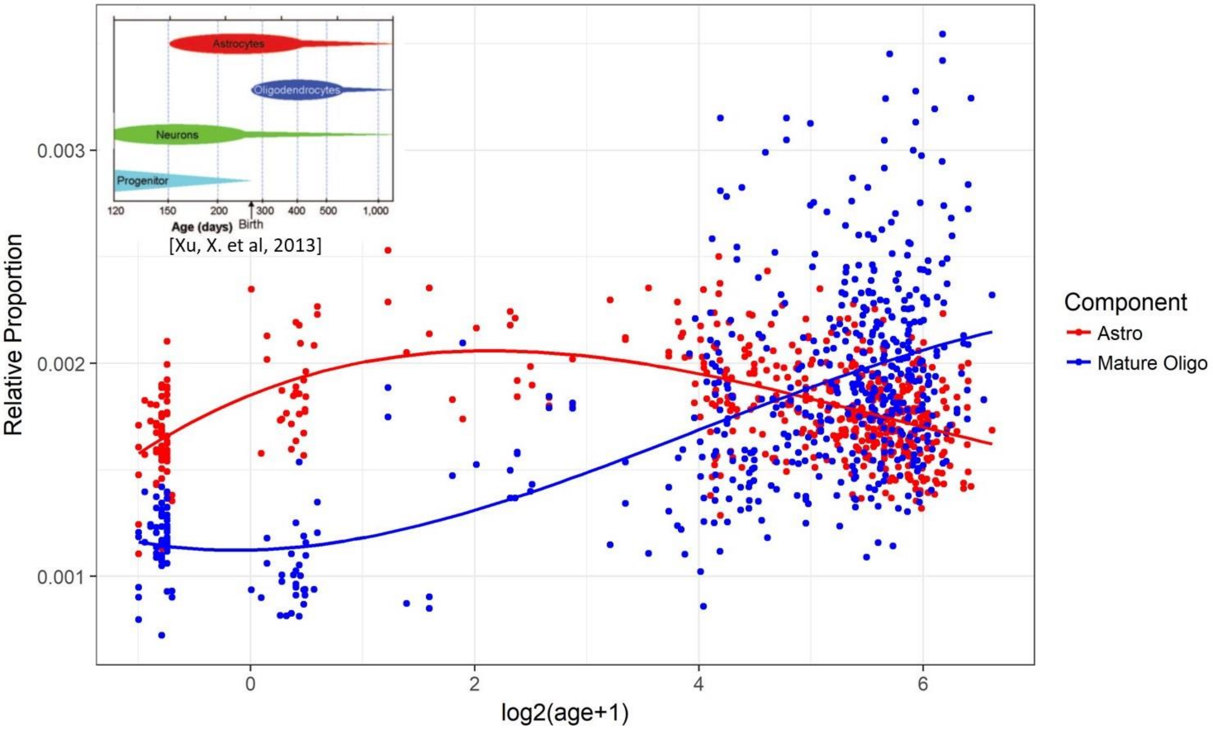


Figure 2.16 Temporal cellular dynamics of two glia subtypes estimated from their markers' expressions across samples, from fetal to aging. Biological knowledge from literatures indicates the subtype that rises first is astrocyte and the other subtype is mature oligodendrocyte.

by the positive correlation with age and select the top 50 (Table 2.4). The space median of their scaled expression vectors, $\tilde{\boldsymbol{a}}_{glia1}$ and $\tilde{\boldsymbol{a}}_{glia2}$, reflected the composition changes across samples and thus across ages. According to temporal cellular dynamics in cortex around birth date (debCAM estimation from gene expression datasets shown in Subsection 2.4.2 and literature [43]), we concluded that the subtype that rose first was astrocyte and the other subtype was that of mature oligodendrocyte (Fig. 2.16).

Table 2.4 Methylation markers to estimate cellular dynamics

	NEURON	ASTRO	MATURE OLIGO
		(GLIA SUBTYPE1)	(GLIA SUBTYPE2)
1	cg11881754	cg12289251	cg16732616
2	cg08742288	cg26939946	cg10755058
3	cg22409276	cg02130905	cg00960700
4	cg08490791	cg09008080	cg02056653
5	ch.4.80693657F	cg01343363	cg00853733
6	ch.13.54497256F	cg27587033	cg04751035
7	ch.16.17742437R	cg10377153	cg04267691
8	ch.5.52085246F	cg17795695	cg00630164
9	cg12927498	cg26853855	cg13438961
10	ch.11.28744729R	cg01933073	cg02782426
11	ch.12.7819240R	cg10273072	cg20560075
12	ch.19.7675672F	cg13921444	cg03245733
13	ch.4.3837399F	cg20244327	cg25365990
14	ch.7.116679792R	cg26835683	cg25813864
15	ch.16.5395104F	cg17058383	cg00612595
16	ch.2.88629272F	cg10907866	cg16907885
17	ch.14.43720575R	cg12066473	cg06377635
18	ch.8.120063967R	cg09414535	cg17315426
19	ch.13.90716753R	cg24812523	cg14999291
20	cg15844835	cg16050468	cg16336494
21	ch.4.73355803R	cg03403507	cg12150784
22	ch.2.144399654R	cg23704362	cg04569152
23	ch.12.58754139R	cg23184518	cg21100518
24	cg07495664	cg05005343	cg02143487
25	ch.6.67663644F	cg10351284	cg09745430
26	ch.11.96117892F	cg17665652	cg17797591
27	ch.8.68016094R	cg06055229	cg06550177
28	ch.7.10564690F	cg16683572	cg03231163
29	ch.4.90357449R	cg12166520	cg04055819
30	ch.2.113320801R	cg01368900	cg26681847
31	ch.18.734816F	cg23165164	cg17468773
32	ch.6.57776846F	cg10517535	cg01002030
33	ch.13.88042588R	cg01882539	cg14467281
34	ch.19.23069863R	cg13667782	cg16310491

35	cg21056723	cg12859046	cg25034557
36	ch.13.58256388R	cg17808910	cg08572315
37	cg12359904	cg20546778	cg15564619
38	ch.6.48208379F	cg14374432	cg20511089
39	cg09458272	cg06446408	cg16031065
40	ch.7.54221757R	cg22274117	cg08206536
41	ch.17.49865880R	cg01375262	cg23189410
42	cg18533883	cg13772849	cg13847070
43	cg23477460	cg09894683	cg01827726
44	ch.4.155269049F	cg17133045	cg10813980
45	cg18732064	cg22276811	cg01164344
46	cg23121335	cg24213115	cg06636541
47	cg00992297	cg21356998	cg04155862
48	ch.4.167276096F	cg03532673	cg17840178
49	cg15091407	cg19163395	cg04042106
50	ch.4.73355811R	cg12082271	cg12943363

Batch effects in GSE74193

In GSE74193, methylation level variations across the lifespan are likely caused by relevant variables, *e.g.*, age and unknown cellular proportion, together with batch effects, *e.g.*, position and slide. Before supervised deconvolution, we filtered samples measured in the slides containing less than 8 samples (at most 12 samples in one slide) because enough samples in each batch can produce a more accurate parameter estimation in ComBat. We also filtered CpG sites located on the X or Y chromosome. We set age as a variable of interest to be kept when removing batch effects caused by position and slide. 559 samples were involved in further analysis of temporal cellular dynamics. We ignored other confounding factors, *e.g.*, sex, race, and disease, which did not obscure our analysis in age or cellular dynamics. Applying unsupervised/supervised deconvolution to samples of one sex (female or male), one race (AA or CAUC), and one condition (normal or schizophrenia) generated similar results.

Without batch effect removal, the simplex is likely to present vertices associated with batches. GSE74193 is such an example. We projected 143 samples (normal, AA, age>13) to the simplex and found three obvious vertices that could not be explained by neuron or glia cell types. We selected samples with age >13 to avoid detecting the progenitor cell type, which otherwise would have formed a significant vertex and made it difficult to present other vertices in low dimensions. To confirm that the three unexplainable vertices were 'artificial' due to batch effects, their markers' expression values across samples were compared with batch categorical variables (Fig. 2.17). The

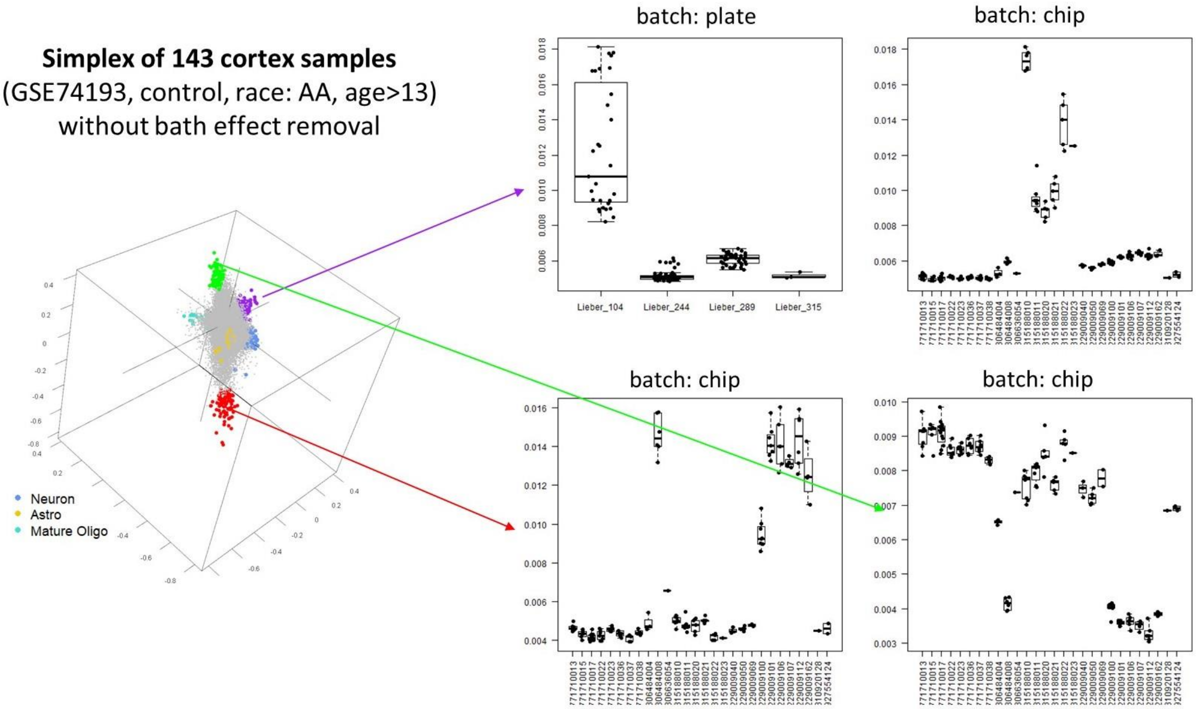


Figure 2.17 Scatter simplex of 143 cortex samples (GSE74193, control, race: AA, age>13) without batch effect removal and boxplot of methylation level distributions in each slide/plate for CpG sites locating around three 'artificial' vertices..

boxplot showed one 'artificial' vertex corresponded to a set of CpG sites over-expressed in a plate of slides, while the other two held CpG sites overexpressed and underexpressed in a subset of slides, respectively. Even with these 'artificial' vertices, debCAM still could detect true vertices holding markers of true cell types. Generally speaking, debCAM can detect all kinds of confounding factors, biological or technical, as long as there are sufficient molecules exclusively affected by these factors. However, for an accurate estimation of proportion and cell-type-specific expression profiles, it is better to try removing all the confounding factors except those most relevant to the system under evaluation.

2.4.4 Validation on DNA mixtures of immune cells

The last public benchmark dataset to test debCAM contains 12 *in vitro* DNA mixtures of immune cells, profiled by HumanMethylationEPIC Beadchip that interrogates > 860,000 CpG sites (GSE110554) [47]. debCAM is not affected by the number of expression units but by the shape of

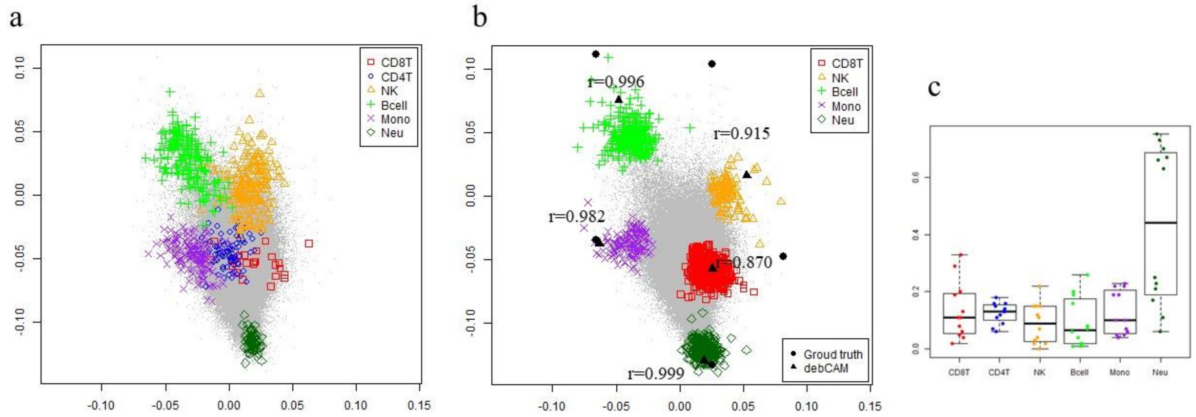


Figure 2.18 Simplex of unmethylation levels of 12 in vitro DNA mixtures profiled by HumanMethylationEPIC Beadchip, with colored points being supervisedly derived markers (OVESEG-test p-value ranking top 200 among those with OVE-FC > 4) (a) or being debCAM-detected markers (b) and boxplot of ground truth composition distribution of 6 immune cells (c).

the projected simplex, especially the data point distribution around vertices. Given that there are more unmethylation markers than methylation markers observed in the simplex, we applied debCAM to (1-beta) values. Meanwhile, we also applied OVESEG-test [48] to identify supervised markers from purified population profiles (also available in GSE110554) as known markers to verify debCAM's detection. The mixture is composed of 6 immune cell types but debCAM only identified 5 cell types. The simplex shows that known CD4T markers cannot form a vertex detectable by debCAM (Fig. 2.18a), while others known markers locate around debCAM-detected vertices (Fig. 2.18a, Fig. 2.18b). The miss of CD4T is due to low variation of its proportion across samples; CD4T markers are expressed almost as stably as housekeeping CpG sites and so are difficult to detect (Fig. 2.18c). Neutrophils have dramatic dynamics and are easily detectable with the most accurate estimation of the proportion changes (r=0.999). The estimated proportions of CD8T and NK likely deviate from ground truth because these two cell types lack exclusively expressed markers as shown in the simplex. In conclusion, the application in this dataset reveals when debCAM's performance may be degraded. Nonetheless, debCAM can discover several of the underlying cell types or estimate proportions close to the truth.

2.4.5 Application on heterogeneous brain samples (In-house data)

Brain is made up of hundreds of different cell types. While the major classifications are neuronal cells and glial cells, each of these has many subcategories based on their morphology or functions [49]. Our in-house brain data were gene expression profiles measured from 129 parietal cortex (PC) tissues and 129 cerebellum (CB) tissues. The major cell types in PC and CB regions are shown in Table 2.5, with *a priori* markers available from the literatures [43]. debCAM identified three and four subtypes, respectively. The enrichment of *a priori* markers in each deconvoluted subtype indicates debCAM successfully finds three major brain cell types – astrocyte, mature oligodendrocyte, and neuron – in the human parietal cortex (Fig. 2.19a). Among much more complex compositions in the cerebellum, debCAM also distinguish four major cell types (Fig. 2.19b). debCAM-identified markers located in simplex vertices match well with part of *a priori* markers (Fig. 2.19, Table 2.6~2.7). Two simplex scatter plots reflect astrocyte and mature oligodendrocyte, as two major glia cell types, have distinct patterns at molecular expression level from others in both cortex and cerebellum. The debCAM-estimated proportions of cell types (Fig. 2.20) indicate neuron cell types account for a larger proportion than the sum of two major glia cell types.

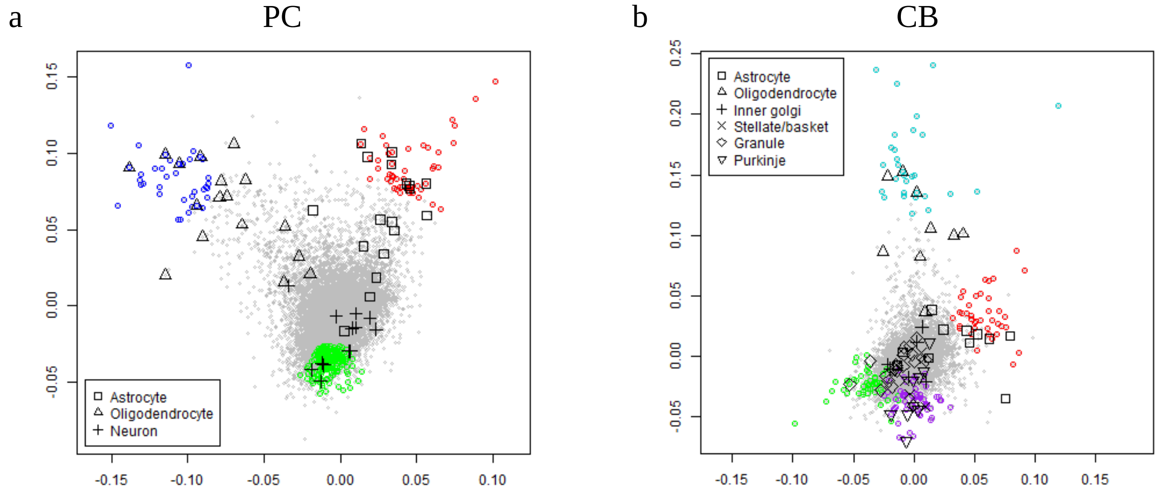


Figure 2.19 Scatter simplex of complex tissues from parietal cortex and cerebellum (Gray dots - all genes forming scatter simplex; colored dots – markers identified by debCAM; black symbols - a priori markers for known cell types)

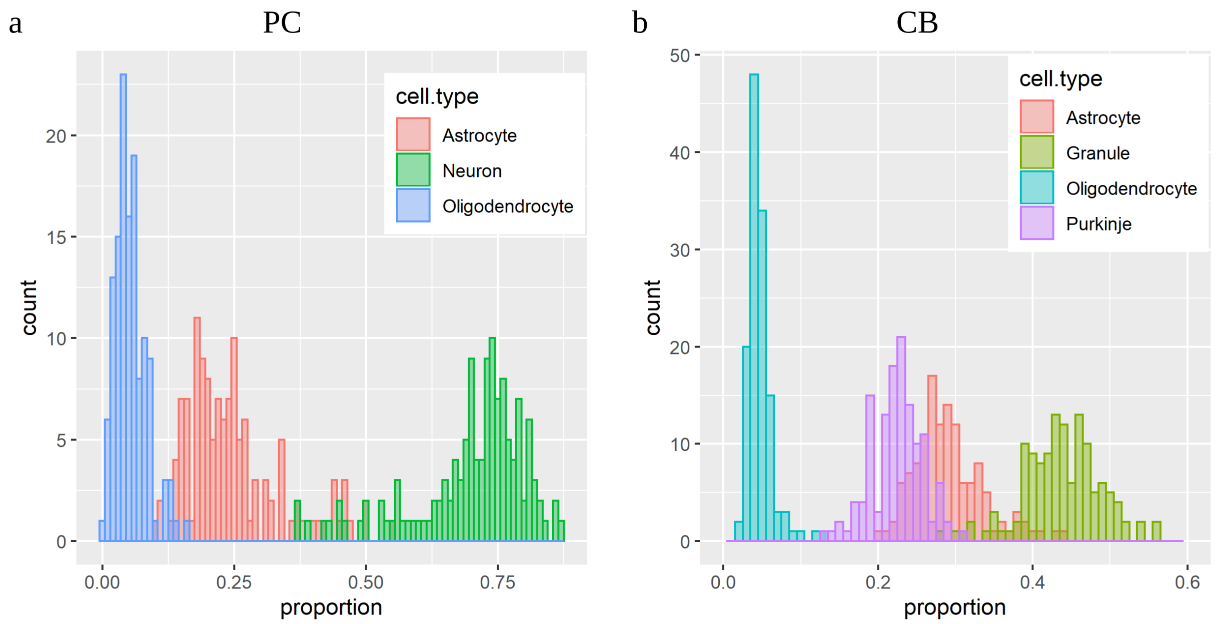


Figure 2.20 Histogram of estimated proportions of debCAM-identified cell types in parietal cortex and cerebellum.

Table 2.5 Subtypes (counts of *a priori* markers) in brain tissues detected by [43]

	Glia	Neuron
FC(PC)	Astrocyte(18) Mature oligodendrocyte(17)	Neuron(13)

CB	Astrocyte and Bergman glia(11)
	Mature oligodendrocyte(9)
		Inner golgi neuron(6)
		Stellate/basket(5)
		Granule(19)
		Purkinje(9)

Table 2.6 Counts of *a priori* markers enriched in debCAM-identified cell types (PC)

	S1	*S2*	*S3*
Astrocyte(18)	13	0	0
Mature oligodendrocyte(17)	0	17	0
Neuron (13)	0	0	7

Table 2.7 Counts *of a priori* markers enriched in debCAM-identified cell types (CB)

	S1	*S2*	*S3*	*S4*
Astrocyte(11)	5	0	0	0
Mature oligodendrocyte(9)	0	9	0	0
Inner golgi (6)	0	0	0	0
Stellate/basket(5)	0	0	0	1
Granule(19)	0	0	3	0
Purkinje(9)	0	0	0	8

While the two major types of cells in the brain are known to be glia and neuron, the true ratio of glia to neurons in the brain remains a mystery. One of the recent studies using an efficient cell counting method provides compelling evidence for 1:1 ratio on four whole human brains [50]. The same study also reveals that the ratio of glia to neurons in the brain varies from one region to another, sometimes dramatically, e.g., 3.76:1 in the cerebral cortex versus 1:4.3 in the cerebellum [50, 51]. However, other scientists have argued that more rigorous studies are needed in which just about every known or unknown marker for both neurons and glia is used to capture as many of the different cell types as possible.

We apply UNDO2.0 (or debCAM with K=2) to the gene expression data of 129 samples acquired from cerebellum (CB). The deconvolution results indicate that the overall neuron proportion is about 70% and non-neuron proportion about 30%, consistent with what previously reported [50, 51]. We then apply UNDO2.0 to the gene expression data of 129 samples acquired from parietal cortex (PC). The neuro/glia marker genes blindly detected by UNDO2.0 are geometrically close to, in the scatter simplex, the *a priori* marker genes given by literatures [6], as shown in Fig. 2.21. More interestingly, the deconvolution also reveals that the overall glia to neuron ratios are different between CB and PC, indicating that neuron/glia distributions may be region-dependent.

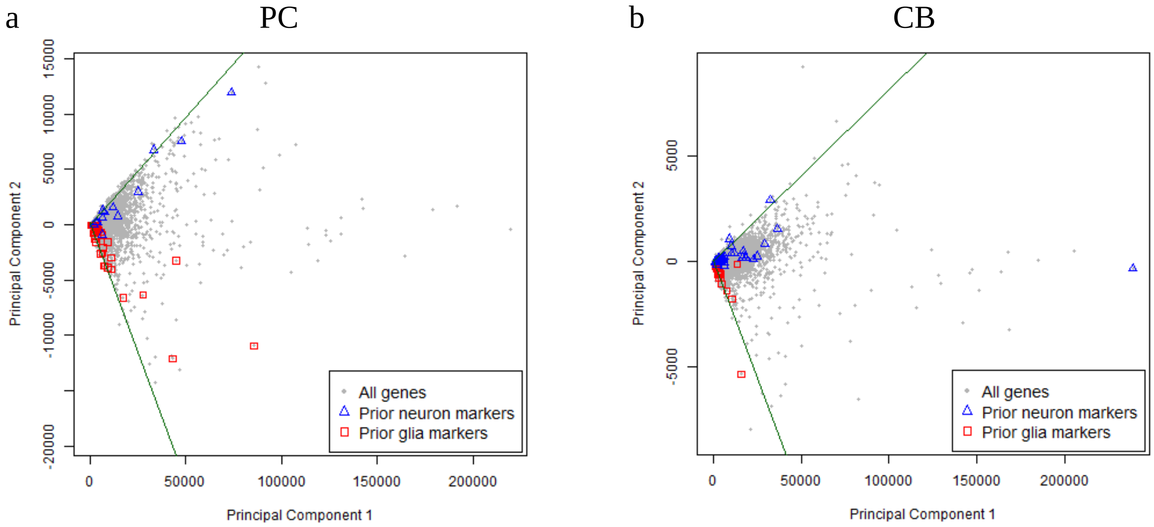


Figure 2.21 Convex cone in scatter space identified by debCAM (UNDO2.0), where *a priori* marker genes given by literatures are geometrically close to the edges of convex cone.

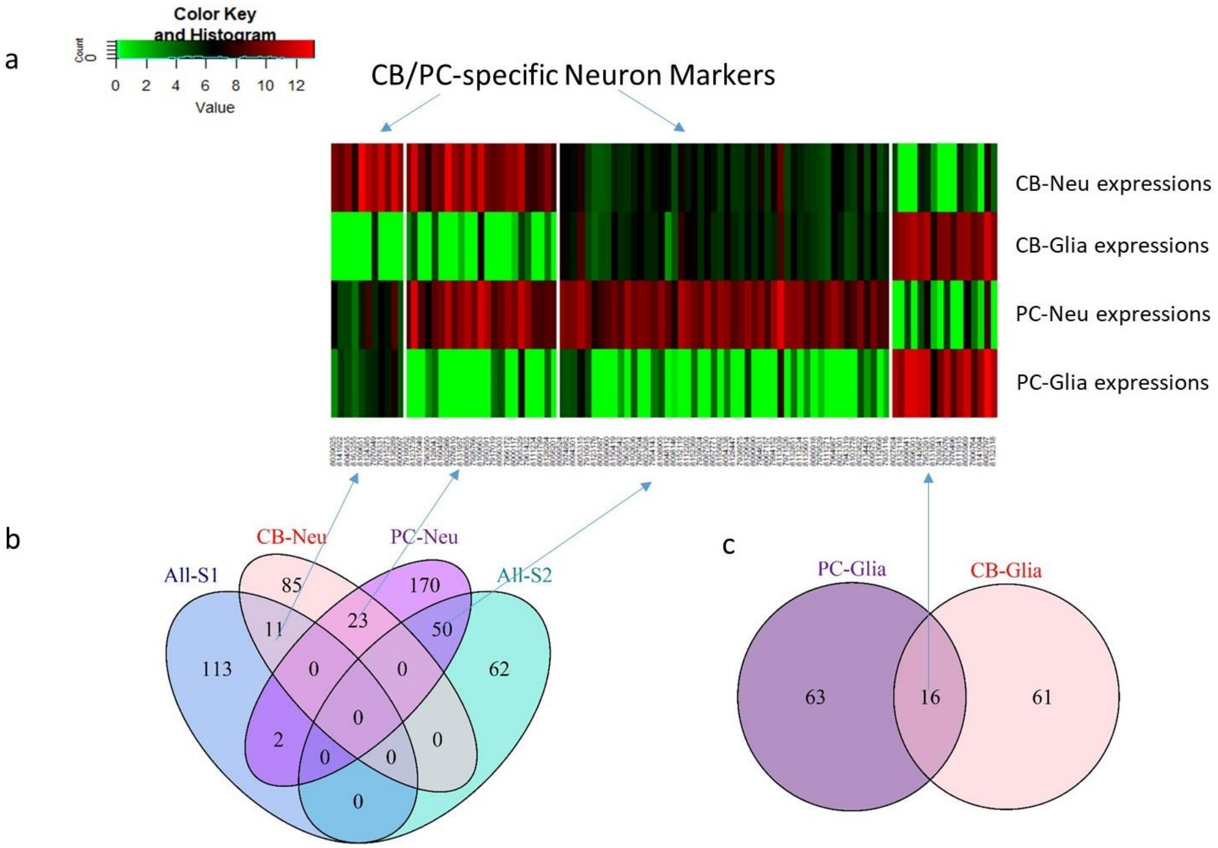


Figure 2.22 Heatmap of neuron/glia-specific expressions in CB and PC (region-dependent/independent marker genes). (b) Overlap of neuron markers (FC > 20) blindly detected in CB samples, PC samples, and all samples. (c) Overlap of glia markers (FC > 20) blindly detected in CB samples, and PC samples.

We further apply UNDO2.0 to the paired 258 samples all together, and detect the two dominant yet most differential neuron cell types that are concordantly associated with CB versus PC brain regions. Strikingly, our experiment detects many region-specific neuron markers that are exclusively enriched in CB or PC neurons, shown in Fig. 2.22, molecularly defining two region-dependent neuron subtypes.

2.4.6 Application on heterogeneous vascular proteomes (In-house data)

At the molecular level, atherosclerosis can be defined as an assembly of hundreds of intra- and extra-cellular proteins that jointly alter cellular processes and produce characteristic remodeling

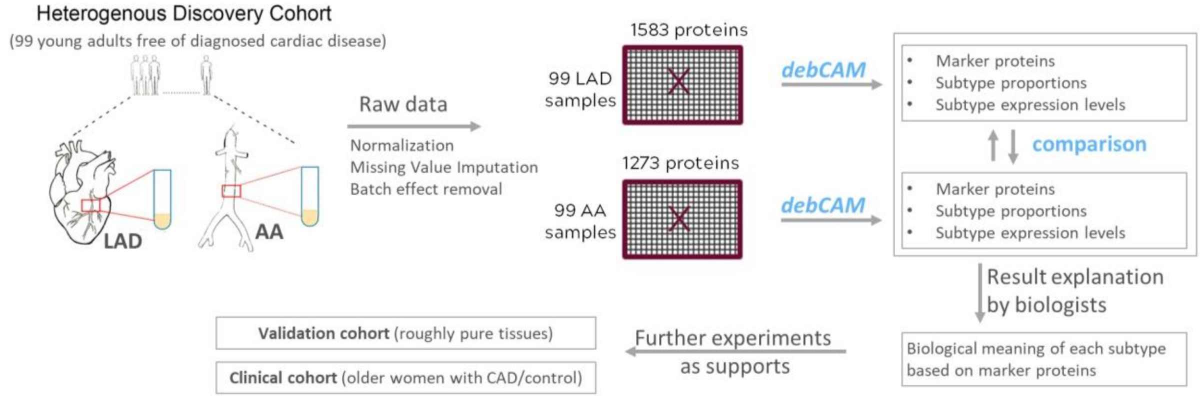

Heterogenous Discovery Cohort
(99 young adults free of diagnosed cardiac disease)
LAD
AA
Raw data
Normalization
Missing Value Imputation
Batch effect removal
1583 proteins
99 LAD samples
1273 proteins
99 AA samples
debCAM
debCAM
Marker proteins
Subtype proportions
Subtype expression levels
comparison
Marker proteins
Subtype proportions
Subtype expression levels
Result explanation by biologists
Biological meaning of each subtype based on marker proteins
Further experiments as supports
Validation cohort (roughly pure tissues)
Clinical cohort (older women with CAD/control)

Figure 2.23 deCAM analysis workflow for proteomics data acquired from two vascular regions (LAD and AA). deCAM results are cross-validated by comparing marker proteins detected in two regions, interpreted by biologists and further supported by biological validation cohort and clinical cohort.

of the local vascular environment. Ultimately, these proteomic changes produce the lesions responsible for most ischemic cardiovascular events. The ability to identify marker proteins characteristic of early- and late-stage pathological tissue states would have meaningful clinical impact, but challenged by the heterogeneous nature of whole tissue proteomic profiling involving varying proportions of different tissue phenotypes.

In one of our projects for Global Analysis of Coronary and Abdominal Aortic Proteomes, specimens were collected from left anterior descending coronary artery (LAD) and abdominal aorta (AA) regions in 99 donors free of clinical/diagnosed cardiovascular disease (Fig. 2.23). A trained pathologist scored each specimen for surface involvement of fatty streak (FS), fibrous plaque (FP) and normal (NL). 1583 proteins in LAD and 1273 proteins in AA are quantified by the DIA-MS technique with less than 50% missingness. After data normalization and imputation, debCAM algorithm was performed and the results indicated four and two distinct expression subtypes in the LAD and AA specimen, respectively (Fig. 2.24). We also applied a supervised method called csSAM, which deconvolute $\mathbf{S}$ from $\mathbf{X}$ with prior $\mathbf{A}$, to estimate tissue subtype-specific expression profiles by pathologist-scored proportions of three tissue subtypes and further obtained associated markers using OVE-FC scheme. Although subtype proportions observed by

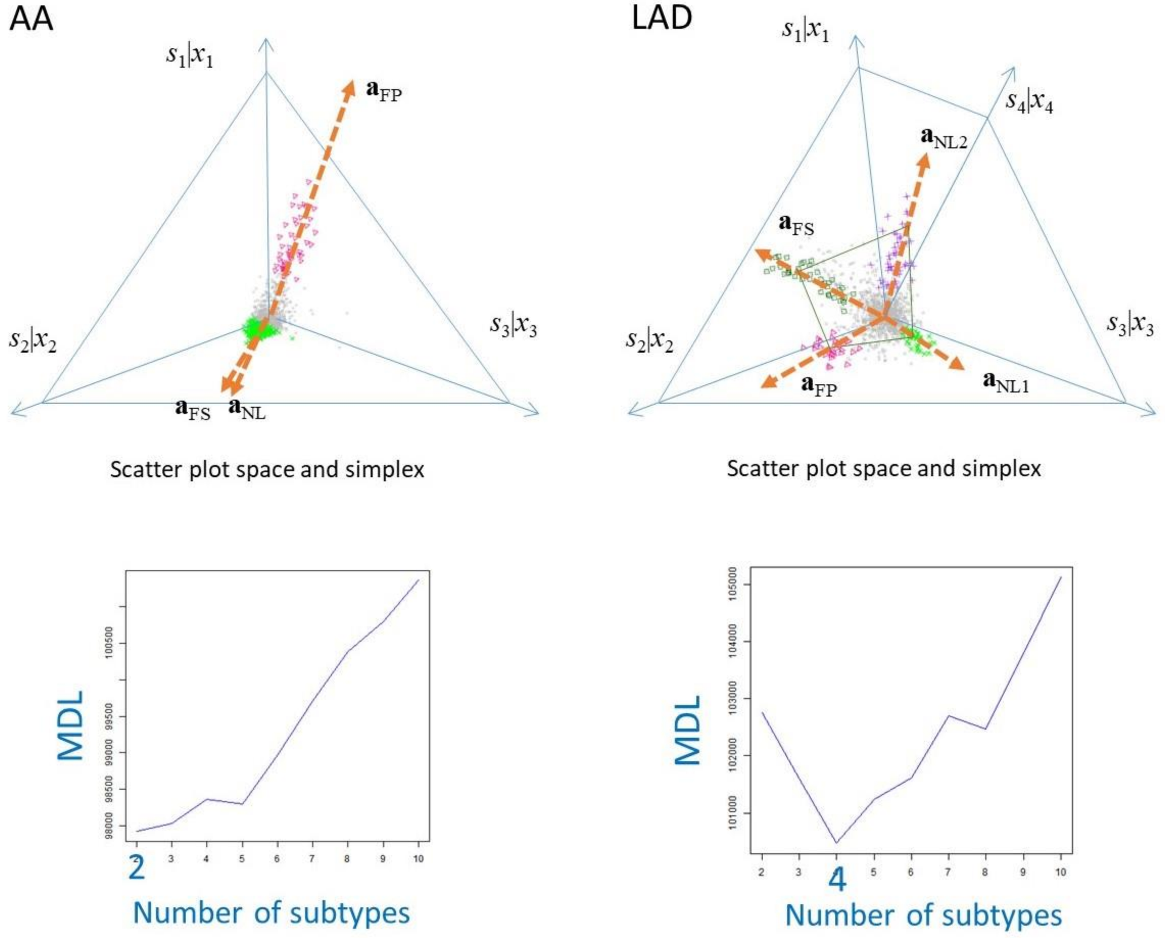


Figure 2.24 The debCAM workflow with four major functional modules: data preprocessing, simplex and marker detection, deconvolution of mixed expression profiles, and model selection.

the pathologist were quite crude, the overlap of debCAM-identified markers and csSAM-identified markers (Table 2.8, 2.9, numbers in parentheses are counts of markers) provided clues about the biological interpretation of debCAM-identified subtypes in LAD/AA. While FS and NL subtype in AA were not separated by debCAM due to insufficient mixture diversity, LAD detected an extra subtype (debCAM-NL1), which may reflect complex compositions of normal tissues.

