Accurate estimation of rare cell type fractions from tissue omics data via hierarchical deconvolution

Penghui Huang[1], Manqi Cai[1], Xinghua Lu[2], Chris McKennan[3], and Jiebiao Wang[1,*]

[1]Deparment of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA
[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
[3]Deparment of Statistics, University of Pittsburgh, Pittsburgh, PA, USA
[*]Correspondence: jbwang@pitt.edu

Abstract

Bulk transcriptomics in tissue samples reflects the average expression levels across different cell types and is highly influenced by cellular fractions. As such, it is critical to estimate cellular fractions to both deconfound differential expression analyses and infer cell type-specific differential expression. Since experimentally counting cells is infeasible in most tissues and studies, *in silico* cellular deconvolution methods have been developed as an alternative. However, existing methods are designed for tissues consisting of clearly distinguishable cell types and have difficulties estimating highly correlated or rare cell types. To address this challenge, we propose Hierarchical Deconvolution (HiDecon) that uses single-cell RNA sequencing references and a hierarchical cell type tree, which models the similarities among cell types and cell differentiation relationships, to estimate cellular fractions in bulk data. By coordinating cell fractions across layers of the hierarchical tree, cellular fraction information is passed up and down the tree, which helps correct estimation biases by pooling information across related cell types. The flexible hierarchical tree structure also enables estimating rare cell fractions by splitting the tree to higher resolutions. Through simulations and real data applications with the ground truth of measured cellular fractions, we demonstrate that HiDecon significantly outperforms existing methods and accurately estimates cellular fractions.

Keywords: Cellular deconvolution, Single-cell data, RNA sequencing, Hierarchical tree, Penalized regression

1 Introduction

Tissue-level gene expression data quantify the average expression across cell types, which is largely affected by the heterogeneity of cell type proportions. The varying cellular fractions

across tissue samples can confound tissue-level analyses (Jaffe et al., 2014), potentiate the estimation of cell type-specific differential expression (J. Wang et al., 2021), and can help understand the etiology of disease (Mostafavi et al., 2018). Although biochemical methods like flow cytometry and immunohistochemistry can measure a tissue sample's cell counts, they are labor-intensive and costly. Thus, *in silico* cellular deconvolution methods have been developed to estimate cellular fractions from bulk tissue data as a cost-effective alternative.

Existing deconvolution methods can be grouped into three categories: unsupervised, semi-supervised, and supervised. Unsupervised deconvolution methods do not require a reference and mimic factor analyses, but the resulting factors are usually hard to annotate and interpret (Avila Cobos et al., 2020). Semi-supervised deconvolution methods depend on marker genes that are only expressed in certain cell types and are used to infer cell type references solely using bulk data (Zhong et al., 2013).With the development of sorted-cell or single-cell RNA sequencing (scRNA-seq) reference data, supervised deconvolution has become a powerful alternative and has precipitated the development of methods that leverage these references to better estimate cell type proportions (Newman et al., 2015; Hunt et al., 2019; X. Wang et al., 2019; Wilson et al., 2020).

As the number of cell clusters obtained from numerous single-cell studies increases, cell type hierarchy has become important for understanding the topology of cell types across datasets (Wu et al., 2020; Peng et al., 2021; L. Chen et al., 2022). Large efforts have been made to enhance the interpretation of cell types, such as cell ontology (Miller et al., 2020) and hierarchically organized cell types (Hodge et al., 2019). In practice, tissues with various differentiated cell types that share the same origin of cell differentiation, like peripheral blood mononuclear cells (PBMCs), bring great difficulties to reference-based deconvolution methods because cell types from the same origin have similar expression levels. This begets co-linearity in cellular deconvolution regression models, which results in highly variable estimates. Some methods attempt to alleviate this issue by selecting better cell type marker genes to reduce the correlation between the reference gene expression levels in different cell types. However, these methods give highly biased results when the quality of cell subtype marker genes is not high (Fischer et al., 2021). In addition, most deconvolution methods work better for more abundant cell types and thus are limited to applications of major cell types. Their estimated fractions of rare cell types are often zero, which precludes downstream analyses of rare but biologically important cell types.

To address the issue of co-linearity and rare types, existing methods HEpiDISH (Zheng et al., 2018) and MuSiC (X. Wang et al., 2019) implemented a top-down recursive deconvolution process guided by a hierarchical cell-type tree. After estimating cell fractions of major cell types in the first layer, they calculate artificial omics of major cell types and use it as the response in reference-based deconvolution to estimate subtype fractions in the second layer, assuming the fraction of a major cell type equals the sum of its subtype fractions. Although this process can be extended to hierarchical trees with more than two layers, the top-down approach may fail when the "parent" cell types are poorly estimated. Importantly, the bias of each layer's estimation will accumulate and increase the estimation bias of cellular fractions in subsequent layers. Therefore, it is pressing to develop methods that can utilize more complicated tree structures to provide accurate estimates of cellular fractions.

Here we present Hierarchical Deconvolution (HiDecon), a penalized approach with constraints from both "parent" and "children" cell types to make full use of a hierarchical tree

structure. The hierarchical tree is readily available from well-studied cell lineages or can be learned from hierarchical clustering of scRNA-seq data (Peng et al., 2021). The tree reflects the similarities between cell types and their differential trajectories. The basic intuition of HiDecon is that there exists a summation relationship between the estimation results of adjacent layers. For instance, after estimating cellular fractions at different resolutions with two deconvolution layers, say lymphocytes (layer 1) and B cells and T cells (layer 2), it will be ideal if the estimated proportion of lymphocytes is the sum of B cell and T cell proportions. If these layers' estimates do not follow the sum constraints implied by the hierarchical tree, it suggests that estimation bias occurs in certain layers and should be corrected by the estimation results of other layers. To fully use the cell type hierarchy and marker information in different layers, HiDecon implements the sum constraint penalties from the upper and lower layers to aggregate estimates across layers for more accurate cellular fraction estimates. This is especially useful for rare cell types, which may be poorly estimated in other methods that do not use the hierarchical tree.

The remainder of the manuscript is organized as follows. We first introduce our model and estimation algorithm in Section 2. Then in Section 3, we compare HiDecon with existing methods via simulations based on a real scRNA-seq dataset of PBMC from COVID-19 patients and controls. In Section 4, we further benchmark HiDecon in large-scale human blood datasets with measured cell counts. We conclude and summarize our findings in Section 5.

