NAR Genomics and Bioinformatics, 2025, **7**, lqaf046
https://doi.org/10.1093/nargab/lqaf046
Computational Methods for Transcriptome Analysis

ARTdeConv: adaptive regularized tri-factor non-negative matrix factorization for cell type deconvolution

Tianyi Liu [1,*], Chuwen Liu [1], Quefeng Li [1,*], Xiaojing Zheng [1,2,*], Fei Zou [1,3,*]

[1]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
[2]Department of Pediatrics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
[3]Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

[*]To whom correspondence should be addressed. Email: tianyi.liu.academic@gmail.com
Correspondence may also be addressed to Quefeng Li. Email: quefeng@email.unc.edu
Correspondence may also be addressed to Xiaojing Zheng. Email: xiaojinz@email.unc.edu
Correspondence may also be addressed to Fei Zou. Email: feizou@email.unc.edu

Abstract

Accurate deconvolution of cell types from bulk gene expression is crucial for understanding cellular compositions and uncovering cell-type specific differential expression and physiological states of diseased tissues. Existing deconvolution methods have limitations, such as requiring complete cellular gene expression signatures or neglecting partial biological information. Moreover, these methods often overlook varying cell-type messenger RNA amounts, leading to biased proportion estimates. Additionally, they do not effectively utilize valuable reference information from external studies, such as means and ranges of population cell-type proportions. To address these challenges, we introduce an adaptive regularized tri-factor non-negative matrix factorization approach for deconvolution (ARTdeConv). We rigorously establish the numerical convergence of our algorithm. Through benchmark simulations, we demonstrate the superior performance of ARTdeConv compared to state-of-the-art semi-reference-based and reference-free methods as well as its robustness under challenges to its assumptions. In a real-world application to a dataset from a trivalent influenza vaccine study, our method accurately estimates cellular proportions, as evidenced by the nearly perfect Pearson's correlation between ARTdeConv estimates and flow cytometry measurements. Moreover, our analysis of ARTdeConv estimates in COVID-19 patients reveals patterns consistent with important immunological phenomena observed in other studies. The proposed method, ARTdeConv, is implemented as an R package and can be accessed on GitHub for researchers and practitioners.

Introduction

Heterogeneity in cell-type proportions exists across biological samples, and neglecting this heterogeneity in bulk gene expression can introduce biases into subsequent analyses, such as those of differential gene expression. Conversely, acknowledging and accounting for this heterogeneity has shown clear advantages, yielding more accurate survival time predictions and tumor type classifications [1, 2].

Laboratory techniques such as flow cytometry or immunohistochemistry are available for physically sorting cells into cell types and quantifying their abundances. However, these methods are often limited by the availability of cell samples, the specificity of antibodies for separating cells, and the substantial labor and time investments required [3, 4]. Cell type deconvolution (simply referred to as "deconvolution" in this work), a computational process aimed at digitally separating heterogeneous mixture signals into their constituent components, has been critical in expediting the estimation of cell-type proportions from bulk gene expression data such as those from RNA sequencing (RNA-seq). In recent years, several deconvolution methods have emerged, with extensive applications in the field of computational biology. These methods can generally be categorized into reference-based and reference-free methods, depending on whether they require individual cell-type gene expression signatures, sometimes referred to as a signature matrix, as prior knowledge [2, 5, 6].

For reference-based deconvolution, gene signatures can be derived from either single-cell RNA-seq (scRNA-seq) data or sorted bulk RNA profiles of individual cell types [7]. Although gene signatures have been successfully established for some cell types, acquiring those for other cell types might be labor-intensive or even infeasible [8, 9]. Neglecting to incorporate gene signatures for a prevalent cell type can induce biases in the proportion estimates of other cell types within reference-based deconvolution [2]. On the other hand, reference-free techniques enable the unsupervised estimation of cell-type proportions but at the cost of disregarding information embedded in known gene signatures [10, 11]. A deconvolution method that can utilize partial reference gene signatures presents a plausible compromise.

Additionally, external information on population cell-type proportions may also aid in estimating cell-type proportions. For example, in blood samples, the proportion means (or medians) and ranges of major cell types are available from complete blood count tests. However, deconvolution techniques that effectively incorporate such information are currently lacking in the field.

Received: July 17, 2024. **Revised:** March 2, 2025. **Editorial Decision:** April 2, 2025. **Accepted:** April 14, 2025

During data pre-processing, information specific to the amount of RNA molecules packed within an individual cell is often lost after library normalization. Zaitsev *et al.* found that failures to account for messenger RNA (mRNA) amounts produced biased cell-type proportion estimates [10]. However, many deconvolution methods assume a generative model where the loss of this information cannot be accounted for [2]. Therefore, incorporating these quantities into the generative model can be pivotal to deconvolution accuracy.

In recent years, several semi-reference-based methods have been developed to perform deconvolution using only partial references. EPIC addresses the lack of representation of cancer cells in common gene signature matrices for the deconvolution of the tumor microenvironment (TME) [12, 13]. Specifically, EPIC treats all types of cancer cells as a single super cell type in deconvolution, allowing for a detailed characterization of immune cell proportions in the TME while accounting for the fraction of cancer cells. This super-cell-type rationale can also be extended to the deconvolution of other tissue types, such as blood. Moreover, EPIC assigns weights to each gene in the signature matrix based on its variability across cell types; genes with lower expression variance are given higher weights for deconvolution. EPIC also adjusts for differing mRNA amounts between cell types, which must be specified prior to deconvolution. However, this becomes infeasible if the tissue is complex, especially when there is a super cell type with multiple constituent cell types whose reference expression is missing. EPIC further requires that deconvolution be performed using signature genes not specific to the uncharacterized super cell type and estimates the proportions of cell types with reference using an inequality-constrained optimization procedure. Consequently, the estimated proportions of the cell type with a missing reference are biased toward zero [14].

Two other tools, quanTIseq and SECRET, have similar functions (i.e. they can accommodate one cell or super cell type with a missing reference), with the former essentially employing the same mathematical model as EPIC [15, 16]. An innovation of quanTIseq over EPIC is its carefully defined set of reference gene signatures for immune cells and the calculation of cell-type mRNA amount scaling factors using housekeeping genes [15], but it does not weigh the genes for deconvolution. These innovations also render quanTIseq unable to use customized signature matrices when needed. On the other hand, SECRET can incorporate customized gene signatures from scRNA-seq experiments and uses an outlier-insensitive L_1 norm on the residuals instead of EPIC's L_2 norm [16]. However, it does not truly utilize any reference information about the uncharacterized cell type, making it prone to the same bias toward zero as EPIC does.

BayICE is another semi-reference-based method serving the same purpose as the other methods above, but it adopts a hierarchical Bayes design with stochastic gene signature selection [14]. Due to its probabilistic nature, BayICE can quantify the uncertainty of cellular abundance estimates and claims to mitigate the biases of EPIC. These benefits come at the expense of increased computational complexity, sensitivity to priors, and difficulty in interpreting the parameters involved. Moreover, it requires bulk samples of purified cells as reference inputs, which, like quanTIseq, precludes the use of other types of reference data, such as customized signature matrices.

In this article, we propose a novel semi-reference-based method called ARTdeConv (short for "adaptive regularized tri-factor non-negative matrix factorization method for deConvolution"). It addresses the three outstanding issues discussed above: utilizing partial reference gene signatures, incorporating external information on cell-type proportions, and accounting for cell-type mRNA amounts. Compared to EPIC, ARTdeConv does not require the specificity of the gene signatures to cell types that are characterized. It also uses external proportion information to correct for EPIC's observed biases toward zero. Moreover, compared to EPIC and quanTIseq, ARTdeConv learns the relative cell-type mRNA abundances automatically without the need to use housekeeping genes or other information while maintaining the straightforward interpretation of the resulting estimates. It can also incorporate scRNA-seq data as SECRET does and uses a different optimization procedure. Our further contributions include deriving and implementing a multiplicative update (MU) algorithm for solving ARTdeConv, proving the algorithm's convergence, and demonstrating its merits through simulations and real data analysis. A schematic representation of ARTdeConv's workflow is shown in Fig. 1. An R package implementing ARTdeConv is also available on GitHub.