As pathologist-scored proportions were just visual inspection of the arterial samples and thus less reliable than debCAM estimations, our collaborator biologists started focusing on debCAM-identified subtypes, especially FP subtype whose abundance represents the vascular FP burden and indicates more severe status than NL and FS. Biologists searched the literatures and public database to confirm that debCAM-identified FP markers are associated with lesion/FP. Besides, the 34 FP markers blindly detected from LAD and 49 FP markers from AA have an overlap of 21(33.9%), cross-validating the unsupervised discovery from each other.

Table 2.8 Counts of marker proteins detected in LAD by csSAM, by debCAM and their overlaps

Overlap	csSAM-NL (83)	csSAM-FS (102)	csSAM-FP (115)
debCAM-NL1 (41)	2	2	0
debCAM-NL2 (50)	30	1	0
debCAM-FS (46)	0	26	0
debCAM-FP (42)	0	1	26

Table 2.9 Counts of marker proteins detected in AA by csSAM, by debCAM and their overlaps

Overlap	csSAM-NL (8)	csSAM-FS (44)	csSAM-FP (55)
debCAM-NL/FS (283)	3	16	0
Deb-CAM-FP (50)	0	0	46

As the enrichment levels of FP markers will be good candidates for representing the vascular FP burden and marking early atherosclerosis, biologists conduct two more experiments to further support putative FP markers from debCAM. The first validation cohort collected a set of specimens isolated from an orthogonal, separate cohort of donors. These pure aortic tissues are isolated from regions with no pathology (pure normal, n=3) and regions completely comprised of fatty streak (pure FS, n=3) or fibrous plaque (pure FP, n=4) luminal surface involvement. 1478 proteins were quantified from the pure specimens by the same DIA-MS technique as LAD/AA tissue profiling used so that expression levels could be comparable from two independent experiments. 58 of 62 putative FP marker proteins are quantified in a validation cohort and most are enriched in pure FP specimens compared to pure FS and pure normal specimens. Statistical comparison (OVESEG-test [48]) of the expression profiles among pure FP, pure FS and pure normal tissue specimen shows 18 of 58 putative markers have a significant enrichment with q-value < 0.02. The expression values of FP-marker estimated from either LAD or AA correlated well with those measured from pure specimens, with correlation coefficients of 0.661~0.919 (Table 2.10). The second validation cohort collected clinical data to examine putative FP markers' performance in classifying coronary artery disease (CAD) patients and healthy women. High intercorrelation was observed among FP marker proteins. Thus, we applied elastic net variable selection within a logistic regression analysis to select a panel of 10 proteins that achieved high AUC and low misclassification rate [52].

Table 2.10 Correlation coefficient between debCAM-estimation and pure measurement

Correlation coefficient	FP Specimen 1	FP Specimen 2	FP Specimen 3	FP Specimen 4
debCAM in LAD	0.8586666	0.6613852	0.8438751	0.8634448
debCAM in AA	0.9193555	0.7554338	0.8634865	0.8820235

2.5 Discussion

The debCAM provides a completely unsupervised deconvolution tool, complementary to numerous existing supervised deconvolution methods [26, 53, 54]. Moreover, the debCAM software can readily perform semi-supervised deconvolution by incorporating relevant *a priori* information such as known or reference-derived subtype-specific markers. We expect that debCAM method, with a Bioconductor R package, to be a very useful software tool for unbiased molecular analysis of complex tissues in their native environment. Though the case studies are illustrated only in scatter space, debCAM is principally applicable to sample space when significantly dense sampling is available such as single-cell profiling. Historically this technique was motivated by the need to identify expressed marker genes for tissue/cell subtypes within a mixed sample; however, the approach described here is applicable to any molecular profiling technique for identifying the unique molecule marker or other features that are affiliated with distinct subtypes within the studied samples.

In real applications, debCAM could be sensitive to parameter settings. Changing the random generator seed for K-means clustering or using different convex-hull-to-data fitting criteria (Eq. 2.7) is likely to detect a different set of vertices in the simplex, especially when forcing debCAM to detect less or more subtypes than ground truth. Smaller K will make debCAM randomly merge some subtypes into one, while a larger K will model some random noise to be a predicted subtype. If the optimal K predicts stable results, the solution is accurate. However, in real data analysis, noises or hidden confounding factors could generate fake subtypes and some of latent subtypes are hard-to-separate (*e.g.,* no good markers, proportions are highly correlated with other subtypes or with housekeeping molecules). The MDL curve may find different optimal K under different parameter settings. Treating the MDL curve as a reference only, we can assume that the true K are within a range and hard to determine. If we check the detected markers for each possible K, some cell types can always be detected and some detected occasionally. The former cell types can be treated as a stable discovery, while the latter types could be either noise or hard-to-detect cell types.

As a fully unsupervised machine learning method, debCAM is expectedly not immune to hidden confounding factors, e.g. batch effect, collinearity (Fig. S7-8). Where possible, we can use prior information (*e.g.,* markers from literature, markers from test over purified samples) to help guide the final decisions, which is also implemented in debCAM package as semi-supervised deconvolution.

Chapter 3

Robust detection of subtype-specific markers

3.1 Introduction

Molecular characterization (e.g. gene expression profile) of a complex biological system often includes features that are ubiquitously expressed by all cell or tissue types in the system (e.g. housekeeping genes) [55], and expressed features that are specific to one or more cell or tissue types present in the system (marker genes or differentially-expressed genes) [5, 7, 56]. An important but frequently underappreciated issue is how a "cell or tissue specific expression pattern" is defined. Ideally, a specific expression pattern would be composed of individual features that are exclusively and consistently expressed in the cognate cell or tissue subtype of interest and in no others across various conditions – so called marker genes (MGs) [6, 11, 13, 57] (or marker protein, marker CPG site, etc.).

When MGs are known *a priori*, they are often used to facilitate supervised deconvolution [6, 11, 13, 57]. However, detecting MGs using purified tissue or cell-specific molecular expression profiles remains a challenging task [58]. For example, the most frequently used methods rely on an extension of ANOVA that identifies genes differentially expressed across all the relevant cell or tissue subtypes. In this case, the null hypothesis is that samples in all subtypes are drawn from the same population, resulting in the selection of genes that may not conform to the ideal MG definition. One-Versus-Rest Fold Change (OVR-FC) is another popular method based on the ratio of the average expression in a particular subtype to that of the average expression in the rest samples [58-60]. OVR t-test is occasionally used to assess the statistical significance of MGs [61]. However, a gene with low average expression in the rest is not necessarily expressed at a low level in every subtype in the rest, clearly violating the definition of MGs. Conversely, One-Versus-Everyone Fold Change (OVE-FC) [17, 62] has been proposed to specifically detect MGs that has led to some novel discoveries [22, 27] and much improved classification [17, 62]. OVE-FC checks whether the mean of one subtype is significantly higher/lower than the mean from each of the other subtypes, consistent with the biologic definition of MGs [6, 22]. Supportively, simulation studies show that Marker Gene Finder in Microarray data (MGFM), a method similar to OVE-FC, outperforms OVR t-test [63]. Similar strategies include One-Versus-One (OVO) t-test and Multiple Comparisons with the Best (MCB) [64] that use additional pairwise significance testing or the confidence intervals of OVO statistics [7, 65], but without rigorous modeling the null distribution in relation to the definition of MGs.

To address the critical problem of the absence of a detection method explicitly matched to the definition of MGs, we developed a statistically-principled marker detection method (One Versus Everyone Subtype Exclusively-expressed Genes – OVESEG-test) that can detect tissue or cell-specific MGs among many subtypes. OVESEG-test is based on our earlier work on detecting One Versus Everyone Phenotype Upregulated Genes – OVEPUG [17, 22, 62]. To assess the statistical significance of MGs, OVESEG-test uses a specifically designed test statistic that mathematically matches the definition of MGs. A novel permutation scheme is used to estimate the corresponding distribution under the null hypothesis, where the expression patterns of non-MGs can be highly complex. In this chapter, we will describe the proposed OVESEG-test and then validate its performance on extensive simulation data, in terms of type 1 error rate, False Discovery Rate (FDR), partial area under the receiver operating characteristic curve (pAUC), and in comparison with top peer methods. We will also demonstrate the utility of OVESEG-test by applying it to benchmark public data, and assessing the performance by comparing with known MGs and by the accuracy of supervised deconvolution that uses the *de novo* MGs detected by OVESEG-test.

3.2 Method

3.2.1 OVESEG-test statistic

Consider the measured expression level $s_k(i,j)$ of gene j in sample i across $k = 1, \dots, \dots K$ subtypes. We assume that log $s_k(i,j) \sim N(\mu_k(j), \sigma^2(j)\ )$, where $\mu_k(j)$ and $\sigma^2(j)$ are the mean and variance of the logarithmic expression values of gene j in subtype k. Motivated by our earlier work on OVE-FC [17] , we define One-Versus-Everyone Log Fold Change (OVE-LFC) as

$$d_{jk} = \mu_k(j) - \max_{l \neq k} \mu_l(j) = \min_{l \neq k}\{\mu_k(j) - \mu_l(j)\}, k = 1, \dots, K, \tag{3.1}$$

where $\mu_k(j)$ is the mean of logarithmic expressions of gene j in subtype k. Conceptually, the null hypothesis for non-MGs of subtype k and alternative hypothesis for MGs of subtype k can be described as

$$\begin{aligned} H_{non\text{-}MG(k)}&: d_{jk} \le 0; \\ H_{MG(k)}&: d_{jk} > 0; \end{aligned} \tag{3.2}$$

Standardizing OVE-LFC by variance, we obtain the subtype-specific OVESEG-test statistic

$$t_{jk} = \min_{l \neq k} \left\{ \frac{\mu_k(j) - \mu_l(j)}{\sigma(j)\sqrt{\frac{1}{N_k} + \frac{1}{N_l}}} \right\} = \min_{l \neq k} \left\{ t\text{-}stat_{k,l}(j) \right\}, k = 1, \dots, K, \tag{3.3}$$

where $\sigma^2(j)$ is genewise variance of logarithmic expressions within one subtype, $t\text{-}stat_{k,l}(j)$ is t-statistic between subtype k and l. N_k and N_l are the numbers of samples in subtypes k and l, respectively. Using t_{jk} to select MGs for subtype k, denoted by $\mathrm{MG}(k)$, is equivalent to performing pairwise t-tests between subtype k and each of the remaining subtypes and then taking their intersections, since the null hypothesis and alternative hypothesis can be rewritten as $H_{\text{non-MG}(k)} = \bigvee H_{\text{non-MG}(k,l)}$ and $H_{\mathrm{MG}(k)} = \bigwedge H_{\mathrm{MG}(k,l)}$ respectively, where $\mathrm{MG}(k, l)$ denotes MGs enriched in subtype k against subtype l. However, naively combining the pairwise t-test p-values to assess significance of MGs did not model the null distributions rigorously according to the definition of MGs.

Notice that we only need to compute t_{jk} when $\mu_k(j)$ is the largest value among $\mu_l(j), l = 1, \dots, K$, as we only need to determine whether the highest expressed subtype has significant enrichment against other subtypes. Thus, we can define OVESEG-test statistic, regardless of a specific subtype, as

$$t_j = \max_{k=1,\dots,K} \{t_{jk}\} = t_{j(K)} = \min_{l \neq (K)} \left\{ \frac{\mu_{(K)}(j) - \mu_l(j)}{\sigma(j)\sqrt{\frac{1}{N_{(K)}} + \frac{1}{N_l}}} \right\} \tag{3.4}$$

to test the significance level of gene j as a MG, where subscript (K) indicates the Kth order of ranked sequence $[\mu_l(j), l = 1, \dots, K]$, i.e. the subtype with maximum mean (corresponding to the ranked subtype sample means in an increasing order). $N_{(K)}$ and N_l are the numbers of samples in subtypes (K) and l, respectively. Conceptually, the null hypothesis for non-MGs, and alternative hypothesis for MGs, can be described as

$$\begin{aligned} H_{\text{non-MG}}&: d_j = 0; \\ H_{\text{MG}}&: d_j > 0; \end{aligned} \tag{3.5}$$

where the OVESEG-test uses the gap between the highest expressed subtype and the second highest expressed subtype $d_j = \min_{l \neq (K)} \{\mu_{(K)}(j) - \mu_l(j)\}$. Note that this strategy is consistent with the approach proposed in our earlier work on OVE-FC.

We should know the subtype with maximum estimated mean is not necessary to have the largest true population mean due to disturbing variance, especially when sample sizes are unbalanced among all subtypes. However, we still focus on the significance level of upregulation in the subtype with maximum estimated mean. Even through a gene is actually expressed highest in one of the remaining subtypes and thus corresponds to more significant upregulation level in other subtypes, this molecule is not likely to be an interested MG with low p-value.

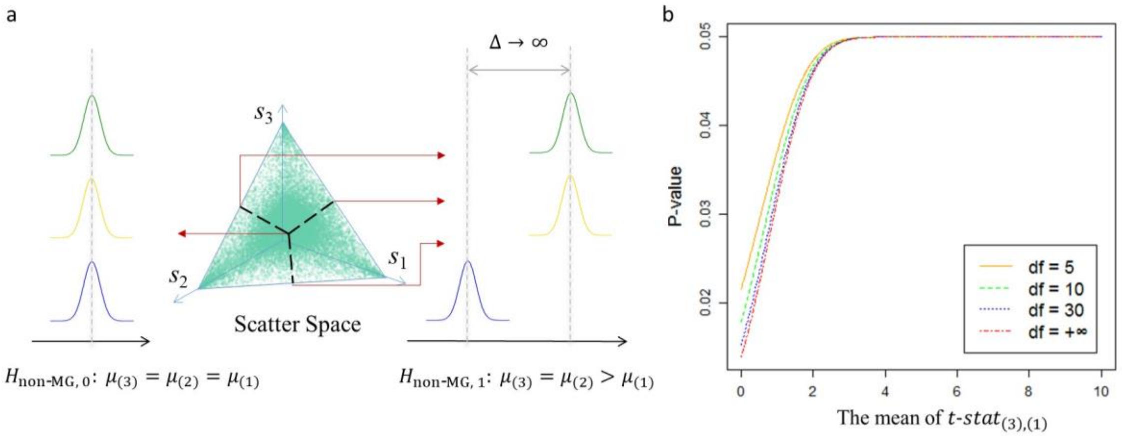


Figure 3.1 P-values approach the upper bound when two subtypes are relatively larger than the third and decrease to be minimum when all three subtypes are drawn from the same population (equal sample size for three subtypes).

3.2.2 Modeling OVESEG-test statistics under null hypothesis

Complex expression patterns of non-MGs

For more than two subtypes $K \geq 3$, modeling the null distribution of OVESEG-test statistics is challenging because of the highly complex expression patterns of non-MGs. Under the null hypothesis $H_{\text{non-MG}}\colon d_j = 0$, non-MGs include all the counterparts of MGs (housekeeping genes and differentially-expressed genes -- DEGs of various combinatorial forms). As $d_j = \min_{l \neq (K)}\{\mu_{j(K)} - \mu_{jl}\}$, $H_{\text{non-MG}}$ includes the following $K-1$ components:

$$H_{\text{non-MG, }0}\colon \mu_{(K)}(j) = \mu_{(K-1)}(j) = \cdots = \mu_{(1)}(j);$$

$$H_{\text{non-MG, }1}\colon \mu_{(K)}(j) = \mu_{(K-1)}(j) = \cdots = \mu_{(2)}(j) > \mu_{(1)}(j);$$

...

$$H_{\text{non-MG, }K-2}\colon \mu_{(K)}(j) = \mu_{(K-1)}(j) > \mu_{(K-2)}(j), \dots, \mu_{(1)}(j);$$

where subscript (k) enclosed in parentheses indicates the kth order of $[\mu_l(j), l = 1, \dots, K]$. The null distribution of OVESEG-test statistics under $H_{\text{non-MG, }m}, 0 < m \leq K-2$, is asymptotically equivalent to that under $H'_{\text{non-MG, }m}\colon \mu_{(K)}(j) = \cdots = \mu_{(m+1)}(j) \gg \mu_{(m)}(j), \dots, \mu_{(1)}(j)$ when $\Delta_m(j) = \mu_{(K)}(j) - \mu_{(m)}(j)$ is sufficiently large that only the highest expressed $(K-m)$ subtypes affect OVESEG-test statistics. On the contrary, when $\Delta_m(j)$ approaches zero, $H_{\text{non-MG, }m}$ will tend to become $H_{\text{non-MG, }m-1}$ with the null distribution of OVESEG-test statistics becoming less dispersed, as shown in the following simulation.

To illustrate the change of null distribution with varied $\Delta_m(j)$, we can calculate theoretical p-values based on a multivariate normal/student's t distribution of $(K-1)$ random variables $t\text{-}stat_{(K),l}, l = 1, \dots, K, l \neq (K)$ who have mean vector $[\mathbf{0}_{K-m-1}, \Delta_m(j), \dots, \Delta_1(j)]$, unit variances, and correlation coefficients

$$\rho_{l_1,l_2} = \frac{1}{\sqrt{\left(\frac{N_{(K)}}{N_{l_1}} + 1\right)\left(\frac{N_{(K)}}{N_{l_2}} + 1\right)}}. \tag{3.6}$$

Decreasing $\Delta_m(j)$ tends to shift the distribution of $\min_{l\neq(K)}\{t\text{-}stat_{(K),l}\}$ in the negative direction, increasing the significance level (smaller p-values). For example, when $K = 3$ with an equal sample size for all subtypes, the correlation coefficient between $Z_0 = t\text{-}stat_{(3),(2)}$ and $Z_1 = t\text{-}stat_{(3),(1)}$ is $\rho = 0.5$. Under $H_{\text{non-MG},\,1}$, Z_0 has zero mean and Z_1 could have a positive mean. Figure 3.1 shows the change of p-values with the critical value equal to that of two-group t-test p = 0.05. If the mean of Z_1 is large, three-group OVESEG-test statistics can be approximated by two-group t-statistics, yielding the same significance level as a t-test. Otherwise, the p-value will decrease and reach as low as 0.01387 when the mean of Z_1 also equals zero.

In fact, if we know which null hypothesis each gene comes from as well as the true $\Delta_m(j)$, we can calculate the theoretical p-value. Of course, these are unknown and have to be modeled. We propose a mixture model of null distributions under the component hypotheses to estimate the p-values of candidate MGs.

Proposed Mixture model

We propose the following mixture distribution to model OVESEG-test statistics under the null hypothesis (Fig. 3.2)

$$f\{t|H_{non-MG}\} = \sum_{m=0}^{K-2} f\{t|H_{\text{non-MG},\,m}\}P\{H_{\text{non-MG},\,m}|H_{\text{non-MG}}\}, \tag{3.7}$$

where t is the OVESEG-test statistic, and $H_{\text{non-MG},\,m}$ is the mth component of the mixture null hypothesis $H_{\text{non-MG}}$. We design a novel nested permutation scheme that both approximates the complex null distribution and is consistent with the definition of MGs. Principally, $H_{\text{non-MG},\,m}$ is constructed by permuting the samples in the top $(K - m)$ subtypes with higher mean expressions; that is, the samples in the bottom m subtypes with lower mean expressions are removed from the permutation. Note that $H_{\text{non-MG},\,0}$ corresponds to the same null distribution used in ANOVA where all samples participate in the permutation.

This highly-capable mixture null distribution model is proposed to model unknown yet potentially complex expression patterns of non-MGs under null hypothesis. Accordingly, the specifically-designed permutation scheme(s) estimates such a mixture null distribution against MGs. The main advantage of the proposed permutation scheme(s) is its flexibility and comprehensiveness, which well match the mixture null distribution of various types and combinations. With varying the proportions of different non-MG types, OVESEG-test is able to maintain the type 1 error rate close to the expected level with the help of the proposed permutation scheme(s) and the conditional probability of each non-MG type.

Note that $H_{non\text{-}MG,m}$, $m = 0, \ldots, K-2$ represent $(K-1)$ different null hypotheses, each with an individualized null distribution that can be estimated by specific permutation scheme(s), *e.g.*, permuting samples in the top $(K-m)$ subtypes. Collectively, a mixture null distribution is constructed via combinations of different null hypotheses in various proportions. In contrast, without conditioning on $H_{non\text{-}MG,m}$, all null distributions are aggregated equally into the mixture null distribution in the same proportion. Consequently, this simpler permutation scheme produces an equal-weight mixture model that cannot represent the complexity of the null distribution. Thus, the null distribution of the OVESEG-test statistic could become distorted, *e.g.*, a uniform distribution of p-values in null data is not guaranteed, and the observed False Discovery Rate may not match the expected level.

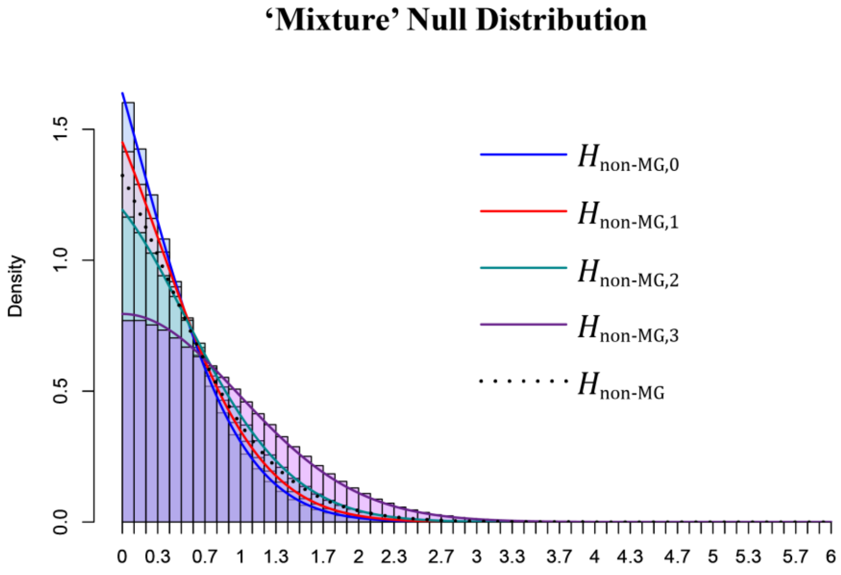


Figure 3.2 Mixture null distribution of OVESEG-test statistics for detecting MGs. The mixture distribution consists of $(K-1)$ null components, each estimated from permuting samples in the top $(K-m)$ subtypes of high mean expressions and weighted by the posterior probabilities of component null hypotheses.

Specifically, the null distribution of OVESEG-test statistics under $H_{\text{non-MG},\,m}$ is estimated from permuted samples and aggregated from different genes with weights. Let $\boldsymbol{s}(j) = [s(1,j), \dots, s(N,j)]$ denote the measured expression vector of gene j across samples, where N is the total number of samples. These weights are the posterior probabilities of a component null hypothesis given the observation $\Pr\{H_{\text{non-MG},\,m} \middle| \boldsymbol{s}(j)\}$, estimated by the local FDR $\text{fdr}_{\text{non-MG},\,m}(j)$ [66], given by

$$w_{\text{non-MG},\,0}(j) = \Pr\{H_{\text{non-MG},\,0} \middle| \boldsymbol{s}(j)\} = \text{fdr}_{\text{non-MG},\,0}(j), \tag{3.8a}$$

$$\begin{aligned} w_{\text{non-MG},\,m}(j) &= \Pr\{H_{\text{non-MG},\,m} \middle| \boldsymbol{s}(j)\} \\ &= \left\{1 - \sum_{n=0}^{m-1} \Pr\{H_{\text{non-MG},\,n} \middle| \boldsymbol{s}(j)\}\right\} \Pr\{H_{0m} | \boldsymbol{s}(j), \neg H_{0n}, for\ all\ 0 \le n < m\} \\ &= \left\{1 - \sum_{n=0}^{m-1} w_{\text{non-MG},\,n}(j)\right\} \text{fdr}_{\text{non-MG},\,m}(j), 0 < m < K - 2, \end{aligned} \tag{3.8b}$$

where $\text{fdr}_{\text{non-MG},\,0}(j)$ is the local FDR associated with ANOVA on all subtypes, and $\text{fdr}_{\text{non-MG},\,m}(j)$ is the local FDR associated with ANOVA on the top $(K - m)$ subtypes, estimated using R package “fdrtool” [67].

3.2.3 Assessing statistical significance of candidate MGs

The p-values of candidate MGs are estimated using the learned ‘mixture’ null distribution

$$p\text{-}value = \Pr\{T > t_{obs} | H_{\text{non-MG}}\} = \sum_{m=0}^{K-2} \Pr\{T > t_{obs} \middle| H_{\text{non-MG},\,m}\} P\{H_{\text{non-MG},\,m} \middle| H_{\text{non-MG}}\}, \tag{3.9}$$

where t_{obs} is the observed OVESEG-test statistic, and T is the continuous dummy random variable. Specifically, $\Pr\{T > t_{obs} \middle| H_{\text{non-MG},\,m}\}$ is calculated by the weighted permutation scores

$$\Pr\{T > t_{obs} \middle| H_{\text{non-MG},\,m}\} = \frac{\sum_{p=1}^{P} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j) I(T_{j,p} > t_{obs})}{P \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)}, \tag{3.10}$$

where P is the number of permutations, J is the number of participating genes, $I(.)$ is the indicator function, and $T_{j,p}$ is the OVESEG-test statistic in the pth permutation on jth gene. Furthermore, the component weight in the mixture null distribution is estimated by the membership expectation of the posterior probabilities over all genes

$$
\begin{aligned}
P\{H_{\text{non-MG},\,m}|H_{\text{non-MG}}\} &= \frac{\Pr\{H_{\text{non-MG},\,m}\}}{\sum_{n=0}^{K-2}\Pr\{H_{\text{non-MG},\,n}\}} \\
&= \frac{\frac{1}{J}\sum_{j=1}^{J}\Pr\{H_{\text{non-MG},\,m}|\boldsymbol{s}(j)\}}{\frac{1}{J}\sum_{j=1}^{J}\sum_{n=0}^{K-2}\Pr\{H_{\text{non-MG},\,n}|\boldsymbol{s}(j)\}} \\
&= \frac{\sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)}{\sum_{j=1}^{J}\sum_{n=0}^{K-2} w_{\text{non-MG},\,n}(j)},
\end{aligned}
\tag{3.11}
$$

Lastly, substituting Eq. 3.10~3.11into Eq. 3.9, the p-value associated with gene j is calculated by:

$$
p\text{-}value = \frac{\sum_{m=0}^{K-2}\sum_{p=1}^{P}\sum_{j=1}^{J} w_{\text{non-MG},\,m}(j) I\left(T_{j,p} > t_{obs}\right)}{P\sum_{m=0}^{K-2}\sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)},
\tag{3.12}
$$

with a lower bound of $\min_j\{\sum_{m=0}^{K-2} w_{\text{non-MG},\,m}(j)\} / P\sum_{m=0}^{K-2}\sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)$.

Note the permutation distributions from different genes could be pooled together with certain weights where: **(a)** the gene with large posterior probabilities of $H_{\text{non-MG},\,m}$ should contribute more to the null distribution under $H_{\text{non-MG},\,m}$ and **(b)** genes with relatively small $\Delta_m(j)$ may affect null distribution estimation under both $H_{\text{non-MG},\,m}$ and $H_{\text{non-MG},\,m-1}$. As local FDR is a good estimator of posterior probability of a null hypothesis [66], the weight of each gene contributing to a component null hypothesis is assigned according to the local FDR of ANOVA on the observed expression values across certain subtypes (Eq. 3.8). Note that we include the genes in the $fdr_{\text{non-MG, m}}$ calculation with a probability of $\{1 - \sum_{n=0}^{m-1} w_{\text{non-MG},\,n}(j)\}$ so that the components of the mixture distribution are mutually exclusive. The computation of $w_{\text{non-MG},\,m}$ is almost not affected by genes associated with $H_{\text{non-MG},\,0}, \ldots, H_{\text{non-MG},\,m-1}$.

3.2.4 Subtype-specific OVESEG-test

Using $t_j = \max_{k=1,\ldots,K}\{t_{jk}\}$ as a test statistic in K-group comparison is equivalent to using the absolute value of the statistic in two-group differential test, which gives a concise way to define extremes in the K directions. However, in two-group differential test, if the test statistics have an asymmetric null distribution for two sides, another popular choice is to double the smaller of the two tail regions [68, 69]. Extending this method to $K > 2$ cases, we can calculate subtype-specific p-values based on the null distribution of each $t_{jk}, k = 1, \dots, K$, by

$$\begin{aligned} p\text{-}value(k) &= \Pr\{T_k > t_{obs,k} | H_{\text{non-MG},\, m}\} \\ &= \sum_{m=0}^{K-2} \Pr\{T_k > t_{obs,k} | H_{\text{non-MG},\, m}\} P\{H_{\text{non-MG},\, m} | H_{\text{non-MG}}\} \\ &= \frac{\sum_{m=0}^{K-2} \sum_{p=1}^{P} \sum_{j=1}^{J} w_{\text{non-MG},\, m}(j) I(T_{jk,p} \geq t_{k,obs})}{P \sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-MG},\, m}(j)}, \end{aligned} \tag{3.13}$$

as one-tailed p-values. We then multiply the smallest tail by K. $t_{k,obs}$ is the observed OVESEG-test statistic for a gene being tested under subtype k, and $T_{jk,p}$ is OVESEG-test statistic in the pth permutation on jth gene under subtype k. We also need to avoid the extra computational burden brought about by calculating OVESEG-test statistic for every subtype. Thus, a gene's p-value is calculated only for the subtype with highest mean, $p\text{-}value((K))$, corresponding to a positive $t_{k,obs}$. The OVESEG-test statistics of permutated samples, $T_{jk,p}$, are also calculated only for the subtype with the highest mean. The OVESEG-test statistics for other subtypes must be negative and could be set to be any negative constant because $t_{k,obs} > T_{jk,p}$ always holds in this case. Then the multiple-tailed p-value becomes

$$p\text{-}value = \min_{k=1,\dots,K} \{p\text{-}value(k)\} * K \doteq p\text{-}value((K)) * K. \tag{3.14}$$

The right hand of Eq. 3.14 does not always hold, resulting in p-values > 1, which is also a drawback of tail-doubling for two-tailed p-values. However, p-values for subtypes other than (K) are unimportant. Even if the smallest tail appears under other subtypes, this gene is less likely to be an adequate MG. "Multiplying a one-sided p-value by K" can also be explained as multiple testing correction. Each molecule is actually tested K times but we only record the p-value associated with the highest expressed subtype.

Subtype-specific OVESEG-test p-values differ from the overall p-values in Eq. 3.12 when the OVESEG-test statistic is asymmetric across subtypes due to unbalanced sample sizes and/or unbalanced null hypothesis compositions. Null distribution of $H_{\text{non-MG},\,m}, m > 0$, is varied for each subtype either when sample sizes are unequal across subtypes (Eq. 3.6) or when the non-MGs is unevenly distributed in the scatter plot. One example for the latter case is that m subtypes are closer to each other than any others, causing these m subtypes to have a larger condition probability of $H_{\text{non-MG},\,K-m}$.

3.2.5 By-products of aggregating genewise posterior probability

Estimating condition probability of one subtype being upregulated under null

More subtypes will increase the complexity of null hypotheses. $P\{H_{\text{non-MG},\,m} | H_{\text{non-MG}}\}$, $m = 0, \ldots, K-2$, only reflects the component weights of $(K-1)$ null hypothesis types, identifying no specific subtype. By aggregating genewise posterior probability, we can obtain the probability of one subtype being upregulated conditioned on $H_{\text{non-MG}}$, which will affect the number of False Positive MGs of this subtype.

Suppose $H_{\text{non-MG},\,m}, m = 0, \ldots, K-2$, has an equal prior probability for $(K-m)$ directions. Under the composition of multiple null hypotheses, we get an estimated probability of one subtype being upregulated under null:

$$\Pr\{up\ in\ k | H_{\text{non-MG}}\} = \frac{\sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j) I\big(reOrder_{kj} \leq K-m\big)/(K-m)}{\sum_{m=0}^{K-2} \sum_{j=1}^{J} w_{\text{non-MG},\,m}(j)}, \qquad (3.15)$$

where $reOrder_{kj}$ is the reversed order of subtype k sorted by expressions of gene j across all subtypes. $I(*)$ indicates whether, for gene j, subtype k is among the highest expressed $(K-m)$ subtypes and thus being upregulated in probability $1/(K-m)$.

Landscape of gene expression patterns

There are $(2^K - 1)$ types of gene expression patterns in a real dataset: genes exclusively expressed in $1, 2, \ldots$, or K of subtypes. Genes unexpressed in any of subtypes have been eliminated during pre-processing. Aggregating the genewise posterior probability of null hypotheses, $w_{\text{non-MG},\,m}$, and of alternative hypotheses, $w_{\text{MG}}(j) = 1 - \sum_{n=0}^{K-2} w_{\text{non-MG},\,n}(j)$, can present us a landscape of complex expression patterns as

$$P\{H_{\text{non-MG(A)},\, m}\} = \frac{1}{J}\sum_{j=1}^{J} w_{\text{non-MG},\, m}(j) I\left(\left\{k \middle| reOrder_{kj} \le K - m\right\} = A\right),$$
$$A \subseteq \{1, \dots, K\}, |A| = K - m, 0 \le m < K - 2, \qquad (3.16a)$$

$$P\{H_{\text{MG}(k)}\} = \frac{1}{J}\sum_{j=1}^{J} w_{\text{MG}}(j) I\left(reOrder_{kj} = 1\right), 1 \le k \le K, \qquad (3.16b)$$

where $reOrder_{kj}$ is the reversed order of subtype k sorted by expressions of gene j across all subtypes. A could be any subset of $\{1, \dots, K\}$ with $(K - m)$ elements. $\left\{k \middle| order_{kj} \le K - m\right\}$ is a subset with elements being the indexes of $(K - m)$ highest expressed subtypes for gene j.

3.2.6 Within-subtype variance estimation

The importance of an accurate estimator on pooled within-subtype variance $\sigma^2(j)$ is twofold - calculating the OVESEG-test statistic t_j and determining the local false discovery rate $\text{fdr}_{\text{non-MG},\, m}(j)$, particularly with small sample size. We assume a scaled inverse chi-square prior distribution $\sigma^2(j) \sim \nu_0 \sigma_0^2 / \mathcal{X}_{\nu_0}^2$, where ν_0 and σ_0^2 are the prior degrees of freedom and scaling parameter, respectively [70]. We then adopt the empirical Bayes moderated variance estimator $\tilde{\sigma}^2(j)$ that leverages information across all genes, as used in *limma* and given by

$$\tilde{\sigma}^2(j) = \frac{\nu_0 \hat{\sigma}_0^2 + (N - K)\hat{\sigma}^2(j)}{\nu_0 + N - K}, \qquad (3.17)$$

where *N* is the total number of samples, and $\hat{\sigma}^2(j)$ is the pooled variance estimator, given by

$$\hat{\sigma}^2(j) = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\log s_k(i, j) - \mu_k(j)\right)^2}{N - K}. \qquad (3.18)$$

The prior parameters ν_0 and σ_0^2 are estimated from the pooled variances. The moderated variances shrink the pooled variances towards the prior values depending on the prior degrees of freedom and the number of samples. Note that $t\text{-}stat(j)$ with moderated variance estimator $\tilde{\sigma}^2(j)$ follows a *t*-distribution with $\nu_0 + N - K$ degrees of freedom. If mean-variance relationship exists in expression data, e.g. RNAseq data, “limma-voom” weights are incorporated into the linear modeling procedures to stabilize variance [71]. Other variance estimators designed for two-group differential analysis can also easily modified and integrated into the multiple-group comparisons by OVESEG-test. For example, ROTS [72] adds a constant to the pooled variance estimator to optimize reproducibility across bootstrap resamplings.