2 Methods

2.1 A model for deconvolving bulk gene expression

Gene expression levels in bulk tissue samples can be modeled as the weighted sum of cell type-specific expression. A reference-based cellular deconvolution model can be written as

$$\boldsymbol{X} = \boldsymbol{ASP} + \boldsymbol{E}, \tag{2.1}$$

where $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{m \times n}$ denotes the gene expression levels of m marker genes in n tissue samples; $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{m \times K}$ represents the observed reference signature matrix of K cell types derived from scRNA-seq or sorted-cell data; $\boldsymbol{S} \in \mathbb{R}_{\geq 0}^{K \times K}$ is an observed diagonal matrix with diagonal elements representing the cell size of different cell types (Jia et al., 2017), that is, the average abundance of observed transcripts in each cell type; $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{K \times n}$ is the cellular fractions for the K cell types that need to be estimated; and $\boldsymbol{E} \in \mathbb{R}^{m \times n}$ is the error term. Most existing deconvolution algorithms rely on a pre-specified fixed number of cell types (K), but cell types are usually hierarchically organized by cell differentiation or lineage. It is critical to estimate cellular fractions at different resolutions and improve the estimation by borrowing information from related "parent" and "children" cell types. We term the precise cellular deconvolution guided by a hierarchical cell-type tree as hierarchical deconvolution and describe it in detail below.

2.2 Hierarchical deconvolution

Assume a known hierarchical tree (either from biology literature or hierarchical clustering) with L layers such that cells are split into finer resolutions as $l \in \{1, 2, \ldots, L\}$ increases, where $l = 0$ denotes all cells as a single cluster and K_l is the number of cell types in layer l. We first describe how to map cell types across layers l and $l+1$ using a simple example from PBMCs. Assume layer l has two clusters representing monocytes and lymphocytes and layer $l+1$ has three cell types consisting of monocytes, B cells, and T cells, where lymphocytes in layer l are divided into B cells and T cells in layer $l+1$. Let $\boldsymbol{p}_{il}$ and $\boldsymbol{p}_{i(l+1)}$ be sample i's cellular fractions in layers l and $l+1$. Since the fraction of lymphocytes should be similar to the sum of fractions of B cells and T cells, we should have

$$\boldsymbol{p}_{il} \approx \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \boldsymbol{p}_{i(l+1)} = \boldsymbol{B}_{l,(l+1)} \boldsymbol{p}_{i(l+1)}.$$

The matrices $\boldsymbol{B}_{l,(l+1)} \in \{0,1\}^{K_l \times K_{l+1}}$ define a mapping between cell types across adjacent layers and parameterize the hierarchical tree, where row k's non-zero elements are exactly cell type k's "children" in layer $l+1$.

We define an estimator for sample i's cellular fractions from all layers $(\boldsymbol{p}_{i1}, \ldots, \boldsymbol{p}_{iL})$ to be

$$\text{argmin}_{(\boldsymbol{p}_{i1},\ldots,\boldsymbol{p}_{iL})} \left\{ \frac{1}{2} \sum_{l=1}^{L} \frac{1}{m_l} \|\boldsymbol{x}_{il} - \boldsymbol{A}_l \boldsymbol{S}_l \boldsymbol{p}_{il}\|_2^2 \right\}, \tag{2.2}$$

subject to

$$\sum_{l=1}^{L-1} \frac{\|\boldsymbol{p}_{il} - \boldsymbol{B}_{l,(l+1)} \boldsymbol{p}_{i(l+1)}\|_2^2}{K_l} \leq \zeta, \quad \boldsymbol{p}_{i1}, \ldots, \boldsymbol{p}_{iL} \geq \boldsymbol{0}, \quad \text{and} \quad \|\boldsymbol{p}_{iL}\|_1 = 1, \tag{2.3}$$

where $\boldsymbol{x}_{il} \in \mathbb{R}_{\geq 0}^{m_l}$ denotes sample i's bulk gene expression at m_l marker genes in layer l; $\boldsymbol{A}_l \in \mathbb{R}_{\geq 0}^{m_l \times K_l}$ represents the reference signature matrix of K_l cell types derived from scRNA-seq data; and $\boldsymbol{S}_l \in \mathbb{R}_{\geq 0}^{K_l \times K_l}$ is a diagonal size factor matrix for layer l with diagonal entries representing the cell size of the K_l cell types in layer l. Nonnegativity of all parts in the model originates from the nature of gene expression data.

The first constraint in (2.3) reflects the hierarchical tree, and ensures "parent" and "children" cellular fraction estimates are similar. To ensure the interpretation of proportional estimates, we further require the last layer's fractions to sum to one. The optimization in (2.2) and (2.3) is convex and can be re-written as the following penalized regression problem:

$$\underset{\substack{\boldsymbol{p}_{i1},\ldots,\boldsymbol{p}_{iL} \\ \boldsymbol{p}_{il} \geq 0,\ \|\boldsymbol{p}_{iL}\|_1 = 1}}{\text{argmin}} \left\{ \frac{1}{2} \sum_{l=1}^{L} \frac{1}{m_l} \|\boldsymbol{x}_{il} - \boldsymbol{A}_l \boldsymbol{S}_l \boldsymbol{p}_{il}\|_2^2 + \frac{\lambda}{2} \sum_{l=1}^{L-1} \frac{\|\boldsymbol{p}_{il} - \boldsymbol{B}_{l,(l+1)} \boldsymbol{p}_{i(l+1)}\|_2^2}{K_l} \right\}, \tag{2.4}$$

where the tuning parameter $\lambda \geq 0$ is implicitly a decreasing function of ζ. That is, $\lambda = 0$ implies the tree has no impact on estimates and $\lambda = \infty$ means parent cellular fractions are completely determined by their children's fractions.

2.3 Estimation algorithm

To simplify our algorithm and follow the common practice of cellular deconvolution (Mohammadi et al., 2016), we optimize the objective function under the nonnegative constraint and then normalize the fraction estimates for the Lth layer so they sum to 1. To describe our algorithm, we first note that for $\boldsymbol{p}_i = (\boldsymbol{p}_{i1}^\top, \ldots, \boldsymbol{p}_{iL}^\top)^\top$, (2.4) can be re-written as the following quadratic problem:

$$\underset{\boldsymbol{p}_i \in \mathbb{R}^K_{\geq 0}}{\operatorname{argmin}} f(\boldsymbol{p}_i), \quad f(\boldsymbol{p}_i) = \frac{1}{2}\|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{A}}\boldsymbol{p}_i\|_2^2 + \frac{\lambda}{2}\|\tilde{\boldsymbol{B}}\boldsymbol{p}_i\|_2^2, \tag{2.5}$$

where $K = \sum_{l=1}^L K_l$, $\tilde{\boldsymbol{x}}_i = (m_1^{-1/2}\boldsymbol{x}_{i1}^\top, \ldots, m_L^{-1/2}\boldsymbol{x}_{iL}^\top)^\top \in \mathbb{R}^m_{\geq 0}$, and $\tilde{\boldsymbol{A}} = \bigoplus_{l=1}^L (m_l^{-1/2}\boldsymbol{A}_l\boldsymbol{S}_l) \in \mathbb{R}^{m\times K}_{\geq 0}$ for $m = \sum_{l=1}^L m_l$. The matrix $\tilde{\boldsymbol{B}} \in \mathbb{R}^{(K-K_L)\times K}_{\geq 0}$ is an upper-triangular difference operator taking the form

$$\tilde{\boldsymbol{B}} = \left(\bigoplus_{l=1}^{L-1} K_l^{-1/2} I_{K_l}, \boldsymbol{0}_{(K-K_L)\times K_L}\right) - \left(\boldsymbol{0}_{(K-K_L)\times K_1}, \bigoplus_{l=1}^{L-1} K_l^{-1/2}\boldsymbol{B}_{l,(l+1)}\right).$$