Materials and methods

Notation

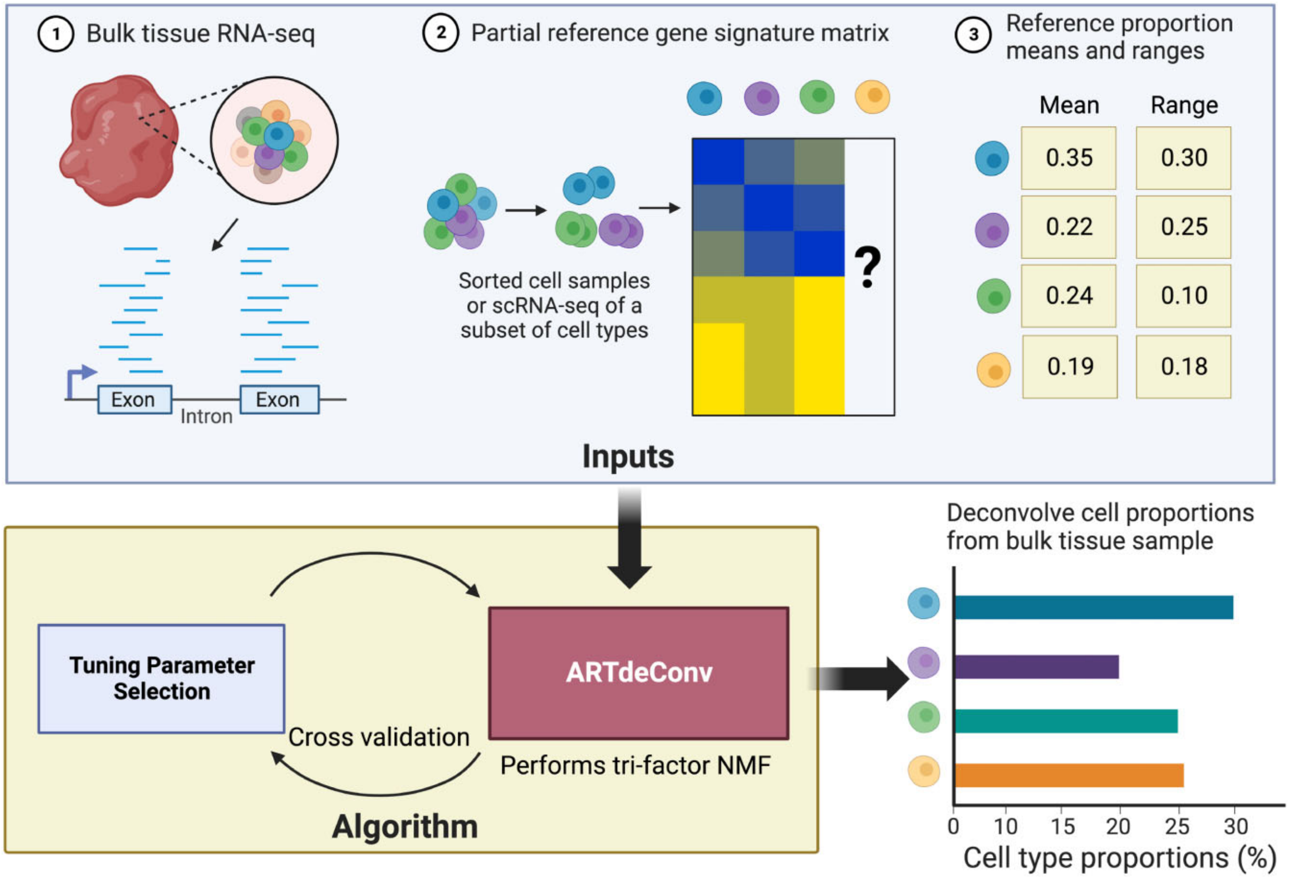
Let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of non-negative and positive real numbers, respectively. m, n, and K are positive integers used to denote the number of genes, samples, and cell types with $K \leq \min(m, n)$ in deconvolution. Let a positive integer K_0 denote the number of cell types for which we have reference gene expression available. Unless otherwise mentioned, we set $K = K_0 + 1$. Let $\boldsymbol{Y} \in \mathbb{R}_+^{m\times n}$ be the matrix of bulk gene expression and $\boldsymbol{\Theta} \in \mathbb{R}_+^{m\times K}$ be the full gene signature matrix. Let $\boldsymbol{s} \in \mathbb{R}_{++}^{K}$ be a vector of relative cell-type mRNA amounts, and $\boldsymbol{P} \in \mathbb{R}_+^{K\times n}$ the proportion matrix. Let $\boldsymbol{\epsilon} \in \mathbb{R}^{m\times n}$ be the random error matrix. Unless otherwise specified, we denote $\boldsymbol{\theta}_k$ as the k-th column of $\boldsymbol{\Theta}$. Let $\boldsymbol{A}$ be a generic $m \times n$ matrix. We use $\boldsymbol{A}^\top$ to denote the transpose of a matrix. If $\boldsymbol{A}$ is further a square matrix ($m = n$), we denote its trace $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} A_{ii}$. For any $m \times n$ matrix $\boldsymbol{A}$, its Frobenius norm is denoted as $\|\boldsymbol{A}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{A})} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}$. For two matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same dimensions, $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the element-wise product of $\boldsymbol{A}$ and $\boldsymbol{B}$ and $\frac{\boldsymbol{A}}{\boldsymbol{B}}$ denotes their element-wise quotient.

Model and problem setup

We first propose the following tri-factor generative model

$$\boldsymbol{Y} = \boldsymbol{\Theta}\mathrm{diag}(\boldsymbol{s})\boldsymbol{P} + \boldsymbol{\epsilon} \tag{1}$$

for ARTdeConv, which extends the canonical model $\boldsymbol{Y} = \boldsymbol{\Theta}\boldsymbol{P} + \boldsymbol{\epsilon}$ [2]. The difference between them is $\mathrm{diag}(\boldsymbol{s})$, a diagonal matrix of cell-type mRNA amounts. In practice, normalization (e.g. by library sizes, etc.) is frequently employed to alleviate between-sample technical artifacts in sequencing data [17, 18]. During this process, information regarding the quantities of RNA molecules packed within different kinds of cells sometimes gets lost. Failures to recover these quantities lead to documented biases in estimating cell-type proportions [10, 12]. Should this occur, $\mathrm{diag}(\boldsymbol{s})$ can account for the RNA molecule quantities. Otherwise, $\mathrm{diag}(\boldsymbol{s})$ would be close to the identity matrix $\boldsymbol{I}_K$ (an option to fix $\mathrm{diag}(\boldsymbol{s}) = \boldsymbol{I}_K$ is given by ARTdeConv). The values in $\mathrm{diag}(\boldsymbol{s})$ should be interpreted in



1 Bulk tissue RNA-seq
Exon
Intron
Exon
2 Partial reference gene signature matrix
Sorted cell samples or scRNA-seq of a subset of cell types
?
3 Reference proportion means and ranges
Mean
Range
0.35
0.30
0.22
0.25
0.24
0.10
0.19
0.18
Inputs
Tuning Parameter Selection
Cross validation
ARTdeConv
Performs tri-factor NMF
Algorithm
Deconvolve cell proportions from bulk tissue sample
0
10
15
20
25
30
Cell type proportions (%)

Figure 1. A schema of the ARTdeConv workflow. Specifically, ARTdeConv takes in three pieces of information as inputs: the gene expression data of the bulk tissue to be deconvolved, the reference gene expression data with one cell type or super cell type uncharacterized, and the reference means/medians and ranges of the proportions of all cell types in deconvolution. It then passes these inputs to a cross-validation algorithm to select the optimal hyperparameters. Lastly, these hyperparameters are passed along with the rest of the inputs for a deconvolution run that estimates the proportions of each cell type or super cell type involved.

a relative sense. For example, s_2/s_1 is the ratio between the amounts of mRNA molecules packed by the second and the first cell type. We do not require any prior specification of $\mathrm{diag}(\boldsymbol{s})$ and notice that Equation (1) is similar to the model of MuSiC, a reference-based method where each sample is accorded an additional scaling factor c_i to account for between-sample measurement variation in bulk gene expression [19]. Moreover, this setup differs from that of EPIC, which requires $\boldsymbol{s}$ be manually specified [12].

For performing deconvolution based on Equation (1), ARTdeConv assumes prior knowledge of a partial signature matrix $\boldsymbol{\Theta}_{K_0} \in \mathbb{R}_+^{m\times K_0}$ as well as $\boldsymbol{Y}$. Then, an objective function is defined:

$$f(\boldsymbol{\Theta}, \boldsymbol{s}, \boldsymbol{P}) = \frac{1}{2mn}\|\boldsymbol{Y} - \boldsymbol{\Theta}\mathrm{diag}(\boldsymbol{s})\boldsymbol{P}\|_F^2 + \frac{\alpha_1}{2}R_1 + \frac{\alpha_2}{2}R_2 + \frac{\beta}{2}R_3, \quad (2)$$

where R_1, R_2, and R_3 are regularizers to be explained later and α_1, α_2, and β are their tuning parameters. We discuss the selection of optimal tuning parameters in Section 2.4.$\boldsymbol{\Theta}$, $\boldsymbol{s}$, and $\boldsymbol{P}$ are then estimated via

$$(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{P}}) = \underset{\boldsymbol{\Theta}\in\mathbb{R}_+^{m\times K},\ \boldsymbol{s}\in\mathbb{R}_{++}^{K},\ \boldsymbol{P}\in\mathbb{R}_+^{K\times n}}{\operatorname{argmin}} f(\boldsymbol{\Theta}, \boldsymbol{s}, \boldsymbol{P}). \quad (3)$$

The main objective here is to obtain an estimate of the proportion matrix $\hat{\boldsymbol{P}}$. To follow a common practice of deconvolution and simplify the algorithm [20], ARTdeConv does not directly constrain each column of $\boldsymbol{P}$ to sum to one during the estimation process. Instead, it obtains an unconstrained estimate $\hat{\boldsymbol{P}}$ and then re-normalizes its columns to have the unit sum.

Let $\boldsymbol{\Theta}_0 = \left[\boldsymbol{\Theta}_{K_0}\ 0 \cdots 0\right] \in \mathbb{R}_+^{m\times K}$ and $\boldsymbol{\Delta}$ be an $m \times K$ matrix such that $\Delta_{jk} = I(k \leq K_0)$ for $1 \leq j \leq m$. The squared Frobenius distance between the estimated $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_0$ is penalized through $R_1 = \|\boldsymbol{\Delta} \odot (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)\|_F^2$. Though we present a special case where $\boldsymbol{\Theta}_{K_0}$ occupies the first K_0 columns of $\boldsymbol{\Theta}$, by redefining $\boldsymbol{\Delta}$ correspondingly, $\boldsymbol{\Theta}_{K_0}$ can occupy any K_0 columns, covering all structures of prior knowledge on the signature matrix. Same as other well established semi-reference-based methods (i.e. EPIC, quanTIseq, etc.), we set $K = K_0 + 1$. We also recommend that the unrepresented cell types in $\boldsymbol{\Theta}_{K_0}$ be grouped into a single artificial "cell type." Albeit that ARTdeConv technically allows $K > K_0 + 1$, we have found that the resulting estimates from ARTdeConv became less reliable as $K - K_0$ grew larger in preliminary analysis (results not shown). This makes ARTdeConv suitable for the situations where a major cell type is missing from the reference or that distinguishing the multiple cell types that are missing is not salient. We shall see such examples at work late in Sections 3.2 and 3.3.

On the other hand, R_1 is not a strictly convex function of $\boldsymbol{\Theta}$, making the guarantee of the numerical convergence of ARTdeConv's algorithm difficult (more details on this later). To atone for this, we make $R_2 = \|(\boldsymbol{J}_{m\times K} - \boldsymbol{\Delta}) \odot \boldsymbol{\Theta}\|_F^2$, where $\boldsymbol{J}_{m\times K}$ is an $m \times K$ matrix of 1s. It can be shown that any positive linear combination of R_1 and R_2 is strictly convex with respect to $\boldsymbol{\Theta}$. It is of notes that R_2 will force the estimated gene signatures of the uncharacterized close to zero. While this is the case for genes that are specific to the cell types in $\boldsymbol{\Theta}_{K_0}$, ARTdeConv does not restrict the gene signatures to only those genes. Thus, it is recommended to reduce the penalization effects R_2 by setting its tuning parameter α_2 to a very small value.