3.2.7 Brief review on the most relevant peer MG selection methods

The OVR-FC uses a simple test defined by

$$\text{OVR-FC}_k(j) = \frac{\bar{s}_k(j)}{\bar{s}_{\text{-}k}(j)}, \tag{3.19}$$

where $\bar{s}_k(j)$ and $\bar{s}_{\text{-}k}(j)$ are the geometric means of the jth gene expressions within subtype k and associated with the combined remaining subtypes, respectively.

The OVR t-test uses a statistical test given by

$$\text{OVR t-stat}_k(j) = \frac{\hat{\mu}_k(j) - \hat{\mu}_{-k}(j)}{\sqrt{\frac{\hat{\sigma}_k(j)}{N_k} + \frac{\hat{\sigma}_{-k}(j)}{N - N_k}}}, \tag{3.20}$$

where $\hat{\mu}_k(j)$ and $\hat{\mu}_{-k}(j)$ are the sample means of the jth gene expressions within subtype k and associated with the combined remaining subtypes, respectively; and $\hat{\sigma}_k(j)$ and $\hat{\sigma}_{-k}(j)$ are the sample variance of the jth gene expressions within subtype k and associated with the combined remaining subtypes, respectively. Note the degrees of freedom in OVR t-test (Welch's t-test) vary across genes, generating different null distributions for each gene’s test. Therefore, ranking genes based on OVR t-stat is different from that based on OVR t-test p-values.

The OVE-FC (our earlier version of OVESEG-test) is defined as

$$\text{OVE-FC}_k(j) = \frac{\bar{s}_k(j)}{\max\limits_{l \neq k} \bar{s}_l(j)}, \tag{3.21}$$

with various variations [17, 62]. As aforementioned, OVESEG-test evolves logically from our earlier work on OVE-FC for multiclass classification. It is worth mentioning that the ideal MGs detected by OVE-FC with a stringent threshold for supervised deconvolution [6, 13], is principally similar to what detected by the Convex Analysis of Mixtures (CAM) for fully unsupervised deconvolution [22, 27], i.e. the marker genes resided near the vertices of the scatter simplex.

The OVO t-test conducts t-tests among all subtype pairs and selects genes upregulated in one subtype for all the involved tests, where the variances are estimated only from every pair of subtypes [65]. The OVO t-test p-value can be calculated as

$$p\text{-}value = \max_{l \neq (K)} \{ p_{(K)l} \} \tag{3.22}$$

where $p_{(K)l}$ is the p-value of differential test between subtype (K) and subtype l. For a fair comparison between OVO t-test and OVESEG-test, moderated variance estimator "limma" is applied to pairwise differential analyses in OVO t-test. As moderated variance estimator depends on the information across all genes for each subtype pair, all the $K(K-1)/2$ subtype pairs for all genes have to be tested. While OVO t-test leverages information from each subtype pair for variance modeling, OVESEG-test leverages information from all subtypes. The benefit of using all subtypes for modeling is significant in challenging cases with higher variance, smaller sample size, and more subtypes. Moreover, OVESEG-test re-estimates p-value by novel permutation scheme to model the complex null distribution.

3.3 Results

3.3.1 Validation on type 1 error

To test whether our OVESEG-test statistics can detect MGs at appropriate significance levels, we assessed the type 1 error using simulation studies under the null hypothesis. Accuracy of type 1 error is crucial for any hypothesis testing methods that detect MGs based on their p-values because if the type 1 error is either too conservative or too liberal, the p-value is inflated by either too many false positive or false negative estimates, loses its intended meaning, and becomes difficult to interpret.

The simulation data contain 10,000 genes whose baseline expression levels are sampled from the real benchmark microarray gene expression data with replicates of purified cell populations (GSE19380 [6]). Among the simulated genes, a significant portion are housekeeping genes that take the baseline expression levels across all subtypes under $H_{\text{non-MG},\,0}$. Expression levels of the remaining genes are adjusted to exhibiting similar upregulations in at least two subtypes, mimicking all types of participating null hypotheses. The upregulations are modeled by uniform distribution(s) in scatter space (black dash line in Fig. 3.1), with variance following an inverse chi-square distribution $\sigma^2(j) \sim \nu_0 \sigma_0^2 / \mathcal{X}_{\nu_0}^2$, where the prior degree of freedom ν_0 takes 5 or 40, and σ_0 takes 0.2, 0.5, or 0.8.

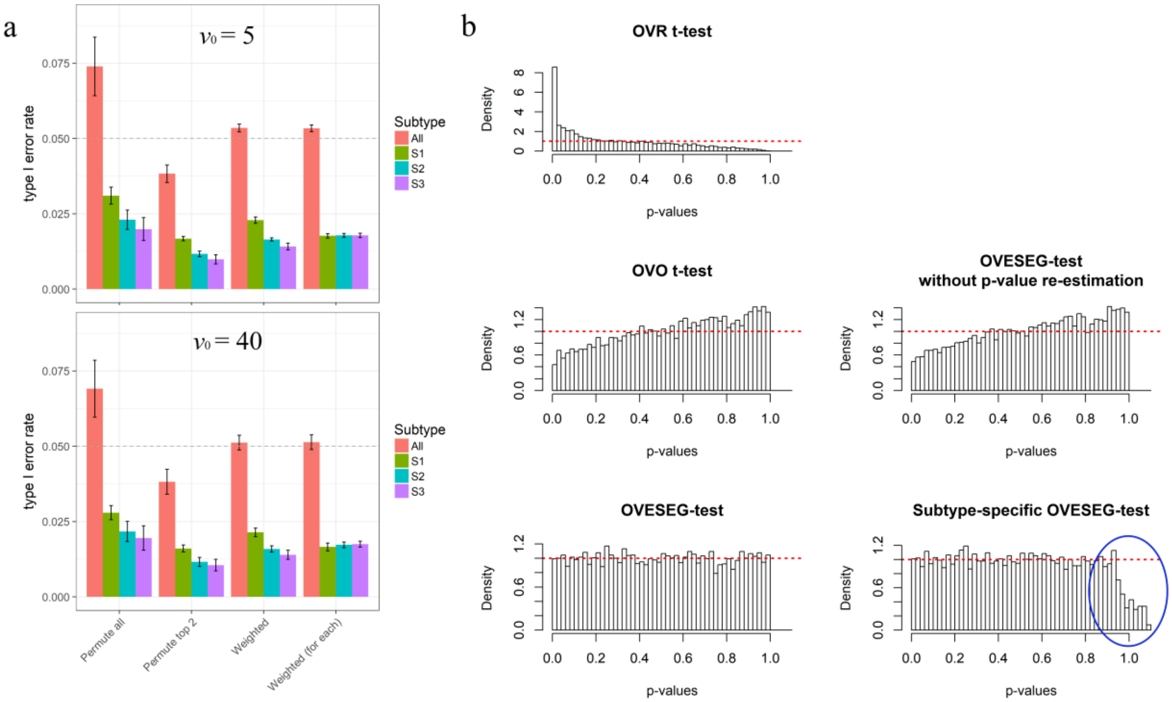


Figure 3.3 Assessment on Type 1 error rates and p-value distributions using simulated data sets under null hypothesis, involving three subtypes with unbalanced sample sizes ($N_1 = 3, N_2 = 6, N_3 = 9$). (a) Bar chart for the mean and 95% confidence interval of type 1 error rates with p-value cutoff at 0.05 over 30 simulated experiments, showing both overall and subtype-specific false positive rates corresponding to different permutation schemes. (b) Histograms of p-value distributions associated with five MG detection methods, where simulation data consisted of 60% housekeeping genes, $\sigma_0 = 0.5$ and $\nu_0 = 40$. Note that subtype-specific p-values can be higher than 1.0 after multiple testing correction and thus will be truncated (indicated by the blue circle; see Supplementary Information for details).

Fig. 3.3 and Fig. 3.4 show type 1 error control in the setting of three subtypes with balanced/unbalanced sample sizes. While permuting all subtypes generates a less dispersed null distribution, and permuting only the top two subtypes results in a more compact distribution, our posterior weighted permutation scheme can achieve an automatic balance. Moreover, when the percentage of housekeeping genes (i.e. $H_{\text{non-MG},\,0}$) increases, the null distribution of OVESEG-test statistics tends to be that generated from permuting all three subtypes. Conversely, those calculated by permuting the top two subtypes approximate true p-values in $H_{\text{non-MG},\,1}$-dominated experiments. Comparison of results with different σ_0 demonstrates our test can control the type 1 error rate even in very noisy scenarios. However, small ν_0 generates a modestly inflated type 1 error control, especially when sample sizes are small and the true $H_{\text{non-MG},\,1}$ dominates. This effect may be due to less reliable estimation of the moderated variance caused by small ν_0 and sample sizes.

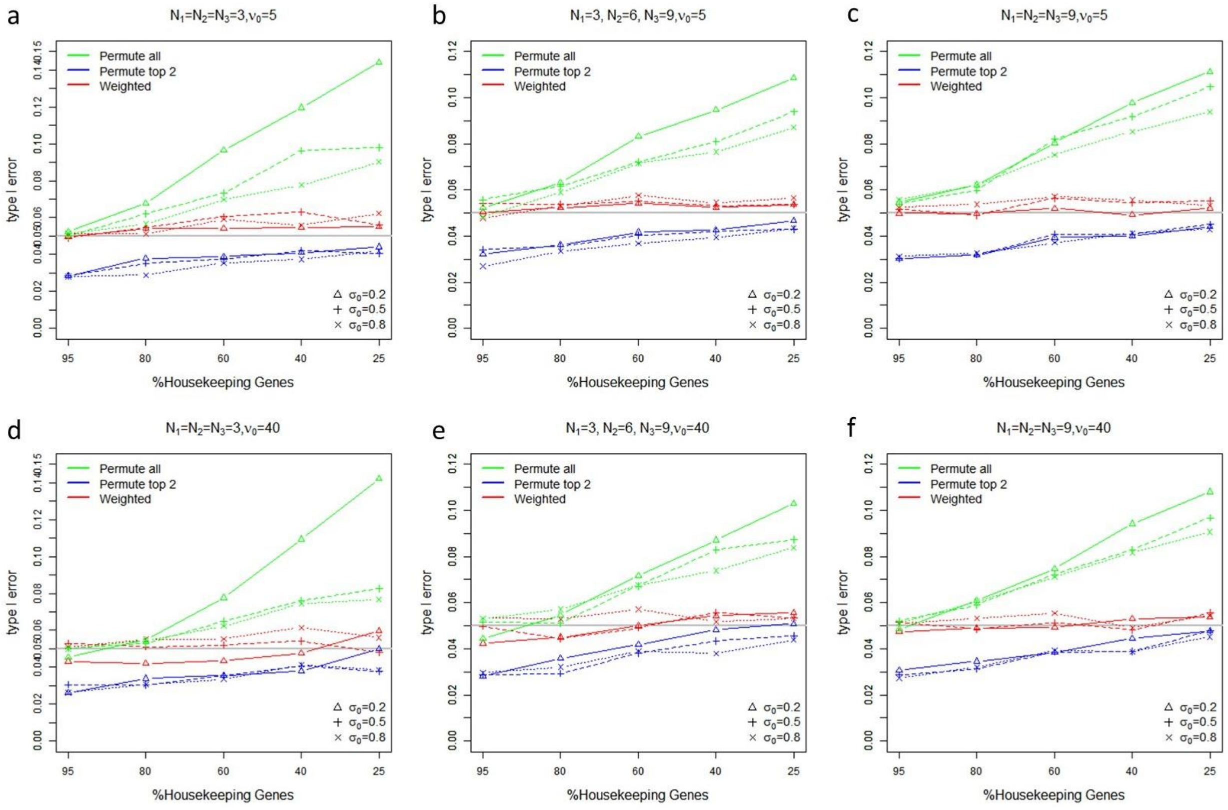


Figure 3.4 Comparisons of type 1 error rates from different permutation schemes under different settings of noisy scenarios and sample sizes (housekeeping genes: 95%, 80%, 60%, 40%, or 25%; σ_0: 0.2, 0.5, or 0.8; ν_0: 5(top), or 40(bottom)).

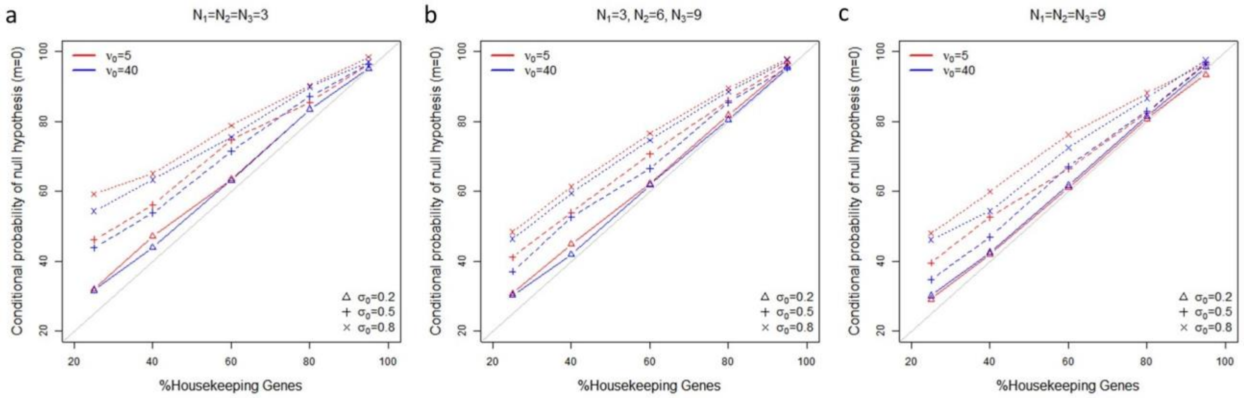


Figure 3.5 Comparisons of estimated conditional probabilities of null hypothesis $H_{\text{non-MG, 0}}$, $P\{H_{\text{non-MG, 0}} | H_{\text{non-MG}}\}$, versus the true proportions of housekeeping genes, under different settings of noisy scenarios and sample sizes.

The estimated conditional probabilities of $H_{\text{non-MG, 0}}$, $P\{H_{\text{non-MG, 0}} | H_{\text{non-MG}}\}$ in our model are expected to match the true proportions of housekeeping genes. As seen from **Fig. 3.5**, in noisy scenarios with high σ_0, $P\{H_{\text{non-MG, 0}} | H_{\text{non-MG}}\}$ tends to get over-estimated, implying that many genes sampled from the true $H_{\text{non-MG, 1}}$ are treated as being from $H_{\text{non-MG, 0}}$ in certain weights. In fact, these true $H_{\text{non-MG, 1}}$ genes have no significantly large $\Delta_1(j)$, which is the difference between the top two subtypes and the third subtype. Thus, the null distributions generated from them are expected to be intermediate between those from true $H_{\text{non-MG, 0}}$ (permuting three subtypes) and from true $H'_{\text{non-MG, 1}}$ (permuting top 2 subtypes).

Subtype-specific OVESEG-test tests MGs of each subtype separately, having the benefit that all subtypes a broadly similar type 1 error rate (**Fig. 3.3a**). Otherwise, the subtype consisting of smaller sample size will contribute more False Positive MGs. Note that subtype-specific p-values could be larger than 1 after multiple-tail scaling (**Fig. 3.3b**) and need to be truncated at 1. While we only calculate the p-value for the subtype with the largest group mean to reduce the computational burden, we ignore the possibility that genes with large p-values (around one) could have smaller p-values associated with another subtype. Since these genes are not likely to be valid MGs, it is not necessary to provide their precise p-values.

All the above analyses are repeated for MG identification involving five subtypes, producing p-values that control error rates correctly over a wide range of simulation scenarios (**Fig. 3.6a**). Since we randomly assigned genes under $H_{\text{non-MG},\,m}, m > 0$ without considering subtypes in the simulations above, the null hypothesis compositions are almost the same for all subtypes. To check the type 1 error rate under unbalanced null hypothesis compositions, five baseline profiles are generated by simulating two cell lines but assigning one to two subtypes and the other to three subtypes, so that the first cell line's up-regulated genes become two subtypes' $H_{\text{non-MG},\,3}$ genes, while the second cell line's up-regulated genes become three subtypes' $H_{\text{non-MG},\,2}$ genes. No true MGs exist for any of the five subtypes. OVESEG-test (overall or subtype-specific) controls type 1 error rates well but with more False Positive MGs in the first two subtypes (**Fig. 3.6b**). The unbalance of null hypotheses leads to different probability estimates of a specific subtype being upregulated (**Eq. 3.15**). As simulated up-regulated genes in two cell lines are the same but divided into two or three subtypes evenly, the first two subtypes have a higher probability of being upregulated, thus more False Positive MGs are expected (**Fig. 3.7**). In our simulations, when the

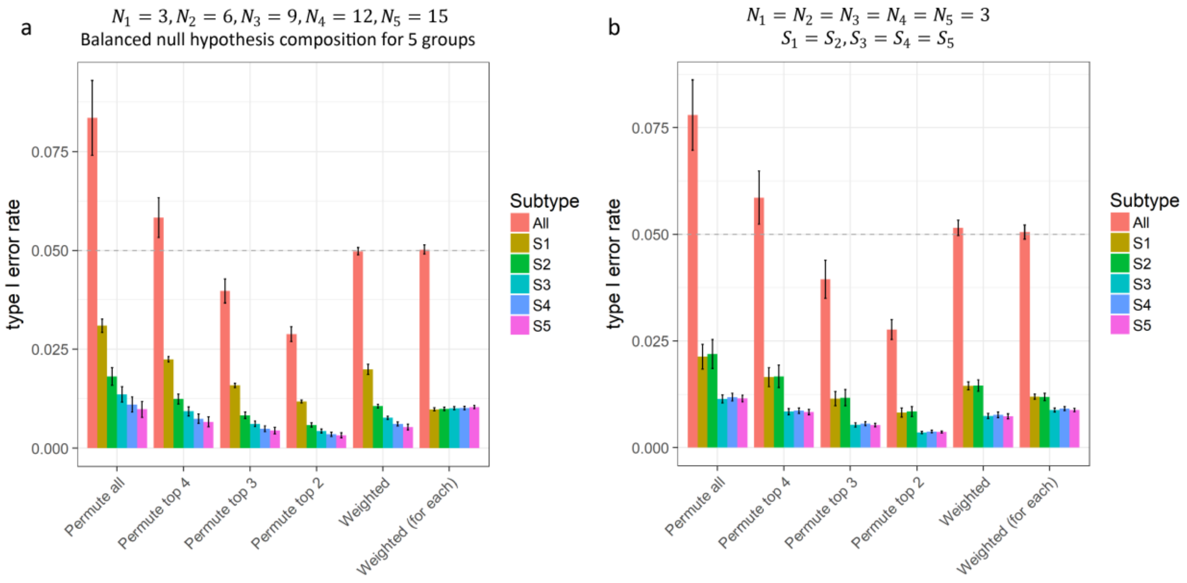


Figure 3.6 Assessment on Type 1 error rates using simulated data sets involving five subtypes. The results are obtained using p-value cutoff at 0.05 over 30 experiments. (a) Bar chart of the mean and 95% confidence interval of type 1 error rates with unbalanced sample sizes. (b) Bar chart of the mean and 95% confidence interval of type 1 error rates with unbalanced compositions of mixture null distribution.

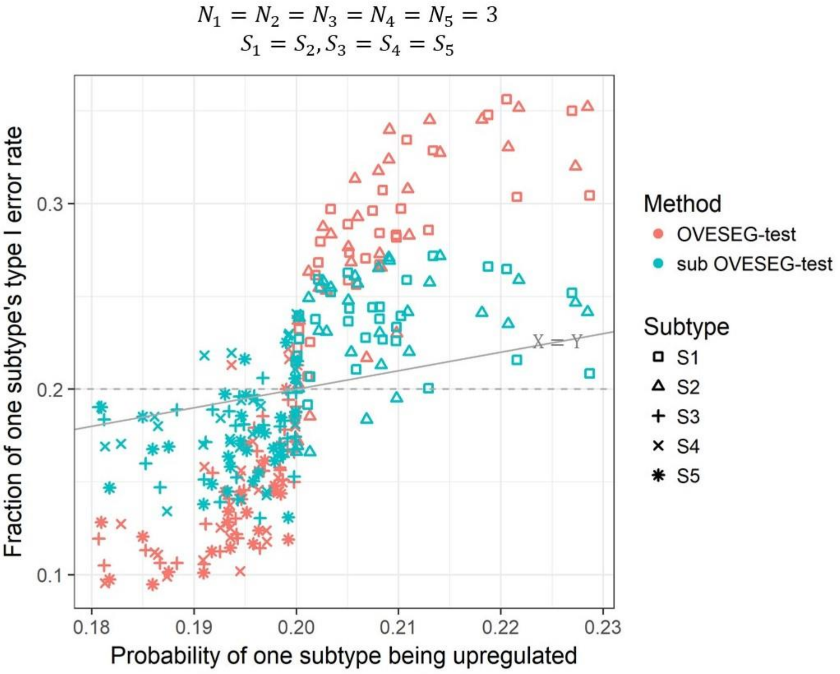


Figure 3.7 Fraction of type 1 error in each of five subtypes, versus probability of each subtype being upregulated under $H_{\text{non-MG}}$ (estimated by Eq. S9). Each point is associated with one of simulation settings (housekeeping genes: 95%, 80%, 60%, 40%, or 25%; σ_0: 0.2, 0.5, or 0.8; ν_0: 5, or 40). Sample size is three per subtype. The first two subtypes are drawn from the same one population and the remaining three drawn from another.

ratio of housekeeping genes is high, the unbalance becomes slight and approximately 1/5 of False Positive MGs are allocated to each subtype. Conversely, a low ratio of housekeeping genes intensifies the uneven distribution of non-MGs in the scatter plot, increasing the number of False Positive MGs detected in the first two subtypes. Subtype-specific tests can reduce the impact of this effect.

Overall, using realistic simulation data sets with various parameter settings, we show that in all scenarios the empirical type 1 error produced by OVESEG-test statistics closely approximates the expected type 1 error (Fig. 3.3a, Fig. 3.4, Fig. 3.6a-b). The p-values associated with OVESEG-test statistics exhibit a uniform distribution as expected (Fig. 3.3b). Specifically, even with unbalanced sample sizes among the subtypes, the mixture null distribution estimated by our posterior weighted permutation scheme produces the expected empirical type 1 error rate (Fig. 3.4 and Fig. 3.6a). In contrast, the empirical type 1 error produced by OVR t-test and OVO t-test either over-estimates or under-estimates the expected type 1 error. The p-values associated with OVR t-test and OVO t-test deviate from a uniform distribution (Fig. 3.3b). We also evaluate the type 1 error under high noise levels and small sample sizes using subtype-specific p-value estimates. For each of the subtypes, experimental results again show that the empirical type 1 error produced by OVESEG-test statistics closely matches the expected type 1 error (Fig. 3.3). Besides, subtype-specific p-value estimates can even effectively balance the uneven type 1 error rates among the subtypes with different numbers of upregulated genes (Fig. 3.6b).

3.3.2 Comparative assessment on FDR and detection power

FDR control is an important factor when assessing the detection power involving true MGs. For a well-designed significance test, the objective is to maximize power while controlling FDR below the allowable level. To test whether the q-value reflects the true FDR, ‘fdrtool’ package is used to estimate the q-value for each gene [67], where the FDR with an estimated q-value of 0.05 is expected to be around 0.05. Another informative criterion is the partial area under curve (pAUC) that emphasizes the leftmost portion of the receiver operating characteristic (ROC) curve, focusing on the sensitivity at a low False Positive Rate(FPR) [73]. As ROC curve reflects whether the methods are able to rank true MGs above true non-MGs, pAUC can assess the sensitivity or power of MG detection methods when the FPR is significantly low, such as 0.05/0.01 cutoff. Hence, we use pAUC with specificity larger than 0.95/0.99 (equivalent to FPR less than 0.05/0.01) to evaluate the power of each method. Another reason for preferring pAUC rather than AUC is that the emergence of mismatched detections (True Positive but associated with incorrect subtypes) with low specificity may exhibit a false high power.

For power considerations, we simulated a comprehensive set of scenarios to compare the power of OVESEG-test statistics and peer methods in detecting subtype-specific MGs. Simulation data are generated, similarly as above, by modifying the expression levels of real gene expression data. Variance is either drawn from an inverse chi-square distribution with different ν_0 and σ_0, or sampled from real data (microarray data GSE28490, RNAseq data GSE60424) with keeping mean-variance trend. A portion of the genes are designated as MGs with exclusive and consistent upregulation in each of the participating subtypes, with fold change drawn randomly in certain ranges. To recapitulate the characteristics of real expression data, various parameter settings are considered including unbalanced sample sizes or diverse mixture null distribution cross subtypes, each with 20 replications. We conducted three simulation sets: (1) microarray simulations with variance drawn from an inverse chi-square distribution (K=3); (2) microarray simulations with variance sampled from real data (K=7); (3) RNASeq simulations with variance sampled from real data (K=7). The latter two simulation sets are more challenging with more subtypes and noisy RNASeq data, which help demonstrate the benefit of using all subtypes for modeling by the proposed OVESEG-test.

In the first simulation set, genewise variation was drawn from $\sigma^2(j) \sim \nu_0 \sigma_0^2 / \mathcal{X}_{\nu_0}^2$, with the prior degree of freedom ν_0 being 5 or 40 related to less or more stabilized variances, respectively. σ_0 is set to be 0.2, 0.5, or 0.8 to check the performance under different noisy scenarios. 20% of genes are upregulated in one subtype with fold change following a uniform distribution in the Ternary plot (Fig. 3.1a). Non-MGs under $H_{\text{non-MG, 0}}$ and under $H_{\text{non-MG, 1}}$ account equally for the remaining genes. Two scenarios are tested with unbalanced sample sizes or unbalanced null hypothesis

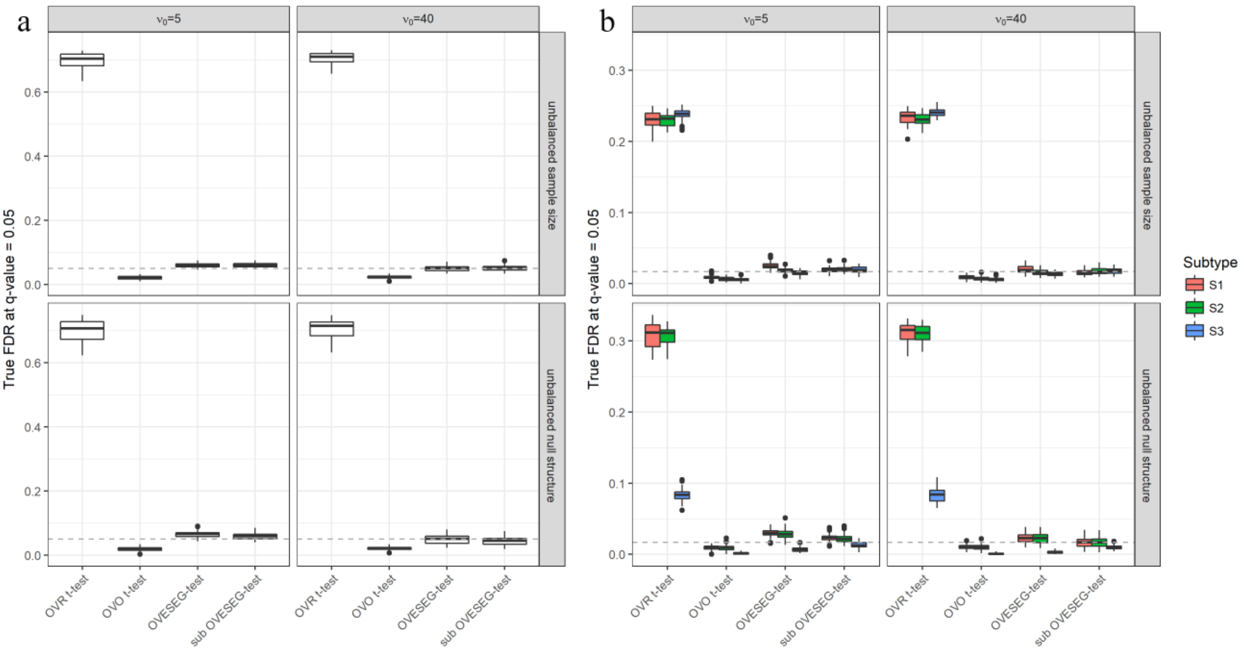


Figure 3.8 FDR control under the multiple simulation settings with three unbalanced subtypes. (a) True FDR at q-value =0.05 across all subtypes (dash line is at 0.05); (b) True FDR at q-value =0.05 in each subtype (dash line is at 0.05/3).

compositions. In the former, the sample size in each subtype is 3, 6, and 9, while $H_{\text{non-MG},\,1}$ genes are distributed evenly. In the latter, each subtype has three samples and $H_{\text{non-MG},\,1}$ genes only appear between the first two subtypes, which is a common case in real data: two subtypes are closer to each other than to the other subtypes. Each simulation setting is repeated 20 times. Both the overall and subtype-specific OVESEG-tests can control FDR around the expected 0.05 level at q-value cutoff 0.05, with modestly weaker control in the case of small ν_0 (Fig. 3.8). OVR t-test is too liberal, while OVO t-test is conservative. Furthermore, a subtype-specific test can alleviate the unbalance of False Positive MGs among subtypes. Other peer methods detect more False Positive MGs in subtypes with a small sample size or with a large probability of being upregulated under $H_{\text{non-MG}}$ (Eq. 3.15). In terms of pAUC, OVESEG-test achieves better performance than peer methods (Table 3.1). OVE-FC and OVO t-test show a comparable detection power in less noisy cases. Clearly and expectedly, all three OVR methods exhibit a weaker detection power while ANOVA has the lowest detection power.

Table 3.1 pAUC (FPR<0.05) obtained from Microarray simulations involving 3 subtypes and with various experimental settings

		$\nu 0 = 5$			$\nu 0 = 40$		
		$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.8$	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.8$
Balanced Null Hypothesis Structure Unbalanced Sample Size (n=3,6,9)	ANOVA	0.53747	0.5402	0.54266	0.53794	0.54081	0.54344
	OVR-FC	0.67778	0.65644	0.62664	0.68155	0.66657	0.64723
	OVR t-stat	0.9093	0.77705	0.6904	0.92751	0.80715	0.71382
	OVR t-test	0.82649	0.70774	0.64513	0.84036	0.73137	0.66212
	OVO t-test	0.94982	0.8569	0.76983	0.96193	0.89067	0.80318
	OVE-FC	0.94292	0.84033	0.74264	0.96179	0.88991	0.80195
	OVESEG-test	**0.95116**	**0.85907**	**0.77208**	**0.96196**	**0.89074**	**0.80403**
	sub OVESEG-test	**0.95119**	**0.85915**	**0.77301**	**0.96211**	**0.89078**	**0.80424**
Unbalanced Null Hypothesis Structure (n=3,3,3)	ANOVA	0.54128	0.54473	0.54533	0.54105	0.54472	0.54411
	OVR-FC	0.70822	0.67034	0.62272	0.71086	0.68668	0.65558
	OVR t-stat	0.92793	0.78802	0.68392	0.94068	0.81609	0.70628
	OVR t-test	0.89682	0.73482	0.6429	0.91385	0.75827	0.65584
	OVO t-test	0.93685	0.8218	0.72267	0.95218	0.86223	0.76863
	OVE-FC	0.93113	0.80302	0.69664	0.9522	0.86167	0.7676
	OVESEG-test	**0.93887**	**0.82354**	**0.72451**	**0.95211**	**0.86229**	**0.76811**
	sub OVESEG-test	**0.94075**	**0.82749**	**0.72821**	**0.95383**	**0.86651**	**0.77229**

Table 3.2 pAUC (FPR<0.05 and 0.01) obtained from Microarray simulations involving 7 subtypes and with various experimental settings

		pAUC (FPR<0.05)			pAUC (FPR<0.01)		
		FC ∈ [2,20]	FC ∈ [5,20]	FC ∈ [10,20]	FC ∈ [2,20]	FC ∈ [5,20]	FC ∈ [10,20]
Unbalanced Null Hypothesis Structure 3 samples /per subtype	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	NA	0.50997	0.56310	NA	NA	0.50035
	OVR t-stat	0.81362	0.92408	0.96226	0.80365	0.91644	0.95595
	OVR t-test	0.55572	0.57203	0.59205	0.56113	0.57531	0.59397
	OVO t-test	0.93403	0.98283	0.98955	0.89324	0.96797	0.97626
	OVE-FC	0.93612	**0.99533**	**0.99878**	0.81569	**0.98423**	**0.99759**
	OVESEG-test	**0.94797**	0.98833	0.99348	**0.92393**	0.97953	0.98765
	sub OVESEG-test	**0.95052**	0.98897	0.99384	**0.92372**	0.98028	0.98764
Balanced Null Hypothesis Structure 3 samples /per subtype	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	NA	0.50671	0.55154	NA	NA	0.50275
	OVR t-stat	0.82253	0.92019	0.94936	0.80918	0.91560	0.94403
	OVR t-test	0.56269	0.58333	0.58408	0.56490	0.58558	0.58595
	OVO t-test	0.92870	0.98165	0.98961	0.88065	0.96383	0.97822
	OVE-FC	0.93763	**0.99568**	**0.99770**	0.82500	**0.98386**	**0.98897**
	OVESEG-test	**0.95044**	0.98929	0.99317	**0.91775**	0.98127	**0.98948**
	sub OVESEG-test	**0.95062**	0.98935	0.99316	**0.91830**	0.98163	**0.98947**
Balanced Null Hypothesis Structure Unbalanced Sample Size (3,3,3,4,5,5,5)	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	NA	0.50686	0.54375	NA	NA	0.50004
	OVR t-stat	0.79462	0.90719	0.94392	0.78328	0.89854	0.93593
	OVR t-test	0.52553	0.54197	0.55172	0.52574	0.54076	0.55029
	OVO t-test	0.95338	0.99001	0.99563	0.92252	0.98104	0.98687
	OVE-FC	0.95395	**0.99763**	**0.99882**	0.85616	**0.98841**	**0.99423**
	OVESEG-test	**0.96077**	0.99411	0.99711	**0.94228**	0.98681	0.99301
	sub OVESEG-test	**0.96101**	0.99418	0.99711	**0.94222**	0.98693	0.99317

Table 3.3 pAUC (FPR<0.05 and 0.01) obtained from RNAseq simulations involving 7 subtypes and with various experimental settings

		pAUC (FPR<0.05)			pAUC (FPR<0.01)		
		FC ∈ [2,20]	FC ∈ [5,20]	FC ∈ [10,20]	FC ∈ [2,20]	FC ∈ [5,20]	FC ∈ [10,20]
Unbalanced Null Hypothesis Structure 3 samples /per subtype	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	0.50007	0.51649	0.59034	0.50120	0.50260	0.52331
	OVR t-stat	0.63528	0.76419	0.81669	0.60403	0.70409	0.77145
	OVR t-test	0.54461	0.60287	0.63803	0.53923	0.58136	0.62174
	OVO t-test	0.77600	0.88615	0.93747	0.68080	0.76939	0.85403
	OVE-FC	0.75622	**0.93221**	**0.97183**	0.62926	0.80988	**0.90791**
	OVESEG-test	**0.79344**	0.91728	0.96178	**0.68795**	**0.81828**	0.89212

	sub OVESEG-test	**0.79869**	0.91810	0.96212	**0.69353**	**0.81877**	0.89454
Unbalanced Null Hypotheses 20 samples /per subtype	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	NA	0.53594	0.62450	NA	0.50883	0.53201
	OVR t-stat	0.61767	0.75120	0.82295	0.60970	0.72330	0.80238
	OVR t-test	0.50731	0.55117	0.60501	0.51059	0.54524	0.59214
	OVO t-test	0.97075	0.99193	0.99578	**0.96192**	0.98840	0.99354
	OVE-FC	**0.97611**	0.99369	0.99660	0.95103	0.98974	0.99263
	OVESEG-test	0.97232	**0.99448**	**0.99782**	0.95100	**0.99131**	**0.99632**
	sub OVESEG-test	0.97363	**0.99438**	**0.99770**	0.95323	**0.99093**	**0.99645**
Balanced Null Hypothesis Structure 3 samples /per subtype	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	NA	0.52764	0.59584	NA	0.50463	0.52588
	OVR t-stat	0.63400	0.76959	0.82773	0.60631	0.72285	0.77652
	OVR t-test	0.54489	0.61288	0.63604	0.54151	0.59807	0.62051
	OVO t-test	0.77225	0.89831	0.93346	0.67555	0.79982	0.86136
	OVE-FC	0.74739	**0.92323**	**0.97642**	0.61980	0.78167	**0.91866**
	OVESEG-test	**0.77921**	0.91873	0.96874	**0.68426**	**0.81631**	0.91196
	sub OVESEG-test	**0.77931**	0.91857	0.96867	**0.68397**	**0.81483**	0.91164
Balanced Null Hypothesis Structure Unbalanced Sample Size (3,3,3,4,5,5,5)	ANOVA	NA	NA	NA	NA	NA	NA
	OVR-FC	NA	0.51603	0.57669	NA	0.50377	0.51511
	OVR t-stat	0.61345	0.74834	0.79636	0.59512	0.70675	0.76168
	OVR t-test	0.53701	0.58239	0.62472	0.53539	0.56349	0.60675
	OVO t-test	0.79078	0.91174	0.93552	0.69959	0.82953	0.87458
	OVE-FC	0.75689	**0.95120**	**0.97319**	0.61997	**0.85831**	**0.92000**
	OVESEG-test	**0.79951**	0.92714	0.95949	**0.70584**	0.84436	0.89120
	sub OVESEG-test	**0.79978**	0.92689	0.95925	**0.70544**	0.84413	0.88982

In the second and third simulation set, genewise variation was sampled from microarray data GSE28490 or RNAseq data GSE60424. To keep potential mean-variance trend, we divided genes in real data into 100 buckets based on their logarithmic expression means and randomly selected a real variance for each simulated gene from the bucket which the simulated expression mean fell in. 100 MGs were simulated for each subtype while the remaining genes were simulated as different yet realistic types of non-MGs, $H_{\text{non-MG},\,m}, m = 0,1,\ldots,5$. The null hypothesis structure, i.e. the percentage of different types of non-MGs for each subtype, was generated based on the estimated probability of gene expression patterns (Eq. 3.16a) from real data. Since this null hypothesis structure from real data is somewhat unbalanced, we also simulated a balanced null hypothesis structure with each subtype having the same percentage of different types of non-MGs and same size being balanced or unbalanced. To compare the performance of detecting ideal/strict

MGs (significantly large fold change) or detecting more realistic MGs (sufficiently large fold change), the fold changes of simulated MGs are drawn from a uniform distribution with different range: [2,20], [5,20], or [10,20]. For pAUC, the OVESEG-test strategy and its earlier version OVE-FC approach the highest power in detecting true MGs (Figure 3.9, Figure 3.10, Table 3.2-3.3), where OVESEG-test performs the best for more realistic MGs (sufficiently large fold change). For more ideal/strict MGs (significantly large fold change), OVESEG-test and OVE-FC both achieve the best performance with slight outperformance by OVE-FC.