To solve (2.5), we note that the minimizer for $\boldsymbol{p}_i$'s kth coordinate, while fixing all other coordinates, is exactly

$$\boldsymbol{p}_{ik}^{(\min)} = \max[0, \{\boldsymbol{b}_k - \boldsymbol{p}_{i(-k)}^\top \boldsymbol{H}_{k(-k)}\}/\boldsymbol{H}_{kk}], \quad \boldsymbol{b} = \tilde{\boldsymbol{A}}^\top\tilde{\boldsymbol{x}}, \quad \boldsymbol{H} = \tilde{\boldsymbol{A}}^\top\tilde{\boldsymbol{A}} + \lambda\tilde{\boldsymbol{B}}^\top\tilde{\boldsymbol{B}},$$

where $\boldsymbol{p}_{i(-k)}$ and $\boldsymbol{H}_{(-k)k}$ are the sub-vectors of $\boldsymbol{p}_i$ and the kth column of $\boldsymbol{H}$ obtained after deleting their kth elements. This naturally leads to Algorithm 1, which employs coordinate-wise descent to solve (2.5). While not explicitly stated in Algorithm 1, we normalize our estimate for $\boldsymbol{p}_{iL}$, the tree's last layer's cellular fractions, so that its entries sum to 1.

Data: $\boldsymbol{b} = \tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{x}}_i$, $\boldsymbol{H} = \tilde{\boldsymbol{A}}^\top\tilde{\boldsymbol{A}} + \lambda\tilde{\boldsymbol{B}}^\top\tilde{\boldsymbol{B}}$, and $\epsilon > 0$
Result: $\boldsymbol{p}_i$
Initialize $\boldsymbol{p}_i = \boldsymbol{H}^{-1}\boldsymbol{b}$;
if $\boldsymbol{p}_i \geq 0$ **then**
  return $\boldsymbol{p}_i$;
else
  $\boldsymbol{p}_{ik} = \max(0, \boldsymbol{p}_{ik})$, $k \in \{1, \cdots, K\}$;
  repeat
    $\boldsymbol{p}_{ik} = \max[0, \{\boldsymbol{b}_k - \boldsymbol{p}_{i(-k)}^\top \boldsymbol{H}_{k(-k)}\}/\boldsymbol{H}_{kk}]$, $k \in \{1, \ldots, K\}$;
  until $|(\boldsymbol{H}\boldsymbol{p}_i - \boldsymbol{b})_k| \leq \epsilon$ *OR* $\{(\boldsymbol{H}\boldsymbol{p}_i - \boldsymbol{b})_k \geq 0$ *AND* $\boldsymbol{p}_{ik} = 0\}$ *for all* $k \in \{1, \ldots, K\}$ *//Karush–Kuhn–Tucker (KKT) conditions*;
  return $\boldsymbol{p}_i$;
end

Algorithm 1: HiDecon optimization algorithm.

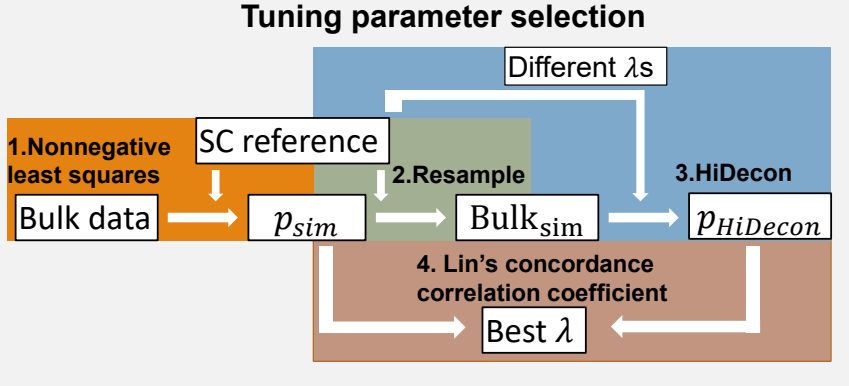

Figure 1: Flow chart for tuning parameter selection. Different steps are marked with different colors.

2.4 Tuning parameter selection

We propose a novel procedure (Figure 1) to select HiDecon's tuning parameter λ. The idea is to select the optimal λ using a bulk data surrogate with "ground truth" fractions. In order to generate a bulk data surrogate, we apply nonnegative least squares (NNLS) to the observed bulk data $\boldsymbol{X}_L \in \mathbb{R}_{\geq 0}^{m_L \times n}$ with given signature matrix $\boldsymbol{A}_L \in \mathbb{R}_{\geq 0}^{m_L \times K_L}$,

$$\boldsymbol{P}_{sim} = \underset{\boldsymbol{P} \geq 0}{\operatorname{argmin}} \|\boldsymbol{X}_L - \boldsymbol{A}_L \boldsymbol{P}\|_F \tag{2.6}$$

to get rough estimates for cellular fractions $\boldsymbol{P}_{sim}$ of bulk data to imitate the cellular composition structure of tissue samples. Then, we simulate bulk data surrogate $\boldsymbol{X}_{sim}$ with cellular fractions of $\boldsymbol{P}_{sim}$ by resampling individual cells from single-cell reference with replacement. Finally, we compare the performance of HiDecon when deconvolving bulk data surrogate $\boldsymbol{X}_{sim}$ under a series of tuning parameter λ's. We use Lin's concordance correlation coefficient (Lin, 1989) between HiDecon estimates $\boldsymbol{P}_{HiDecon}$ and the "ground truth" $\boldsymbol{P}_{sim}$ for each cell type as the evaluation metric. The λ with the highest mean concordance correlation coefficient across cell types is considered the optimal tuning parameter for HiDecon.

2.5 Data normalization and marker selection methods

We normalize the expression of each tissue sample and cell to the same scale and avoid extreme values. To adjust for library size, the expression of each tissue sample and cell in the raw count matrix is divided by its total count and multiplied by 1,000,000 as the count per million (CPM). Then, data is $\log_2$ transformed with a pseudo count of 1 to ensure all elements are nonnegative.