Information on the cell-type proportions, such as means (or medians) and ranges in a population, frequently exists in external data and is accessible online, for example, through complete blood counts of leukocytes [21]. Let m_k denote the reference mean (or median) of cell type k's population proportion and let r_k denote its range. To incorporate it into our deconvolution method, we propose adding another penalty $R_3 = \sum_{k=1}^{K}(1/r_k)\|\boldsymbol{p}_k^\top - m_k\mathbf{1}_n\|_2^2$, where $\boldsymbol{p}_k^\top$ is the k-th row of $\boldsymbol{P}$. In R_3, m_k acts as a pivot, and the deviation from which is penalized for the estimated $\boldsymbol{p}_k^\top$. Meanwhile, $1/r_k$ acts as a weight so the proportion estimates of cell types with wider ranges are less penalized for departing from their respective m_k. Letting $\boldsymbol{M}$ be a $K \times n$ matrix such that $\boldsymbol{M} = \left[m_1\mathbf{1}_n\ m_2\mathbf{1}_n \cdots m_K\mathbf{1}_n\right]^\top$ and $\boldsymbol{\rho} = \operatorname{diag}(r_1, r_2, \cdots, r_K)$, we can write R_3 in a matrix form $R_3 = \|\boldsymbol{\rho}^{-1/2}(\boldsymbol{P} - \boldsymbol{M})\|_F^2$.

MU algorithm

A MU algorithm is proposed to solve Equation (3). The MU algorithm was originally designed to solve the canonical bi-factor non-negative matrix factorization (NMF) problem [22] and was extensively studied [23]. Technically, it is similar to the majorization–minimization (MM) algorithm of [24]. It can be readily extended to solving a multi-factor regularized NMF problem like the one for ARTdeConv.

Let t be a non-negative integer denoting the current number of iteration. The MU steps are derived by finding the gradients of a set of auxiliary functions for each row of $\boldsymbol{\Theta}^t$, each column of $\boldsymbol{P}^t$, and $\boldsymbol{s}^t$. These update steps are:

$$\boldsymbol{\Theta}^{t+1} = \boldsymbol{\Theta}^t \odot \frac{\boldsymbol{Y}(\boldsymbol{P}^t)^\top \operatorname{diag}(\boldsymbol{s}^t) + mn\alpha_1 \boldsymbol{\Delta} \odot \boldsymbol{\Theta}_0}{\boldsymbol{\Theta}^t \operatorname{diag}(\boldsymbol{s}^t)\boldsymbol{P}^t(\boldsymbol{P}^t)^\top \operatorname{diag}(\boldsymbol{s}^t) + mn\left\{\alpha_1\boldsymbol{\Delta} \odot \boldsymbol{\Theta}^t + \alpha_2(\boldsymbol{J}_{m\times K} - \boldsymbol{\Delta}) \odot \boldsymbol{\Theta}^t\right\}}; \tag{4}$$

$$\boldsymbol{P}^{t+1} = \boldsymbol{P}^t \odot \frac{\operatorname{diag}(\boldsymbol{s}^t)(\boldsymbol{\Theta}^{t+1})^\top \boldsymbol{Y} + mn\beta\boldsymbol{\rho}^{-1}\boldsymbol{M}}{\operatorname{diag}(\boldsymbol{s}^t)(\boldsymbol{\Theta}^{t+1})^\top \boldsymbol{\Theta}^{t+1}\operatorname{diag}(\boldsymbol{s}^t)\boldsymbol{P}^t + mn\beta\boldsymbol{\rho}^{-1}\boldsymbol{P}^t}; \tag{5}$$

$$\boldsymbol{s}^{t+1} = \boldsymbol{s}^t \odot \frac{\boldsymbol{u}^t}{\boldsymbol{Z}^t\boldsymbol{s}^t}, \tag{6}$$

where $u_k^t = \operatorname{tr}\left\{\boldsymbol{Y}^\top\boldsymbol{\theta}_k^{t+1}\left(\boldsymbol{p}^{t+1}\right)_k^\top\right\}$ and $Z_{kl}^t = \left\{(\boldsymbol{\theta}^{t+1})_k^\top\boldsymbol{\theta}_l^{t+1}\right\}\left\{(\boldsymbol{p}^{t+1})_k^\top\boldsymbol{p}_l^{t+1}\right\}$ for $k, l = 1, 2, \cdots, K$.

Mathematical details for finding the auxiliary functions and the update steps can be found in Supplementary Material Section B. The pseudo-code of the MU algorithm is shown in Algorithm 1.

Algorithm 1 The MU Algorithm For Solving ARTdeConv

```
  Initialize $\boldsymbol{\Theta}^0 \in \mathbb{R}_+^{m\times K}$ such that $R_1 = 0, \boldsymbol{P}^0 \in \mathbb{R}_+^{K\times n}, \boldsymbol{s}^0 \in \mathbb{R}_{++}^K$. Fix $\alpha_1$, $\alpha_2$, and $\beta > 0$.
  Repeat
      Update $\boldsymbol{\Theta}^t$ to $\boldsymbol{\Theta}^{t+1}$ via (4);
      Update $\boldsymbol{P}^t$ to $\boldsymbol{P}^{t+1}$ via (5);
      Update $\boldsymbol{s}^t$ to $\boldsymbol{s}^{t+1}$ via (6);
  Until
      $\frac{|f(\boldsymbol{\Theta}^{t+1}, \boldsymbol{s}^{t+1}, \boldsymbol{P}^{t+1}) - f(\boldsymbol{\Theta}^t, \boldsymbol{s}^t, \boldsymbol{P}^t)|}{f(\boldsymbol{\Theta}^t, \boldsymbol{s}^t, \boldsymbol{P}^t)} < \delta,$
  where $\delta$ is a small tolerance parameter and $f$ defined as in (2). Define the final outputs as $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{P}})$. Re-normalize $\hat{\boldsymbol{P}}$ by dividing each column of by its sum.
```

The MU algorithm has two advantages (under the assumptions discussed in Supplementary Material Section A). First, $\boldsymbol{\Theta}^t$, $\boldsymbol{P}^t$, and $\boldsymbol{s}^t$ remain non-negative throughout all iterations if their initial values are non-negative. It is recommended that $\boldsymbol{\Theta}^0$ is set to satisfy $R_1 = 0$, so any zero in the known partial signature matrix will remain zero. Second, the MU algorithm can be shown to achieve numerical convergence. If we enforce $\operatorname{diag}(\boldsymbol{s}^t) = \boldsymbol{I}_K$ and set $\alpha_1 = \alpha_2 = \beta = 0$, our MU algorithm is reduced to that of [22] for the canonical NMF. Due to the non-linearity of the objective function Equation (2), it is not guaranteed that a global minimizer can be produced. Following the precedence of other NMF-based deconvolution software, we recommend restarting the MU algorithm multiple times and choosing the run with the smallest Frobenius norm of the residuals on the estimated bulk expression [25, 26]. We set 10 restarts as the default in our implementation.

We now describe the default initialization procedure of ARTdeConv. First, the first K_0 columns of $\boldsymbol{\Theta}^0$ were used as $\boldsymbol{\Theta}_{K_0}^*$ to satisfy $R_1 = 0$. Each initial value in the last column θ_{jK}^0 is generated independently by $\theta_{jK}^0 \sim |N(\bar{y}_j, \hat{\sigma}_j^2)|$, where $\bar{y}_j$ and $\hat{\sigma}_j^2$ are the mean and variance of the bulk expression values for gene j across all n samples. This differs from the procedure of SECRET [16], which initiates $\theta_{jK}^0 = 0$. This is because the MU algorithm would produce $\theta_{jK}^t = 0$ for any $t = 1, 2, \ldots, T$ following this initiation. Then, for each cell type k of sample i, p_{ki}^0 is generated from a $N(m_k, 0.1)$ distribution. After this, negative entries in $\boldsymbol{P}^0$ are corrected to 0.01. Lastly, $\boldsymbol{s}^0$ is initialized by solving

$$\boldsymbol{s}^0 = \operatorname*{argmin}_{\boldsymbol{s}\in\mathbb{R}^K} \frac{1}{2}\|\boldsymbol{Y}^* - \boldsymbol{\Theta}^0 \operatorname{diag}(\boldsymbol{s})\boldsymbol{P}^0\|_F^2.$$

Negative values in $\boldsymbol{s}^0$ will be replaced by the mean of positive values.

Tuning parameter selection

A grid search using a B-fold cross-validation (CV) is designed for selecting tuning parameters. Since R_2 is designed to make the objective function in Equation (2) strictly convex, α_2 is advised to be fixed at a minuscule value. In the implementation of ARTdeConv, $\alpha_2 = 10^{-12}$ is the default value.

To begin, $\boldsymbol{Y}$ is randomly divided into B different folds, each containing $1/B$ columns of $\boldsymbol{Y}$. Two grids $\mathfrak{A}_1, \mathfrak{B} \subset \mathbb{R}_{++}$ containing respective candidates values for α_1 and β are declared. Next, for one $\alpha_1 \in \mathfrak{A}_1$, one $\beta \in \mathfrak{B}$ and a fold $b \in \{1, 2, \ldots, B\}$, the columns in the fold are held out as the test set $\boldsymbol{Y}_{\text{test}}^{(b)}$. The rest of the columns in $\boldsymbol{Y}$ are used as the training set $\boldsymbol{Y}_{\text{train}}^{(b)}$, upon which an ARTdeConv solution $(\tilde{\boldsymbol{\Theta}}^{(b)}, \tilde{\boldsymbol{s}}^{(b)}, \tilde{\boldsymbol{P}}^{(b)})$ is obtained.

Then, since it is assumed that n samples should share the same $\boldsymbol{\Theta}$ and $\boldsymbol{s}$ in Equation (1), the estimated proportions $\hat{\boldsymbol{P}}^{(b)}$ on the test set $\boldsymbol{Y}_{\text{test}}^{(b)}$ is obtained by

$$\hat{\boldsymbol{P}}^{(b)} = \operatorname*{argmin}_{\boldsymbol{P}\in\mathbb{R}_+^{K\times n/B}} \|\boldsymbol{Y}_{\text{test}}^{(b)} - \tilde{\boldsymbol{\Theta}}^{(b)} \operatorname{diag}(\tilde{\boldsymbol{s}})^{(b)}\boldsymbol{P}\|_F^2.$$

Computationally, this is accomplished by solving a non-negative least square problem using the R package `nnls`. The CV error given α_1 and β is

$$\operatorname{Err}(\alpha_1, \beta) = \frac{1}{B}\sum_{b=1}^{B} \|\boldsymbol{Y}_{\text{test}}^{(b)} - \tilde{\boldsymbol{\Theta}}^{(b)} \operatorname{diag}(\tilde{\boldsymbol{s}})^{(b)}\hat{\boldsymbol{P}}^{(b)}\|_F^2.$$

Finally, we select the best tuning parameters α_1^* and β^* that minimize $\text{Err}(\alpha_1, \beta)$. We offer some additional discussion on the CV steps in Supplementary Material Section D. Moreover, in order to facilitate effective regularization and tuning parameter selection, it is important to have correct reference parameters set up. We shall discuss this further in Section 3.1.