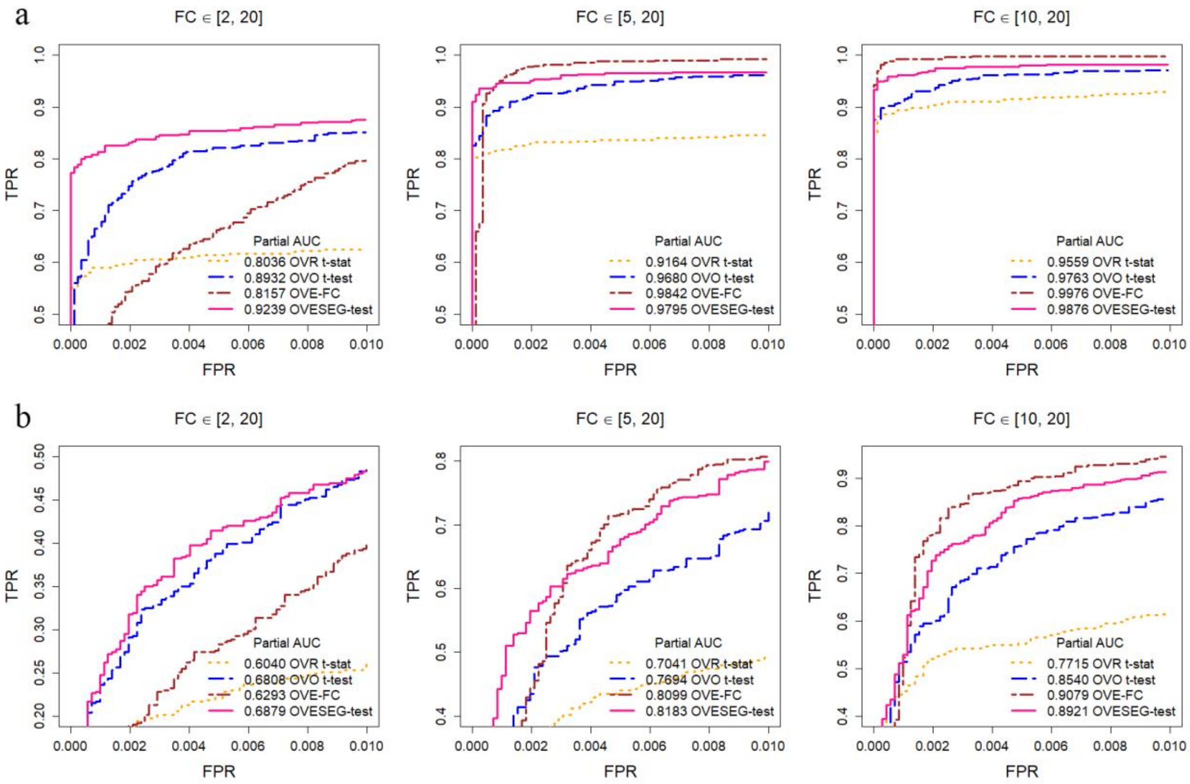


Figure 3.9 Assessment on detection power (partial ROC curves, FPR < 0.01) using real data derived simulations (data distribution is consistent with the base real dataset under null hypothesis) involving seven unbalanced subtypes with various parameter settings. (a) partial ROC curves across different FPR points on microarray-derived data. (b) partial ROC curves across different FPR points on RNAseq-derived data. (OVR-FC and OVR t-test are not shown here due to low pAUC; subtype-specific OVESEG-test's performance is quite similar to OVESEG-test; more complete ROC curves can be found in Fig. 3.10)

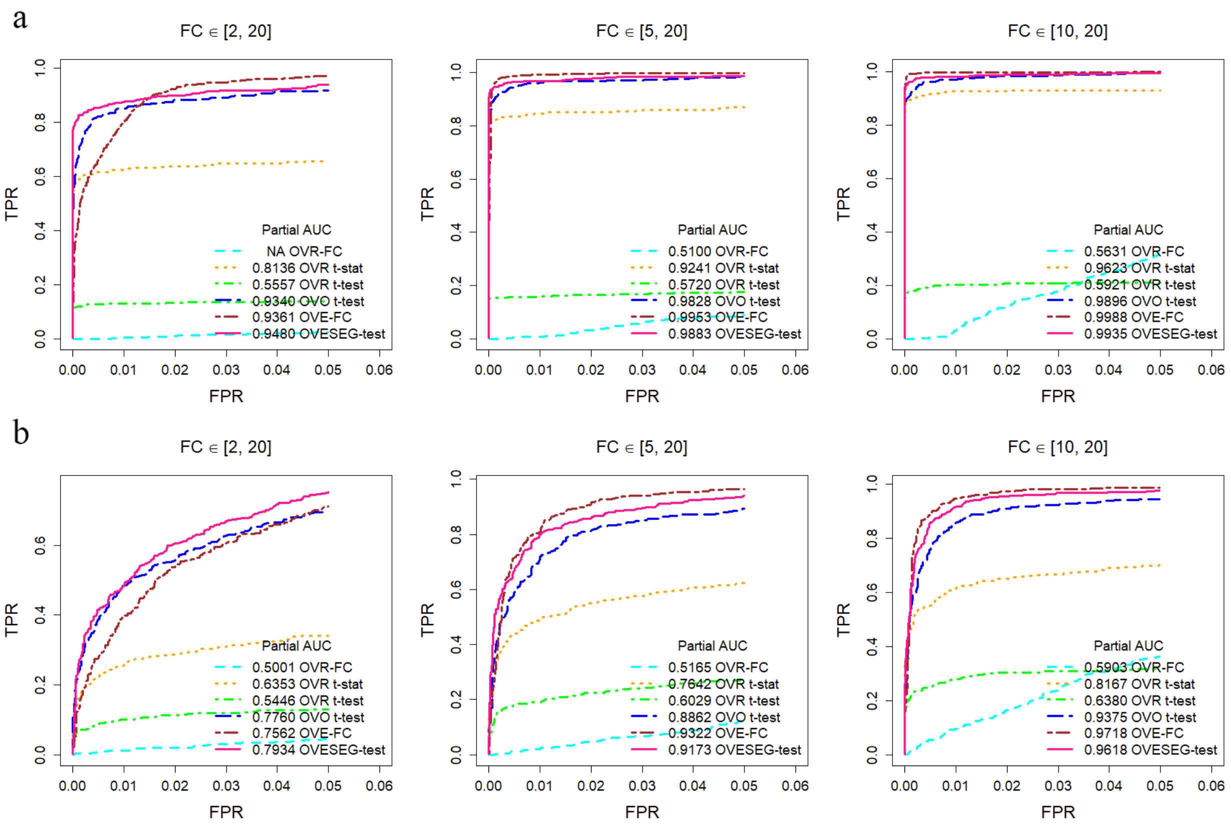


Figure 3.10 Assessment on detection power (partial ROC curves, FPR < 0.05) using real data derived simulations (data distribution is consistent with the base real dataset under null hypothesis) involving seven unbalanced subtypes with various parameter settings. (a) partial ROC curves across different FPR points on microarray-derived data. (b) partial ROC curves across different FPR points on RNAseq-derived data.

Specifically, the OVESEG-test consistently outperforms the OVO t-test in these experiments that represent more challenging cases involving more subtypes and using RNAseq data. More importantly, the outperformance of the proposed OVESEG-test (and its earlier version OVE-FC) over peer methods at a stringent FPR range in ROC analysis is significant and important because FDR is problematic in many real-world applications where a large number of multiple comparisons are involved. Furthermore, all three OVR methods exhibit lower detection power, and ANOVA has the lowest detection power as expected.

When sample size is small, the OVESEG-test leverages the limma method to borrow information across genes in estimating *a priori* variance, thus stabilizing the estimated variance for each gene. In addition, the OVESEG-test estimates the parameters of the limma model from all subtypes, producing better results than that applying t-test independently with the limma model for each subtype pair. Indeed, with a small sample size, our experimental results show that the proposed OVESEG-test clearly outperforms the OVO t-test (Tables 3.2-3.3).

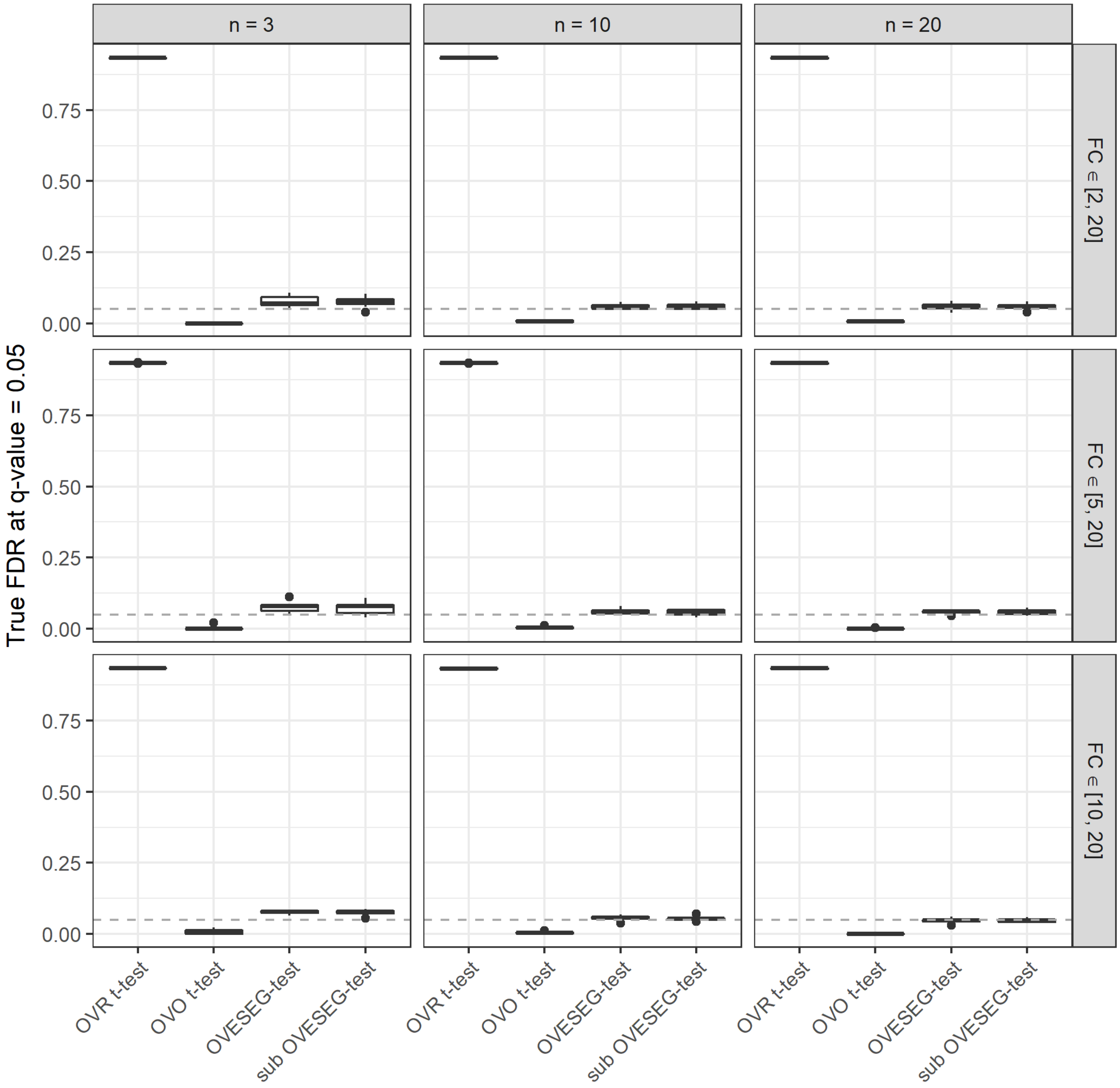


Figure 3.11 FDR control under the multiple simulation settings in RNAseq-derived simulations (dash line is at 0.05). Sample size is 3, 10 or 20 per subtype. Fold change range of MGs is [2,20], [5,20], or [10,20]. Non-MG distribution is consistent with the base real dataset under null hypothesis. Each simulation setting is repeated 20 times.

In the OVESEG test, the proposed permutation scheme does not require the data to be normally distributed. To ensure that the null distributions across genes can be aggregated together, variances require standardization and stabilization. We conducted experiments for RNAseq-data with different samples sizes and fold change ranges. The results demonstrate the OVESEG-test can maintain the expected type 1 error rates or specified FDR, with the mean-variance relationship estimated by limma-voom on RNASeq data (Fig. 3.11).

3.3.3 Application on gene expression data of purified human immune cells

We applied OVESEG-test to two real microarray gene expression data sets, GSE28490 (Roche) and GSE28491 (HUG), to detect subtype-specific markers associated with human immune cells [74]. In these data sets, the constituent subtypes are seven human immune cells isolated from healthy human blood: B cells, CD4+ T cells, CD8+ T cells, NK cells, monocytes, neutrophils, and eosinophils. Each cell subtype consists of at least five samples, excluding a few outliers (Table 3.4). Following preprocessing of the raw measurements and elimination of low expressed probesets (average $\log_2$ RMA signal value < 6 in any of the cell type groups), 12,022 probesets in Roche and 11,339 probesets in HUG were retained and used in the analyses. Because Roche and HUG used the same protocols for cell isolation and sample processing from two independent panels of donors, the derived gene expression enable the use of a cross-validation strategy.

Table 3.4 Sample size of each cell type in four datasets

	GSE28490	GSE28491	GSE60424	GSE72056*
B cells	5	5	20	628
CD4+ T cells	5	5	20	873
CD8+ T cells	5	5	20	1099
Regulatory T cells	NA	NA	NA	141
NK cells	5	5	14	89
Eosinophils	4	3	NA	NA
Macrophages/Monocytes	10	5	20	170
Neutrophils	3	5	20	NA
Plasmacytoid dendritic cells	5	NA	NA	26
Myeloid dendritic cells	5	NA	NA	NA
Endothelial cells	NA	NA	NA	71
Cancer associated fibroblasts	NA	NA	NA	96

*Cell type labels are from single-cell classification results [75]

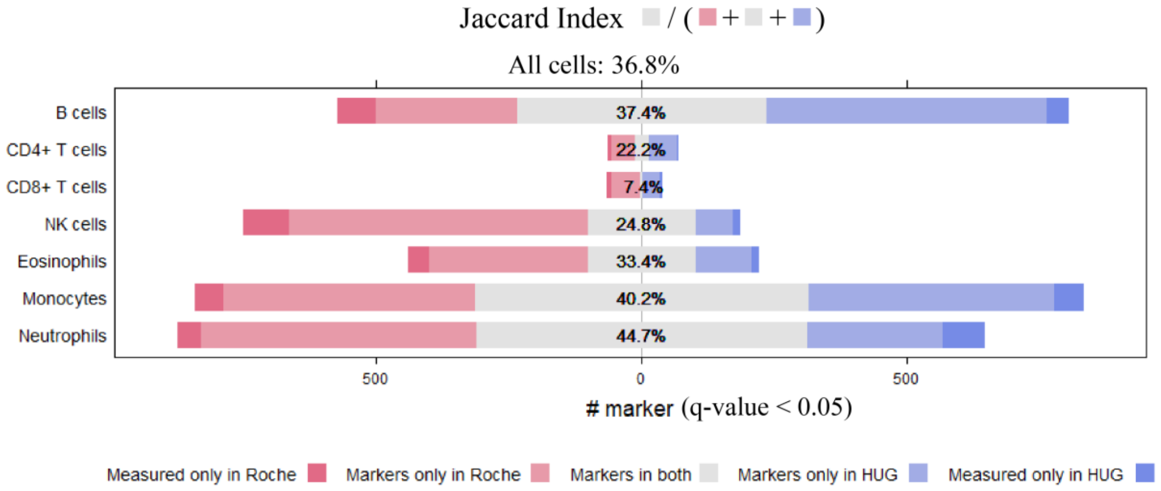


Figure 3.12 Percentile overlap of cell-type specific MGs between Roche and HUG datasets, quantified by Jaccard index (intersection over union). MGs are detected by subtype-specific OVESEG-test with q-value < 0.05).

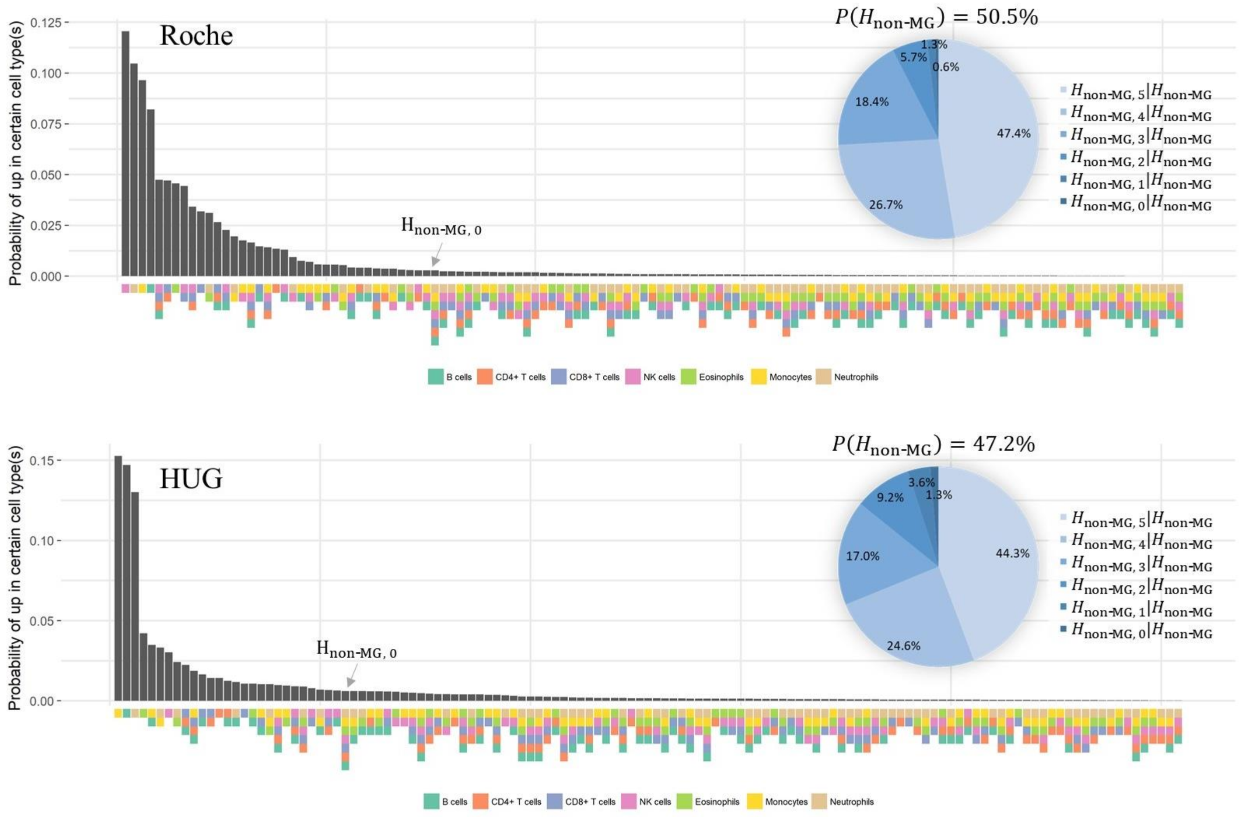


Figure 3.13 Landscape of 127 ($= 2^7 - 1$) gene expression patterns. Estimated probabilities of exclusively expression in any certain cell type(s) are ordered in decreasing value. The pie charts show conditional probability of each null hypothesis.

Table 3.5 Counts of cell-type specific markers under certain threshold

Threshold	Subtype	Measured only in Roche	Markers only in Roche	Markers in both	Markers only in HUG	Measured only in HUG	Total
q-value <0.05	B cells	70	266	474	528	39	1377
	CD4+ T cells	6	44	28	54	1	133
	CD8+ T cells	8	55	7	33	2	105
	NK cells	83	563	208	68	13	935
	Eosinophils	37	301	204	106	11	659
	Monocytes	51	475	630	463	55	1674
	Neutrophils	43	519	626	256	76	1520
q-value <0.001	B cells	56	209	264	102	13	644
	CD4+ T cells	2	12	4	4	0	22
	CD8+ T cells	3	20	3	1	0	27
	NK cells	52	255	85	8	4	404
	Eosinophils	27	208	55	13	4	307
	Monocytes	46	427	260	67	18	818
	Neutrophils	24	452	173	31	32	712
Corrected p-value <0.05	B cells	44	134	181	55	9	423
	CD4+ T cells	1	4	2	0	0	7
	CD8+ T cells	0	14	1	0	0	15
	NK cells	40	185	37	1	3	266
	Eosinophils	19	49	24	15	0	107
	Monocytes	36	289	137	20	5	487
	Neutrophils	13	203	56	13	15	300
Corrected p-value <0.001	B cells	41	116	148	44	9	358
	CD4+ T cells	1	5	1	0	0	7
	CD8+ T cells	0	13	1	0	0	14
	NK cells	32	158	18	2	3	213
	Eosinophils	19	61	12	2	0	94
	Monocytes	33	296	60	1	0	390
	Neutrophils	13	255	3	1	3	275

With an FDR control of q-value < 0.05 applied to both data sets, OVESEG-test detects n=28 CD4+ T cell markers, n=7 CD8+ T cell markers, and multiple markers for other more distinctive cell types (Table 3.5~3.7). Between the two data sets, we obtain a Jaccard index (intersection over union) of 36.8% for all MGs across all cell types. Overlap of monocyte and neutrophil markers detected from the two datasets is >40% (Fig. 3.12**Error! Reference source not found.**). Subtype-specific MGs account for about one third of all probesets (Roche: 39%, HUG: 34%). This result is expected because these subtypes are pure cell types and so more distinctive than would be seen with samples from multicellular tissues [22, 27, 76]. We also applied a Bonferroni multiple testing correction and a more stringent p-value < 0.001; the number of MGs account for 10.7% and 2.7% of all probesets in Roche and HUG data sets, respectively (Table 3.5), with only one common CD4+ T cell marker (FHIT) and one common CD8+ T cell marker (CD8B).

Table 3.6 Statistics of CD4+ T cell markers detected in both Roche and HUG (q<0.05)

	Probe	Gene	Roche			HUG		
	/Probeset	Symbol	OVE-FC	p-value	q-value	OVE-FC	p-value	q-value
1	206492_at	FHIT	3.302	3.89E-17	8.87E-16	4.42	2.08E-14	7.36E-13
2	229070_at	ADTRP	7.068	4.25E-13	4.60E-12	7.365	3.21E-06	3.02E-05
3	209442_x_at	ANK3	2.397	6.33E-13	6.48E-12	1.747	2.79E-04	1.21E-03
4	206385_s_at	ANK3	3.227	3.90E-10	2.71E-09	2.696	2.54E-04	1.13E-03
5	236341_at	CTLA4	2.919	4.88E-06	2.10E-05	3.233	1.65E-04	7.95E-04
6	1562731_s_at	MDS2	1.728	1.03E-05	4.12E-05	1.688	7.43E-04	2.61E-03
7	203410_at	AP3M2	2.061	1.91E-05	7.11E-05	3.621	4.33E-05	2.69E-04
8	1729_at	TRADD	1.376	1.15E-04	3.35E-04	1.409	4.63E-03	1.07E-02
9	227641_at	FBXL16	1.388	4.61E-04	1.14E-03	1.911	1.07E-04	5.64E-04
10	217147_s_at	TRAT1	1.842	6.26E-04	1.48E-03	1.73	7.75E-03	1.59E-02
11	213135_at	TIAM1	1.47	7.25E-04	1.68E-03	1.798	2.19E-03	5.98E-03
12	203386_at	TBC1D4	1.398	8.25E-04	1.88E-03	2.034	2.06E-03	5.71E-03
13	227580_s_at	TECPR1	1.364	9.64E-04	2.16E-03	1.512	9.29E-03	1.82E-02
14	204773_at	IL11RA	1.416	1.12E-03	2.46E-03	1.455	7.31E-03	1.52E-02
15	204777_s_at	MAL	1.541	1.39E-03	2.98E-03	2.496	5.44E-05	3.22E-04
16	1557733_a_at	CHRM3-AS2	1.643	1.60E-03	3.38E-03	3.211	5.12E-05	3.06E-04
17	40016_g_at	MAST4	1.351	4.64E-03	8.55E-03	1.679	1.68E-03	4.88E-03
18	225613_at	MAST4	1.524	4.88E-03	8.94E-03	2.116	2.22E-03	6.04E-03
19	224832_at	DUSP16	1.274	8.84E-03	1.51E-02	1.712	8.09E-03	1.64E-02
20	213028_at	NFRKB	1.171	8.85E-03	1.51E-02	1.243	3.44E-02	4.74E-02
21	203717_at	DPP4	1.394	1.46E-02	2.35E-02	1.834	4.14E-04	1.66E-03
22	225611_at	MAST4	1.366	1.88E-02	2.93E-02	2.131	1.33E-04	6.64E-04
23	206545_at	CD28	1.307	2.16E-02	3.28E-02	1.644	2.00E-02	3.22E-02
24	56197_at	PLSCR3	1.176	2.16E-02	3.29E-02	1.332	8.65E-03	1.73E-02
25	210439_at	ICOS	1.338	2.29E-02	3.45E-02	2.262	1.31E-03	3.99E-03
26	220048_at	EDAR	1.289	2.84E-02	4.15E-02	1.565	2.24E-03	6.07E-03
27	211675_s_at	MDFIC	1.378	2.88E-02	4.20E-02	1.663	5.97E-03	1.30E-02
28	225478_at	MFHAS1	1.327	3.23E-02	4.63E-02	1.96	1.22E-03	3.78E-03

Table 3.7 Statistics of CD8+ T cell markers detected in both Roche and HUG (q<0.05)

	Probe	Gene	Roche			HUG		
	/Probeset	Symbol	OVE-FC	p-value	q-value	OVE-FC	p-value	q-value
1	215332_s_at	CD8B	38.888	1.15E-31	8.90E-30	19.368	1.54E-20	2.77E-18
2	205758_at	CD8A	5.635	4.81E-13	5.12E-12	3.485	7.02E-05	3.99E-04
3	203413_at	NELL2	2.741	7.08E-09	4.02E-08	1.898	9.87E-04	3.23E-03
4	206666_at	GZMK	3.963	4.18E-07	2.04E-06	4.644	1.12E-05	8.73E-05
5	209871_s_at	APBA2	1.815	8.94E-05	2.70E-04	1.575	9.05E-03	1.79E-02
6	226474_at	NLRC5	1.497	3.16E-03	6.13E-03	1.439	2.02E-02	3.24E-02
7	225803_at	FBXO32	1.736	1.52E-02	2.43E-02	2.066	1.14E-03	3.60E-03

To present all kinds of upregulation patterns among cell types (Fig. 3.13), probeset-wise posterior probabilities of component hypotheses in the null mixtures (Eq. 3.8) are accumulated and normalized to estimate the counterpart probabilities of the alternative hypotheses (Eq. 3.16), where the patterns of upregulation in B cells, monocytes, or neutrophils rank the top in both datasets, followed by upregulation in lymphoid cells (B cells, CD4+ T cells, CD8+ T cells, NK cells) and T cells (CD4+ T cells, CD8+ T cells) in the Roche dataset.

3.3.4 Evaluation via supervised deconvolution performance

Accurate and reliable detection of MGs has a significant impact on the performance of many supervised deconvolution methods that use the expression patterns of MGs to score constituent subtypes in heterogeneous samples [22, 77, 78]. There are two major classes of supervised deconvolution approaches to quantify proportions of subtypes in mixtures: fitting coefficients for the linearly modeled relationship between the mixture and the pure expression levels or scoring each subtype by expression levels of markers in mixtures [79]. The former approach estimates the absolute fraction of each subtype in a heterogeneous sample. The latter approach provides relative scores that are comparable across samples but not comparable between subtypes. However, the second class has its own advantages. **(1)** It only assumes the indexes of markers are unchangeable, so that the markers selected based on pure expression levels can help deconvolute mixtures in different microenvironments, even when measured on other platforms. **(2)** Scores are computed for each subtype individually and thus less affected by each other and unknown subtypes. **(3)** Scores can achieve a higher correlation with ground truth proportions, since they focus on catching the dynamic trend across samples, not the absolute proportion values. Generally, the factors affecting coefficient fitting are complex, whereas scoring one subtype relies mostly on the precision of the indexes of selected markers. Therefore, we adopted a debCAM score (Eq 2.11) derived from MGs-guided supervised deconvolution to quantify each subtype. A correlation coefficient between estimated scores and true proportions was used to assess the accuracy of various MGs selection methods.

OVESEG-test statistics were applied to three independent data sets acquired from the purified subtype expression profiles (GSE28490 Roche), purified subtype RNAseq profiles (GSE60424), and classified single-cell RNAseq profiles (GSE72056), respectively. The subtype-specific MGs were detected by six different methods including OVR-FC, OVR t-stat, OVR t-test, OVO t-test, OVE-FC, and OVESEG-test, and used to supervise the deconvolution of realistically synthesized mixtures with ground truth. Five subtypes (B cell, CD4+ T cell, CD8+ T cell, NK cell, monocytes) were included in synthesizing n=50 *in silico* mixtures, where purified subtype mean expressions

(GSE28491 HUG) were combined according to pre-determined proportions with additive noise, simulating heterogeneous biological samples:

$$x_{ij} = \sum_{k=1}^{K} a_{ik}\left(s_k(j) + \Delta s_{ik}(j)\right) + \varepsilon_{ij}, \qquad (S16)$$

where $\Delta s_{ik}(j)$ reflects subtype-specific biological variation across samples and ε_{ij} is technical noise of measurement. In the first and the third simulation (**Fig. 3.14a, 3.14c**), $s_k(j)$ is cell-type specific mRNA expression mean (averaged across samples in log2 scale and transformed to original scale) in GSE28491. $\Delta s_{ik}(j)$ and ε_{ij} follows zero-mean Gaussian distribution, with variance drawn from inverse chi-square distribution with σ_0 being 0.5 and 0.1, ν_0 being 10 and 5, respectively. In the second simulation (**Fig. 3.14b**), as GSE60424 has almost 20 samples per cell type, one sample is randomly selected for each cell type and its RNAseq count profile can be treated as $s_k(j)$ with biological and technical variance already included. Mixing proportions, $a_{ik}, k = 1, \dots, K$, are drawn randomly from a flat Dirichlet distribution and used in the three simulations.

The proportions of constituent subtypes are estimated by the debCAM scores derived from expression levels of top-ranked markers for each subtype. Supervised deconvolution results show that OVESEG-test, OVE-FC and OVO t-test methods achieved the highest correlation coefficients between debCAM score and true proportions when compared with the performance of other methods (**Fig. 3.14a, Fig. 3.15**). As a more biologically realistic case involving between-sample variations, we synthesize a set of n=50 *in silico* mixtures by combining the subtype expression profiles from bootstrapped samples in the RNAseq data set according to pre-determined proportions. Again, supervised deconvolution results show that the subtype-specific MGs detected by OVESEG-test or OVE-FC or OVO t-test achieved superior deconvolution performance (**Fig. 3.14b, Fig. 3.16**). As a more challenging case of noisy RNAseq data and small sample size, we repeated the simulations where *in silico* mixtures were synthesized by combining subtype mean expressions (GSE28491 HUG) and markers were detected from downsampled RNAseq profiles (GSE60424, n=3). Three purified samples were randomly selected for each subtype and analyzed by six methods. The experimental results show that, in terms of MG-based deconvolution performance, the OVESEG-test strongly beats the OVO t-test. Moreover, OVESEG-test outperforms OVE-FC for phenotypically closer cell types (CD4+ T and CD8+ T cell types) (**Fig. 3.14c**).