In HiDecon, we select marker genes for each layer of the hierarchical tree with the following considerations. First, marker genes selected for coarser clusters are not necessarily marker genes for finer clusters and cannot reduce co-linearity efficiently. Second, in single cell references, coarser clusters might have many cells from subtypes that might not be major in bulk data. Then, the reference mainly represents major cells in the single cell data. Selecting another set of markers for coarser layers can reduce the case that the averaged

reference misrepresents the expression level of coarser clusters in bulk data. In both the simulation study and real data application, we use the Wilcoxon rank-sum test to identify the differentially expressed (DE) genes between two cell clusters after normalizing the reference data as described above. To identify DE genes of one cluster compared with all other clusters in this layer, we use the intersection union test (Berger, 1997) to calculate the combined p-value for assigning rankings to genes. We use the top 50 genes with the smallest p-values for each cell cluster as its marker genes.

2.6 Evaluation metrics

Denote the ground truth of cell type fractions by $\boldsymbol{P}_{K\times n}$ and the estimated fractions by $\hat{\boldsymbol{P}}_{K\times n}$. We use two evaluation metrics to evaluate the performances of methods in our simulation study and real data application.

1. Mean absolute error:
$$\text{MAE}\left(\boldsymbol{P}, \hat{\boldsymbol{P}}\right) = avg\left(|\boldsymbol{P} - \hat{\boldsymbol{P}}|\right),$$
where avg of a matrix or a vector is the average over all its entries;

2. Lin's concordance correlation coefficient (Lin, 1989):
$$\text{CCC}\left(\boldsymbol{P}_{k*}, \hat{\boldsymbol{P}}_{k*}\right) = \frac{2cov\left(\boldsymbol{P}_{k*}, \hat{\boldsymbol{P}}_{k*}\right)}{\sigma^2_{\boldsymbol{P}_{k*}} + \sigma^2_{\hat{\boldsymbol{P}}_{k*}} + \left(avg\left(\boldsymbol{P}_{k*} - \hat{\boldsymbol{P}}_{k*}\right)\right)^2},$$
where $\boldsymbol{P}_{k*}$ denotes the row k of $\boldsymbol{P}$, that is, the kth cell type, and σ^2 denotes the variance. CCC $\in (-1, 1)$ is a comprehensive measure sensitive to variability and slope and intercept of the linearity. The concordance improves as the value of CCC approaches 1. It captures the deviation of estimates from ground truth, that is points' degree of departure from the line $y = x$.

3 Simulation studies

3.1 Simulation benchmarking with large PBMC scRNA-seq data

We first compared HiDecon with existing cellular deconvolution methods via simulations. In addition to the two existing top-down hierarchical deconvolution algorithms (HEpiDISH and MuSiC), we further compared HiDecon with other state-of-the-art deconvolution methods without considering the hierarchical cell tree, including CIBERSORT (Newman et al., 2015) and dtangle (Hunt et al., 2019). We used a real large-scale PBMC scRNA-seq dataset (Ren et al., 2021) to simulate pseudo-bulk data. It is a comprehensive COVID-19 study that contains scRNA-seq data from 27,943 genes of 284 samples, among which there are 28 controls, 122 mild/moderate, and 134 severe/critical samples. When generating bulk data, if there are not enough cells from some cell types, it will introduce large single cell specific variance to the gene expression contribution of this cell type in bulk data. In order to reduce cell specific variance when generating bulk data, we only used samples having at least 20 cells in each type.

a

Abbreviation	Full name
Mono	Monocytes
DC	Dendritic cells
B_1	B_c01-TCL1A
B_2	B_c02-MS4A1-CD27
B_3	B_c03-CD27-AIM2
NK_1	NK_c01-FCGR3A
CD4_1	T_CD4_c01-LEF1
CD8_1	T_CD8_c01-LEF1
CD8_5	T_CD8_c05-ZNF683
CD8_7	T_CD8_c07-TYROBP

b

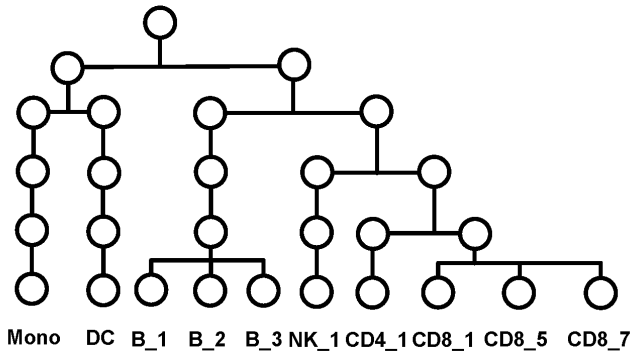

Figure 2: Cell type hierarchical relationship of COVID-19 PBMC data (Ren et al., 2021) used in the simulation. (a) Cell type abbreviation and full name reference. (b) Hierarchical tree constructed from cell lineage relationship and used to guide HiDecon.

Moreover, lymphocytes have complicated differentiation structures. We explored subtypes of lymphocyte cell types to evaluate HiDecon's performance on co-linearity and rare cell types. After filtering samples, we used 608,883 cells from 126 samples to calculate reference and simulate bulk data. Cell types include monocytes (Mono), dendritic cells (DC), B cells (B), natural killer cells (NK), CD4+ T cells (CD4), and CD8+ T cells (CD8). Lymphocyte subtypes (e.g., three subtypes of B cells and CD8+ T cells, respectively) are explicitly shown in Figure 2a, following the cell cluster names in the original paper.

We averaged the expression of all cells from each sample to generate simulated bulk data with known fractions, which are calculated by real cell counts of each sample in the scRNA-seq data. It is computationally intensive to select cell type-specific marker genes when the single cell reference matrix has numerous cells. Moreover, having too many cells from certain samples can cover up the information of some other under-represented samples. Considering these problems, first, we averaged single cell gene expression for each cell type to extract sample-level single cell data. This process was performed over all samples. Then, we pooled these sample level single cell data together to construct the pseudo single cell reference so that all samples are equally represented in this reference. The reference dimension is genes by the product of the number of cell types and the number of samples.

We used a hierarchy tree from biological cell lineage relationship (Figure 2b). Dendritic cells are most similar to monocytes. B cells, natural killer cells, CD4 cells, and CD8 cells are all lymphocytes. B cells are mostly naive or memory cells which sets them apart from natural killer cells, CD4 cells, and CD8 cells which are mostly active immune cells. Finally, CD4 cells and CD8 cells are both subtypes of T cells.