Convergence analysis

In this section, we prove the numerical convergence of Algorithm 1. The main result depends on two reasonable assumptions in Supplementary Material Section A. Consequently, the following theorem holds for Algorithm 1 for solving the problem in Equation (3).

Theorem 1 (Convergence of Algorithm 1 to a Stationary Point). Under the technical assumptions listed in Supplementary Material Section A, the sequence $\{(\boldsymbol{\Theta}^t, \boldsymbol{s}^t, \boldsymbol{P}^t)\}_{t=1}^{\infty}$ in Algorithm 1 converges to a stationary point of $f(\boldsymbol{\Theta}, \boldsymbol{s}, \boldsymbol{P})$.

The proof of Theorem 1 relies on showing that it is a special case of the block successive upper-bound minimization (BSUM) algorithm given in [27]. Thus, the convergence is guaranteed by related BSUM theories. Doing so involved proving that the sub-level set $\mathcal{X}^0 = \{(\boldsymbol{\Theta}, \boldsymbol{s}, \boldsymbol{P}) : f(\boldsymbol{\Theta}, \boldsymbol{s}, \boldsymbol{P}) \le f(\boldsymbol{\Theta}^0, \boldsymbol{s}^0, \boldsymbol{P}^0)\}$ is compact, and the objective function in Equation (3) is coercive, as well as demonstrating the MM properties and strong convexity of the auxiliary functions used to derive the update steps in Algorithm 1. Details on the definitions of coercive functions and MM properties, as well as the proof of Theorem 1 are relegated to Supplementary Material Section C.

Results

Deconvolution performance benchmarks on pseudo-bulk samples

To assess the deconvolution performance of ARTdeConv in comparison to alternative methods, we conducted benchmarking simulation studies by evaluating it against two semi-reference-based methods: EPIC and SECRET, and three reference-free methods: NMF [25], debCAM [11], and LINSEED [10]. We considered the NMF application here semi-supervised and called it "Semi-NMF" due to the prior knowledge on $\boldsymbol{\Theta}$, which was used as the initial values for the basis matrix in NMF. BayICE and quanTIseq were not included due to the incompatibility of their inputs with the simulation setup.

In the first experiments, we generated a set of pseudo-bulk expression matrices using methodologies similar to those outlined in previous studies [2] by the following formula

$$\boldsymbol{Y}^* = \boldsymbol{\Theta}^* \text{diag}(\boldsymbol{s}^*) \boldsymbol{P}^* + \boldsymbol{\epsilon}. \tag{7}$$

Here, simulated cellular gene expression for $K = 5$ hypothetical cell types, CT1–CT5, were constructed with assumed prior knowledge on $K_0 = 4$ of them (CT1–CT4). We also directly simulated the true signature matrix $\boldsymbol{\Theta}^*$, instead of using aggregated purified bulk data or single-cell data, mimicking the common practice of using pre-constructed signature matrices in deconvolution applications [3, 28].

In total, $n = 200$ samples were simulated on $\boldsymbol{P}^*$. To investigate the effects of the true relative abundance of CT5 in bulk tissue samples on deconvolution results, three classes of $\boldsymbol{P}^*$ were generated, representing when CT5 was rare, uniform, and extra compared to other cell types in the tissue. Reference medians $(m_1, \ldots, m_5)$ and ranges $(r_1, \ldots, r_5)$ were obtained directly from each row of $\boldsymbol{P}^*$.

The expression values in the true signature matrix $\boldsymbol{\Theta}^*$ were simulated row-by-row (i.e. gene-by-gene), controlled by a parameter γ, which dictated how cell-type-specific each gene was on average. Each simulated gene was more likely to be cell-type-specific when $\gamma = 1$ compared to when $\gamma = 0$. The expression of $M = 2000$ genes was first created in a matrix $\boldsymbol{\Theta}_{\text{full}}^*$. Then, from these genes, a subset of $m = 1000$ highly cell-type-specific genes called marker genes were selected for ARTdeConv, Semi-NMF, EPIC, and SECRET. In addition, $\boldsymbol{s}^*$ (including s_5^*) was given to EPIC as for adjusting the mRNA amount per cell type. The method for the selection is described in details in Supplementary Material Section E.2. The expression of those selected genes was stored in $\boldsymbol{\Theta}^*$. The matrix $\boldsymbol{\Theta}_{K_0}^*$ consisting of the first four columns of $\boldsymbol{\Theta}^*$ was used as the partial signature matrix for ARTdeConv, EPIC, and SECRET, and the initial value in the basis matrix for Semi-NMF. Each gene was given the same unit weight in EPIC and SECRET during simulation, for all of the simulated gene expression were derived from the identical generative process. On the other hand, debCAM and LINSEED searched for marker genes differently from $\boldsymbol{\Theta}_{\text{full}}^*$ by looking for simplicial vertices of the vector space spanned by normalized bulk gene expression [10, 11].

The matrix of errors $\boldsymbol{\epsilon}$ was first generated for all of the $M = 2000$ simulated genes. Beginning with $(\boldsymbol{\Theta}_{\text{full}}^*, \boldsymbol{s}^*, \boldsymbol{P}^*)$, the error-free full bulk expression matrix was calculated as $\boldsymbol{\Theta}_{\text{full}}^* \text{diag}(\boldsymbol{s}^*) \boldsymbol{P}^*$. The relative strength of the errors (noises) to the mean expression of genes (signals) was controlled by another parameter σ. Two levels of noises with $\sigma = 0.1$ or 10 (low versus high) were introduced to evaluate the robustness of the methods to added noises. The final bulk expression matrix of all simulated genes was calculated using $\boldsymbol{Y}_{\text{full}}^* = \boldsymbol{\Theta}_{\text{full}}^* \text{diag}(\boldsymbol{s}^*) \boldsymbol{P}^* + \boldsymbol{\epsilon}$. From this, a sub-matrix $\boldsymbol{Y}^*$ corresponding to the genes in $\boldsymbol{\Theta}^*$ was selected. A detailed description of the probability distributions and their parameters in the generation of $\boldsymbol{\Theta}^*$, $\boldsymbol{s}^*$, $\boldsymbol{P}^*$, and $\boldsymbol{\epsilon}$ can be found in Supplementary Material Section E.1.

With necessary parts generated, 100 simulations were conducted for each combination of γ, σ, and CT-5 abundance class, where the deconvolution performance of all benchmarked methods was reported. The tuning grid and algorithm parameters for ARTdeConv are described in Supplementary Material Section E.3. Due to a lack of identifiability, estimated proportions from Semi-NMF, debCAM, and LINSEED were manually matched to different cell types, whose details are relegated to Supplementary Material Section E.4.

Sometimes, the assumptions of ARTdeConv can be challenged in practice. While the above setup in Equation (7) is in line with the model assumption in Equation (1) and is popular among previous studies on testing the reference-based methods [2, 3], recent findings suggest that the derived signature matrix may also come with non-negligible errors from the true underlying matrix [29, 30]. To assess the robustness of our model assumption under a combination of errors in the signature matrix and the bulk matrix, we conducted a second series of experiments, where the observed partial signature matrix $\boldsymbol{\Theta}_{K_0}^\circ$ was generated in the same fashion as that of the first experiments. The true partial signature matrix $\boldsymbol{\Theta}_{K_0}^*$

was perturbed from $\boldsymbol{\Theta}^{\circ}_{K_0}$ with an error term, whose mean and variance were related and controlled by a single positive parameter η. The larger the η, the more inaccurate the observed signature matrix is from the underlying true signature matrix. The pseudo-bulk $\boldsymbol{Y}^*$ was then calculated using Equation (7). In the simulations, we considered $\eta = 1, 5, 10, 100, 200$ together with $\sigma = 0.1, 10$ to investigate the deconvolution performance of the above-mentioned methods under various degrees of errors stemming from the signature matrix acquisition and the bulk data. The parameter γ was fixed at 1, for many signature matrices with highly cell-type-specific marker genes could be found in practice [2, 31]. Other simulation setups were kept the same as in the first experiments. Details on generating $\boldsymbol{\Theta}^{\circ}_{K_0}$ and $\boldsymbol{\Theta}^{*}_{K_0}$ can also be found in Supplementary Material Section E.1.

Another possible deviation from the assumptions for ARTdeConv comes from inaccurate reference cell type proportions, which can result in biased $\boldsymbol{m}$ and $\boldsymbol{r}$. As experiments measuring proportions of cell subsets grew abundant and more specific to tissue disease scenarios, it has become unlikely that the relative abundances of the K known cell types would be misrepresented with prior knowledge, which in turn would misguide ARTdeConv (i.e. a typically abundant cell type in a tissue would unlikely be measured as a rare cell type in prior experiments unless in extreme cases; passing such hugely inaccurate reference parameters to ARTdeConv would certainly bias the results) [3, 7]. Rather, the most likely sources of uncertainty are the absolute proportions of these K cell types and that of the unknown "super cell type" in a new tissue sample [32]. We investigated the impact of this uncertainty in reference parameters by assuming the relative proportions of the K known cell types were correct, but their absolute proportions and that of the $(K + 1)$-th cell type were not. To do so, we first generated all parts in Equation (7) as described above when $\boldsymbol{\Theta}^*$ was accurate. Then, we obtained the true medians $(m_1, \ldots, m_5)$ and ranges $(r_1, \ldots, r_5)$ from each row of $\boldsymbol{P}^*$. Next, we designated a parameter $\xi \in (0, 1)$ to denote the size of deviation from the true medians and ranges. For CT1–CT4, we let their observed medians and ranges to be shrunken from the truths by ξ and those of CT5 to be relatively inflated. The new observed $\boldsymbol{m}^* = (m_1^*, \ldots, m_5^*)$ and $\boldsymbol{r}^* = (r_1^*, \ldots, r_5^*)$ were passed to ARTdeConv during the simulations in place of the true reference medians and ranges. We considered $\xi = 0.05, 0.1, 0.2, 0.35, 0.5$ in our simulations. Obviously, the larger the ξ, the more our information about the absolute abundances of CT1–CT5 was different from the truths. We defer the exact details of the setup of as well as some light discussion on these simulation to Supplementary Material Section E.1.