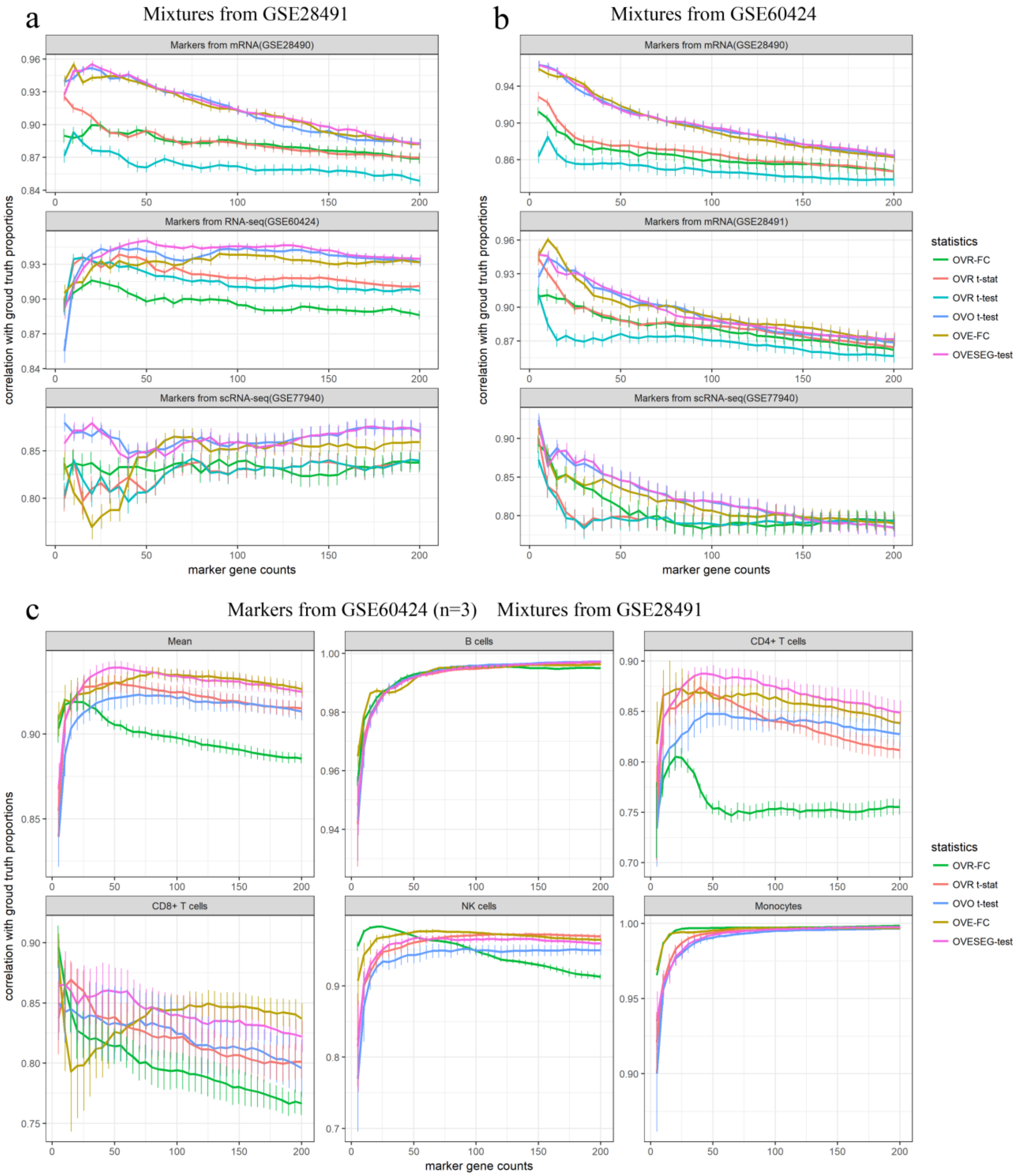


Figure 3.14 Correlation coefficients between debCAM scores and ground truth proportions in simulated heterogeneous samples of mixed subtype mRNA expression profiles or RNAseq counts (a-c based on three different real gene expression datasets). The mean and 95% confidence interval are computed over 20 repeated experiments.

Across the varying number of MGs (5~200) being selected, Fig. 3.14 shows the impact of MGs (both at a fixed number and the corresponding content) selected by different methods on the performance of supervised deconvolution. While different subtypes are expected to have different number of MGs practically and biologically, e.g., B cell or monocyte versus CD4+ T cell or CD8+ T cell, the fundamental working principle of various tissue deconvolution methods is that there is a proper small number of MGs to be expressed exclusively in only a particular subtype. Thus applying a stringent OVESEG-test p-value threshold, e.g., < 0.001 after correction (Table 3.5) is a good option, because suitable number of MGs for CD4+ or CD8+ T cell is 5~20, while B cells or monocytes often allows more MGs to be involved in supervised deconvolution.

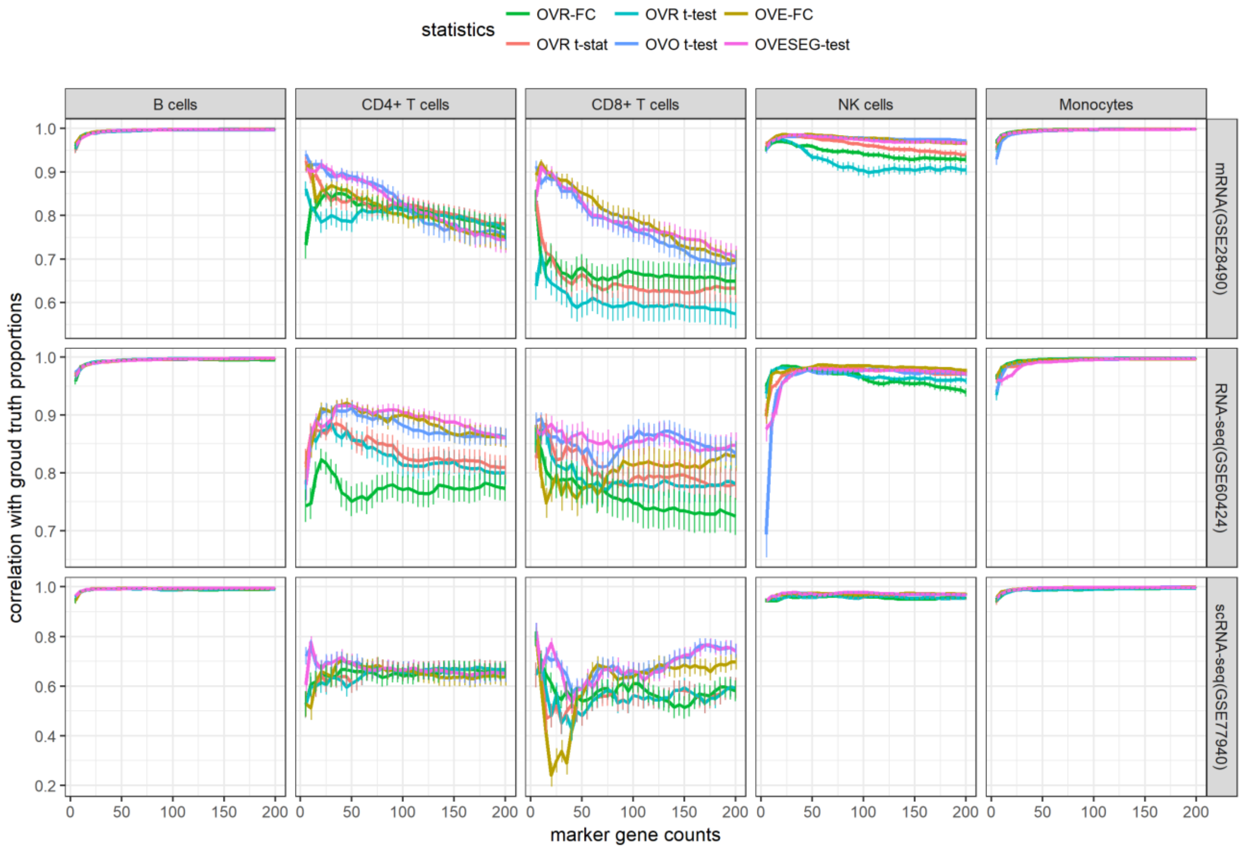


Figure 3.15 Correlation coefficients between CAM score and ground truth proportion for each cell type, with score estimated by a fixed number of markers from independent dataset to quantify subtypes in heterogeneous samples simulated by mixing purified mRNA expression levels in GSE28491. Mean and 95% confidence interval are computed among 20 repeated experiments.

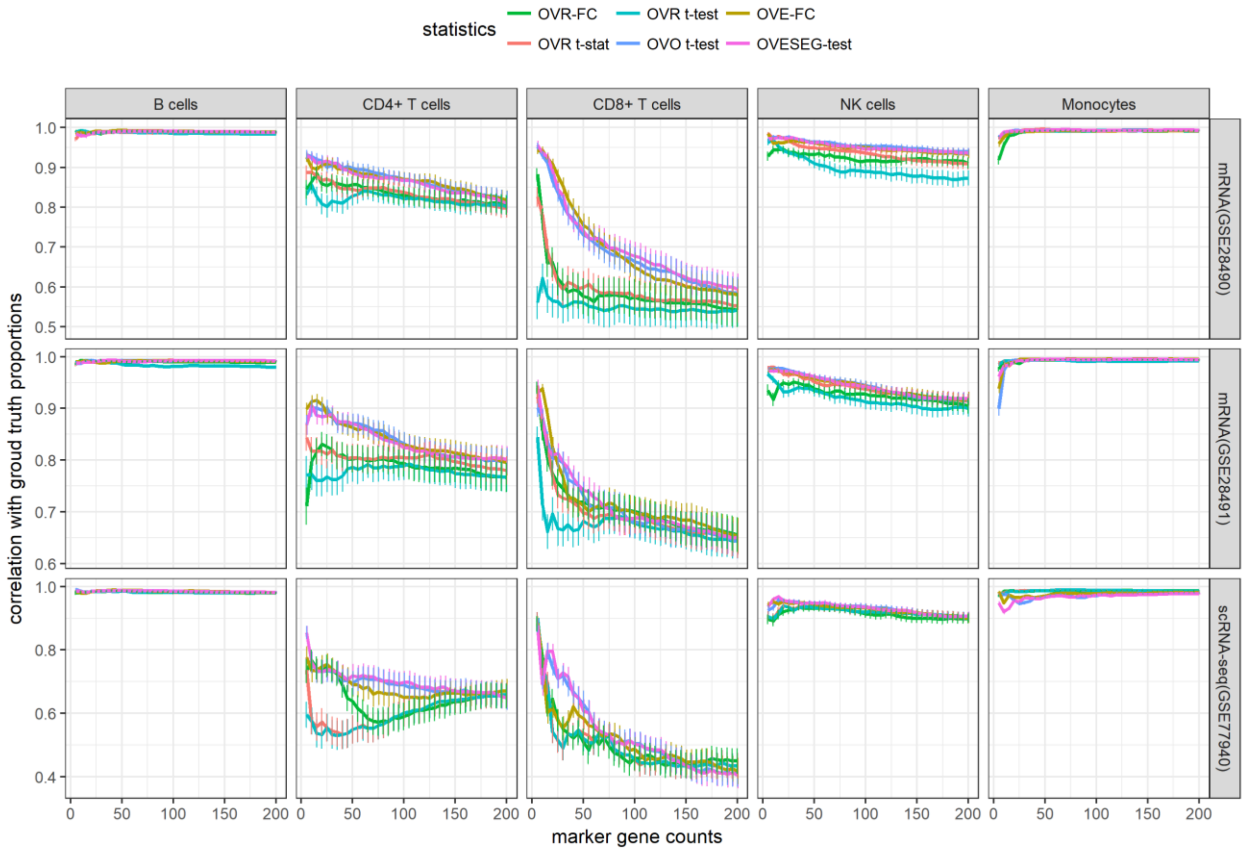


Figure 3.16 Correlation coefficients between CAM score and ground truth proportion for each cell type, with score estimated by a fixed number of markers from independent dataset to quantify subtypes in heterogeneous samples simulated by mixing purified RNAseq counts in GSE60424. Mean and 95% confidence interval are computed among 20 repeated experiments.

3.3.5 Application on heterogeneous vascular proteomes (In-house data)

In one of our projects for Global Analysis of Coronary and Abdominal Aortic Proteomes (Subsection 2.4.6), pure aortic tissues are sampled by the visual inspection of an experienced pathologist. The expression profiles of 1478 proteins were obtained in three pure normal specimens (NL, n=3), three pure fatty streak specimens (FS, n=3) and four pure fibrous plaque specimens (FP, n=4). With subtype-specific OVESEG-test q-value cutoff at 0.2, we detect 59 significant markers (54 for FP, 4 for FS and 1 for normal tissue type), as illustrated in Figure 3.17.

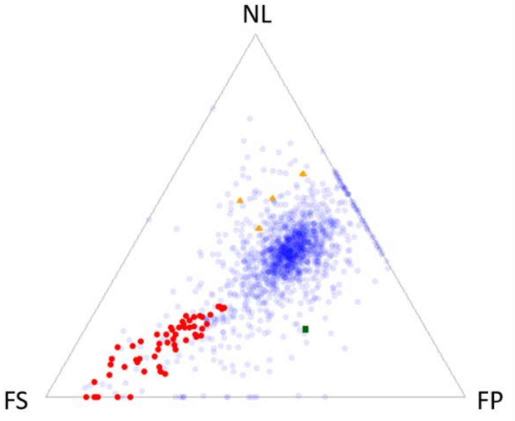


Figure 3.17 Ternary plot that depicts the ratios of purified expressions in three tissue types (red, green, and orange: significant FS, FP, and NL markers).

Table 3.8 Marker proteins detected by OVESEG-test

	Gene	OVE-FC	OVESEG-test q-value	Subtype-specific OVESEG-test q-value	debCAM-FP(LAD)	debCAM-FP(AA)
1	ITIH2	2.960756	0.018325	0.055059	FALSE	TRUE
2	A2MG	7.286197	0.025386	0.061585	FALSE	TRUE
3	ALS	3.342612	0.02693	0.063442	FALSE	TRUE
4	IGJ	6.729321	0.027019	0.063546	TRUE	TRUE
5	SAA4	4.339697	0.027119	0.063617	FALSE	TRUE
6	PLMN	2.131126	0.027185	0.063679	FALSE	TRUE
7	HPT	5.513884	0.032966	0.063864	FALSE	TRUE
8	FIBG	4.46506	0.060506	0.084407	TRUE	TRUE
9	APOD	2.576451	0.062964	0.086177	FALSE	TRUE
10	CBPB2	2.343175	0.066386	0.089159	FALSE	TRUE
11	FIBB	4.228847	0.066693	0.089627	TRUE	TRUE
12	APOE	2.883494	0.083827	0.094946	TRUE	TRUE
13	FIBA	3.312959	0.10773	0.114221	TRUE	FALSE
14	VTNC	3.100292	0.110287	0.115004	TRUE	TRUE
15	C4BPA	3.006885	0.120786	0.120449	FALSE	TRUE
16	APOM	5.280693	0.134778	0.138265	FALSE	TRUE
17	APOB	2.712174	0.193937	0.168478	TRUE	TRUE
18	LYSC	1.583381	0.225457	0.184199	TRUE	FALSE

The significance level of markers in pure profiles could help to validate markers from completely unsupervised analysis of mixtures by debCAM. 18 of 58 debCAM-FP marker proteins have subtype-specific OVESEG-test q-value less than 0.2, among which 8 and 16 proteins are detected by debCAM in LAD and AA tissues, respectively (Table 3.8). These proteins recognized by both supervised and unsupervised methods are very likely to be reliable and context-specific markers, which deserve further biological investigation.

3.4 Discussion

Interpreting an expression profile requires both the knowledge of the relative abundance of the different cell or tissue types and their individual expression patterns. Understanding the relative contribution of individual cell or tissue types in individual samples may illuminate pathophysiologic mechanisms, biologic responses to various stimuli, or transitions in tissue phenotype - especially when the cell-cell and cell-matrix interactions in a complex system are necessary conditions for biological relevance. Expression patterns of MGs for the relevant cells or tissues can be used to support supervised deconvolution to estimate the relative prevalence of the contributing cell or tissue subtypes. Our present work is focused on subtype-exclusively expressed genes, *i.e.* restricted to the MG definition widely adopted [6, 11, 13, 57]. Indeed, the work here is motivated by the need to obtain subtype-specific marker genes for supervised *in silico* tissue deconvolution [22] and tissue subtype characterization [27], where the measured data are the mixtures of the expressions from multiple underlying subtypes and the MGs are used to estimate both the proportions of each subtype in individual heterogeneous samples as well as the averaged subtype-specific expression profiles.

Though ideal MGs are defined as being exclusively and consistently expressed in a particular tissue or cell subtype across varying conditions, biological reality dictates a more relaxed definition that allows MGs of a particular tissue or cell subtype having low or insignificant expressions in all other subtypes. Experimental results show that MGs detected by OVESEG-test with small p-values can accurately estimate both subtype proportions and expression profiles, serving as effective molecular markers (Fig. 3.14c, Fig. 3.15 and Fig. 3.16).

Accuracy of OVESEG-test based MG detection may be affected by both batch effect and normalization methods, and reliability would depend on the variance estimate particularly when sample size is small. One solution adopted in our method is to leverage the ability of "limma" that can borrow information across genes in estimating *a priori* variance, thus stabilizing the estimated variance for each gene. In addition, the OVESEG-test estimates the parameters of the limma model from all subtypes, producing better results than that applying t-test independently with the limma model for each subtype pair (Fig. 3.9b, Fig. 3.14**Error! Reference source not found.**c and Table 3.2-3.3)

The OVESEG-test makes a few assumptions and works best when all assumptions are valid. In the OVESEG test, the proposed permutation scheme does not require the data to be normally distributed. Under the null hypothesis, the OVESEG-test assumes that samples are drawn from the same distribution for a given gene and are drawn from the distribution of the same 'shape' for different genes. These assumptions are made to ensure that the null distributions across genes can be aggregated together with variance-based standardization. Practically, when data distributions deviate significantly from a common shape, the limma-voom/vooma/voomaByGroup's variance models is used to accommodate unequal variances by appropriate observational-level weights (Ritchie, Phipson et al. 2015); when data distributions deviate significantly from normality, a permutational ANOVA is used to estimate the null hypothesis components of the mixture distribution. Our experimental results show that with the mean-variance relationship estimated by limma-voom on RNASeq data, the OVESEG-test can maintain the expected type 1 error rates or specified FDR (Fig. 3.11). For outliers and drop-out zero values in RNAseq data, when needed, state-of-the-art two-group test methods designed specifically for RNAseq will be exploited and adopted, *e.g.* edgeR [80], DESeq2 [81].

While OVE-FC is the earlier version of our OVE strategy and drives the present work on OVESEG-test, the major differences between the OVE-FC and OVESEG-test are twofold: (1) OVE-FC does not assess statistical significance (calculating p-values) while OVESEG-test specifically aims to provide a significance assessment and to potentially improve both FDR control and detection power in more challenging situations; (2) OVE-FC is originally for subtype classification while the OVESEG-test is mainly for subtype deconvolution. Theoretically, detecting MGs by evaluating the significance with accurate p-values is an attractive feature of

OVESEG-test that can help control FDR at the expected level. Indeed, our additional experimental results show that the OVESEG-test outperforms its earlier version OVE-FC in the more challenging cases involving non-significantly large fold change (Fig. 3.9) or phenotypically closer cell types (Fig. 3.14c). Nevertheless, we acknowledge that the OVE-FC and OVESEG-test are complementary to each other and both are good methods for MG detection.

While ANOVA has been the most commonly used method to test differences among the means of multiple subtypes, often in conjunction with a post-hoc Tukey HSD comparing all possible pairs of means [82], it is not suitable for detecting subtype-specific markers because the null hypothesis used by ANOVA does not truly enforce the definition of MGs. ANOVA detects differentially expressed genes rather than MGs and therefore produces too many false positives to be generally useful here.

In addition to the MGs defined here (genes exclusively up-regulated in a particular subtype), the counterpart subtype-specific down-regulated genes (genes exclusively down-regulated in a particular subtype) are also of great biological interest [17]. OVESEG-test principle is readily applicable to detecting down-regulated MGs by reversing the comparison rule. There are certainly other alternative definitions of ‘informative genes’ (or broadly defined marker genes) for different analytical purposes, *e.g.*, sample classification. In our earlier work on multiclass classification [17, 62], we have shown that genes exclusively upregulated in each of the subtypes selected by OVE-FC are sufficient to achieve multiclass classification and can often improve classifier performance over alternative informative gene subsets of the same size.

Lastly, when the relevant expression patterns are unknown, unsupervised deconvolution techniques (*e.g.*, debCAM) are required. A theoretical advantage of unsupervised deconvolution is that it can identify both the cell or tissue subtype distributions and their specific expression patterns, albeit with possibly less fidelity than when one or the other is known a priori or independently measured from the same sample.

Chapter 4

Sample-specific deconvolution

4.1 Introduction

Decades of research on molecule regulatory mechanisms have provided a rich framework with which we can extract molecule expression patterns to gain insight into the organization and structure of the large biological networks [83-85]. However, most discoveries are concluded from measured molecule expressions in heterogeneous tissues, in which the underlying changes of constituent components could obscure molecule regulations and corrupt network inference that only occur in particular tissue subtypes. The inference of subtype-specific molecule expression patterns becomes an essential problem for understanding complex molecule functions and the role of each subtype during the dynamic biological process, such as cell fate specification [86-88]. The ability to obtain sample-wise expression variation in each subtype is critical prior to infer subtype-specific molecular networks [5, 89]. Single-cell expression profiling techniques have become popular to investigate cell-type-specific network but may lose critical information of cell-cell interactions and is prone to cell-cycle/state confounders [87, 90].

While the current debCAM tool can dissect mixed signals of multiple samples into the 'averaged' expression profiles of subtypes, many subsequent molecular analyses of complex tissues require sample-specific signal deconvolution where each sample is a mixture of 'individualized' subtype-specific expression profiles. Here we propose a new algorithm called debCAM2.0 as an extension of debCAM to solve sample-specific Blind Source Separation (sBSS) problem. The sBSS problem, because the number of variables is much larger than the number of observations, is ill-posed and underdetermined. As a result, simple yet highly regularized approaches often become the methods of choice [91]. The sBSS problems have received increasing interest in hyperspectral imagery area where the spatially smooth and variation sparsity regularization can be exploited to unmix spectral signals [92]. In the context of biological process, transcriptional regulatory networks connect regulatory proteins, such as transcription factors (TFs) and signaling proteins, to target genes and thus form co-expressed gene sets as function modules in each subtype. Based on such underlying cellular mechanisms, we impose and exploit the low-rank assumption on between-sample variations of molecule expressions in each subtype. While estimating subtype-specific signals from a single mixture for each gene independently is an underdetermined problem, the low-rank assumption can aggregate information from gene sets within function modules to help find a biologically plausible solution.

General rank minimization is a challenging nonconvex optimization problem for which all existing finite-time algorithms have at least doubly exponential running times in both theory and practice [93]. Minimizing the nuclear norm, or the sum of the singular values of the matrix, over the affine subset, have multiple advantages. The nuclear norm is a convex function and can be optimized efficiently, thus can provide the best convex approximation of the rank function over the unit ball of matrices with norm less than one [93, 94]. Nuclear norm regularization has been successfully applied in many practical applications with low-rank modeling, such as image denoising [94] and matrix completion [95]. debCAM2.0 will adopt nuclear norm regularization to optimize the estimation of between-sample variations in each subtype to recover sample-specific signals for each subtype. This Chapter introduces mathematical modeling of sample-specific deconvolution and optimization solver used in debCAM2.0 algorithm, followed by validation in simulations and discussion on further possible improvement by sparsity regularization.

4.2 Method

4.2.1 Problem formulation and nuclear norm regularization

A fundamental assumption for the conventional linear mixing model, $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{E}$, is that all the mixture samples share a common source matrix $\boldsymbol{S}$, leading to BSS model studied in Chapter 2. However, each sample may have its ‘individualized’ sample-specific sources as the sampled realizations in additional sample-specific subtype proportions (Fig. 4.1):

$$\boldsymbol{S}_i = \overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i, i = 1, \dots, M.$$

The associated sample-specific BSS (sBSS) model is given by

$$\boldsymbol{x}_i = \boldsymbol{a}_i(\overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i) + \boldsymbol{n}_i \in \mathbb{R}^L, \forall i = 1, \dots, M \tag{4.1}$$

where $\boldsymbol{a}_i$ is the row vector in proportion matrix $\boldsymbol{A}$ (different from Chapter 2), and $\boldsymbol{n}_i$ is noise term. $\Delta\boldsymbol{S}_i$ is expected to have small-valued entries so that source matrices among different samples are similar. If some expression units (molecular features) are highly correlated in a particular source, the rank of the matrix consisting of $\Delta\boldsymbol{S}_i$ entries in this source will be low, leading to the following debCAM2.0 objective function:

$$\min_{\boldsymbol{A},\{\Delta\boldsymbol{S}_i\}_{i=1}^{M}} \frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{x}_i - \boldsymbol{a}_i(\overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i)\|_2^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{T}_k\|_* \tag{4.2}$$

$$s.t.\ \boldsymbol{a}_i \succcurlyeq \boldsymbol{0}_K, \boldsymbol{a}_i\boldsymbol{1}_K = 1, \overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i \succcurlyeq \boldsymbol{0}_{K\times L},$$

$$\boldsymbol{T}_k = [\Delta\boldsymbol{S}_1^T(k), \dots, \Delta\boldsymbol{S}_M^T(k)] \in \mathbb{R}^{L\times M},\ k = 1, \dots, K,$$

where $\boldsymbol{T}_k$ consists of the kth column in all $\Delta\boldsymbol{S}_i, i = 1, \dots, M$, representing between-sample variation in source k, namely the variation matrix for the kth subtype. $\|\boldsymbol{T}_k\|_*$ is the nuclear norm of $\boldsymbol{T}_k$; and $\lambda > 0$ is the regularization parameter of nuclear norms. The hyperparameter λ actually

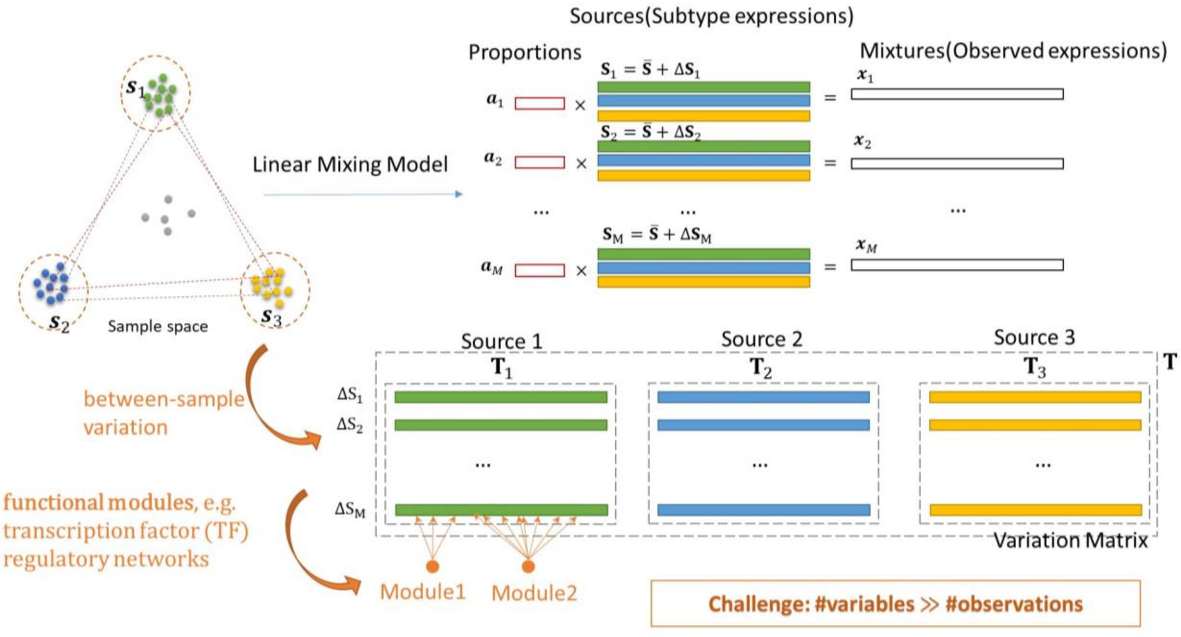


Figure 4.1 Sample-specific deconvolution problem formulation and the assumption of hidden low-rank pattern in each source. (For convenient illustration, $\boldsymbol{T}$ matrix in all figures are the transposed version of those in the text and equations.)

can be different among different source k and thus the regularizer in (4.2) can be replaced by $\sum_{k=1}^{K}\lambda_k\|\boldsymbol{T}_k\|_*$ if necessary. The supplementary note (Subsection 4.5.1) shows the reason to select subtype-specific regularization terms.

4.2.2 Optimization of debCAM2.0 objective function

The objective function in (4.2) is bi-convex w.r.t. the two block-wise variables, i.e. $\boldsymbol{A}\triangleq[\boldsymbol{a}_1^T,\dots,\boldsymbol{a}_M^T]^T$ and $\boldsymbol{T}\triangleq[\boldsymbol{T}_1^T,\dots,\boldsymbol{T}_M^T]^T\in\mathbb{R}^{KL\times M}$. Accordingly, we can solve (4.2) by alternatively solving the following two convex subproblems until convergence:

$$\boldsymbol{T}^{p+1}\in\underset{\Delta\boldsymbol{S}_{\mathrm{i}}\succcurlyeq-\boldsymbol{S},\forall\boldsymbol{i}}{\operatorname{argmin}}\,\mathcal{J}(\boldsymbol{A}^p,\boldsymbol{T}) \tag{4.3}$$

$$\boldsymbol{A}^{p+1}\in\underset{\boldsymbol{A}\succcurlyeq 0_{M\times K},\boldsymbol{A}\boldsymbol{1}_K=\boldsymbol{1}_M}{\operatorname{argmin}}\,\mathcal{J}(\boldsymbol{A},\boldsymbol{T}^{p+1}) \tag{4.4}$$

where

$$\mathcal{J}(\boldsymbol{A},\boldsymbol{T})\triangleq\frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{x}_i-\boldsymbol{a}_i(\overline{\boldsymbol{S}}+\Delta\boldsymbol{S}_{\mathrm{i}})\|_2^2+\lambda\sum_{k=1}^{K}\|\boldsymbol{T}_k\|_*$$

debCAM-estimated subtype-specific expression matrix serves as the initial reference $\overline{\boldsymbol{S}}$. Note that in (4.3) (4.4), we have implicitly used the following relationship for concise representation:

$$\boldsymbol{T}\triangleq[vec(\Delta\boldsymbol{S}_1^T),\dots,vec(\Delta\boldsymbol{S}_M^T)],$$

where (4.4) can be decoupled w.r.t each row of $\boldsymbol{A}$:

$$\boldsymbol{a}_i^{p+1}\in\underset{\boldsymbol{a}_i\succcurlyeq\boldsymbol{0}_K,\boldsymbol{a}_i\boldsymbol{1}_K=1}{\operatorname{argmin}}\frac{1}{2}\left\|\boldsymbol{x}_i-\boldsymbol{a}_i\left(\overline{\boldsymbol{S}}+\Delta\boldsymbol{S}_i^{p+1}\right)\right\|_2^2$$

which can be solved using quadratic programming. If *a prior* proportion matrix or debCAM-estimated proportion matrix has already been of high quality, we can skip the alternative optimization on $\boldsymbol{A}$ matrix, and obtain $\boldsymbol{T}$ matrix by optimizing the subproblem (4.3) only once.

To solve (4.3), we notice that the main bottleneck is its huge dimension of variables (typically, L is several ten thousands), preventing conventional convex solvers from being readily applicable here. We propose to solve (4.3) by adapting the alternating direction method of multipliers (ADMM), which has been widely applied to many large-scale problems in areas such as statistical learning, image processing and computational biology [96].