We used the tuning parameter selection method to find the best λ for HiDecon. With the ground truth of cellular fractions in the simulated dataset, we calculated CCC between the ground truth fraction $\boldsymbol{P}_{K\times n}$ and deconvolution estimated fraction $\hat{\boldsymbol{P}}_{K\times n}$ for each cell type across all samples to measure the concordance. We also calculated MAE to assess the accuracy of estimates (Table 1). HEpiDISH can only process the split of one cluster

Table 1: Comparing the true and estimated cellular fractions in the simulation. The first ten columns show Lin's concordance correlation coefficient (CCC) calculated for each cell type across all samples. The last two columns present the mean CCC across cell types and MAE (mean absolute errors) for each method. The boldface highlights the best method in each column. Average true fractions for each cell type are shown in parentheses under cell type names. Full names of cell types are listed in Figure 2a.

	Mono (0.38)	DC (0.02)	B_1 (0.07)	B_2 (0.03)	B_3 (0.03)	NK_1 (0.07)
HiDecon	0.85	0.25	**0.80**	**0.54**	**0.50**	0.34
CIBERSORT	0.88	**0.31**	0.66	0.31	0.37	0.01
dtangle	0.35	0.08	0.56	0.18	0.30	**0.47**
MuSiC	**0.89**	0.01	NA	NA	0.16	NA

	CD4_1 (0.13)	CD8_1 (0.09)	CD8_5 (0.09)	CD8_7 (0.08)	Mean CCC	MAE
HiDecon	**0.62**	0.66	0.52	0.64	**0.57**	**0.05**
CIBERSORT	0.42	**0.76**	0.25	0.25	0.42	0.07
dtangle	0.51	0.51	**0.60**	**0.69**	0.43	0.06
MuSiC	NA	0.35	0.22	0.39	0.20	0.08

in the first layer, which is not applicable for complex structures and thus not compared in the simulation. HiDecon shows a comprehensively accurate estimation performance with the highest mean CCC and lowest MAE. It is worth noticing that HiDecon is powerful in estimating rare cell type fractions. For the eight cell types with average true abundance lower than 10%, the mean CCC across these types of HiDecon estimates is 0.53 while these of CIBERSORT and dtangle are only 0.36 and 0.42 respectively.

We further checked the estimation details with scatter plots of measured and estimated cellular fractions (Figure 3). HiDecon estimated fractions are well aligned by the diagonal line $y = x$, while other methods fail to estimate some cell types. CIBERSORT estimates natural killer cells NK_1 (NK_c01-FCGR3A) as zero in 95.2% of samples. dtangle has generally flat estimates for lymphocyte subtypes. MuSiC has exaggerated estimates for dendritic cells and CD8_1 (T_CD8_c01-LEF1) cells and flat estimates for CD8_5 (T_CD8_c05-ZNF683) cells. In contrast to other methods that have many zero estimates, HiDecon saves many rare cell type fractions from being estimated as zero, providing evidence of HiDecon's ability to estimate rare cell types.

3.2 Robustness evaluation

In real settings, there is a deviation from weighted signature sum and bulk data, that is, the error term $\boldsymbol{E}$ shown in Equation 2.1. Here, we added noises to simulated bulk data and kept track of estimation performances as the noise level gets greater. We set noise $e \sim N(0, sd^2)$, where the error standard deviation (sd) ranges from 0 to the standard deviation of the simulated pseudo-bulk data when there is no noise, with a step size of 0.1. For each level of noise, we repeated the experiments 50 times under different random seeds to eliminate the

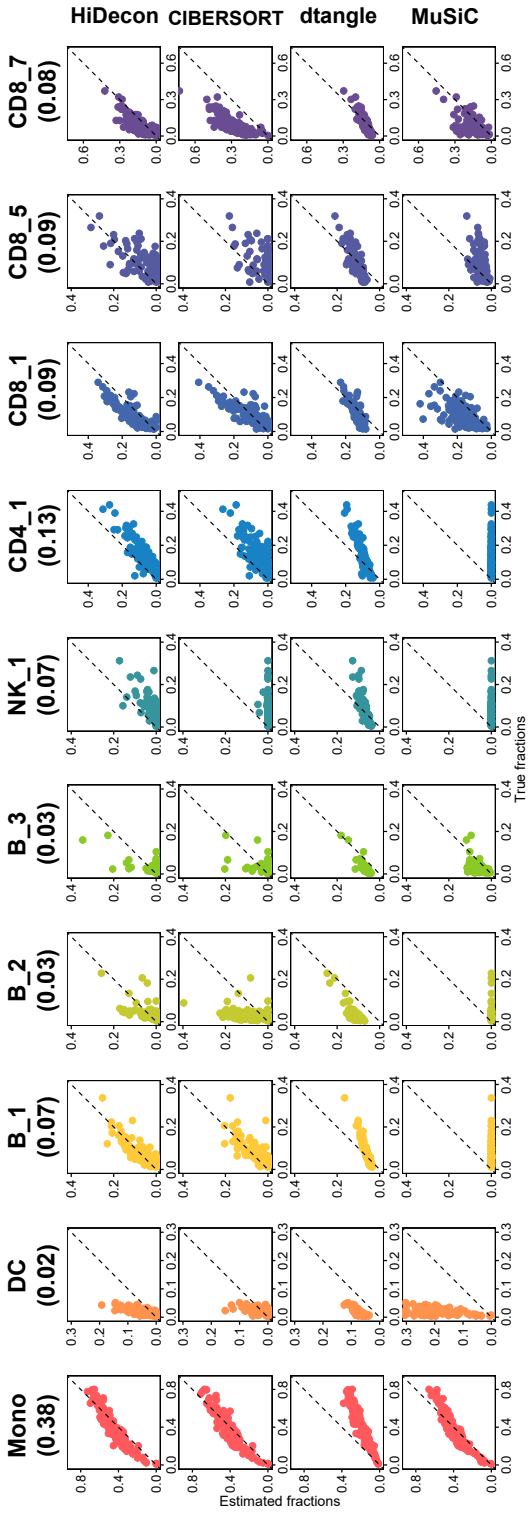

Figure 3: Scatter plots of cellular fractions in the COVID-19 data simulation study for different deconvolution methods. The x-axis represents the ground truth of PBMC cell fractions, and the y-axis shows the estimated cell type fractions. Cell type names and their mean fractions calculated by ground truth are shown at the top.

randomness and we averaged evaluation metrics from the 50 repetitions. We calculated the mean concordance correlation coefficient (CCC) and mean absolute error (MAE) trajectories and made box plots under different levels of noise (Figure 4). HiDecon has higher CCC and lower MAE than CIBERSORT, dtangle, and MuSiC. The CCC curve for MuSiC is not shown due to all zero estimates of some cell types, probably caused by the top-down recursive tree-guided process in MuSiC. The experiment shows that HiDecon has consistently outstanding performances under different noise levels.