Performance metrics for evaluating deconvolution results included the following: (a) $\frac{1}{K}\sum_{l=1}^{K}\mathrm{CCC}(\boldsymbol{p}_l^{*\top}, \hat{\boldsymbol{p}}_l^{\top})$, (b) $\frac{1}{5n}\sum_{i=1}^{n}\sum_{k=1}^{5}|p_{ki}^* - \hat{p}_{ki}|$, (c) $\mathrm{CCC}(\boldsymbol{p}_5^{*\top}, \hat{\boldsymbol{p}}_5^{\top})$, and (d) $\frac{1}{n}\sum_{i=1}^{n}|p_{5i}^* - \hat{p}_{5i}|$, where CCC denotes the concordance correlation coefficient (CCC) [33, 34] between two vectors. An advantage of CCC over Pearson's correlation is that CCC directly measures the agreement between two sets of values by penalizing deviations from the 45-degree line in a scatterplot. Among the four metrics, (a) and (b) delineated the overall deconvolution accuracy, while (c) and (d) described the deconvolution accuracy for CT5, the missing cell type.

Overall, when the assumptions are satisfied, ARTdeConv demonstrated superior performance compared to other semi-reference-based and reference-free methods in accurately recovering cell-type proportions, irrespective of the cell-type specificity of signature genes, the level of additive noise, and the relative abundance of the missing cell type compared to other cell types (Figs 2A and 3A). Notably, ARTdeConv also exhibited robust performance in estimating the proportions of CT5, especially when CT5 was relatively prevalent in the pseudo-bulk samples under high noise conditions, where other methods, except debCAM, showed reduced precision (Figs 2B and 3B).

Among the semi-reference-based methods, EPIC achieved high overall accuracy when CT5 was relatively rare in the pseudo-bulk samples. However, its performance in estimating the proportion of CT5 alone was surpassed by SECRET (Fig. 2). While the overall performance of EPIC was comparable to that of SECRET when CT5 had similar abundance to other cell types, EPIC's accuracy diminished when CT5 was relatively common among the pseudo-bulks. This is consistent with previous findings that EPIC tends to underestimate the proportions of cell types not characterized in the reference [14]. Although SECRET showed higher accuracy than EPIC in scenarios where CT5 was common, its overall performance remained inferior to ARTdeConv and did not significantly enhance the accuracy of CT5 estimates over EPIC, except when CT5 was relatively rare, where it performed slightly better than ARTdeConv (Figs 2 and 3).

Among the reference-free methods, debCAM demonstrated accuracy that exceeded those of semi-reference-based methods when CT5 was relatively common (Figs 2 and 3). As expected, debCAM's performance improved with highly cell-type-specific signature genes ($\gamma = 1$). Conversely, Semi-NMF performed well when CT5 had similar abundance to other cell types in the pseudo-bulk samples. LINSEED, known for its insensitivity to cell-type mRNA amounts, consistently showed the worst performance among the methods, except in a few specific cases.

ARTdeConv also showed decent robustness when its assumptions were challenged, particularly when inaccurate signature matrices were observed and applied. In our simulations, the overall performance of ARTdeConv did not decrease as η grew from 1 to 200, except for under the uniform abundance of CT5 and when η reached 200 (Supplementary Figs S1 and S2). In contrast, the other two semi-reference-based methods were somewhat prone to the misspecification of signature matrices, as we could observe a slight drop in performance as η grew in most scenarios (and a big drop in the case of EPIC when CT5 was relatively rare), although a performance gap between them and ARTdeConv still existed when η was small (Supplementary Figs S1 and S2). Notice that the performance of debCAM, a reference-free method, was also affected when η grew. This was due to the loss in the cell-type-specificity of genes in $\boldsymbol{\Theta}^*$ as η increases in magnitude per our simulation setup.

On the other hand, when the observed reference medians and ranges of relative proportions between CT1 and CT4 were accurate but the absolute proportions of all five cell types became more inaccurate as ξ grew from 0.05 to 0.5, ARTdeConv did suffer from a visible slide in performance. The slide was more severe when the true underlying proportions of CT5 were relatively small, but milder when true CT5 proportions were relatively large. In all cases, the loss in performance was not as pronounced when $\xi < 0.1$ (Supplementary Fig. S3). These results called for efforts in finding references that are accurate in the absolute scale, although more room for

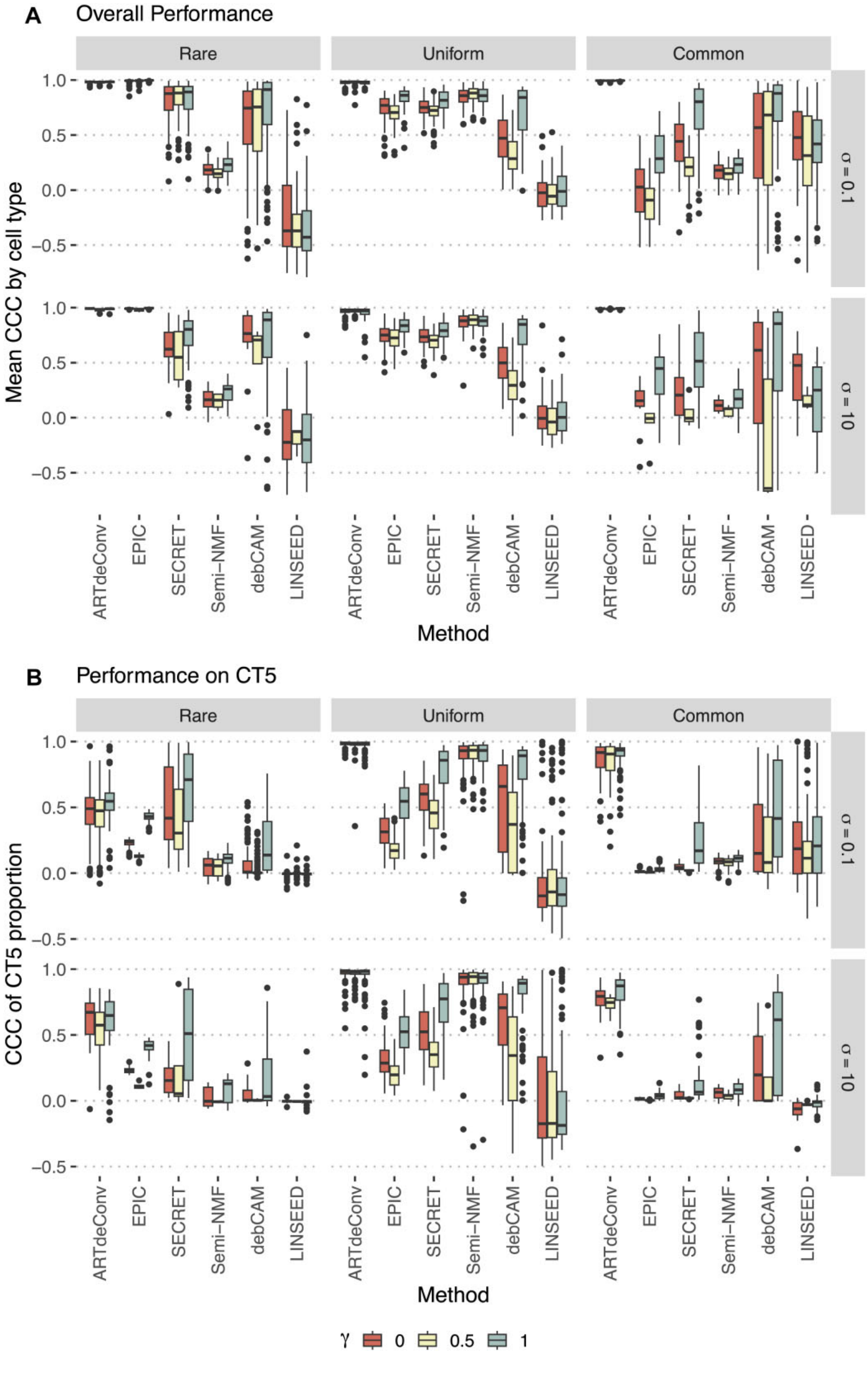


Figure 2. (**A**) Mean CCC by cell type between true proportions and estimated cell-type proportions from simulated pseudo-bulks. (**B**) CCC between estimated and true proportions of CT5 in the pseudo-bulks for the benchmark simulations. Each column represents a case of the relative abundance of CT5 against other cell types in the simulated pseudo-bulks. Each row represents the level of noises controlled by σ. Colors represent different levels of cell-type specificity of genes in $\mathbf{\Theta}^*$ regulated by different γ.