ADMM naturally allows decoupling the non-smooth regularization term from the smooth loss term, which is computationally advantageous. Specifically, we reformulate (4.3) in the form that the primal variable can be "split" into several parts, with the associated objective function "separable" across this splitting [96]. We will use the following definitions:

$$\boldsymbol{T} \triangleq [vec(\Delta\boldsymbol{S}_1^T), \dots, vec(\Delta\boldsymbol{S}_M^T)] = \begin{bmatrix} \boldsymbol{T}_1 \\ \dots \\ \boldsymbol{T}_K \end{bmatrix} \in \mathbb{R}^{KL\times M}$$

$$\boldsymbol{S} \triangleq [vec(\boldsymbol{S}_1^T), \dots, vec(\boldsymbol{S}_M^T)] \in \mathbb{R}^{KL\times M}$$

$$\boldsymbol{V} \triangleq \boldsymbol{X}^T \in \mathbb{R}^{L\times M}$$

$$\boldsymbol{W} \triangleq \begin{bmatrix} \boldsymbol{T} \\ \boldsymbol{S} \end{bmatrix} \in \mathbb{R}^{2KL\times M}$$

$$\boldsymbol{C}_1 \triangleq \begin{bmatrix} \boldsymbol{I}_{KL} \\ \boldsymbol{I}_{KL} \end{bmatrix} \in \mathbb{R}^{2KL\times KL}$$

$$\boldsymbol{C}_2 \triangleq -\boldsymbol{I}_{2KL} \in \mathbb{R}^{2KL\times 2KL}$$

$$\boldsymbol{C}_3 \triangleq \begin{bmatrix} \boldsymbol{1}_M^T \otimes vec(\overline{\boldsymbol{S}}^T) \\ \boldsymbol{0}_{KL\times M} \end{bmatrix} \in \mathbb{R}^{2KL\times M}$$

$$\boldsymbol{B}_0 \triangleq [\boldsymbol{0}_{KL\times KL}, \boldsymbol{I}_{KL}] \in \mathbb{R}^{KL\times 2KL}$$

$$\boldsymbol{B}_k \triangleq \left[\boldsymbol{0}_{L\times(k-1)L}, \boldsymbol{I}_L, \boldsymbol{0}_{L\times(K-k)L}, \boldsymbol{0}_{L\times KL}\right] \in \mathbb{R}^{L\times 2KL}, k = 0, \dots, K$$

Then we can simplify (4.3) as the equivalent form:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{KL\times M}, \boldsymbol{W}\in\mathbb{R}^{2KL\times M}} \frac{1}{2}\|\mathcal{A}(\boldsymbol{U}) - \boldsymbol{V}\|_F^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{B}_k\boldsymbol{W}\|_* + I_+(\boldsymbol{B}_0\boldsymbol{W}) \qquad (4.5)$$

$$s.t.\, \boldsymbol{C}_1\boldsymbol{U} + \boldsymbol{C}_2\boldsymbol{W} = \boldsymbol{C}_3,$$

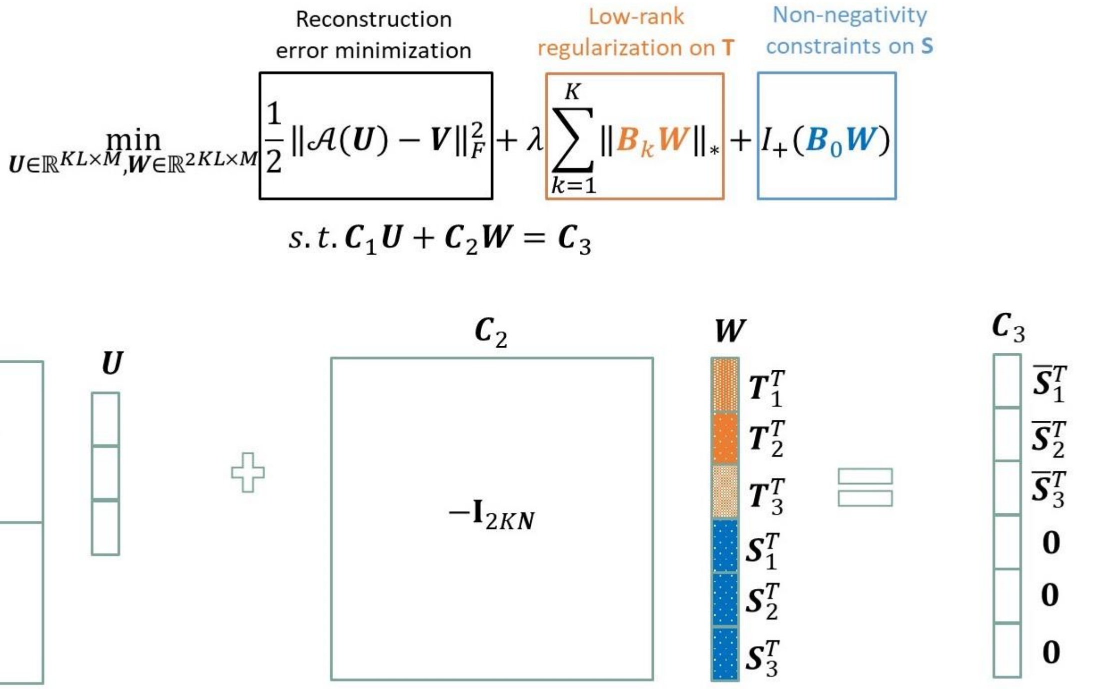

Reconstruction error minimization
Low-rank regularization on T
Non-negativity constraints on S
$\min_{\boldsymbol{U}\in\mathbb{R}^{KL\times M},\boldsymbol{W}\in\mathbb{R}^{2KL\times M}} \frac{1}{2}\|\mathcal{A}(\boldsymbol{U})-\boldsymbol{V}\|_F^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{B}_k\boldsymbol{W}\|_* + I_+(\boldsymbol{B}_0\boldsymbol{W})$
$s.t.\ \boldsymbol{C}_1\boldsymbol{U}+\boldsymbol{C}_2\boldsymbol{W}=\boldsymbol{C}_3$
$\boldsymbol{C}_1$
$\mathbf{I}_{KN}$
$\mathbf{I}_{KN}$
$\boldsymbol{U}$
$\boldsymbol{C}_2$
$-\mathbf{I}_{2KN}$
$\boldsymbol{W}$
$\boldsymbol{T}_1^T$
$\boldsymbol{T}_2^T$
$\boldsymbol{T}_3^T$
$\boldsymbol{S}_1^T$
$\boldsymbol{S}_2^T$
$\boldsymbol{S}_3^T$
$\boldsymbol{C}_3$
$\overline{\boldsymbol{S}}_1^T$
$\overline{\boldsymbol{S}}_2^T$
$\overline{\boldsymbol{S}}_3^T$
$\mathbf{0}$
$\mathbf{0}$
$\mathbf{0}$

Figure 4.2 The objective function of debCAM2.0 for sample-specific deconvolution problem and its reformulation by ADMM. (For convenient illustration, $\boldsymbol{T}$ matrix in all figures are the transposed version of those in the text and equations.)

where $I_+(\cdot)$ is the indicator function for the non-negative orthant; $I_+(\boldsymbol{B}_0\boldsymbol{W}) = I_+(\boldsymbol{S}) = 0$ if $\boldsymbol{S} \succcurlyeq \mathbf{0}_{KL\times M}$ ($I_+(\boldsymbol{U}) = +\infty$, otherwise). The linear transformation in the first term is $\mathcal{A}(\boldsymbol{U}) = \mathcal{A}([\boldsymbol{u}_1, \dots, \boldsymbol{u}_M]) = [\boldsymbol{H}_1\boldsymbol{u}_1, \dots, \boldsymbol{H}_M\boldsymbol{u}_M]$ with $\boldsymbol{H}_i = \left[\boldsymbol{a}_i^p \otimes I_L\right], i = 1, \dots, M$. Note that (4.5) has been with the ADMM form w.r.t. the two split block variables $\boldsymbol{U}$ and $\boldsymbol{W}$, and, as (4.5) is solved, the solution of (4.3) can be obtained by $\boldsymbol{T}^{p+1} = [\,\boldsymbol{I}_{KL}, \mathbf{0}_{KL\times KL}]\boldsymbol{W}^*$.

Given a penalty parameter $\gamma > 0$ (empirically, $\gamma \coloneqq 1$ generally guarantees good convergence speed), the augmented Lagrangian (ignoring some irrelevant terms) of problem (4.5) is defined by

$$\mathcal{L}(\boldsymbol{U},\boldsymbol{W},\boldsymbol{Z}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{U}) - \boldsymbol{V}\|_F^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{B}_k\boldsymbol{W}\|_* + I_+(\boldsymbol{B}_0\boldsymbol{W}) + \frac{\gamma}{2}\|\boldsymbol{C}_1\boldsymbol{U} + \boldsymbol{C}_2\boldsymbol{W} - \boldsymbol{C}_3 - \boldsymbol{Z}\|_F^2$$

where “$-\gamma\boldsymbol{Z}$”$\in \mathbb{R}^{2KL\times M}$ is the dual variable (or Lagrange multiplier) associated with the constraint $\boldsymbol{C}_1\boldsymbol{U} + \boldsymbol{C}_2\boldsymbol{W} = \boldsymbol{C}_3$. Then, ADMM solves (4.5) via the following iterative procedure:

$$\boldsymbol{U}^{q+1} \epsilon \operatorname*{argmin}_{\boldsymbol{U}\in\mathbb{R}^{KL\times M}} \mathcal{L}(\boldsymbol{U}, \boldsymbol{W}^{q}, \boldsymbol{Z}^{q}) \tag{4.6a}$$

$$\boldsymbol{W}^{q+1} \epsilon \operatorname*{argmin}_{\boldsymbol{W}\in\mathbb{R}^{2KL\times M}} \mathcal{L}(\boldsymbol{U}^{q+1}, \boldsymbol{W}, \boldsymbol{Z}^{q}) \tag{4.6b}$$

$$\boldsymbol{Z}^{q+1} = \boldsymbol{Z}^{q} - (\boldsymbol{C}_1\boldsymbol{U}^{q+1} + \boldsymbol{C}_2\boldsymbol{W}^{q+1} - \boldsymbol{C}_3) \tag{4.6c}$$

where $\boldsymbol{W}^0$ can be initialized by $[\boldsymbol{T}_0^T, \boldsymbol{U}_0^T]^T$ with $\boldsymbol{T}_0 = \boldsymbol{0}_{KL\times M}$ and $\boldsymbol{U}_0 = \boldsymbol{1}_M^T \otimes vec(\overline{\boldsymbol{S}}^T)$; $\boldsymbol{Z}^0$ can be simply initialized by $\boldsymbol{0}_{2KL\times M}$**.** As we will show, both (4.6a) and (4.6b) can be solved with closed-form expressions, thanks to the decomposability of ADMM.

Notice that (4.6a) is a column-wise separable optimization problem. So we can decouple w.r.t each column of $\boldsymbol{U}$:

$$\boldsymbol{u}_i^{q+1} \in \operatorname*{argmin}_{\boldsymbol{u}_i\in\mathbb{R}^{KL}} \frac{1}{2}\|\boldsymbol{H}_i\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2 + \frac{\gamma}{2}\left\|\boldsymbol{C}_1\boldsymbol{u}_i + \boldsymbol{y}_{\boldsymbol{i}}^q\right\|_F^2 \tag{4.7}$$

where $\left[\boldsymbol{y}_1^q, \dots, \boldsymbol{y}_M^q\right] \triangleq \boldsymbol{C}_2\boldsymbol{W}^q - \boldsymbol{C}_3 - \boldsymbol{Z}^q$. The subproblem (3.7) is an unconstrained quadratic problem, which can be solved by

$$\boldsymbol{u}_i^{q+1} = (\boldsymbol{H}_i^T\boldsymbol{H}_i + \gamma\boldsymbol{C}_1^T\boldsymbol{C}_1)^{-1}\left(\boldsymbol{H}_i^T\boldsymbol{v}_i - \gamma\boldsymbol{C}_1^T\boldsymbol{y}_{\boldsymbol{i}}^q\right). \tag{4.8}$$

The matrix inversion can speed up by

$$(\boldsymbol{H}_i^T\boldsymbol{H}_i + \gamma\boldsymbol{C}_1^T\boldsymbol{C}_1)^{-1} = \left(\left(\boldsymbol{a}_i^p\right)^T\boldsymbol{a}_i^p + 2\gamma\boldsymbol{I}_K\right)^{-1} \otimes \boldsymbol{I}_L.$$

The right term in (4.8) can also be simplified as

$$\boldsymbol{H}_i^T\boldsymbol{v}_i - \gamma\boldsymbol{C}_1^T\boldsymbol{y}_{\boldsymbol{i}}^q = \left(\boldsymbol{a}_i^p\right)^T \otimes \boldsymbol{x}_i^T - \gamma\left(\overline{\boldsymbol{y}_{\boldsymbol{i}}^q} + \underline{\boldsymbol{y}_{\boldsymbol{i}}^q}\right),$$

where $\boldsymbol{y}_{\boldsymbol{i}}^q = \left[\left(\overline{\boldsymbol{y}_{\boldsymbol{i}}^q}\right)^T, \left(\underline{\boldsymbol{y}_{\boldsymbol{i}}^q}\right)^T\right]^T$ with $\overline{\boldsymbol{y}_{\boldsymbol{i}}^q} \in \mathbb{R}^{KL}$ and $\underline{\boldsymbol{y}_{\boldsymbol{i}}^q} \in \mathbb{R}^{KL}$ being the first and second half vector of $\boldsymbol{y}_{\boldsymbol{i}}^q$, respectively.

Finally, the column vectors of $\boldsymbol{U}^{q+1}$ in (4.6a) can be computed fast by

$$\boldsymbol{u}_i^{q+1} = vec\left\{devec\left\{\left(\boldsymbol{a}_i^p\right)^T \otimes \boldsymbol{x}_i^T - \gamma\left(\overline{\boldsymbol{y}_{\boldsymbol{i}}^q} + \underline{\boldsymbol{y}_{\boldsymbol{i}}^q}\right) | L, K\right\}\left(\left(\boldsymbol{a}_i^p\right)^T \boldsymbol{a}_i^p + 2\gamma \boldsymbol{I}_K\right)^{-1}\right\} \tag{4.9}$$

To solve (4.6b), we remove some irrelevant terms from its objective function:

$$\min_{\boldsymbol{W} \in \mathbb{R}^{2KL \times M}} \lambda \sum_{k=1}^{K} \|\boldsymbol{B}_k \boldsymbol{W}\|_* + I_+(\boldsymbol{B}_0 \boldsymbol{W}) + \frac{\gamma}{2}\|\boldsymbol{C}_1 \boldsymbol{U}^{q+1} + \boldsymbol{C}_2 \boldsymbol{W} - \boldsymbol{C}_3 - \boldsymbol{Z}^q\|_F^2, \tag{4.10}$$

And then, by defining $\boldsymbol{U}_k^{q+1} \in \mathbb{R}^{L \times M}, k = 1, \dots, K$ as block matrices from top to bottom in $\boldsymbol{U}^{q+1} \in \mathbb{R}^{KL \times M}$, $\boldsymbol{Z}_k \in \mathbb{R}^{L \times M}, k = 1, \dots, K$ and $\boldsymbol{Z}_0 \in \mathbb{R}^{KL \times M}$ as block matrices from top to bottom in $\boldsymbol{Z} \in \mathbb{R}^{2KL \times M}$, respectively (i.e., $\boldsymbol{Z} \triangleq [\boldsymbol{Z}_1^T, \dots, \boldsymbol{Z}_K^T, \boldsymbol{Z}_0^T]^T$), we decouple the objective function (4.10) as functions of $\boldsymbol{T}_k, k = 1, \dots, K$ and $\boldsymbol{S}$:

$$\min_{\boldsymbol{W} \in \mathbb{R}^{2KL \times M}} \sum_{k=1}^{K} \left\{\lambda \|\boldsymbol{T}_k\|_* + \frac{\gamma}{2}\left\|\boldsymbol{U}_k^{q+1} - \boldsymbol{T}_k - \boldsymbol{1}_M^T \otimes \bar{\boldsymbol{s}}_k - \boldsymbol{Z}_k^q\right\|_F^2\right\} + \left\{I_+(\boldsymbol{S}) + \frac{\gamma}{2}\left\|\boldsymbol{U}^{q+1} - \boldsymbol{S} - \boldsymbol{Z}_0^q\right\|_F^2\right\}$$

Therefore, $\boldsymbol{W}^{q+1}$ can be solved by the proximal point algorithm (PPA) [97]. Specifically, we have

$$\boldsymbol{W}^{q+1} = \left[\left(\boldsymbol{T}_1^{q+1}\right)^T, \dots, \left(\boldsymbol{T}_K^{q+1}\right)^T, (\boldsymbol{S}^{q+1})^T\right]^T$$

in which

$$\boldsymbol{T}_k^{q+1} \in \underset{\boldsymbol{T} \in \mathbb{R}^{KL \times M}}{\operatorname{argmin}} \lambda \|\boldsymbol{T}_k\|_* + \frac{\gamma}{2}\left\|\boldsymbol{U}_k^{q+1} - \boldsymbol{T}_k - \boldsymbol{1}_M^T \otimes \bar{\boldsymbol{s}}_k - \boldsymbol{Z}_k^q\right\|_F^2 \tag{4.11a}$$

$$\boldsymbol{S}^{q+1} \in \underset{\boldsymbol{T} \in \mathbb{R}^{KL \times M}}{\operatorname{argmin}} I_+(\boldsymbol{S}) + \frac{\gamma}{2}\left\|\boldsymbol{U}^{q+1} - \boldsymbol{S} - \boldsymbol{Z}_0^q\right\|_F^2 \tag{4.11b}$$

Note that (4.11a) and (4.11b) are exactly the proximal operators of $\|\boldsymbol{T}_k\|_*$ and $I_+(\boldsymbol{S})$, respectively [97], and their closed-form solutions are given by

$$\boldsymbol{T}_k^{q+1} = \sum_{\ell=1}^{r} \left(\sigma_{k\ell} - \frac{\lambda}{\gamma}\right)_+ \boldsymbol{\mu}_{k\ell} \boldsymbol{v}_{k\ell}^T, k = 1, \dots, K, \tag{4.12}$$

$$\boldsymbol{S}^{q+1} = \left[\boldsymbol{U}^{q+1} - \boldsymbol{Z}_0^q\right]_+, \tag{4.13}$$

where the singular value decomposition (SVD) of is performed ahead of the computation of (4.12), i.e. $\boldsymbol{U}_k^{q+1} - \boldsymbol{T}_k - \mathbf{1}_M^T \otimes \bar{\boldsymbol{s}}_k - \boldsymbol{Z}_k^q = \sum_{\ell=1}^{r} \sigma_{k\ell} \boldsymbol{\mu}_{k\ell} \boldsymbol{v}_{k\ell}^T$.

A reasonable termination criterion is that the primal residual, $\varepsilon^{pri} = \|\boldsymbol{C}_1\boldsymbol{U} + \boldsymbol{C}_2\boldsymbol{W} - \boldsymbol{C}_3\|_2$, and dual residual, $\varepsilon^{dual} = \|\gamma \boldsymbol{C}_1^T \boldsymbol{C}_2(\boldsymbol{W}^{q+1} - \boldsymbol{W}^q)\|_2$, are smaller than a predefined tolerance.

4.2.3 Model parameter tuning

In noisy scenarios, the penalty parameter λ setting is critical to determine how much variation is persevered as patterns of interest or ignored as noise. An extremely large λ will coerce the individual variation to be zero. Decreasing λ will allow more subtype-specific patterns to be detected until overfitting.

Cross-validation is a popular strategy in parameter tuning for the balance of underfitting and overfitting. One round of cross-validation excludes a certain portion of samples and uses the model learned from other samples to predict the excluded ones. Then every model is assessed by summarizing prediction performances across multiple rounds. However, our sample-specific deconvolution estimates the individual expression of each sample in each subtype, which cannot be used to predict the excluded samples directly. So we proposed to randomly exclude entries rather than samples in $\boldsymbol{X}$ matrix (Fig. 4.3), similar to the strategy used in missing value imputation. The foundation of success is that the low-rank patterns in $\boldsymbol{T}_k$ matrix are detectable by only a portion of $\boldsymbol{X}$ entries and able to predict the excluded $\boldsymbol{X}$ entries.

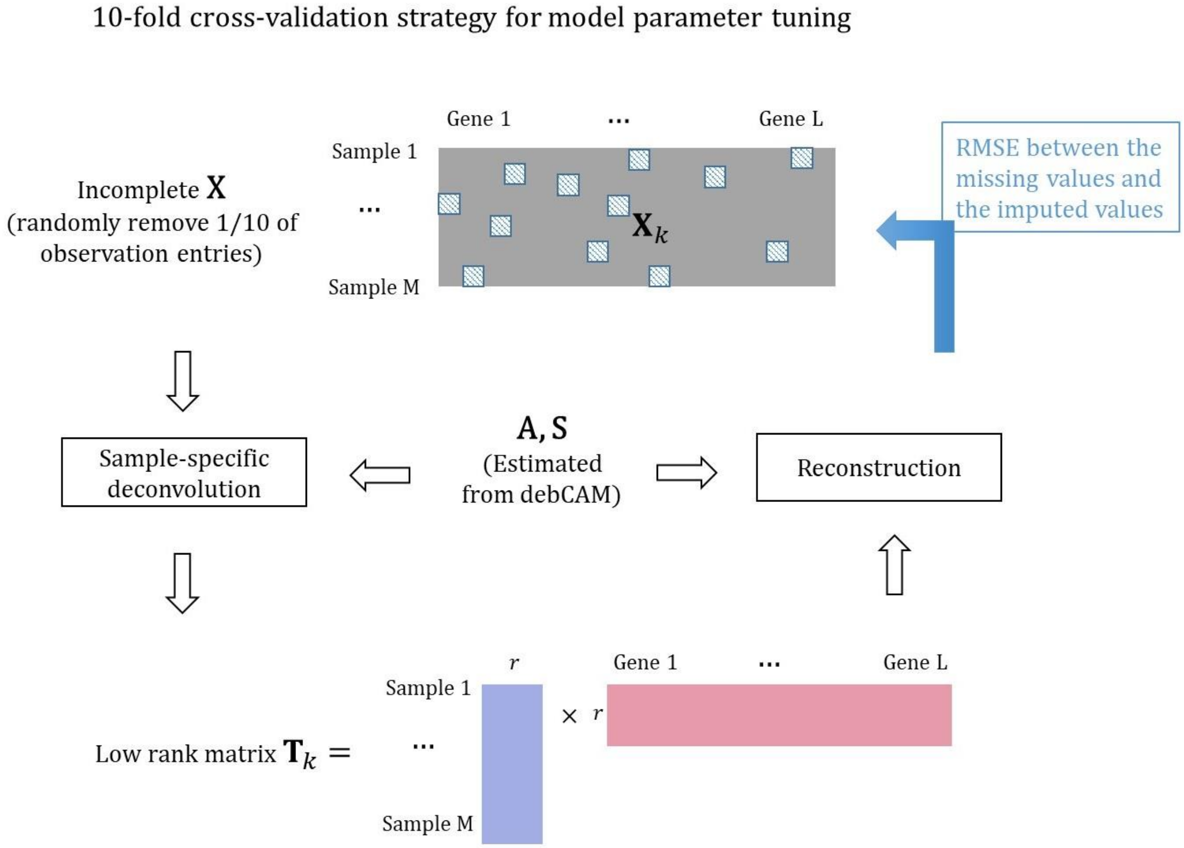

10-fold cross-validation strategy for model parameter tuning
Gene 1
…
Gene L
Sample 1
…
Sample M
Incomplete X
(randomly remove 1/10 of observation entries)
$\mathbf{X}_k$
RMSE between the missing values and the imputed values
Sample-specific deconvolution
A, S
(Estimated from debCAM)
Reconstruction
Low rank matrix $\mathbf{T}_k$ =
r
Gene 1
…
Gene L
Sample 1
…
Sample M
× r

Figure 4.3 10-fold cross-validation strategy for model parameter tuning. A part of entries are randomly removed before applying debCAM2.0. The removed entries are reconstructed by estimated $\boldsymbol{T}$ matrix and compared to observed expressions for computing RMSE to decide the optimal parameter λ.

Specifically, we fix the $\boldsymbol{A}$ and $\overline{\boldsymbol{S}}$ at the initialization values (from debCAM-estimation or *a priori* knowledge) and randomly remove entries in $\boldsymbol{X}$ matrix, leading to the objective function w.r.t $\Delta\boldsymbol{S}_i, i = 1, \dots, M$:

$$\min_{\{\Delta\boldsymbol{S}_i\}_{i=1}^{M}} \frac{1}{2}\sum_{i=1}^{M}\left\|P_{\Omega_i}(\boldsymbol{x}_i) - P_{\Omega_i}\left(\boldsymbol{a}_i(\overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i)\right)\right\|_2^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{T}_k\|_* \tag{4.14}$$

$$s.t. \overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i \succcurlyeq \mathbf{0}_{K\times L},$$

$$\boldsymbol{T}_k = [\Delta\boldsymbol{S}_1^T(k), \dots, \Delta\boldsymbol{S}_M^T(k)] \in \mathbb{R}^{L\times M}, k = 1, \dots, K,$$

where $P_{\Omega_i}(\boldsymbol{x}_i) \in \mathbb{R}^L$ denote a vector with the entries in Ω_i left alone, and all other entries set to zero. The workflow of our proposed 10-fold cross-validation strategy is:

(1) Randomly split all entries into 10 folds;

(2) Remove one fold of entries and use the remaining 9 folds of entries to solve (4.14) with different λ values $[\lambda_1, \lambda_2, ...]$;

(3) Use estimated $\Delta\boldsymbol{S}_i(\lambda_\theta), i = 1, \dots, M,\ \theta = 1,2, ...$, together with fixed $\boldsymbol{A}$ and $\overline{\boldsymbol{S}}$ matrix to reconstruct $\boldsymbol{X}$ matrix and only record the reconstructed values for the removed entries in $\boldsymbol{X}$;

(4) Repeat Step (2)-(3) and obtained a reconstructed $\widetilde{\boldsymbol{X}}(\lambda_\theta)$ matrix in which all entry values are reconstructed when their original values are absent in optimization processes with $\lambda = \lambda_\theta$.

(5) Calculate Root Mean Square Error (RMSE) by

$$RMSE(\lambda_\theta) = \sqrt{\frac{1}{ML}\sum_{i=1}^{M}\sum_{j=1}^{L}\left(\boldsymbol{X}_{ij} - \widetilde{\boldsymbol{X}}_{ij}(\lambda_\theta)\right)^2} \tag{4.15}$$

(6) Choose the λ_θ yielding the minimum RMSE.

Warm start can be used in Step (2) with the decreasing parameter $\lambda_1 > \lambda_2 > \cdots$, which use the estimation with λ_θ as the initialization of next optimization with $\lambda_{\theta+1}$.

The optimization problem (4.14) can be solved using a similar ADMM algorithm in (4.5)~(4.13) that have solved (4.3). The only modification is that (4.7) becomes

$$\boldsymbol{u}_i^{q+1} \in \underset{\boldsymbol{u}_i\in\mathbb{R}^{KL}}{\operatorname{argmin}} \frac{1}{2}\left\|P'_{\Omega_i}(\boldsymbol{H}_i\boldsymbol{u}_i) - P'_{\Omega_i}(\boldsymbol{v}_i)\right\|_2^2 + \frac{\gamma}{2}\left\|\boldsymbol{C}_1\boldsymbol{u}_i + \boldsymbol{y}_{\boldsymbol{i}}^q\right\|_F^2 \tag{4.16}$$

where $P'_{\Omega_i}(\cdot) = \left[\boldsymbol{1}_K^T \otimes P_{\Omega_i}(\cdot)^T\right]^T \in \mathbb{R}^{KL}$ makes all excluded-entry related variables be optimized only by the second term, which is still an unconstrained quadratic problem that can be solved easily. The remaining variables unrelated to excluded entries can still be optimized following (4.8)~(4.9).

4.2.4 Sparsity regularization

In addition to low-rank assumption, we could also reasonably assume only limited genes are involved in functional modules and thus impose a row-sparsity regularization by $\ell_{2,1}$-norm minimization. The alternative debCAM2.0 formulation will be:

$$\min_{\boldsymbol{A},\{\Delta \boldsymbol{S}_i\}_{i=1}^{M}} \frac{1}{2}\sum_{i=1}^{M} \|\boldsymbol{x}_i - \boldsymbol{a}_i(\overline{\boldsymbol{S}} + \Delta \boldsymbol{S}_i)\|_2^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{T}_k\|_* + \delta \|\boldsymbol{T}\|_{2,1} \tag{4.17}$$

where $\delta > 0$ is the regularization parameter of $\ell_{2,1}$ norm of $\boldsymbol{T}$, defined as

$$\|\boldsymbol{T}\|_{2,1} \triangleq \sum_{i=1}^{KL} \|\boldsymbol{t}_i\|_2$$

accounting for the row-sparsity of $\boldsymbol{T}$. If necessary, the parameter δ actually can be varied for different rows based on the character of each gene, such as mean-variance trend. The supplementary note (Subsection 4.5.2) gives more details on the optimization of (4.17) by ADMM method. The ℓ_1 or ℓ_2-norm minimization, as common-used sparsity regularization methods, could impose the entry sparsity in $\boldsymbol{T}$ matrix. We also provide ADMM optimization for sample-specific deconvolution with ℓ_1 or ℓ_2-norm minimization, which could be useful in other sBSS problems.

4.3 Results

As debCAM2.0 focuses on subtype-specific variation estimation, simulating biological variance within each subtype and technical variance for each observation is important for validating debCAM2.0 performance. We conduct two sets of simulations. The first is in an ideal scenario where the variance is not related to mean value. The second is more realistic where genes with larger mean usually have larger variance.

4.3.1 Validation on ideal simulations

In the first simulations, we design twelve function modules, with four in each of three subtypes. The observations for 300 genes in 50 samples were simulated with subtype-specific expression baseline, $\overline{\boldsymbol{S}}$, sampled from the purified cell populations in real benchmark microarray gene expression data GSE19380 [6]. $\boldsymbol{a_i}, i = 1, \dots, M$, are drawn randomly from a flat Dirichlet distribution. Between-sample variation, $\Delta \boldsymbol{S}_i(k, j), i = 1, \dots, M,$ for the *k*th subtype and *j*th gene was drawn from normal distribution $\mathcal{N}(0, \sigma_{kj}^{(s)})$ if the *j*th gene was involved in a function module in the *k*th subtype; otherwise zero (Fig. 4.4a). The genes in the same function module has pairwise correlation coefficient equal to one, thus generating a highly correlated gene set in each module. $\sigma_{kj}^{(s)}$ are drawn from uniform distribution $U[50, 300]$. The technical noise, $\boldsymbol{n}_i, i = 1, \dots, M,$ was drawn from zero-mean normal distribution with the variance $\sigma_{ij}^{(n)}$=10.

The twelve functional modules can be recognized in the variation matrix from debCAM2.0 when λ falls into a certain range (Fig. 4.4b~4.4i). Increasing the penalty parameter of the nuclear norm will filter more noise but at the cost of the possibility of missing the true variation signal. RMSE derived by 10-fold cross-validation strategy is relatively small when $\lambda = 1 \sim 50$ and reach the minimum at $\lambda = 5$ (Fig. 4.5a). The estimated variation matrix looks quite similar when $1 \leq \lambda \leq 50$ (Fig. 4.4e~4.4g), with 12 clear patterns and some artifacts. The artifacts are formed when the signal variation in one subtype spreads to other subtypes for the same genes, which are much lower than detected true signals if λ is not extremely small. (The supplementary note in Subsection 4.5.1 shows the nuclear norm minimization for each subtype's variation matrix is a good option to reduce artifacts compared to other regularization terms.)

It is interesting to find $\lambda = 5$ is also the point where both primal and dual residuals surge in ADMM algorithm (Fig. 4.5c~4.5f). It is because larger λ tends to train an over-simplified model and thus approach the optimum solution more easily in ADMM.

The recovery of sample-specific signals in a subtype is also affected by the mixing proportions of this subtype within the sample. When a subtype accounts for a very small portion in a certain sample, its true signal in this sample will be very weak and thus underestimated (green points in Fig. 4.6). On the contrary, the major subtype in a sample can be estimated very well by debCM-SS (red points in Fig. 4.6).

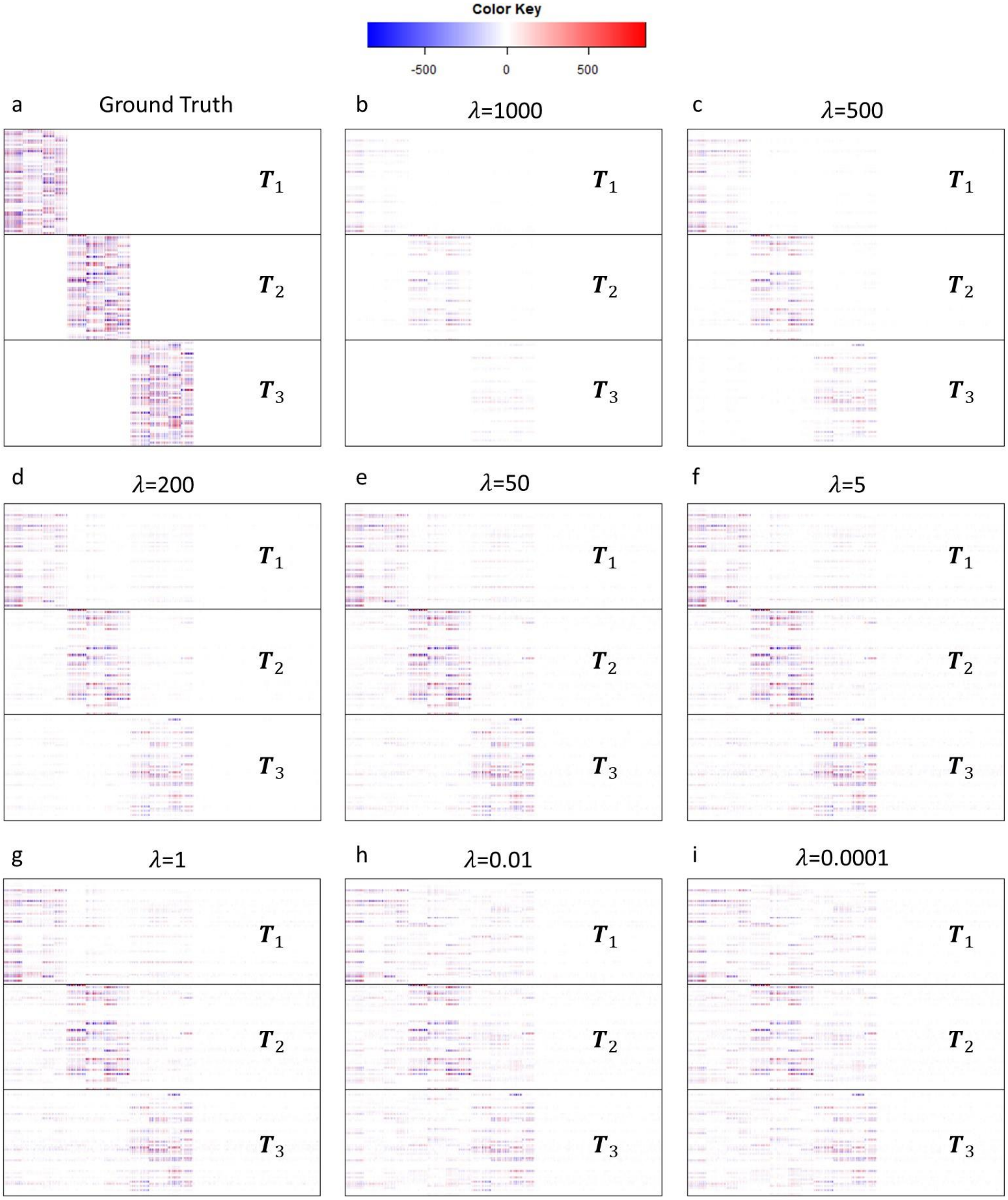


Figure 4.4 Heatmap of estimated $\boldsymbol{T}$ matrix with varied λ parameters compared to ground truth in the ideal simulation. Increasing the penalty parameter of the nuclear norm will filter more noise but at the cost of the possibility of missing true signal variation.

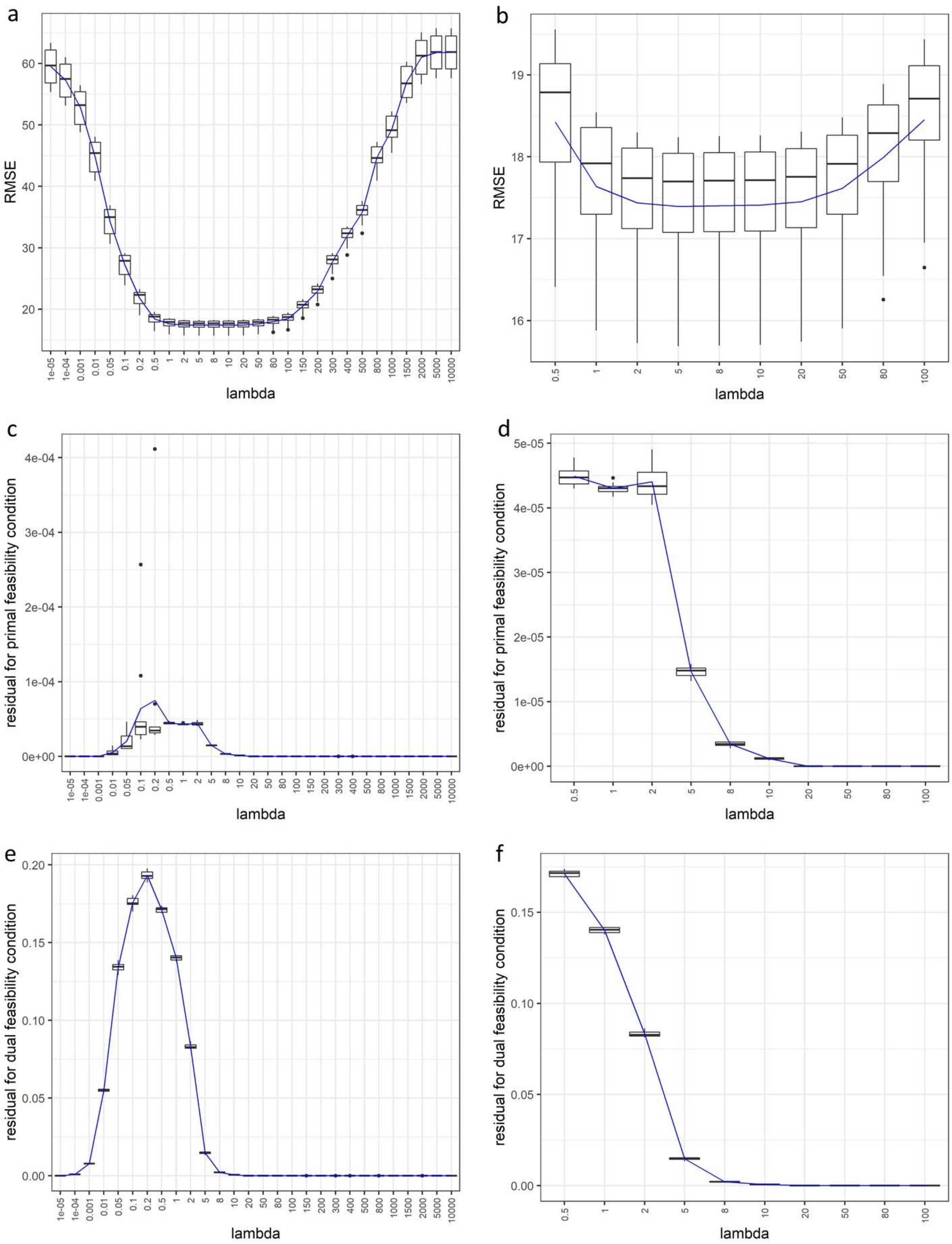


Figure 4.5 10-fold cross-validation results under different λ parameter in the ideal simulation. (a) RMSE; (c) Residuals for primal feasibility condition; (e) Residuals for dual feasibility condition; (b),(d), (f) are zoomed curves of (a),(c),(e).