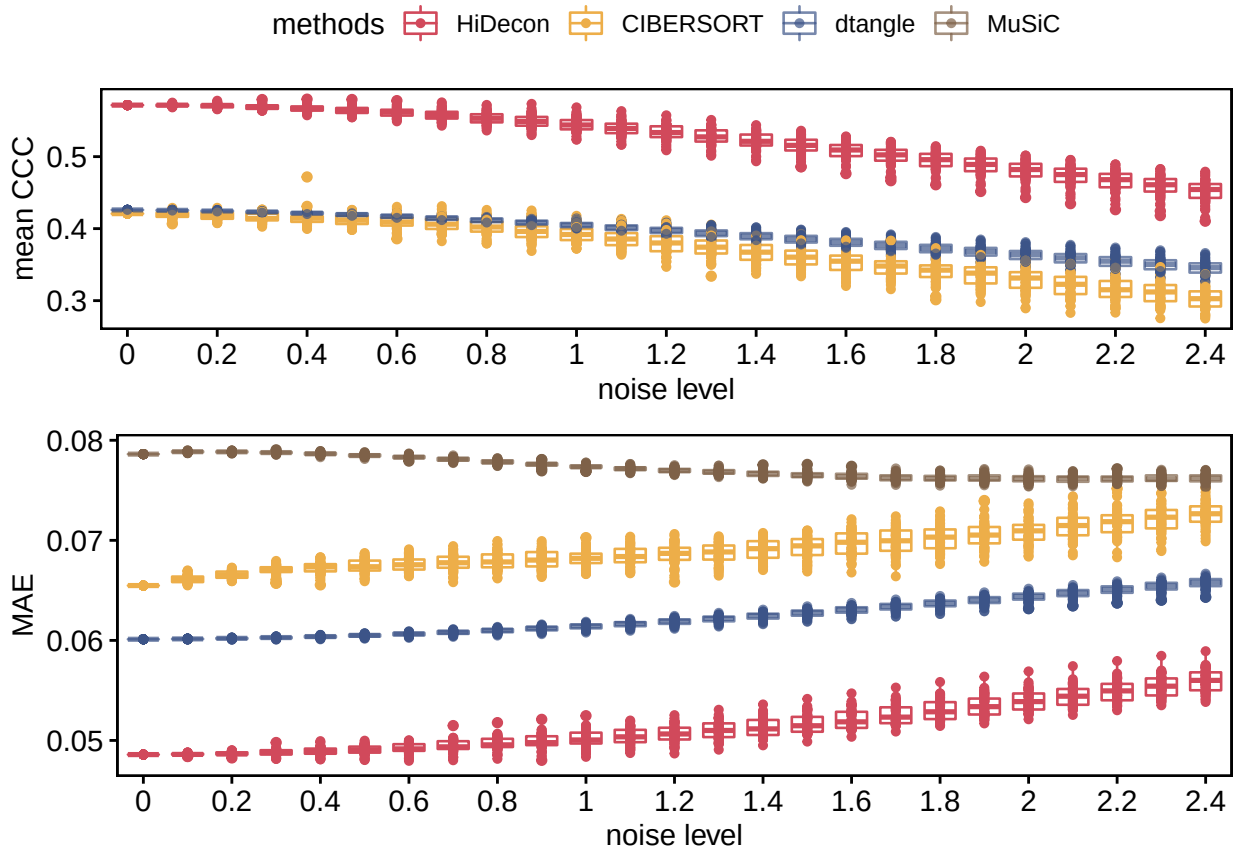

Figure 4: Mean concordance correlation coefficient (CCC) and mean absolute error (MAE) trajectories and box plots under different levels of noises (standard deviation, sd) with 50 repetitions. Noises $e \sim N(0, sd^2)$ are added to simulated bulk data. The CCC curve for MuSiC is not shown due to all zero estimates of some cell types.

4 Real data applications

To assess the performance of HiDecon in real data, we used the Framingham Heart Study (FHS) dataset with measured blood cell counts. FHS is a large-scale longitudinal study with three cohorts: the original (Dawber et al., 1951), offspring (Feinleib et al., 1975), and third-generation cohorts (Splansky et al., 2007). White blood cell (WBC) counts were measured

Table 2: Lin's concordance correlation coefficient (CCC) and MAE between measured and estimated cellular fractions in the FHS data. Missing values (NAs) are caused by all zero estimates of some cell types. When calculating the mean CCC, NAs are treated as zero. The boldface highlights the best method in each column. Average true fractions for each cell type are shown in parentheses under cell type names.

	Neutrophil (0.60)	Lymphocyte (0.28)	Monocyte (0.09)	Eosinophil (0.03)	Mean CCC	MAE
HiDecon	0.13	**0.57**	**0.04**	**0.28**	**0.26**	**0.10**
CIBERSORT	**0.15**	0.31	0.02	0.06	0.13	0.15
dtangle	0.02	0.17	0.01	0.01	0.05	0.17
HEpiDISH	0.12	0.25	0.03	0.04	0.11	0.17
MuSiC	NA	0.08	0.00	NA	0.02	0.32

from a complete blood count using the Coulter HmX Hematology Analyzer. There are 4,110 samples from two FHS cohorts (offspring and third-generation) that have both measured cell counts and high-throughput gene expression data from blood. The counted white blood cell types include neutrophils (Neutro), monocytes (Mono), lymphocytes (Lymph), and eosinophils (Eosino). Based on cell lineage, we use a two-layer hierarchical cell tree to guide deconvolution. The second layer has all four cell types, while in the first layer, the lymphocytes are treated as a separate cell type as lymphoid cells, and the other three cell types form a combined cell type to indicate that they belong to myeloid cells.

Similar to the simulation, we used CCC and MAE as evaluation metrics to compare the estimated and measured fractions. As expected, HiDecon demonstrates superior concordance and the lowest MAE across almost all cell types (Table 2). HiDecon's CCC on eosinophils, a rare type with only 3% abundance, is more than 4 times higher than other methods.

As visualized in the scatter plots, HiDecon estimated fractions have a better concordance with measured fractions along the diagonal line than other methods (Figure 5). Methods like CIBERSORT, HEpiDISH, and MuSiC produce many false zeros or even all zero estimates in some cell types, while dtangle shows flat estimated fractions which is similar to that in the simulation scatter plot (Figure 3). Additionally, HEpiDISH and CIBERSORT estimated fractions have exaggerated variability.