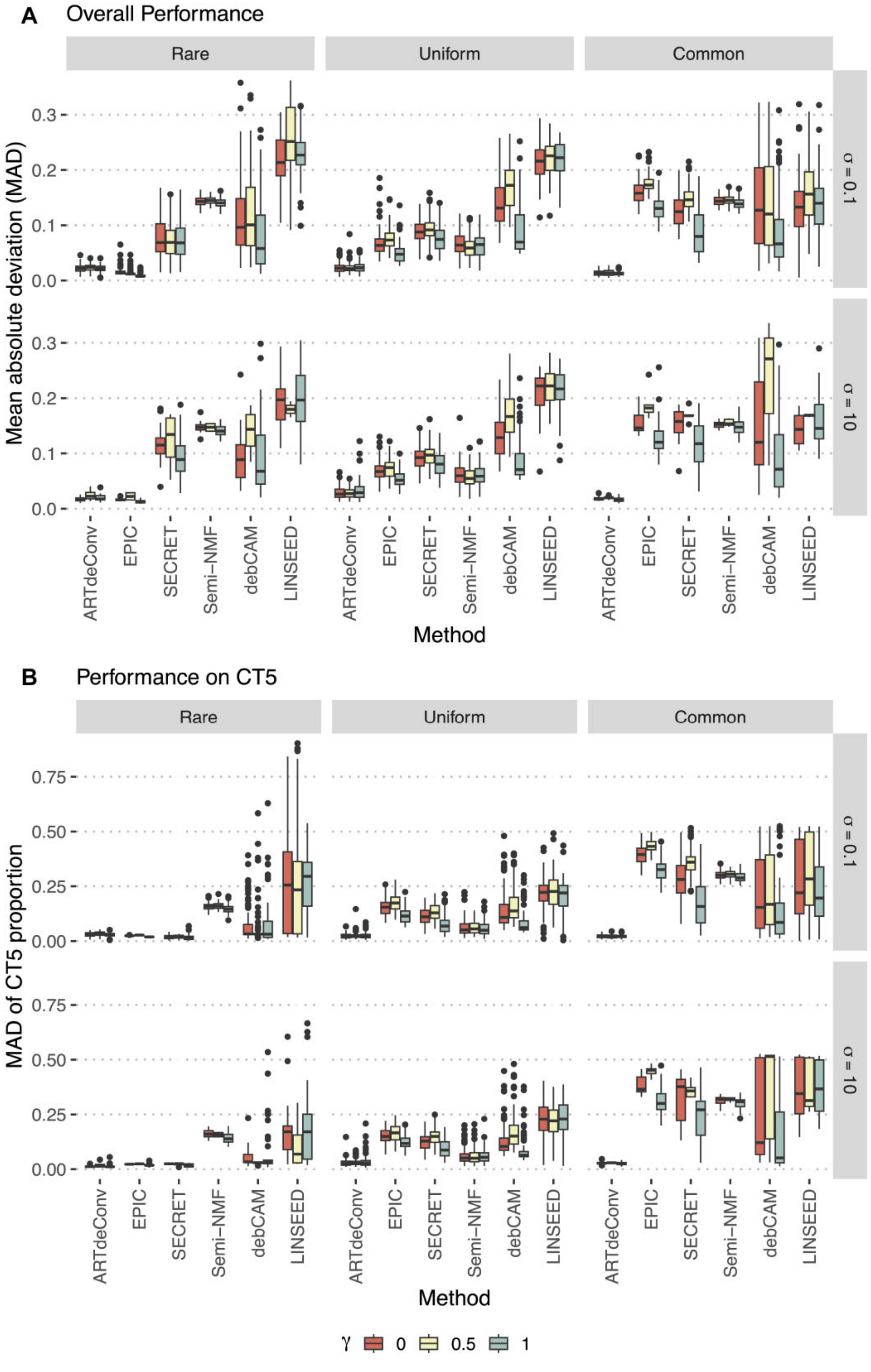


Figure 3. (**A**) Mean absolute deviation (MAD) between true proportions and estimated cell-type proportions from simulated pseudo-bulks. (**B**) MAD between estimated and true proportions of CT5 in the pseudo-bulks for the benchmark simulations. Each column represents a case of the relative abundance of CT5 against other cell types in the simulated pseudo-bulks. Each row represents the level of noises controlled by σ. Colors represent different levels of cell-type specificity of genes in $\mathbf{\Theta}^*$ regulated by different γ.

deviation from the accurate proportions is permitted when the relative proportion of the super cell type is large.

In conclusion, ARTdeConv showed superior performance to its semi-reference-based peers and reference-free methods in our simulations. ARTdeConv also demonstrated the ability to cope with deviation of the observed signature matrix from the true underlying $\mathbf{\Theta}^*$, but would decrease in performance if there are sizable inaccuracies in the reference absolute proportions. We further remark that, in the simulations described above, $\mathbf{\Theta}^*$ included knowledge of marker genes for CT5. In practice, obtaining the marker genes of cell types without reference expression is, while not as straightforward, achievable through methods such as in [35] and [36]. Furthermore, even without well-defined marker genes for those cell types, deconvolution by ARTdeConv is still completely feasible, for instance, the analyses in Sections 3.2 and 3.3.

ARTdeConv accurately estimates cell type proportions of PBMC from bulk gene expression in a human influenza vaccine study

To benchmark the performance of ARTdeConv on real-world data and exemplify its application, we utilized a dataset sourced from [37]. In this study, blood samples were obtained from healthy volunteer subjects who were administered a trivalent inactivated influenza vaccine (TIV). Both bulk peripheral blood mononuclear cell (PBMC) samples and sorted PBMC cell lines from two enrolled subjects ("HD30" and "HD31") vaccinated with a single dose of 2011 and 2012 seasonal TIV were collected at four different time points: before vaccination (Day 0), and on Day 1, 3, and 7 post-vaccination.

Deconvolution was performed on eight bulk PBMC samples whose gene expression was measured in transcript per million (TPM). Gene expression in TPM from the sorted PBMC cell lines of the two subjects on Day 0 was used to construct the partial signature matrix. The study characterized four distinct PBMC cell types: T cells, B cells, natural killer (NK) cells, and monocytes. Additionally, there were other unspecified cell types collectively termed into one super cell type as "others," which could include, for instance, various types of dendritic cells and low-density neutrophils [28, 38]. Detailed data processing steps could be referred to at Supplementary Material Section F.1.

The means and ranges of the cell types in question were obtained through the reference values in [38]. For each cell type, the proportion means were calculated by taking the average of the upper and lower bounds of the reference values, while the ranges were calculated by taking the difference between the two bounds. For the "others" cell type, the mean was calculated by subtracting the means of all other cell types from one, and the range was the difference between the upper and lower possible proportions of the other cell types. The exact values of the means and ranges are reported in Supplementary Table S1 (details of the calculation are given in Supplementary Material Section F.2). The vector of ranges was further normalized such that it has unit L_2 norm before deconvolution. Additional details regarding the tuning grid and algorithm parameter setup of this analysis can be found at Supplementary Material Section F.3.

The PBMC percentages on Day 0 were measured by flow cytometry [12, 37], which were used as the ground truths for evaluating the performance of ARTdeConv. Indeed, ARTdeConv achieved a notable degree of precision in estimating cell-type proportions on Day 0 as compared to the flow cytometry measurements (Fig. 4A), achieving a CCC of 0.974 (Pearson's correlation = 0.974, MAD = 0.036). This performance is on par with that of EPIC and surpassed those of CIBERSORT etc. as reported by [12].

We also estimated proportions of PBMC cell types across all time points utilizing the same partial signature matrix (Fig. 4C). On HD30, a decline of T cell abundances before Day 1 and an increase between Days 1 and 3 were observed, followed by a slight decline until Day 7. An increase of monocyte abundances before and a decrease after Day 1 were also seen. Both trends are consistent with the profile of a virus shedder of the H1N1 virus in [39] with a small time shift. The time shift could be attributed to differences in viral strains and strengths between [39] and [37], as the former is a study on real H1N1 patients. The lack of notable changes in the cell-type proportions of HD31 resembles more to the profile of a non-virus shedder. More information on the viral progression of these two subjects is needed to confirm these observations.

We wish to remark on the utility of adjusting for cell-type mRNA amounts in this case. It is known that normalization by TPM loses such information, which was corrected by ARTdeConv via diag($\hat{s}$). When re-running ARTdeConv on the same samples with diag(s) forced to be the identity matrix, we observed less accurate estimates of proportions (Fig. 4B), corroborating the need for the adjsuetment in this scenario.

ARTdeConv reveals changes in key PBMC cell type proportion in COVID-19 patients

We performed an extensive deconvolution analysis on PBMC bulk samples gathered from 17 healthy controls and 14 patients with COVID-19 diagnosis recruited for a systemic immunity assessment against COVID-19 infections in humans by [40]. The study also classified the COVID-19 severity of infected patients. Patients in the study were designated with three levels of severity: moderate, severe, and intensity care unit-hospitalized (ICU). Details regarding the patient recruitment and the classification of COVID-19 severity are in Supplementary Material Section G.1.

Bulk RNA-seq data downloaded using NCBI GEO accession number GSE152418 were utilized to perform deconvolution for estimating the proportions of the four major PBMC cell types: T cell, B cell, NK cell, and monocyte. The authors also included scRNA-seq data on separate independent blood samples from five healthy controls and seven COVID-19 patients. Of these, all healthy and six COVID-19 infected subjects had matching bulk and single-cell samples that passed quality control. The single-cell data were downloaded using GEO accession GSE155673 and were used to construct the partial signature matrix. Complete descriptions of the bulk and single-cell data, their quality control and pre-processing, as well as the creation of gene signature matrices can be found in Supplementary Material Section G.1.

Deconvolution was performed separately on samples from healthy controls and COVID-19 patients using bulk and signature matrices that matched the disease status. For each set of samples, we assumed knowledge on the reference expression of the four major PBMC cell types: T cell, B cell, NK cell, and monocyte. All other cell types' reference expression were assumed unkown and they were grouped into the super cell type "others." Since obtaining the reference means and ranges of the cell types separately for healthy and diseased populations