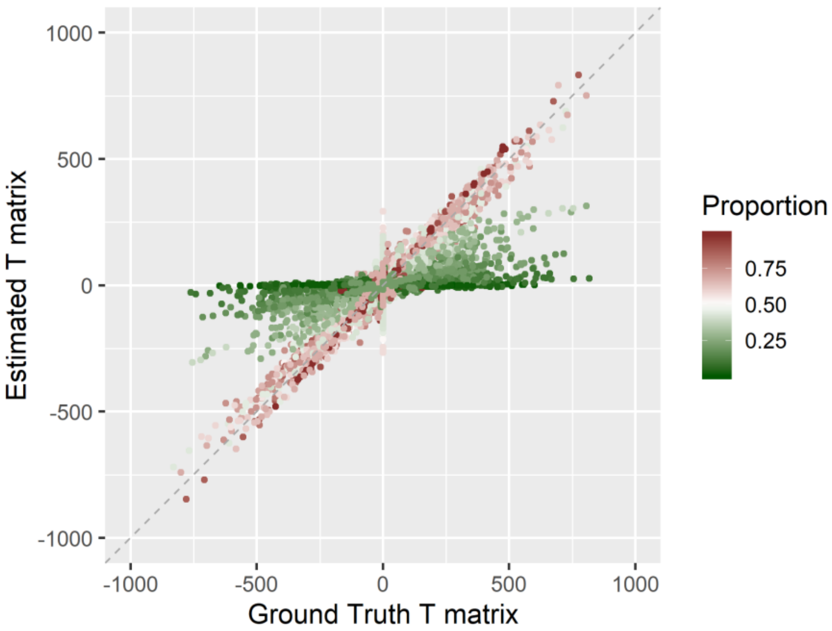


Figure 4.6 Estimated ***T*** matrix versus ground truth when λ=5 in the ideal simulation. The mixing proportions associated with estimated entries are colored to show the sample-specific expression estimations for high-proportion subtypes can be estimated more accurately than those for low-proportion ones

4.3.2 Validation on realistic simulations

Mean-variance trend is widely existing in molecular expression data. In our second simulation, all settings are the same as above except that the variance of subtype-specific expression, $\sigma_{kj}^{(s)}$, and the technical variance of observations, $\sigma_{ij}^{(n)}$, are proportional to the subtype-specific expression mean and mixed expression level, respectively. The coefficient of variation (CV), as the ratio of the standard deviation to the mean, is drawn from uniform distribution $U[0.15, 0.3]$ and $U[0.02, 0.05]$, respectively.

10-fold cross-validation strategy still obtains the minimum RMSE at $\lambda = 5$ (Fig. 4.8a~b) when both primal and dual residuals also surge (Fig. 4.8c~4.8f). However, the estimated variation matrix by debCAM2.0 is blurred by artifacts trained from noise (Fig. 4.7). Some high-expressed genes have relatively large variance, which could be falsely modeled as subtype-specific signal variations. As shown in Fig. 4.9, the entries with zero value in Ground Truth variation matrix could be overestimated.

Though the absolute expression values estimated by debCAM2.0 could deviate from Ground Truth, we can still clearly detect 12 functional modules defined by the Weighted Gene Correlation Network Analysis (WGCNA) [83, 98] on the estimated sample-specific expressions (Fig. 4.10). WGCNA constructs weighted networks based on correlation patterns among genes across samples and thus detects function modules of highly-correlated gene sets. In Fig. 4.10, the second and third subtype finds the exact four true modules with very few genes are missed. The first subtype detects an extra false module, but it is a less significant pattern compared to other modules and can be undetectable with stricter tree height cut threshold. More importantly, without debCAM2.0 based deconvolution (Fig. 4.10d), WGCNA on mixture expression profiles can find none of the true modules, but three false modules that are related to the mixing process of three subtypes.

4.3.3 Incorporation of L21-norm regularization

In the above simulations, the deconvoluted sample-specific signals contain artifacts trained from signals of other subtypes and artifacts trained from noise (Fig. 4.4 and Fig. 4.7). We can use a $\ell_{2,1}$-norm regularization to enforce the sparsity of genes that have signal variation across samples. It is supposed to reduce artifacts while it also follows the assumption that genes contributing to source variation in hidden modules are limited. Figure 4.11 shows the alleviated artifacts with $\lambda = 5$ and $\delta = 10$, 1, or 0.1. The true function modules are correctly detected with $\lambda = 5$ and $\delta = 1$ or 0.1, where the false module in the first subtype is suppressed when $\delta = 1$ (Fig. 4.12).

Increasing the penalty parameter δ will force more genes to have zero variance, which suppresses the artifacts and false function modules but brings the risk of missing the true signals. It is critical to propose a parameter tuning method for δ. However, the cross-validation strategy with randomly excluding entries for tuning parameter λ is based on the low-rank assumption, where the hidden low-rank patterns can be trained from a part of entries and then used to reconstruct the remaining entries. This strategy is not applicable to δ selection, which needs further study.

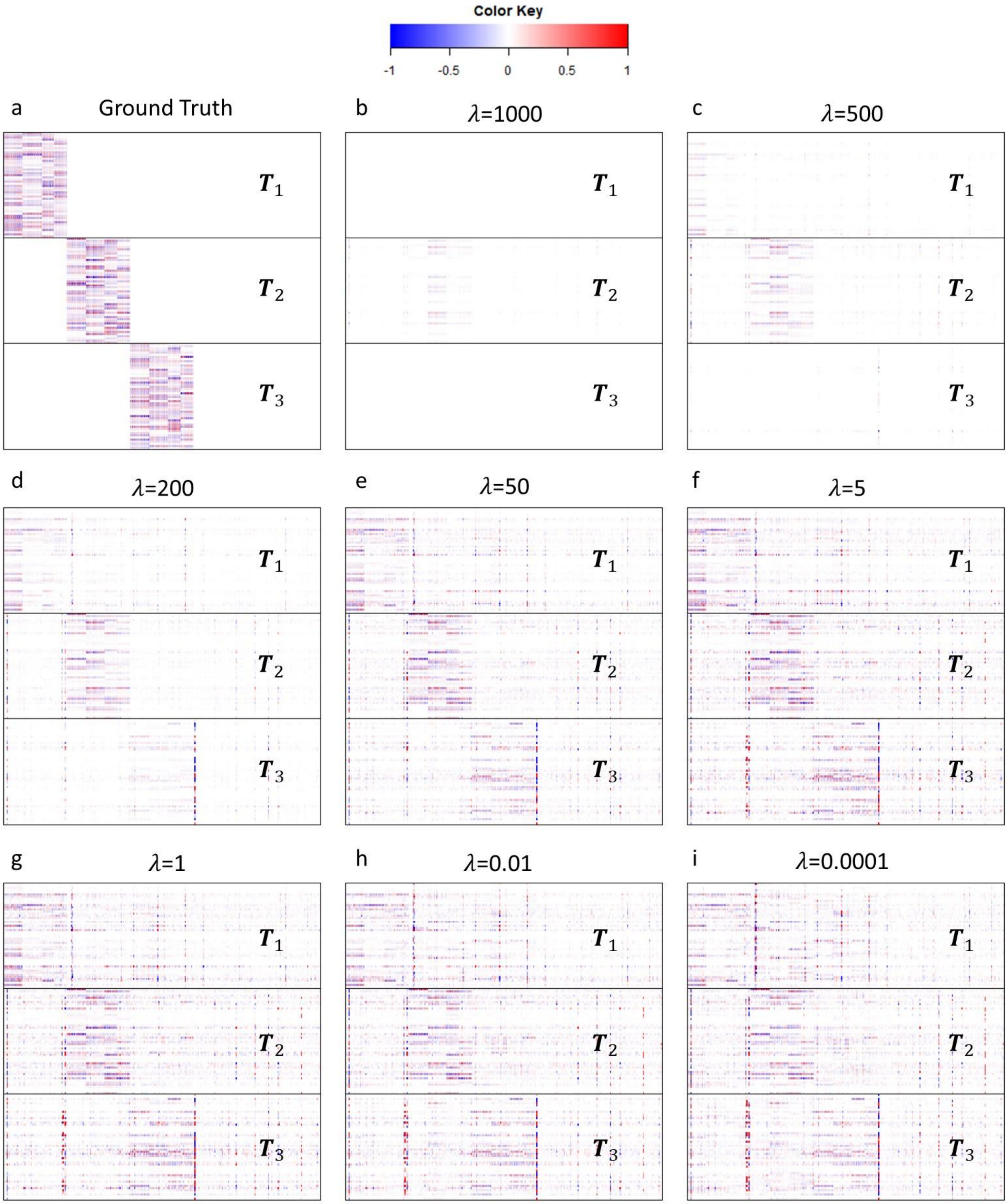


Figure 4.7 Heatmap of estimated $\boldsymbol{T}$ matrix scaled by associated means compared to ground truth in the realistic simulation with varied λ parameters. Increasing the penalty parameter of the nuclear norm will filter more noise but at the cost of the possibility of missing true variation signal.

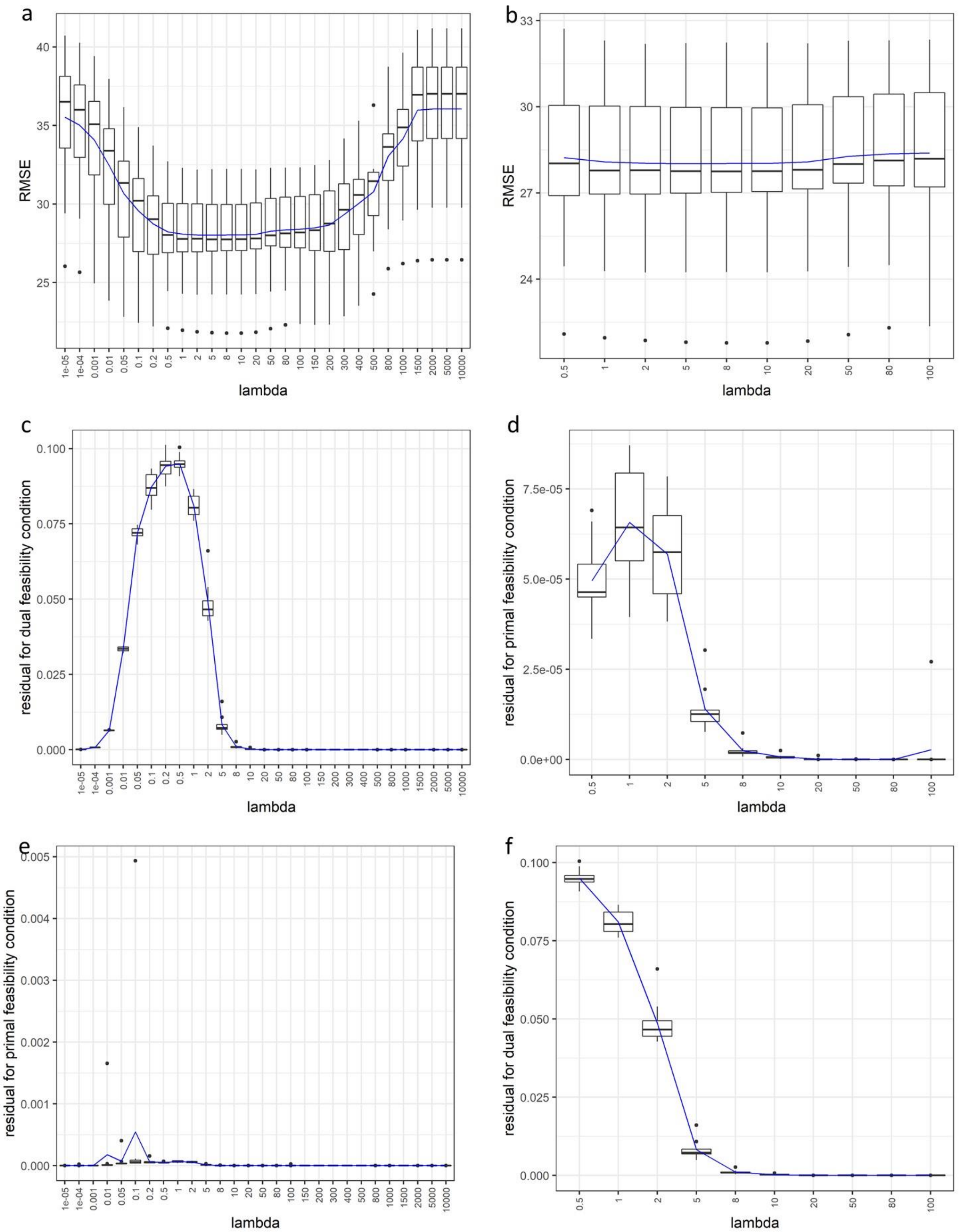


Figure 4.8 10-fold cross-validation results under different λ parameter in the realistic simulation. (a) RMSE; (c) Residuals for primal feasibility condition; (e) Residuals for dual feasibility condition; (b),(d), (f) are zoomed curves of (a),(c),(e).

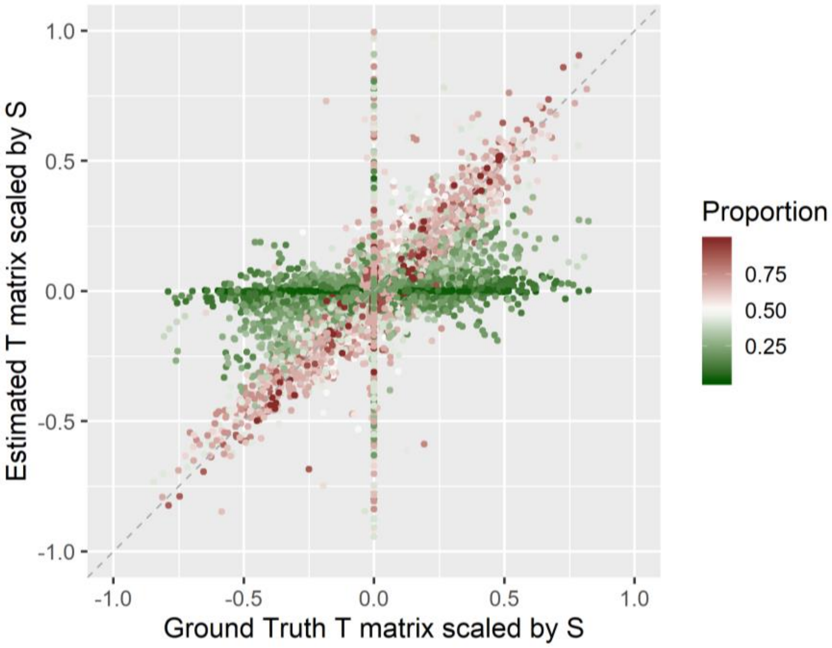


Figure 4.9 Estimated ***T*** matrix scaled by associated means versus ground truth in the realistic simulation (λ=5). The mixing proportions associated with estimated entries are colored to show the sample-specific expression estimations for high-proportion subtypes can be estimated more accurately than those for low-proportion ones

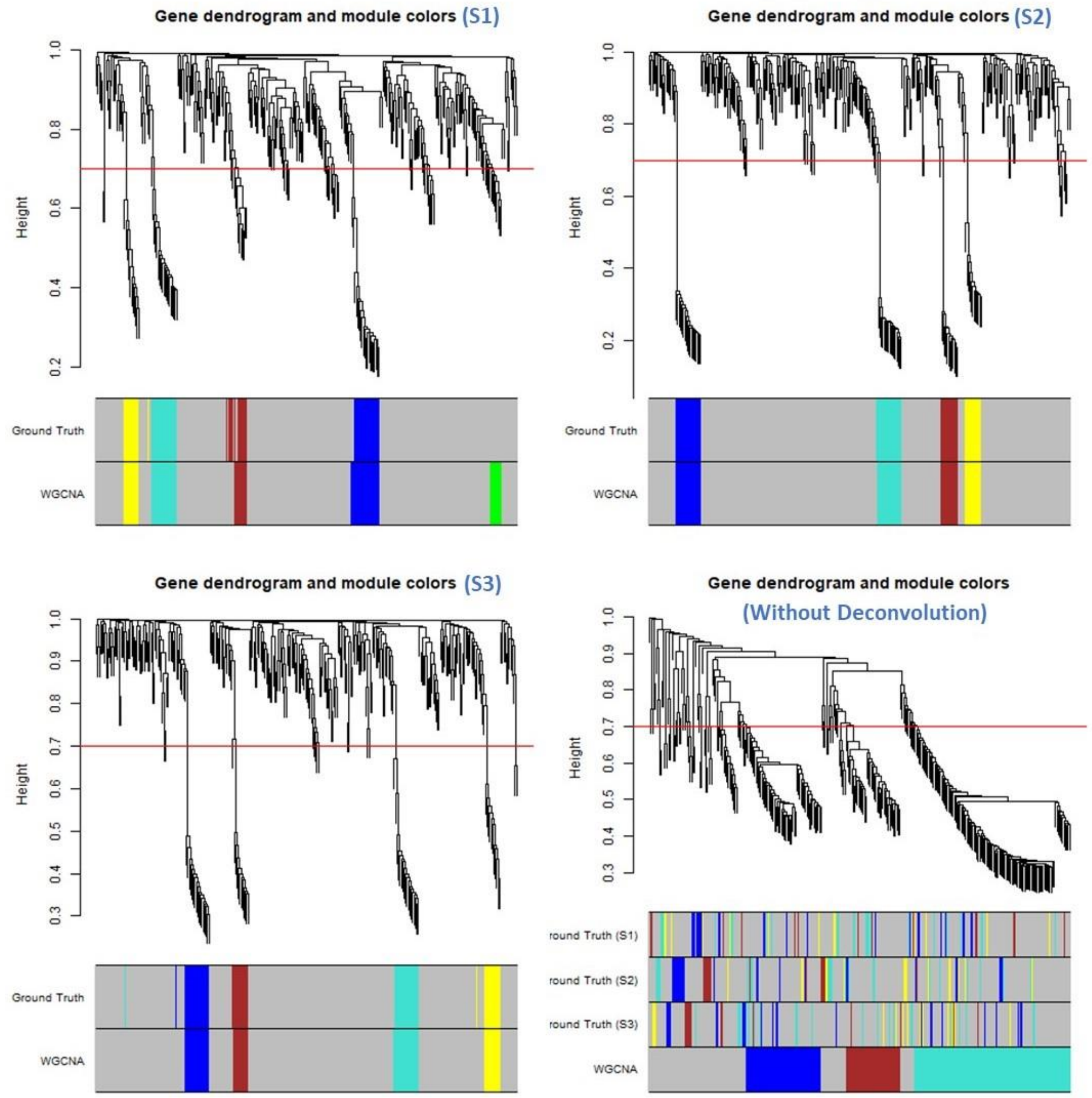


Figure 4.10 Gene co-expressed function modules detected by WGCNA on debCAM2.0 estimated sample-specific expression for each subtype (a~c) or on originally observed expressions without deconvoluton (d). (Network interconnectedness is measured by topological overlap; cutHeight = 0.7; minSize = 8.)

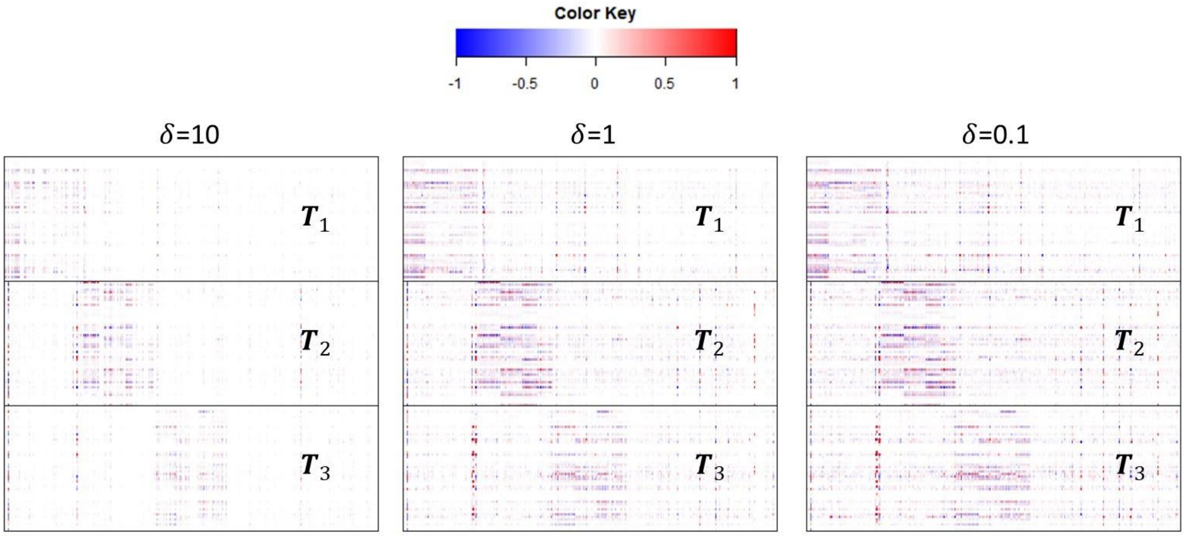


Figure 4.11 Heatmap of estimated T matrix scaled by associated means compared to ground truth in the realistic simulation with $\lambda = 5$ and varied δ. Increasing the penalty of L21 norm will enforce more zero columns in ΔS_k matrix.

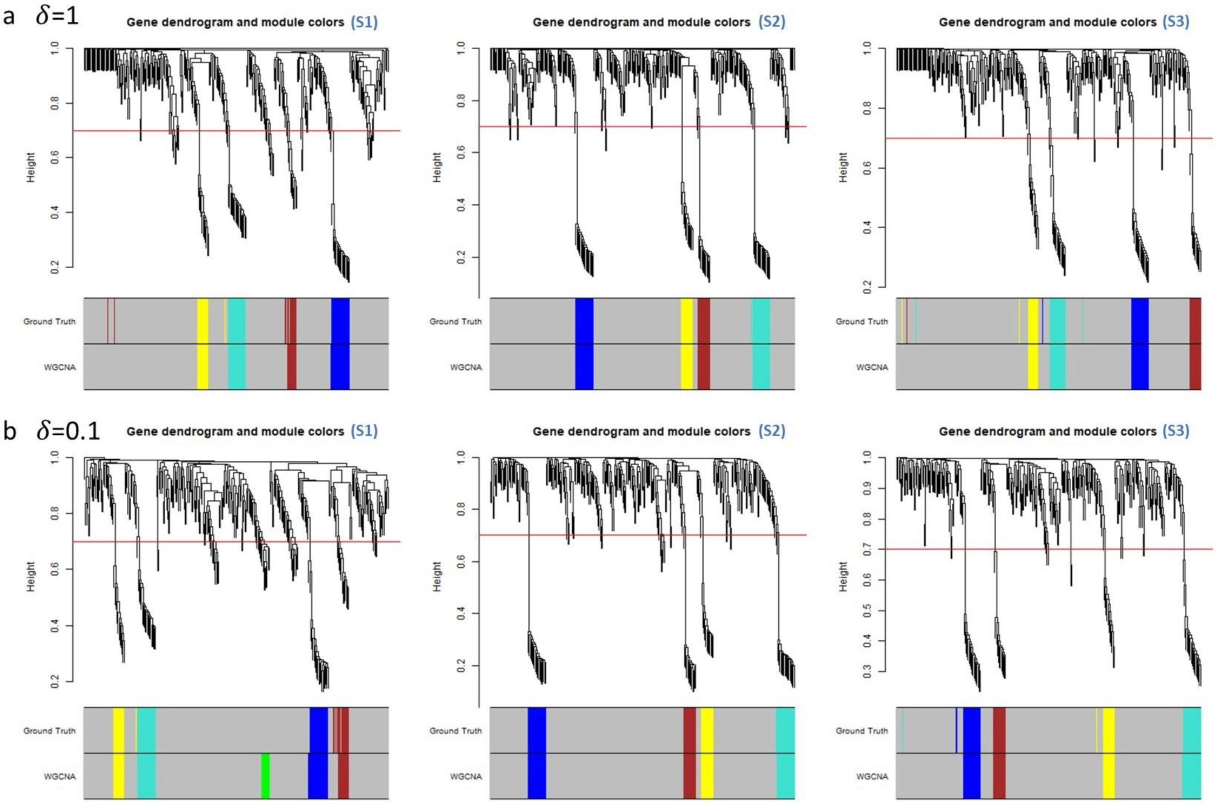


Figure 4.12 Gene co-expressed function modules detected by WGCNA on debCAM2.0 estimated sample-specific expression for each subtype with $\lambda = 5$ and $\delta = 1$ or 0.1. (Network interconnectedness is measured by topological overlap; cutHeight = 0.7; minSize = 8.)

4.4 Discussion

Most existing tissue deconvolution methods ignore the expression variability of subtypes across individual samples. debCAM2.0 will significantly expand the utility of debCAM by producing subtype-specific expression profiles in each sample. The success of debCAM2.0 depends on the low-rank assumption, which takes advantage of biologically expected cooperation among genes and thus sheds light on solving the seemingly underdetermined sample-specific deconvolution problem. The low-rank assumption holds naturally in molecule expression data when there exist activated functional modules required by particular biological processes or pathways in different subtypes. The detection of such subtype-specific associations or networks is one of the major targets in the analysis of molecule expression profiles. After our sample-specific deconvolution by debCAM2.0, conventional network analysis methods can be applied directly to the estimated sample-subtype-specific signals to construct subtype-specific networks, e.g. weighted correlation network analysis (WGCNA [83, 98]) and differential dependency network analysis (DDN [84, 85, 99, 100]).

The cross-validation strategy of excluding entries randomly is inspired by the similar ideas in matrix imputation methods that commonly assume the matrix to be recovered has a low rank. Our results consistently show a U-curve over parameter λ, demonstrating the feasibility of the proposed cross-validation strategy. Meanwhile, debCAM is not sensitive to the choice of λ, as the U-curve has a wide platform where the recovered sample-subtype-specific signals are similar and detected modules are close.

It is also reasonable to assume that genes involved in biological associations or networks are sparse. Therefore, it deserves our further study to use $\ell_{2,1}$-norm regularization for reducing artifacts and improving function module detection.

When group information is available, we can also apply basic debCAM algorithm to each group to obtain group-wise expression profiles of subtypes.Compared to sample-specific deconvolution, group-specific deconvolution aims at a lower resolution of underlying subtype signals and thus could obtain more robust results. If grouping is fine enough, group-specific deconvolution can also acquire signal variation in each subtype and thus help detect function modules and construct biological networks.

4.5 Supplementary Note

4.5.1 Regularization term selection

To observe the effect of the nuclear norm regulation on the estimation of $\boldsymbol{T} \triangleq [\boldsymbol{T}_1^T, \dots, \boldsymbol{T}_M^T]^T \in \mathbb{R}^{KL\times M}$, we design a noise-free scenario where $\mathbf{X} = \mathbf{A}\bar{\mathbf{S}} + \mathcal{A}(\mathbf{T})$, i.e., $\boldsymbol{x}_i = \boldsymbol{a}_i(\bar{\mathbf{S}} + \Delta\mathbf{S}_{\mathrm{i}})$, i $= 1, \dots, \mathrm{M}$, and we also do not consider non-negative constraints $\bar{\boldsymbol{S}} + \Delta\boldsymbol{S}_{\mathrm{i}} \in \mathbb{R}^+$ for the moment. Then the subproblem of optimizing **T** (4.2-4.3) become

$$\min_{\boldsymbol{T}} \sum_{k=1}^{K} \|\boldsymbol{T}_k\|_*, \text{ s.t} \, \mathcal{A}(\boldsymbol{T}) = \boldsymbol{X} - \boldsymbol{A}\bar{\mathbf{S}},$$

where nuclear norm regularization is applied to each subtype independently. We can also apply nuclear norm regularization to the whole $\boldsymbol{T}$ matrix containing all subtypes:

$$\min_{\boldsymbol{T}} \|\boldsymbol{T}\|_*, \text{ s.t} \, \mathcal{A}(\boldsymbol{T}) = \boldsymbol{X} - \boldsymbol{A}\bar{\mathbf{S}}$$

Matlab CVX package can be used to solve this problem when the problem size is small. The results from a simple simulation can help us decide which regulation is more reasonable (Fig. 4.13). Still using the setting in the ideal simulations above, each subtype is set to have one function module of 20 highly-correlated genes. Minimizing the nuclear norm of the whole $\boldsymbol{T}$ matrix generates artifacts that the signal variation in one subtype spread to other subtypes for the same genes. Such artifacts can be reduced by minimizing the nuclear norm for each subtype separately. Note the third subtype always obtains more severe artifacts than the other two subtypes. One reason is that the third subtype accounts for a larger proportion in total and thus could be more over-estimated than others. Nevertheless, we can still extract the meaningful pattern from sample-specific signals as long as the artifacts are less abundant than the true variation of interest.

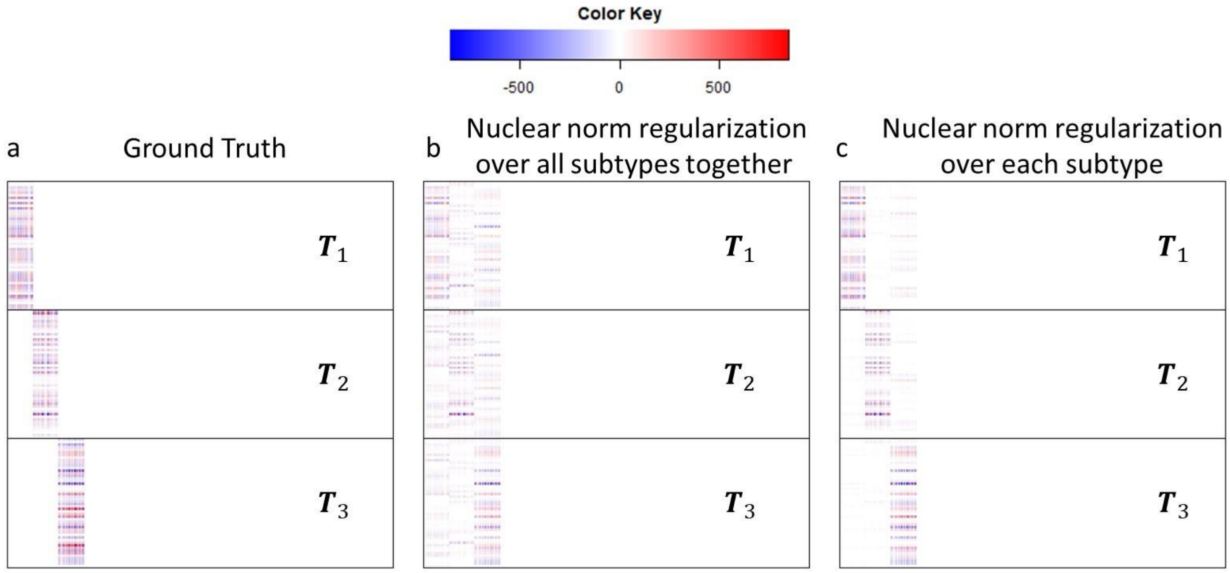


Figure 4.13 Ground truth and estimated $\boldsymbol{T}$ matrix by nuclear norm regularization over all subtypes together or over each subtype independently.

4.5.2 ADMM solutions with extra L21, L2, L1 regularization

Practically, we could assume that there should be only a limited number of genes contributing to source variation in hidden modules, and thus impose a $\ell_{2,1}$ norm regularization. We may also reasonably assume that the source matrices among different samples are similar, and thus assume the Frobenius norm (or ℓ_1 norm) of each $\Delta\boldsymbol{S}_i$is small. The alternative debCAM2.0 formulation with $\ell_{2,1}$ norm and Frobenius norm regularization is:

$$\min_{\boldsymbol{A},\{\Delta\boldsymbol{S}_i\}_{i=1}^{M}} \frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{x}_i - \boldsymbol{a}_i(\overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i)\|_2^2 + \eta\sum_{i=1}^{M}\|\Delta\boldsymbol{S}_i\|_2^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{T}_k\|_* + \delta\|\boldsymbol{T}\|_{2,1} \qquad (4.2)$$

$$s.t.\ \boldsymbol{a}_i \succcurlyeq \boldsymbol{0}_K, \boldsymbol{a}_i\boldsymbol{1}_K = 1, \overline{\boldsymbol{S}} + \Delta\boldsymbol{S}_i \succcurlyeq \boldsymbol{0}_{K\times L},$$

$$\boldsymbol{T}_k = [\Delta\boldsymbol{S}_1^T(k), \dots, \Delta\boldsymbol{S}_M^T(k)] \in \mathbb{R}^{L\times M},\ k = 1, \dots, K,$$

$$\boldsymbol{T} \triangleq [\boldsymbol{T}_1^T, \dots, \boldsymbol{T}_M^T]^T \in \mathbb{R}^{KL\times M},$$

where $\eta > 0$ is the regularization parameter of ℓ_2 norms, and $\delta > 0$ is the regularization parameter of $\ell_{2,1}$ norm of $\boldsymbol{T}$, defined as

$$\|\boldsymbol{T}\|_{2,1} \triangleq \sum_{i=1}^{KL} \|\boldsymbol{t}_i\|_2$$

accounting for the row-sparsity of $\boldsymbol{T}$. The parameter η and δ actually can be varied for different rows if necessary.