5 Discussion

In summary, we proposed hierarchical deconvolution (HiDecon) to incorporate a hierarchical cell type tree into cellular deconvolution to facilitate the estimation of related and rare cell types. To solve the problem of reference-based deconvolution methods that related cell types cause co-linearity in the signature matrix, we developed an algorithm to leverage the constraints from "parent" and "children" cell types when we iteratively estimate the cellular fractions of all layers of a cell type tree. We benchmarked the HiDecon algorithm on simulated COVID-19 PBMC data and a large real human blood dataset from FHS and evaluated HiDecon by comparing estimation results to the true or measured cell counts. Both numerical and real data benchmarking studies indicate that HiDecon shows higher

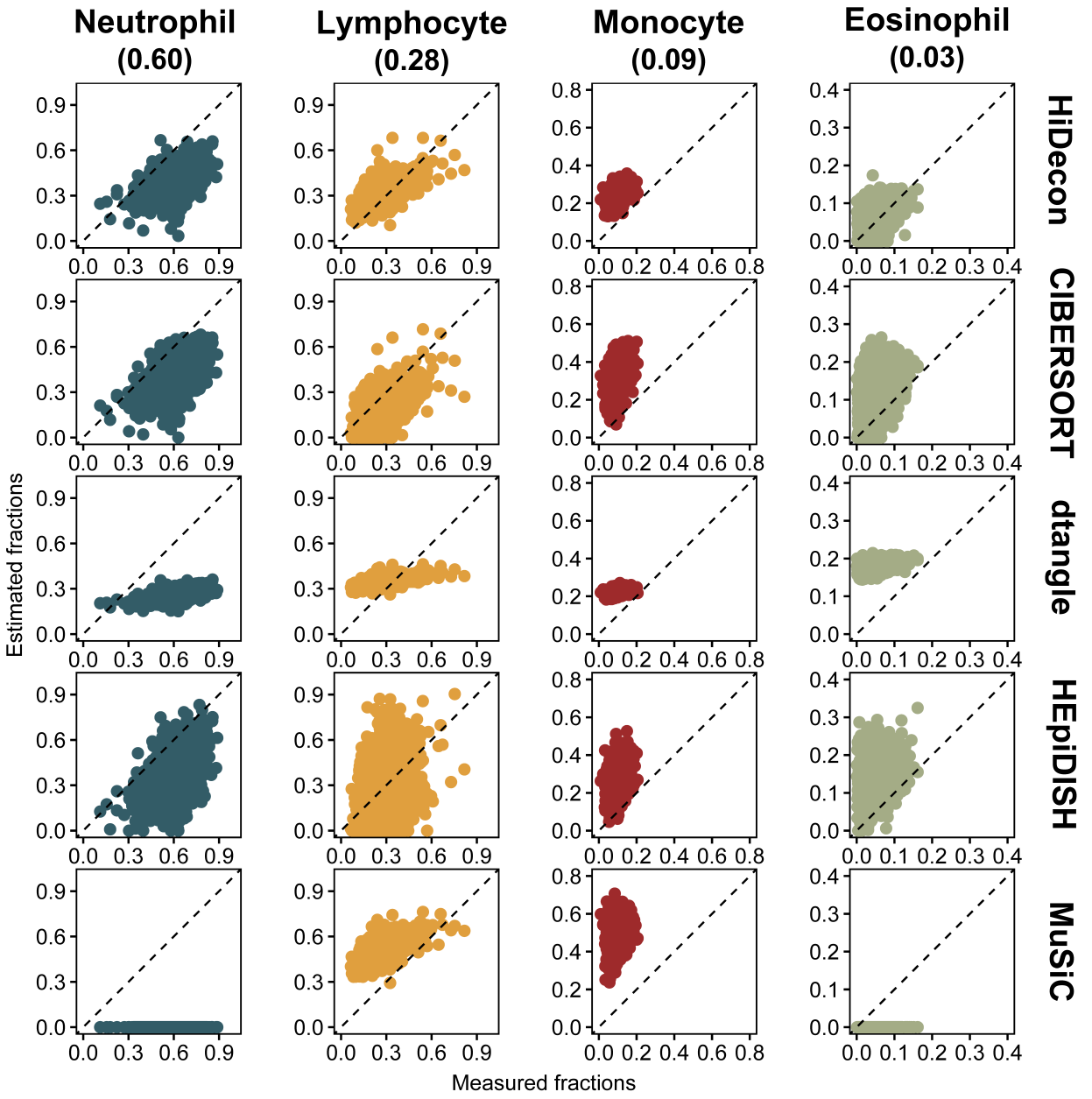

Figure 5: Scatter plots of cellular fractions in the FHS data for different deconvolution methods. The x-axis represents the ground truth of measured white blood cell fractions, and the y-axis shows the estimated cell type fractions. Cell type names and their mean fractions calculated by ground truth are shown at the top.

accuracy than existing methods. We implemented the algorithm as a user-friendly R package HiDecon and hosted it on GitHub (https://github.com/randel/HiDecon). HiDecon can incorporate complex hierarchical cell type tree structure, while the software of the two existing hierarchical deconvolution methods, MuSiC and HEpiDISH, can only incorporate a two-layer tree. Importantly, HiDecon enjoys fast convergence speed brought by the convexity of the objective function. It takes only 12.6 seconds to deconvolve the 4,110 bulk samples of the FHS data.

However, HiDecon also has some limitations. First, in our numerical study, blood data have clearly constructed and biologically and statistically interpretable hierarchical tree. If there does not exist known hierarchical tree for a tissue, we recommend users construct cell-type relationships by hierarchical clustering using single-cell data. Second, when deconvolving samples in which cell types are all highly distinguishable, HiDecon might not outperform existing methods because the hierarchical tree cannot further help in this setting. This is rare in practice especially as the research interest goes into refined cell types.

Accurate estimation of cell type proportions can provide novel insights for many downstream analyses at cell type resolution. Representative analyses include differential fraction analysis (M. Cai et al., 2022), cell type-specific (CTS) differential expression (Z. Li et al., 2019; J. Wang et al., 2020; Jin et al., 2021), CTS eQTLs (expression quantitative trait loci) (Westra et al., 2015; J. Wang et al., 2021) when both gene expression and genetic data are available, and CTS gene co-expression networks (S. Chen et al., 2020). Furthermore, hierarchical structures are not limited to gene expression data but also exist in other omics data such as DNA methylation. We will explore more settings where a hierarchical tree can serve as a useful guideline to deconvolve other omics data types that warrant future work.

Acknowledgement

This research was funded in part through NIH's R01AG080590, R03OD034501, and R01MH123184, and a grant from the University of Pittsburgh UPMC Health System's Competitive Medical Research Fund. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195, HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

References

[1] A. E. Jaffe and R. A. Irizarry. "Accounting for cellular heterogeneity is critical in epigenome-wide association studies". In: *Genome Biology* 15.2 (2014), R31. ISSN: 1465-6906.