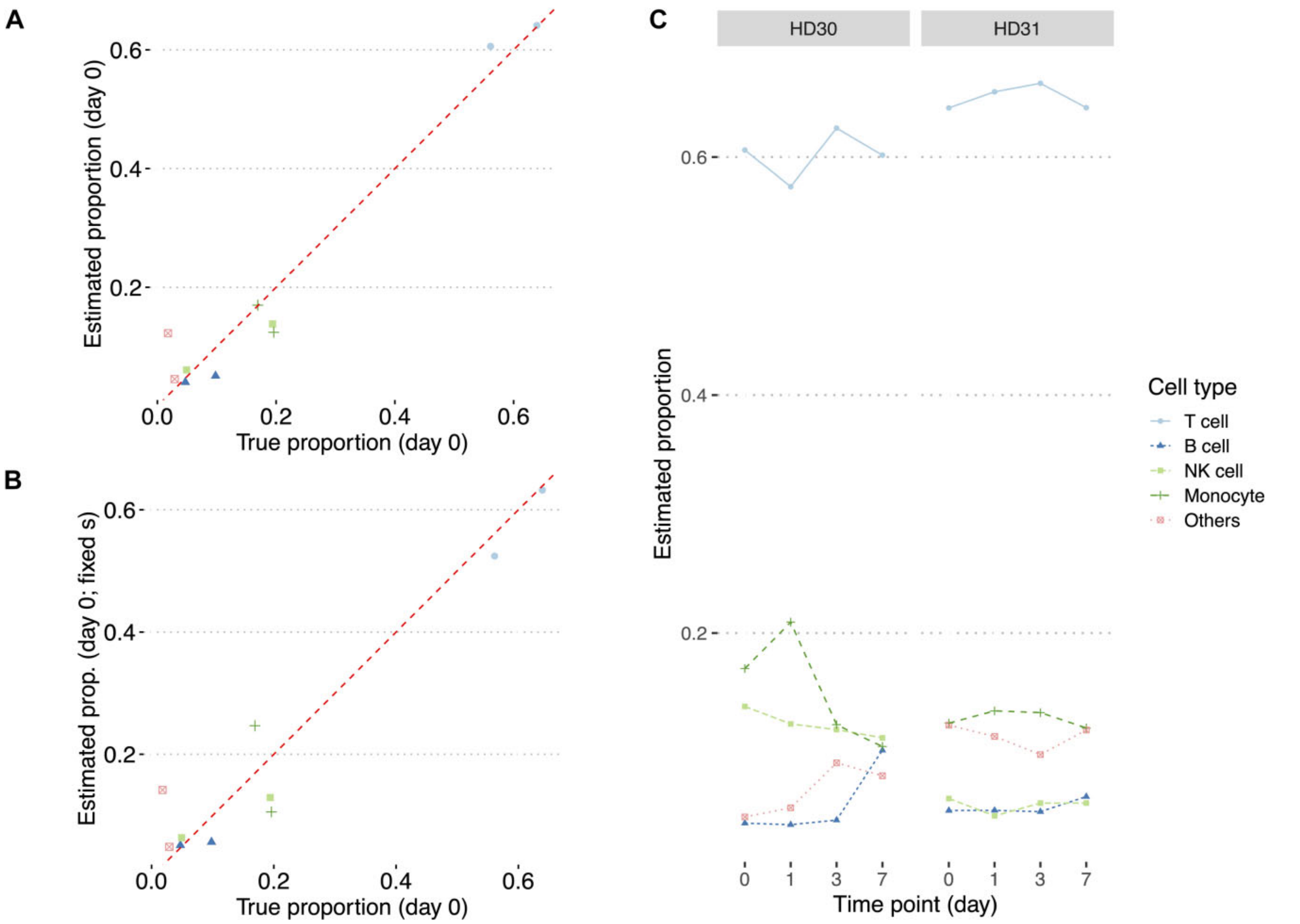


Figure 4. (**A**) Estimated PBMC proportions by ARTdeConv versus true proportions measured by flow cytometry for two PBMC samples collected on Day 0 with flexible mRNA amount parameters. (**B**) Estimated PBMC proportions by ARTdeConv versus true proportions measured by flow cytometry for two PBMC samples collected on Day 0 with mRNA amount parameters coerced to 1 for all cell types. (**C**) Estimated PBMC proportions by ARTdeConv on all of the eight PBMC samples from two subjects across time points with flexible mRNA amount parameters.

proved infeasible, we opted to use the same set of numbers in Supplementary Table S1 for running ARTdeConv. Detailed descriptions of the tuning grid and other ARTdeConv parameters can be found in Supplementary Material Section G.2.

The complete results of deconvolved cell type proportions from ARTdeConv are shown in Fig. 5A (boxplots) as well as Supplementary Figs S4 and S5 (bar charts). We observed that in COVID-19 samples, the proportions of T cells were notably lower than in the healthy control samples (Fig. 5A). Among the diseased samples, T cell depletion was found in several severe and ICU samples (Fig. 5B). This could be explained by T lymphopenia, commonly observed on blood samples of COVID-19 patients with severe symptoms but less frequently on those from patients with mild to moderate symptoms, as a result of the immunological responses of T cells to COVID-19 [41]. It was also observed that severe or ICU samples exhibited higher monocyte abundances in PBMC (Fig. 5C). Similar trends have also been observed by [42] in blood samples of patients with severe COVID-19.

Arunachalam *et al.* used the abundances of cells from the single-cell samples as a surrogate measurement for true abundances (except for dendritic cells, which were manually enriched in the single-cell samples) [40]. The relative deconvolved abundances of T cells, B cells, NK cells, and monocytes were then compared against the relative abundances of single cells. ARTdeConv demonstrated satisfactory accuracy in deconvolving cell-type abundances (Fig. 5D), achieving a CCC of 0.815 among healthy control samples (Pearson's correlation = 0.860, MAD = 0.098) and 0.694 among COVID-19 samples (Pearson's correlation = 0.717, MAD = 0.103).

While we observed a slightly lower deconovlution accuracy among COVID-19 patients compared to healthy controls, the COVID-19 patients spanned several disease severity classes and were under various duration of infection. These factors might increase the variation of gene expression, both at the cell-type and bulk levels, making deconvolution more challenging. The analysis also identified one outlier in healthy controls, S066 (Supplementary Fig. S5). Given that T cells are typically abundant in human PBMC samples, it was surprising that ARTdeConv did not detect any T cell in this sample, while attributing monocytes as the most abundant cell type. After further explorations in the expression of *CD3* and *CD14* genes, two of the experimentally validated marker genes of T cells and classical monocytes respectively [38], we discovered exceptionally low bulk *CD3* expression (Supplementary Fig. S6) and high *CD14* expression (Supplementary Fig. S7) in the sample from S066, which were consistent with and gave reasons to the low detected T cell abundance by ARTdeConv in this sample.

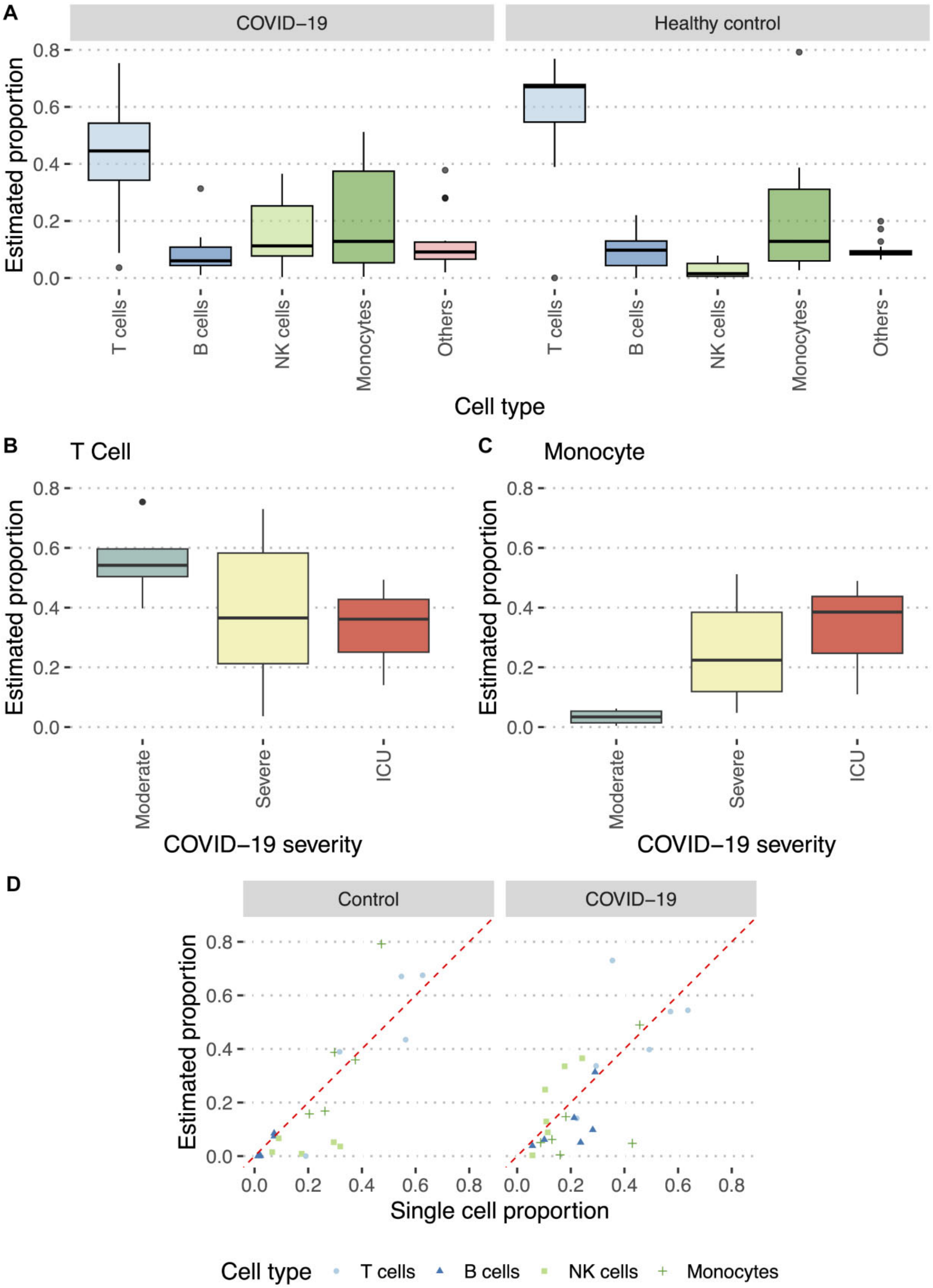


Figure 5. (**A**) Box plots for estimated PBMC proportions by ARTdeConv in separate deconvolution analyses for healthy control and COVID-19-infected samples. (**B**) and (**C**) Box plots for estimated T cell and monocyte proportions on COVID-19-infected samples of different severity. (**D**) Scatter plots for estimated PBMC proportions versus matching tissue PBMC proportions from independent single-cell studies from five healthy controls and six COVID-19-diagnosed subjects.

Discussion

In this paper, we introduce ARTdeConv, an innovative deconvolution approach. An important feature of ARTdeConv is its adoption of a tri-factor model, which integrates the cell-type mRNA amounts during the deconvolution process. As a semi-reference-based method, ARTdeConv offers enhanced flexibility compared to reference-based methods, as it accommodates cell types whose reference gene expression is not known by grouping them into one super cell type, while presenting advantages over reference-free methods by incorporating the partial signature matrix. Moreover, the method makes effective use of reference information on proportion means and ranges derived from external studies.

Additionally, we derive the MU algorithm for ARTdeConv and present a theorem that establishes the convergence of this algorithm to stationary points. This proof is derived by casting ARTdeConv's algorithm as a special case of the BSUM algorithm introduced by [27].

On simulated pseudo-bulks, we demonstrated the advantages of ARTdeConv over other semi-reference-based methods. Notably, ARTdeConv performed better compared to EPIC when the cell types without reference expression became relatively abundant and had an overall edge against SECRET and other well-known reference-free methods. We also showed that ARTdeConv was robust to challenges to its assumptions, particularly when inaccurate signature matrix or reference parameters were applied. Moreover, both EPIC and SECRET require manual re-normalization of the estimated proportions by relatively precise cell-type mRNA amounts, specifications that are not required for ARTdeConv. In practice, such precise amounts can be difficult to obtain when the cell types with missing reference expression become numerous. Compared to quanTIseq and BayICE, ARTdeConv can flexibly utilize customized partial signature matrices, which are common in deconvolution applications. To the best of our knowledge, current semi-reference-based methods only take in one cell type or super cell type whose reference expression is unknown. A direction for future research in methodology is developing novel semi-reference-based methods that can distinguish multiple cell types with missing reference, i.e. truly allowing $K > K_0 + 1$.