Similar to (4.3) and (4.5), we solve the convex subproblem in objective function (4.2) w.r.t $\boldsymbol{T}$ by a ADMM reformulations:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{KL\times M},\boldsymbol{W}\in\mathbb{R}^{2KL\times M}} \frac{1}{2}\|\mathcal{A}(\boldsymbol{U})-\boldsymbol{V}\|_F^2 + \lambda\sum_{k=1}^{K}\|\boldsymbol{B}_k\boldsymbol{W}\|_* + I_+(\boldsymbol{B}_0\boldsymbol{W}) + \delta\|\boldsymbol{B}_{k+1}\boldsymbol{W}\|_{2,1} \qquad (4.5)$$

$$s.t.\ \boldsymbol{C}_1\boldsymbol{U} + \boldsymbol{C}_2\boldsymbol{W} = \boldsymbol{C}_3,$$

ℓ_2 norm regularization is merged into the first term, where the linear transformation $\mathcal{A}(\boldsymbol{U}) = \mathcal{A}([\boldsymbol{u}_1,\dots,\boldsymbol{u}_M]) = [\boldsymbol{H}_1\boldsymbol{u}_1,\dots,\boldsymbol{H}_M\boldsymbol{u}_M]$ with $\boldsymbol{H}_i = \begin{bmatrix}\boldsymbol{a}_i^p \otimes I_L \\ \sqrt{2\eta} I_{LK}\end{bmatrix} \in \mathbb{R}^{L(K+1)\times LK}, i = 1,\dots,M$, and $\boldsymbol{V} = \begin{bmatrix}\boldsymbol{X}^T \\ \sqrt{2\eta}\boldsymbol{1}_M^T \otimes vec(\overline{\boldsymbol{S}}^T)\end{bmatrix} \in \mathbb{R}^{L(K+1)\times M}$. The modification due to the incorporation of $\ell_{2,1}$ norm regularization include:

$$\boldsymbol{W} \triangleq \begin{bmatrix}\boldsymbol{T} \\ \boldsymbol{S} \\ \boldsymbol{R}\end{bmatrix} \in \mathbb{R}^{3KL\times M}$$

$$\boldsymbol{C}_1 \triangleq \begin{bmatrix}\boldsymbol{I}_{KL} \\ \boldsymbol{I}_{KL} \\ \boldsymbol{I}_{KL}\end{bmatrix} \in \mathbb{R}^{3KL\times KL}$$

$$\boldsymbol{C}_2 \triangleq -\boldsymbol{I}_{3KL} \in \mathbb{R}^{3KL\times 3KL}$$

$$\boldsymbol{C}_3 \triangleq \begin{bmatrix}\boldsymbol{1}_M^T \otimes vec(\overline{\boldsymbol{S}}^T) \\ \boldsymbol{0}_{KL\times M} \\ \boldsymbol{1}_M^T \otimes vec(\overline{\boldsymbol{S}}^T)\end{bmatrix} \in \mathbb{R}^{3KL\times M}$$

$$\boldsymbol{B}_0 \triangleq [\boldsymbol{0}_{KL\times KL}, \boldsymbol{I}_{KL}, \boldsymbol{0}_{KL\times KL}] \in \mathbb{R}^{KL\times 3KL}$$

$$\boldsymbol{B}_k \triangleq \left[\mathbf{0}_{L\times(k-1)L}, \boldsymbol{I}_L, \mathbf{0}_{L\times(K-k)L}, \mathbf{0}_{L\times KL}, \mathbf{0}_{L\times KL}\right] \in \mathbb{R}^{L\times 3KL}, k = 0, \dots, K$$

$$\boldsymbol{B}_{K+1} \triangleq [\mathbf{0}_{KL\times KL}, \mathbf{0}_{KL\times KL}, \boldsymbol{I}_{KL}] \in \mathbb{R}^{KL\times 3KL}$$

(4.10) can be modified as

$$\min_{\boldsymbol{W}\in\mathbb{R}^{2KL\times M}} \lambda \sum_{k=1}^{K} \|\boldsymbol{B}_k\boldsymbol{W}\|_* + I_+(\boldsymbol{B}_0\boldsymbol{W}) + \delta\|\boldsymbol{B}_{k+1}\boldsymbol{W}\|_{2,1} + \frac{\gamma}{2}\|\boldsymbol{C}_1\boldsymbol{U}^{q+1} + \boldsymbol{C}_2\boldsymbol{W} - \boldsymbol{C}_3 - \boldsymbol{Z}^q\|_F^2, (4.10)$$

which is decoupled as a function of $\boldsymbol{T}_1, \dots, \boldsymbol{T}_K, \boldsymbol{S}$, and $\boldsymbol{R}$:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{2KL\times M}} \sum_{k=1}^{K} \left\{\lambda\|\boldsymbol{T}_k\|_* + \frac{\gamma}{2}\left\|\boldsymbol{U}_k^{q+1} - \boldsymbol{T}_k - \mathbf{1}_M^T \otimes \bar{\boldsymbol{s}}_k - \boldsymbol{Z}_k^q\right\|_F^2\right\} + \left\{I_+(\boldsymbol{S}) + \frac{\gamma}{2}\left\|\boldsymbol{U}^{q+1} - \boldsymbol{S} - \boldsymbol{Z}_0^q\right\|_F^2\right\}$$
$$+ \left\{\delta\|\boldsymbol{R}\|_{2,1} + \frac{\gamma}{2}\left\|\boldsymbol{U}^{q+1} - \boldsymbol{R} - \mathbf{1}_M^T \otimes vec(\bar{\boldsymbol{S}}^T) - \boldsymbol{Z}_{K+1}^q\right\|_F^2\right\}$$

where $\boldsymbol{Z} \triangleq [\boldsymbol{Z}_1^T, \dots, \boldsymbol{Z}_K^T, \boldsymbol{Z}_0^T, \boldsymbol{Z}_{K+1}^T]^T \in \mathbb{R}^{3KL\times M}$. $\boldsymbol{T}_1, \dots, \boldsymbol{T}_K$, and $\boldsymbol{S}$ can still be updated using (4.11). Let $\boldsymbol{R} \triangleq [\boldsymbol{r}_1, \dots, \boldsymbol{r}_{KL}]^T$ and $\boldsymbol{U}^{q+1} - \mathbf{1}_M^T \otimes vec(\bar{\boldsymbol{S}}^T) - \boldsymbol{Z}_{K+1}^q \triangleq [\boldsymbol{p}_1, \dots, \boldsymbol{p}_{KL}]^T$. With proximal operators of $\|\boldsymbol{R}\|_{2,1}$ [97], $\boldsymbol{R}$ can be updated as

$$\boldsymbol{R}^{q+1} = \left[\boldsymbol{r}_1^{q+1}, \dots, \boldsymbol{r}_{KL}^{q+1}\right]^T$$

where $\boldsymbol{r}_i^{q+1} \in \operatorname{argmin} \delta\|\boldsymbol{r}_i\|_2 + \frac{\gamma}{2}\|\boldsymbol{r}_i - \boldsymbol{p}_i\|_2^2, i = 1, \dots, KL$. So $\boldsymbol{r}_i^{q+1} = I_+\left(1 - \frac{\delta}{\gamma\|\boldsymbol{p}_i\|_2}\right)\boldsymbol{p}_i$.

Imposing a ℓ_2 and/or $\ell_{2,1}$ norm regularization could bring a more reliable solution if the underlying assumptions hold. Note the extra regularization terms also increase the difficulty of model parameter tuning.

The objective function (4.2) can be easily extended to support more regularization terms, e.g. ℓ_1 norm. The modification on ADMM approach for incorporating extra ℓ_1 norm term is similar to that for extra $\ell_{2,1}$ norm term.

Chapter 5

Summary and Future Work

5.1 Summary of Contributions

This dissertation aims to develop data modeling and deconvolution methods for molecular characterization of tissue heterogeneity to aid studies of systems biology. We explore subtype-specific markers, mixing proportions and subtype-specific expressions in a fully unsupervised way, which is further extended to support supervised/semi-supervised learning and enable sample-specific deconvolution. We design a new statistical method for the robust detection of *a priori* markers, which is critical to supervised/semi-supervised deconvolution. We solve the challenging sample-specific deconvolution problem to construct biological networks in each subtype.

5.1.1 Unsupervised deconvolution

We report a completely unsupervised deconvolution algorithm, deconvolution by convex analysis of mixtures – debCAM, that can blindly identify novel markers and subtypes in complex tissues. Specifically, the contributions of this work include:

- We develop a Bioconductor R package for debCAM pipeline, as the first fully unsupervised deconvolution tool to dissect complex tissues into molecularly distinctive tissue or cell subtypes based on bulk expression profiles. In this package, we implement and test the latest functionalities of the debCAM algorithm in the literature, which can automatically detect tissue/cell-specific markers, determine the number of constituent subtypes, calculate subtype proportions in individual samples, and estimate tissue/cell-specific expression profiles.
- We improve the debCAM R package by enhanced data preprocessing and accelerated robust simplex identification, to assure the effective application of debCAM algorithm to the analysis of real data for addressing practical problems in disease study.

- We extend debCAM R package to support supervised or semi-supervised deconvolution by incorporating *a priori* information, *e.g.* known markers, known mixing proportions or subtype-specific expression profiles.
- We design a vertex preserving projection method to visualize the high-dimensional simplex in a 2D plane, which can help users to observe the underlying pattern in mixtures and interpret deconvolution results more easily.
- We demonstrate the wide application and biomedical utility of debCAM using benchmark datasets of multiple molecular omics types (mRNA, methylation, protein) from the study of brain or immune system. We validate novel results by *a priori* information collected from literatures or cross-validate novel results from the analysis of multiple datasets.
- We apply debCAM to heterogeneous brain data, and detect the major neuron and glia subtypes. We use UNDO2.0 (or debCAM with K=2) to explore the glia-neuron in cortex or cerebellum and detect region-specific neuron/glia markers.
- We apply debCAM to heterogeneous vascular proteomic data, and detect distinct expression subtypes in the LAD and AA specimen. Comparison of CAM-estimated proportions and pathologist-observed proportions, both LAD and AA specimens have an underlying subtype linked to fibrous plaque (FP) tissue type, which is further confirmed by pure specimen profiling. The debCAM-FP markers are further investigated as early biomarkers of cardiovascular disease and validated by a clinical cohort.

5.1.2 Robust detection of subtype-specific markers

We describe a statistically-principled marker detection method, One Versus Everyone Subtype Exclusively-expressed Genes – OVESEG-test, that can detect tissue or cell-specific markers among many subtypes from the purified expression profiles. The contributions of this work include:

- We design a test statistic based on the One Versus Everyone (OVE) strategy that mathematically matches the definition of subtype-specific markers. We use a novel permutation scheme to estimate the corresponding distribution under the null hypothesis, where a mixture null distribution is modeled and estimated to match the complex expression patterns of non-markers. The obtained OVESEG-test p-values can assess the significance level of markers and guide marker selection.

- We validate the performance of OVESEG-test with realistic synthetic data sets, where the type 1 error rate or false discovery rate (FDR) can be controlled at the expected level. OVESEG-test and its earlier version OVE-FC outperform top peer methods in terms of detection power, evaluated by partial area under the receiver operating characteristic curve (pAUC). OVESEG-test outperforms OVE-FC in the more challenging cases involving non-significantly large fold change or phenotypically closer cell types.
- We apply OVESEG-test to two benchmark gene expression data sets and detect many known and *de novo* subtype-specific markers. Subsequent supervised deconvolution results, obtained using markers detected by the OVESEG-test or OVE-FC, show superior performance when compared with popular peer methods. OVESEG-test strongly beats OVO t-test methods in a more challenging case of noisy RNAseq data and small sample size, by leveraging information from all subtypes for variance modeling other than from subtype pairs.
- We apply OVESEG-test to proteomics data of pure aortic specimen with certain FDR control, which can help to validate novel marker proteins from completely unsupervised analysis of mixtures by debCAM.

5.1.3 Sample-specific deconvolution

We propose a sample-specific deconvolution algorithm – debCAM2.0, to estimate simple-specific molecule expressions for each subtype, from which between-sample variation can be used to detect biological associations and construct networks in each subtype. The contributions of this work include:

- We formulate the objective function for debCAM2.0 with a penalty term to minimize the nuclear norm of between-sample variation matrix in each subtype, based on our expectation on the existence of subtype-specific networks.
- We design an efficient method based on ADMM to solve debCAM2.0's optimization problem in large-scale biological data.
- We design a 10-fold cross-validation strategy to select the coefficient of nuclear norm term, and demonstrate its feasibility in simulations where a U-curve of RMSE is obtained to determine the optimal selection.

- We validate debCAM2.0 in simulations to demonstrate sample-specific signals can be well estimated when low-rank assumption holds. Even though artificial signal variances exist in debCAM2.0 estimations, the intercorrelations among genes can still be well preserved for function module detection and biological network construction.
- We propose to use extra $\ell_{2,1}$ norm regularization to enforce the sparsity of genes involved in networks and thus reduce the artifacts trained from noise or from signals of other subtypes.

5.2 Future work

Based on the work done in Chapter 2-4, we outline several potential research directions that may deserve to be further explored to extend or refine the current methods for molecular characterization of tissue heterogeneity.

5.2.1 Unsupervised deconvolution

As discussed in Chapter 2, debCAM framework still face many challenges to dissect complex tissues in a fully unsupervised way. Here are some promising directions:

1. Develop more effective methods to incorporate *a priori* knowledge into the debCAM framework. We have proposed to combine *a priori* markers and blindly detected markers together to balance prior knowledge and unsupervised discovery. The new marker sets will be used to re-estimate the mixing proportions and subtype-specific expression profiles. The advantage of this semi-supervised method is that the unsupervised learning process will not be affected by *a priori* markers of likely low quality. However, if *a priori* markers are of high quality, the optimal simplex identification could be improved by the supervision of *a priori* markers. Therefore, in the future, we can develop a semi-supervised simplex identification method, which will incorporate suitable prior markers into the earlier stage of debCAM framework. Before that, we also need to design an evaluation method to select *a priori* markers that are of good quality and match the specific condition of mixtures.
2. Improve debCAM by analyzing convex sets in both sample space and scatter space. The current debCAM focuses on detecting molecule markers in scatter space, which needs the existence of well-grounded points (strictly defined markers). Some reports detect pure

samples in sample space, which needs the presence of pure samples. In real data, markers are more likely to be present than pure samples, therefore making debCAM more widely applicable than sample-space-based supervised deconvolution methods. However, if the existent markers are also not strict enough, the estimation of proportions and subtype-specific expressions will be less accurate. The information from the convex set in sample space could be a supplementary to that in scatter space. Inferring the location of vertices from two imperfect convex sets together might achieve better results than the inference only in one space.

5.2.2 Robust detection of subtype-specific markers

As discussed in Chapter 3, OVESEG-test could be further improved in RNASeq data analysis and in terms of detection power:

1. Improve OVESEG-test for RNASeq data analysis by extending state-of-the-art two-group test methods designed specifically for RNASeq or single-cell RNASeq data. The current OVESEG-test follows “limma-voom” to model the RNASeq read count data by appropriate observational-level weights to accommodate unequal variances. However, edgeR and DESeq2 are more popular methods in RNASeq data analysis, as both use more sophisticated models for read counts. For single-cell RNASeq data, many advanced methods have been reported to address outliers and drop-out zero-value issues. To achieve better marker detections from purified RNASeq profiles or single-cell RNASeq profiles, we can re-design OVESEG-test statistic following the statistic used in edgeR, or DESeq2, or any other two-group test methods, where the mixture null distribution can still be estimated using our novel permutation scheme.
2. Combine OVE-FC and OVESEG-test. OVESEG-test is an extension of OVE-FC to assess the significance level of subtype-specific markers. By considering both fold change and variance, OVESEG-test also achieves better detection power in the more challenging cases involving non-significantly large fold change or phenotypically closer cell types. However, OVE-FC still works better for cases with sufficiently large fold change. We can set both OVE-FC threshold and OVESEG-test p-value threshold for the reliable detection of

markers. However, combining two statistics to be one powerful statistic would be more appealing in the future.

5.2.3 Sample-specific deconvolution

As discussed in Chapter 4, debCAM2.0 can solve a seemingly underdetermined problem theoretically based on a low-rank assumption. It still needs improvement and validations in real data analysis:

1. Improve debCAM2.0 by sparsity regularization. The sparsity assumption is practically reasonable, and we already show some preliminary results after imposing $\ell_{2,1}$ norm regularization. However, introducing one more regularization term will increase the difficulty of parameter tuning. Besides, the current cross-validation strategy with matrix entry sampling is not applicable to selecting the coefficient of $\ell_{2,1}$ norm term. Therefore, the integration of sparsity regularization still needs our further study.
2. Improve function module detection based on debCAM2.0 estimated sample-specific signals in each subtype. Recovering the exact values of sample-specific signals is impossible unless there are more strong assumptions. Luckily, our goal is to detect function module or networks from the between-sample variations in each subtype. So increasing the accuracy of estimated intercorrelations among molecules can be regarded as our target of further efforts.
3. Validate debCAM2.0 in real data analysis. We have demonstrated the capacity of debCAM2.0 to estimate sample-specific signals in each subtype using simulations where the between-sample variation matrices are low-rank. Validation of debCAM2.0 in real molecule expression data would be difficult, as the benchmark datasets with true subtype-specific signals are unavailable. One possible direction is to verify the constructed subtype-specific networks through biological experiments.

Appendix A

Personal Information

A.1 Biographical sketch

Lulu Chen received the B.S. degree in electronic information engineering and the M.S. degree in Signal and Information Processing from the University of Science and Technology of China, Hefei, China, in 2010 and 2013, respectively. Since August 2013, she has been pursuing a Ph.D. degree in computer engineering with the Bradley Department of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University (Virginia Tech), Virginia, United States, under the supervision of Dr. Yue Wang. Her research focuses on applications of machine learning and signal processing to computational biology.

A.2 Publications

Lulu Chen, Niya Wang, David M. Herrington, Robert Clarke, Chiung-Ting Wu, Yue Wang. "debCAM: a Bioconductor R package for fully unsupervised deconvolution of complex tissues," ***Bioinformatics*** application note. (Under revision)

Lulu Chen, David Herrington, Robert Clarke, Guoqiang Yu, Yue Wang, "Data-driven robust detection of tissue/cell-specific markers," *bioRxiv*, 2019.

Lulu Chen, Niya Wang, Chunyu Liu, David Herrington, Robert Clarke, Zhen Zhang, Yue Wang, "Unsupervised deconvolution of molecular heterogeneity uncovers novel signatures and glia-neuron ratio," *bioRxiv*, 2018.

David Herrington, Chunhong Mao, Sarah Parker, Zongming Fu, Guoqiang Yu, **Lulu Chen**, Vidya Venkatraman, Yi Fu, Yizhi Wang, Tim Howard, Jun Goo, Caroline Zhao, Yongming Liu, Georgia Saylor, Grace Athas, Dana Troxclair, James Hixson, Richard Vander Heide, Yue Wang, Jennifer Van Eyk. "Proteomic architecture of human coronary and aortic atherosclerosis," ***Circulation***, 2018. (Impact Factor 23.054)

Niya Wang, **Lulu Chen**, and Yue Wang, "Mathematical modelling and deconvolution of molecular heterogeneity identifies novel subpopulations in complex tissues", Book Chapter, ***Transcriptome Data Analysis: Methods and Protocols***, Springer, 2018.

Chia-Hsiang Lin, Chong-Yung Chi**, Lulu Chen**, David J. Miller, and Yue Wang,"Detection of Sources in Non-negative Blind Source Separation by Minimum Description Length Criterion," ***IEEE Trans Neural Networks and Learning Systems***, 2017.

Niya Wang, Eric P. Hoffman, **Lulu Chen**, Li Chen, Zhen Zhang, Chunyu Liu, Guoqiang Yu, David M. Herrington, Robert Clarke, and Yue Wang, "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues," ***Scientific Reports***, 6:18909, 2016.

Niya Wang, Ting Gong, Robert Clarke, **Lulu Chen**, Ie-Ming Shih, Zhen Zhang, Douglas A. Levine, Jianhua Xuan and Yue Wang, "UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples," ***Bioinformatics***, vol. 31, pp. 137-139, 2015.

Bibliography

[1] E. P. Hoffman *et al.*, "Expression profiling - best practices for data generation and interpretation in clinical trials," *Nature Reviews Genetics,* vol. 5, no. 3, pp. 229-237, 2004.

[2] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology,* journal article vol. 18, no. 1, p. 83, May 05 2017.

[3] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?," (in eng), *Nat Rev Cancer,* vol. 12, no. 5, pp. 323-34, Apr 19 2012.

[4] P. Lu, A. Nakorchevskiy, and E. M. Marcotte, "Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations," (in eng), *Proc Natl Acad Sci U S A,* vol. 100, no. 18, pp. 10370-5, Sep 2 2003.

[5] S. S. Shen-Orr *et al.*, "Cell type-specific gene expression differences in complex tissues," *Nat Methods,* vol. 7, no. 4, pp. 287-9, Apr 2010.

[6] A. Kuhn, D. Thu, H. J. Waldvogel, R. L. Faull, and R. Luthi-Carter, "Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain," (in eng), *Nat Methods,* vol. 8, no. 11, pp. 945-7, Nov 2011.

[7] A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H. F. Clark, "Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus," *PLoS ONE,* vol. 4, no. 7, p. e6098, 2009.

[8] L. A. Herzenberg, D. Parks, B. Sahaf, O. Perez, M. Roederer, and L. A. Herzenberg, "The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford," *Clinical Chemistry,* vol. 48, no. 10, pp. 1819-1827, 2002.

[9] E. Kummari, S. X. Guo-Ross, and J. B. Eells, "Laser Capture Microdissection - A Demonstration of the Isolation of Individual Dopamine Neurons and the Entire Ventral Tegmental Area," *Journal of Visualized Experiments : JoVE,* no. 96, p. 52336, 02/06 2015.

[10] R. O. Stuart *et al.*, "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, no. 2, pp. 615-620, 11/12/received 2004.

[11] A. M. Newman *et al.*, "Robust enumeration of cell subsets from tissue expression profiles," *Nat Methods,* Article vol. 12, no. 5, pp. 453-457, 05//print 2015.

[12] S. Mohammadi, N. Zuckerman, A. Goldsmith, and A. Grama, "A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues," *Proceedings of the IEEE,* vol. 105, no. 2, pp. 340-366, 2017.

[13] N. S. Zuckerman, Y. Noam, A. J. Goldsmith, and P. P. Lee, "A self-directed method for cell-type identification and separation of gene expression microarrays," (in eng), *PLoS Comput Biol,* vol. 9, no. 8, p. e1003189, 2013.

[14] E. Oja and M. Plumbley, "Blind separation of positive sources by globally convergent gradient search," (in eng), *Neural Comput,* vol. 16, no. 9, pp. 1811-25, Sep 2004.

[15] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, p. 788, 10/21/online 1999.

[16] N. Wang *et al.*, "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues," Article vol. 6, p. 18909, 01/07/online 2016.

[17] G. Yu *et al.*, "Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases," *J. Mach. Learn. Res.*, vol. 11, pp. 2141-2167, 2010.

[18] J. Ahn *et al.*, "DeMix: deconvolution for mixed cancer transcriptomes using raw measured data," *Bioinformatics,* vol. 29, no. 15, pp. 1865-71, Aug 1 2013.

[19] P. Creixell, E. M. Schoof, J. T. Erler, and R. Linding, "Navigating cancer network attractors for tumor-specific therapy," *Nature Biotechnology,* vol. 30, p. 842, 09/10/online 2012.

[20] T. H. Chan, W. K. Ma, C. Y. Chi, and Y. Wang, "A Convex Analysis Framework for Blind Separation of Non-Negative Sources," *IEEE Transactions on Signal Processing,* vol. 56, no. 10, pp. 5120-5134, 2008.

[21] Y. Zhu, N. Wang, D. J. Miller, and Y. Wang, "Convex Analysis of Mixtures for Separating Non-negative Well-grounded Sources," *Scientific Reports,* Article vol. 6, p. 38350, 12/06/online 2016.

[22] N. Wang *et al.*, "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues," *Scientific Reports,* Article vol. 6, p. 18909, 2016.

[23] Y. Hart *et al.*, "Inferring biological tasks using Pareto analysis of high-dimensional data," *Nature Methods,* vol. 12, p. 233, 01/26/online 2015.

[24] R. Schwartz and S. E. Shackney, "Applying unmixing to gene expression data for tumor phylogeny inference," *BMC Bioinformatics,* journal article vol. 11, no. 1, p. 42, January 20 2010.

[25] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Trans Signal Processing,* vol. 56, no. 10, pp. 5120-5134, 2008.

[26] F. Avila Cobos, J. Vandesompele, P. Mestdagh, and K. De Preter, "Computational deconvolution of transcriptomics data from mixed cell populations," (in eng), *Bioinformatics,* vol. 34, no. 11, pp. 1969-1979, Jun 1 2018.

[27] D. M. Herrington *et al.*, "Proteomic Architecture of Human Coronary and Aortic Atherosclerosis," (in eng), *Circulation,* vol. 137, no. 25, pp. 2741-2756, 2018.

[28] L. Chen, P. L. Choyke, T. H. Chan, C. Y. Chi, G. Wang, and Y. Wang, "Tissue-specific compartmental analysis for dynamic contrast-enhanced MR imaging of complex tumors," (in eng), *IEEE Trans Med Imaging,* vol. 30, no. 12, pp. 2044-58, Dec 2011.

[29] Y. Hart *et al.*, "Inferring biological tasks using Pareto analysis of high-dimensional data," *Nat Meth,* Brief Communication vol. 12, no. 3, pp. 233-235, 03//print 2015.

[30] R. Schwartz and S. Shackney, "Applying unmixing to gene expression data for tumor phylogeny inference," *BMC Bioinformatics,* vol. 11, no. 1, p. 42, 2010.

[31] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *ACM sigmod record,* 2000, vol. 29, no. 2, pp. 93-104: ACM.

[32] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science,* vol. 315, no. 5814, pp. 972-976, 2007.

[33] C. B. Barber, D. P. Dobkin, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software,* vol. 22, no. 4, pp. 469-483, 1996.

[34] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters,* vol. 15, no. 11, pp. 1119-1125, 1994.

[35] N. Wang *et al.*, "UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples," *Bioinformatics,* vol. 31, no. 1, pp. 137-9, Jan 1 2015.

[36] N. Wang *et al.*, "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues," *Scientific Reports,* vol. 5, p. 18909, 2015.

[37] C. Li, W. E. Johnson, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics,* vol. 8, no. 1, pp. 118-127, 2006.

[38] C. Croux, P. Filzmoser, and M. R. Oliveira, "Algorithms for Projection–Pursuit robust principal component analysis," *Chemometrics and Intelligent Laboratory Systems,* vol. 87, no. 2, pp. 218-225, 2007/06/15/ 2007.

[39] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 33, no. 2, pp. 387-392, 1985.

[40] E. Becht *et al.*, "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression," *Genome biology,* vol. 17, no. 1, pp. 218-218, 2016.

[41] H. J. Kang *et al.*, "Spatio-temporal transcriptome of the human brain," *Nature,* 10.1038/nature10523 vol. 478, no. 7370, pp. 483-489, 10/27/print 2011.

[42] C. Colantuoni *et al.*, "Temporal dynamics and genetic control of transcription in the human prefrontal cortex," *Nature,* 10.1038/nature10524 vol. 478, no. 7370, pp. 519-523, 10/27/print 2011.

[43] X. Xu, A. Nehorai, and J. D. Dougherty, "Cell type-specific analysis of human brain transcriptome data to predict alterations in cellular composition," *Systems Biomedicine,* vol. 1, no. 3, pp. 0--1, 07/01 2013.

[44] J. Stiles and T. L. Jernigan, "The Basics of Brain Development," *Neuropsychology Review,* vol. 20, no. 4, pp. 327-348, 2010.

[45] J. Guintivano, M. J. Aryee, and Z. A. Kaminsky, "A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression," (in eng), *Epigenetics,* vol. 8, no. 3, pp. 290-302, Mar 2013.

[46] C. Jiao *et al.*, "Positional effects revealed in Illumina Methylation Array and the impact on analysis," *Epigenomics,* vol. 10, no. 5, pp. 643-659, 2018.

[47] A. E. Jaffe *et al.*, "Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex," (in eng), *Nat Neurosci,* vol. 19, no. 1, pp. 40-7, Jan 2016.

[48] L. Chen, D. Herrington, R. Clarke, G. Yu, and Y. Wang, "Data-driven robust detection of tissue/cell-specific markers," *bioRxiv,* p. 517961, 2019.

[49] B. O. Mancarci *et al.*, "Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data," (in eng), *eNeuro,* vol. 4, no. 6, Nov-Dec 2017.

[50] F. A. Azevedo *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *J Comp Neurol,* vol. 513, no. 5, pp. 532-41, Apr 10 2009.

[51] F. A. Azevedo *et al.*, "Automatic isotropic fractionation for large-scale quantitative cell analysis of nervous tissue," *J Neurosci Methods,* vol. 212, no. 1, pp. 72-8, Jan 15 2013.

[52] S. J. Parker *et al.*, "Identification of Putative Fibrous Plaque Marker Proteins by Unsupervised Deconvolution of Heterogeneous Vascular Proteomes," vol. 136, no. suppl_1, pp. A17297-A17297, 2017.

[53] Y. Hart *et al.*, "Inferring biological tasks using Pareto analysis of high-dimensional data," *Nat Methods,* vol. 12, no. 3, pp. 233-5, Mar 2015.

[54] A. M. Newman *et al.*, "Robust enumeration of cell subsets from tissue expression profiles," *Nat Methods,* vol. 12, no. 5, pp. 453-7, May 2015.

[55] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative normalization of cDNA microarray data," *IEEE Trans Inf Technol Biomed,* vol. 6, no. 1, pp. 29-37, Mar 2002.

[56] C. Montano *et al.*, "Measuring cell-type specific differential methylation in human brain tissue," *Genome Biology,* vol. 14, no. 8, p. R94, 2013.

[57] W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, and P. W. Zandstra, "PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions," (in eng), *PLoS Comput Biol,* vol. 8, no. 12, p. e1002838, 2012.

[58] M. Chikina, E. Zaslavsky, and S. C. Sealfon, "CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations," *Bioinformatics,* vol. 31, no. 10, pp. 1584-1591, 2015.

[59] Y. Zhang *et al.*, "An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex," (in eng), *J Neurosci,* vol. 34, no. 36, pp. 11929-47, Sep 3 2014.

[60] J. E. Shoemaker, T. J. Lopes, S. Ghosh, Y. Matsuoka, Y. Kawaoka, and H. Kitano, "CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data," *BMC Genomics,* journal article vol. 13, no. 1, p. 460, September 06 2012.

[61] Z. Chen, A. Huang, J. Sun, T. Jiang, F. X. Qin, and A. Wu, "Inference of immune cell composition on the expression profiles of mouse tissue," (in eng), *Sci Rep,* vol. 7, p. 40508, Jan 13 2017.

[62] G. Yu *et al.*, "PUGSVM: a caBIGTM analytical tool for multiclass gene selection and predictive classification," *Bioinformatics,* vol. 27, no. 5, pp. 736-738, 2011.

[63] K. E. Amrani, H. Stachelscheid, F. Lekschas, A. Kurtz, and M. A. Andrade-Navarro, "MGFM: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data," *BMC Genomics,* vol. 16, no. 1, p. 645, 2015.

[64] J. C. Hsu, *Multiple comparisons : theory and methods*. London: Chapman & Hall, 1996.

[65] M. Wang, S. R. Master, and L. A. Chodosh, "Computational expression deconvolution in a complex mammalian organ," *BMC Bioinformatics,* vol. 7, pp. 328-328, 2006.

[66] X. Guo and W. Pan, "Using weighted permutation scores to detect differential gene expression with microarray data," *Journal of Bioinformatics and Computational Biology,* vol. 03, no. 04, pp. 989-1006, 2005.

[67] K. Strimmer, "fdrtool: a versatile R package for estimating local and tail area-based false discovery rates," *Bioinformatics,* vol. 24, no. 12, pp. 1461-1462, 2008.

[68] E. Kulinskaya, "On two-sided p-values for non-symmetric distributions," *arXiv preprint arXiv:0810.2124,* 2008.

[69] Y.-H. Zhou and F. A. Wright, "Hypothesis testing at the extremes: fast and robust association for high-throughput data," *Biostatistics,* vol. 16, no. 3, pp. 611-625, 2015.

[70] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," (in eng), *Stat Appl Genet Mol Biol,* vol. 3, p. Article3, 2004.

[71] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biology,* journal article vol. 15, no. 2, p. R29, February 03 2014.

[72] T. Suomi, F. Seyednasrollah, M. K. Jaakkola, T. Faux, and L. L. Elo, "ROTS: An R package for reproducibility-optimized statistical testing," *PLOS Computational Biology,* vol. 13, no. 5, p. e1005562, 2017.

[73] D. K. McClish, "Analyzing a portion of the ROC curve," (in eng), *Med Decis Making,* vol. 9, no. 3, pp. 190-5, Jul-Sep 1989.

[74] F. Allantaz *et al.*, "Expression Profiling of Human Immune Cell Subsets Identifies miRNA-mRNA Regulatory Relationships Correlated with Cell Type Specific Expression," *PLoS ONE,* vol. 7, no. 1, p. e29979, 2012.

[75] M. Schelker *et al.*, "Estimation of immune cell content in tumour tissue using single-cell RNA-seq data," (in eng), *Nat Commun,* vol. 8, no. 1, p. 2032, 2017.

[76] A. Kuhn, A. Kumar, A. Beilina, A. Dillman, M. Cookson, and A. Singleton, "Cell population-specific expression analysis of human cerebellum," *BMC Genomics,* vol. 13, no. 1, p. 610, 2012.

[77] E. Becht *et al.*, "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression," *Genome Biology,* vol. 17, p. 218, 2016.

[78] D. Aran, Z. Hu, and A. J. Butte, "xCell: digitally portraying the tissue cellular heterogeneity landscape," *Genome Biology,* vol. 18, p. 220, 2017.

[79] F. Finotello and Z. Trajanoski, "Quantifying tumor-infiltrating immune cells from transcriptomics data," *Cancer Immunology, Immunotherapy,* vol. 67, no. 7, pp. 1031-1040, 2018.

[80] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," (in eng), *Nucleic acids research,* vol. 40, no. 10, pp. 4288-4297, 2012.

[81] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," (in eng), *Genome biology,* vol. 15, no. 12, pp. 550-550, 2014.

[82] L. S. Kao and C. E. Green, "Analysis of Variance: Is There a Difference in Means and What Does It Mean?," *The Journal of surgical research,* vol. 144, no. 1, pp. 158-170, 10/22 2008.

[83] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," (in eng), *Stat Appl Genet Mol Biol,* vol. 4, p. Article17, 2005.

[84] B. Zhang *et al.*, "Differential dependency network analysis to identify condition-specific topological changes in biological networks," (in eng), *Bioinformatics,* vol. 25, no. 4, pp. 526-32, Feb 15 2009.

[85] Y. Tian *et al.*, "Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks," *BMC Systems Biology,* journal article vol. 8, no. 1, p. 87, July 24 2014.

[86] D. Chasman and S. Roy, "Inference of cell type specific regulatory networks on mammalian lineages," *Current Opinion in Systems Biology,* vol. 2, no. Supplement C, pp. 130-139, 2017/04/01/ 2017.

[87] E. Gal *et al.*, "Rich cell-type-specific network topology in neocortical microcircuitry," *Nature Neuroscience,* Article vol. 20, p. 1004, 06/05/online 2017.

[88] A. R. Sonawane *et al.*, "Understanding Tissue-Specific Gene Regulation," *Cell Reports,* vol. 21, no. 4, pp. 1077-1088.

[89] M. R. Junttila and F. J. de Sauvage, "Influence of tumour micro-environment heterogeneity on therapeutic response," *Nature,* vol. 501, p. 346, 09/18/online 2013.

[90] F. Buettner *et al.*, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," (in eng), *Nat Biotechnol,* vol. 33, no. 2, pp. 155-60, Feb 2015.

[91] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer New York Inc., 2001.

[92] P. A. Thouvenin, N. Dobigeon, and J. Y. Tourneret, "Hyperspectral Unmixing With Spectral Variability Using a Perturbed Linear Mixing Model," *IEEE Transactions on Signal Processing,* vol. 64, no. 2, pp. 525-538, 2016.

[93] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization," *SIAM Review,* vol. 52, no. 3, pp. 471-501, 2010.

[94] E. J. Candes, C. A. Sing-Long, and J. D. Trzasko, "Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators," *Trans. Sig. Proc.,* vol. 61, no. 19, pp. 4643-4657, 2013.

[95] J.-F. Cai, E. J. Candès, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM Journal on Optimization,* vol. 20, no. 4, pp. 1956-1982, 2010.

[96] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.,* vol. 3, no. 1, pp. 1-122, 2011.

[97] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends® in Optimization,* vol. 1, no. 3, pp. 127-239, 2014.

[98] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," (in eng), *BMC Bioinformatics,* vol. 9, p. 559, Dec 29 2008.

[99] Y. Tian *et al.*, "KDDN: an open-source Cytoscape app for constructing differential dependency networks with significant rewiring," (in eng), *Bioinformatics,* vol. 31, no. 2, pp. 287-9, Jan 15 2015.

[100] B. Zhang *et al.*, "DDN: a caBIG(R) analytical tool for differential network analysis," (in eng), *Bioinformatics,* vol. 27, no. 7, pp. 1036-8, Apr 1 2011.