[2] J. Wang, K. Roeder, and B. Devlin. "Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data". In: *Genome research* 31.10 (2021), pp. 1807–1818.

[3] S. Mostafavi, C. Gaiteri, S. E. Sullivan, C. C. White, S. Tasaki, J. Xu, M. Taga, H.-U. Klein, E. Patrick, V. Komashko, et al. "A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease". In: *Nature neuroscience* 21.6 (2018), pp. 811–819.

[4] F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, and K. De Preter. "Benchmarking of cell type deconvolution pipelines for transcriptomics data". In: *Nature communications* 11.1 (2020), pp. 1–14.

[5] Y. Zhong, Y.-W. Wan, K. Pang, L. M. Chow, and Z. Liu. "Digital sorting of complex tissues for cell type-specific gene expression profiles". In: *BMC bioinformatics* 14.1 (2013), pp. 1–10.

[6] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. "Robust enumeration of cell subsets from tissue expression profiles". In: *Nature methods* 12.5 (2015), pp. 453–457.

[7] G. J. Hunt, S. Freytag, M. Bahlo, and J. A. Gagnon-Bartsch. "Dtangle: accurate and robust cell type deconvolution". In: *Bioinformatics* 35.12 (2019), pp. 2093–2099.

[8] X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li. "Bulk tissue cell type deconvolution with multi-subject single-cell expression reference". In: *Nature communications* 10.1 (2019), pp. 1–9.

[9] D. R. Wilson, C. Jin, J. G. Ibrahim, and W. Sun. "ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns". In: *Journal of the American Statistical Association* 115.531 (2020), pp. 1055–1065.

[10] Z. Wu and H. Wu. "Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering". In: *Genome Biology* 21.1 (2020). ISSN: 1474-760X.

[11] M. Peng, B. Wamsley, A. G. Elkins, D. H. Geschwind, Y. Wei, and K. Roeder. "Cell type hierarchy reconstruction via reconciliation of multi-resolution cluster tree". In: *Nucleic acids research* 49.16 (2021), e91–e91.

[12] L. Chen, Z. Li, and H. Wu. "CeDAR: incorporating cell type hierarchy improves cell type specific differential analyses in bulk omics data". In: *bioRxiv* (2022).

[13] J. A. Miller, N. W. Gouwens, B. Tasic, F. Collman, C. T. van Velthoven, T. E. Bakken, M. J. Hawrylycz, H. Zeng, E. S. Lein, and A. Bernard. "Common cell type nomenclature for the mammalian brain". In: *Elife* 9 (2020), e59928.

[14] R. D. Hodge, T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, L. T. Graybuck, J. L. Close, B. Long, N. Johansen, O. Penn, et al. "Conserved cell types with divergent features in human versus mouse cortex". In: *Nature* 573.7772 (2019), pp. 61–68.

[15] S. Fischer and J. Gillis. "How many markers are needed to robustly determine a cell's type?" In: *Iscience* 24.11 (2021), p. 103292.

[16] S. C. Zheng, A. P. Webster, D. Dong, A. Feber, D. G. Graham, R. Sullivan, S. Jevons, L. B. Lovat, S. Beck, M. Widschwendter, et al. "A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix". In: *Epigenomics* 10.7 (2018), pp. 925–940.

[17] C. Jia, Y. Hu, D. Kelly, J. Kim, M. Li, and N. R. Zhang. "Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data". In: *Nucleic Acids Res* 45.19 (2017), pp. 10978–10988. ISSN: 1362-4962 (Electronic) 0305-1048 (Print) 0305-1048 (Linking).

[18] S. Mohammadi, N. Zuckerman, A. Goldsmith, and A. Grama. "A critical survey of deconvolution methods for separating cell types in complex tissues". In: *Proceedings of the IEEE* 105.2 (2016), pp. 340–366.

[19] L. I.-K. Lin. "A concordance correlation coefficient to evaluate reproducibility". In: *Biometrics* (1989), pp. 255–268.

[20] R. L. Berger. "Likelihood ratio tests and intersection-union tests". In: *Advances in statistical decision theory and applications*. Springer, 1997, pp. 225–237.

[21] X. Ren, W. Wen, X. Fan, W. Hou, B. Su, P. Cai, J. Li, Y. Liu, F. Tang, F. Zhang, et al. "COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas". In: *Cell* 184.7 (2021), pp. 1895–1913.

[22] T. R. Dawber, G. F. Meadors, and F. E. Moore Jr. "Epidemiological approaches to heart disease: the Framingham Study". In: *American Journal of Public Health and the Nations Health* 41.3 (1951), pp. 279–286.

[23] M. Feinleib, W. B. Kannel, R. J. Garrison, P. M. McNamara, and W. P. Castelli. "The Framingham offspring study. Design and preliminary data". In: *Preventive medicine* 4.4 (1975), pp. 518–525.

[24] G. L. Splansky, D. Corey, Q. Yang, L. D. Atwood, L. A. Cupples, E. J. Benjamin, R. B. D'Agostino Sr, C. S. Fox, M. G. Larson, J. M. Murabito, et al. "The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination". In: *American journal of epidemiology* 165.11 (2007), pp. 1328–1335.

[25] M. Cai, M. Yue, T. Chen, J. Liu, E. Forno, X. Lu, T. Billiar, J. Celedón, C. McKennan, W. Chen, et al. "Robust and accurate estimation of cellular fraction from tissue omics data via ensemble deconvolution". In: *Bioinformatics* 38.11 (2022), pp. 3004–3010.

[26] Z. Li, Z. Wu, P. Jin, and H. Wu. "Dissecting differential signals in high-throughput data from complex tissues". In: *Bioinformatics* 35.20 (2019), pp. 3898–3905. ISSN: 1367-4803.

[27] J. Wang, B. Devlin, and K. Roeder. "Using multiple measurements of tissue to estimate subject-and cell-type-specific gene expression". In: *Bioinformatics* 36.3 (2020), pp. 782–788.

[28] C. Jin, M. Chen, D.-Y. Lin, and W. Sun. "Cell-type-aware analysis of RNA-seq data". In: *Nature Computational Science* 1.4 (2021), pp. 253–261.

[29] H.-J. Westra et al. "Cell Specific eQTL Analysis without Sorting Cells". In: *PLoS Genetics* 11.5 (2015), e1005223. ISSN: 1553-7404.

[30] S. Chen, J. Wang, E. Cicek, K. Roeder, H. Yu, and B. Devlin. "De novo missense variants disrupting protein–protein interactions affect risk for autism through gene co-expression and protein networks in neuronal cell types". In: *Molecular autism* 11.1 (2020), pp. 1–16.