An argument can be made that obtaining precise reference information on proportions demands additional efforts in practice. However, the advantages of integrating this information are substantial. Without it, the task of associating estimated proportions with their respective cell types can prove challenging. First, the estimated proportions of the cell types without reference expression could be biased towards zero, as in the case of semi-reference-based methods like EPIC. Secondly, for all reference-free methods, we had to perform manual matching of results (as demonstrated in Section 3.1), which becomes infeasible if multiple cell types lack reference expression.

On the algorithmic side, ARTdeConv provides a guarantee of convergence, closing a gap in the theoretical characterization of deconvolution methods in previous studies. Of note, ARTdeConv does not assure that the estimated proportions are globally optimal. This challenge is not particular to ARTdeConv and is shared by other deconvolution methods with iterative optimization procedures such as SECRET, NMF, and debCAM. To counter this, we recommend employing marker genes for the deconvolution process and considering multiple restarts of ARTdeConv. While methods like debCAM mitigate this issue by using a different mathematical framework, their assumptions are usually breached in practice, and their performance diminishes as a result.

We demonstrated the application of ARTdeConv to two different sets of real data. While the data from [37] had limited sample size, we validated ARTdeConv's performance using the data from the more extensive study of [40]. We remark that both sets of data offer an edge where both the bulk samples and the independent samples for making partial signature matrices were collected from independent samples concurrently, minimizing the influence of most technical artifacts. Although we have illustrated the robustness of ARTdeConv in simulations, in parctical situations where bulk and cell-type reference expression are derived from different studies, additional caution against those technical artifacts should be exercised during data pre-processing prior to deconvolution.

Acknowledgements

The authors gratefully acknowledge the support of National Institute of Allergy and Infectious Diseases at the National Institutes of Health through the grant U19AI144181 and that of National Institute on Aging through the grants R01AG073259 and R01AI170959-01A1. The authors also thank the two anonymous reviewers for their valuable suggestions that greatly help to improve the quality of this article.

Author contributions: TL (Methodology, Investigation, Formal analysis, Data curation, Visualization, Software, Writing—original draft), CL (Formal analysis, Data curation, Writing—review & editing), QL (Methodology, Supervision, Writing—review & editing), XZ (Investigation, Supervision, Writing—review & editing), and FZ (Conceptualization, Methodology, Investigation, Supervision, Writing—review & editing)

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

National Institute of Allergy and Infectious Disease [19AI144181]; National Institute on Aging [01AG073259, 01AI170959-01A1].

Data availability

Raw sequencing data in the fastq format for the analysis in Section 3.2 can be accessed through Sequence Read Archive (SRA) BioProject PRJNA271578. Measured proportions for PBMC cell types on Day 0 samples can be found in the supplementary materials of [12]. Bulk and single-cell data for the analysis in Section 3.3 could be accessed from NCBI GEO using the accession numbers GSE152418 and GSE155673, respectively. ARTdeConv is released as an R package available for download from Zenodo through https://doi.org/10.5281/zenodo.15139554 or from GitHub through https://github.com/gr8lawrence/ARTDeConv.

References

1. Elloumi F, Hu Z, Li Y *et al.* Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. *BMC Med Genomics* 2011;**4**:54. https://doi.org/10.1186/1755-8794-4-54
2. Avila Cobos F, Alquicira-Hernandez J, Powell JE *et al.* Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun* 2020;**11**:5650. https://doi.org/10.1038/s41467-020-19015-1
3. Newman AM, Liu CL, Green MR *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;**12**:453–7. https://doi.org/10.1038/nmeth.3337
4. Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. *Cell* 2016;**165**:780–91. https://doi.org/10.1016/j.cell.2016.04.019
5. Li Z, Wu H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol* 2019;**20**:190. https://doi.org/10.1186/s13059-019-1778-0
6. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol* 2021;**22**:102. https://doi.org/10.1186/s13059-021-02290-6
7. Chen B, Khodadoust MS, Liu CL *et al.* Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol* 2018; **1711**:243–59. https://doi.org/10.1007/978-1-4939-7493-1_12
8. Morris JS, Kopetz S. Tumor microenvironment in gene signatures: critical biology or confounding noise? *Clin Cancer Res* 2016;**22**:3989–91.https://doi.org/10.1158/1078-0432.CCR-16-1044
9. Wigerblad G, Cao Q, Brooks S *et al.* Single-cell analysis reveals the range of transcriptional states of circulating human neutrophils. *J Immunol* 2022;**209**:772–82. https://doi.org/10.4049/jimmunol.2200154
10. Zaitsev K, Bambouskova M, Swain A *et al.* Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat Commun* 2019;**10**:2209. https://doi.org/10.1038/s41467-019-09990-5
11. Chen L, Wu CT, Wang N *et al.* debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. *Bioinformatics* 2020;**36**:3927–9. https://doi.org/10.1093/bioinformatics/btaa205
12. Racle J, de Jonge K, Baumgaertner P *et al.* Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* 2017;**6**:e26476. https://doi.org/10.7554/eLife.26476
13. Racle J, Gfeller D. EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data. *Methods Mol Biol* 2020;**2120**:233–48. https://doi.org/10.1007/978-1-0716-0327-7_17
14. Tai AS, Tseng GC, Hsieh WP. BayICE: a Bayesian hierarchical model for semireference-based deconvolution of bulk transcriptomic data. *Ann Appl Stat* 2021;**15**:391–411.
15. Finotello F, Mayer C, Plattner C *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med* 2019;**11**:34. https://doi.org/10.1186/s13073-019-0638-6
16. Lu Y, Chen QM, An L. Semi-reference based cell type deconvolution with application to human metastatic cancers. *NAR Genom Bioinform* 2023;**5**:lqad109. https://doi.org/10.1093/nargab/lqad109
17. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296. https://doi.org/10.1186/s13059-019-1874-1
18. Lytal N, Ran D, An L. Normalization methods on single-cell RNA-seq data: an empirical survey. *Front Genet* 2020;**11**:41. https://doi.org/10.3389/fgene.2020.00041
19. Wang X, Park J, Susztak K *et al.* Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* 2019;**10**:380. https://doi.org/10.1038/s41467-018-08023-x
20. Mohammadi S, Zuckerman N, Goldsmith A *et al.* A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc IEEE* 2017;**105**:340–66. https://doi.org/10.1109/JPROC.2016.2607121
21. Takami A, Watanabe S, Yamamoto Y *et al.* Reference intervals of white blood cell parameters for healthy adults in japan. *Int J Lab Hematol* 2021;**43**:948–58. https://doi.org/10.1111/ijlh.13486
22. Lee D, Seung HS. Algorithms for non-negative matrix factorization. *NIPS'00: Proceedings of the 14th International Conference on Neural Information Processing Systems*. 2000, **13**, 535–41.
23. Lin CJ. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Trans Neural Netw* 2007;**18**:1589–96. https://doi.org/10.1109/TNN.2007.895831
24. Lange K. MM optimization algorithms. Philadelphia, PA, USA: SIAM, 2016. https://doi.org/10.1137/1.9781611974409
25. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* 2010;**11**:367. https://doi.org/10.1186/1471-2105-11-367
26. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* 2013;**29**:2211–12. https://doi.org/10.1093/bioinformatics/btt351
27. Razaviyayn M, Hong M, Luo ZQ. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J Optim* 2013;**23**:1126–53. https://doi.org/10.1137/120891009
28. Monaco G, Lee B, Xu W *et al.* RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep* 2019;**26**:1627–40. https://doi.org/10.1016/j.celrep.2019.01.041
29. Menden K, Marouf M, Oller S *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv* 2020;**6**:eaba2619. https://doi.org/10.1126/sciadv.aba2619
30. Berson E, Sreenivas A, Phongpreecha T *et al.* Whole genome deconvolution unveils Alzheimer's resilient epigenetic signature. *Nat Commun* 2023;**14**:4947. https://doi.org/10.1038/s41467-023-40611-4
31. Dong M, Thennavan A, Urrutia E *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform* 2021;**22**:416–27. https://doi.org/10.1093/bib/bbz166
32. Newman AM, Steen CB, Liu CL *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**:773–82. https://doi.org/10.1038/s41587-019-0114-2
33. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;**45**:255–68. https://doi.org/10.2307/2532051
34. Huang P, Cai M, Lu X *et al.* Accurate estimation of rare cell type fractions from tissue omics data via hierarchical deconvolution. *Ann Appl Stat* 2024;**18**:1178–94. https://doi.org/10.1214/23-AOAS1829
35. Wang N, Hoffman EP, Chen L *et al.* Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci Rep* 2016;**6**:18909. https://doi.org/10.1038/srep18909
36. Qiu Y, Wang J, Lei J *et al.* Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. *Bioinformatics* 2021;**37**:3228–34. https://doi.org/10.1093/bioinformatics/btab257
37. Hoek KL, Samir P, Howard LM *et al.* A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. *PLoS One* 2015;**10**:e0118528. https://doi.org/10.1371/journal.pone.0118528
38. Kleiveland CR. Peripheral blood mononuclear cells. In: Verhoeckx K *et al.* (eds), *The Impact of Food Bioactives on Health: in vitro and ex vivo models [Internet]*. Springer, 2015, 161–7. https://doi.org/10.1007/978-3-319-16104-4_15

39. Rahil Z, Leylek R, Schürch CM *et al*. Landscape of coordinated immune responses to H1N1 challenge in humans. *J Clin Invest* 2020;**130**:5800–16.https://doi.org/10.1172/JCI137265
40. Arunachalam PS, Wimmers F, Mok CKP *et al*. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* 2020;**369**:1210–20. https://doi.org/10.1126/science.abc6261
41. Chen Z, John Wherry E. T cell responses in patients with COVID-19. *Nat Rev Immunol* 2020;**20**:529–36. https://doi.org/10.1038/s41577-020-0402-6
42. Zhang Y, Wang S, Xia H *et al*. Identification of Monocytes Associated with Severe COVID-19 in the PBMCs of Severely Infected Patients Through Single-Cell Transcriptome Sequencing. *Engineering (Beijing)* 2022;**17**:161–9. https://doi.org/10.1016/j.eng.2021.05.009

Received: July 17, 2024. **Revised:** March 2, 2025. **Editorial Decision:** April 2, 2025. **Accepted:** April 14, 2025