Briefings in Bioinformatics, 2026, 27, bbag184
https://doi.org/10.1093/bib/bbag184
Published: 20 April 2026
Problem Solving Protocol

OXFORD

Evaluating reference-mixture matching in cell-type deconvolution with single-cell RNA-seq references

Yifan Zhao [1], Brian E. Vestal [2,3], Camille M. Moore [1,2,3,*]

[1]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, 13001 E. 17th Place, Mail Stop B119, Aurora, CO 80045, United States
[2]Center for Genes, Environment, and Health, National Jewish Health, 1400 Jackson Street, Denver, CO 80206, United States
[3]Department of Immunology and Genomic Medicine, National Jewish Health, 1400 Jackson Street, Denver, CO 80206, United States
*Corresponding author. Center for Genes, Environment and Health, National Jewish Health, 1400 Jackson Street, Denver, CO, United States. E-mail: mooreca@njhealth.org

Abstract

Accurate estimation of cell-type composition from bulk RNA sequencing (RNA-seq) is critical for dissecting the cellular basis of disease. Single-cell RNA sequencing (scRNA-seq) has emerged as a preferred reference for cell-type deconvolution. However, in practice, scRNA-seq references often differ from the target bulk samples in terms of clinical condition or cohort composition, potentially degrading performance. Here, we systematically evaluate deconvolution methods under varying reference matching conditions, using a scRNA-seq dataset of peripheral blood mononuclear cells (PBMCs) from managed lupus patients and healthy controls. We constructed four scRNA-seq references: (i) 20 lupus patients, (ii) 20 healthy controls, (iii) 10 lupus patients + 10 controls, and (iv) 20 lupus patients + 20 controls. We simulated bulk RNA-seq mixtures with known cell-type proportions from scRNA-seq generated on independent lupus patients and healthy controls from the same study and evaluated the performance of seven cell-type deconvolution algorithms (CIBERSORTx S-mode, CIBERSORTx NS-mode, MuSiC2, InstaPrism, BLADE, DISSECT, Scaden) when using matched, mis-matched, and mixed scRNA-seq references. We also evaluated the performance on a publicly available bulk RNA-seq PBMC dataset with cell-type proportions estimated by flow cytometry. Performance is assessed via root-mean-squared error, Pearson correlation, and Lin's concordance correlation coefficients. Our results show that the choice of method has a greater impact on deconvolution accuracy than the degree of reference matching. DISSECT consistently achieved the best performance. Reference-matching effects were more pronounced for regression-based methods such as CIBERSORTx and MuSiC2. Overall, we recommend using robust methods like DISSECT and employing matched references when available.

Keywords cell-type deconvolution, cell-type proportions, reference mismatch, scRNA-seq, bulk RNA-seq simulation

Introduction

RNA sequencing (RNA-seq) has revolutionized transcriptomics by enabling transcriptome-wide quantification of gene expression in biological samples. While this technology provides a comprehensive snapshot of gene activity, bulk RNA-seq data inherently represent an averaged signal across all cells in a sample. This obscures the underlying cellular heterogeneity, which is critical to understanding many biological contexts, such as the tumor microenvironment, where interactions among diverse cell types drive disease progression and therapeutic response.

Traditionally, researchers have relied on experimental methods, such as flow cytometry and immunohistochemistry [1–7], to resolve cellular compositions within complex tissues. Although these techniques can provide detailed cellular profiles, they are often labor-intensive, costly, and may result in significant sample loss during processing. These limitations have motivated the development of computational approaches that infer cell-type proportions directly from RNA-seq data using existing reference datasets—a strategy known as cell deconvolution.

A wide range of deconvolution methods has been proposed, and these approaches generally fall into two categories [8]: reference-free and reference-based. Reference-free methods infer underlying cell types directly from bulk expression data using techniques such as matrix decomposition or unsupervised factorization. While these methods are attractive when no suitable reference is available, they often yield less accurate or less interpretable estimates. In this study, we focus on reference-based deconvolution methods, which leverage predefined cell-type expression profiles to guide estimation. The advent of single-cell RNA sequencing (scRNA-seq) has made this approach increasingly powerful. Unlike flow cytometry, scRNA-seq provides transcriptomic information at single-cell resolution, enabling the discovery of novel cell types and generating detailed

Received: December 9, 2025. **Revised:** March 18, 2026. **Accepted:** March 20, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

cell-type-specific profiles. These high-resolution reference datasets have catalyzed the development of numerous scRNA-seq–based deconvolution methods [1–7, 9–29]. However, the accuracy of deconvolution depends on the quality and relevance of the reference data. All deconvolution methods rely on the assumption that the reference profiles accurately reflect the cell types present in and the expression characteristics of the target samples. In practice, this assumption may often be violated: differences in health status, environmental exposures, or experimental conditions can lead to mismatches between reference and target datasets, introducing bias and reducing reliability. For example, using scRNA-seq references from healthy individuals to deconvolve bulk RNA-seq samples from lupus patients may fail to account for disease-associated shifts in cell-type-specific gene expression, particularly those driven by chronic inflammation and immune dysregulation.

Although numerous deconvolution methods exist and recent benchmarking studies have provided valuable insights into how scRNA-seq reference composition influences deconvolution performance [11, 21, 30], these evaluations typically assume that reference and mixture data are drawn from comparable biological conditions. As a result, mismatch between the biological conditions represented in the scRNA-seq reference and those present in the bulk RNA-seq data, a common and important real-world challenge, remains largely unexplored.

Here, we systematically evaluate the impact of such reference-mixture condition mismatches on the performance of six deconvolution methods using scRNA-seq data as references. By comparing matched and mismatched conditions, such as using healthy samples as references for lupus samples, we assess deconvolution accuracy even when relevant cell types are nominally shared between reference and bulk data but differ in their underlying biological state. This analysis identifies methods that are more robust to reference condition mismatch and provides practical guidance for applying deconvolution approaches in nonideal scenarios. Addressing these questions advances the reliable application of deconvolution in diverse biological contexts, especially when perfectly matched references are not feasible.

Materials and methods

Mathematical expression of deconvolution problem

Given a bulk gene expression matrix $\mathbf{B}$ of dimension $m \times n$, where m denotes the number of bulk samples (subjects) and n denotes the number of genes, the goal of cell-type deconvolution is to estimate a matrix $\mathbf{X}$ ($m \times c$) representing the cell-type proportions for c distinct cell types within each bulk sample. This relationship can be expressed as follows:

$$\boldsymbol{B}_{m\times n} = \boldsymbol{X}_{m\times c}\ \boldsymbol{S}_{c\times n} + \boldsymbol{\epsilon}_{m\times n}$$

where $\boldsymbol{S}$ is the $c \times n$ signature matrix, representing the gene expression profiles specific to each cell type and $\boldsymbol{\epsilon}$ is the error matrix.

In this framework

- $i \in \{1, \ldots, m\}$ indexes bulk samples,
- $j \in \{1, \ldots, n\}$ indexes genes, and
- $k \in \{1, \ldots, c\}$ indexes cell types.

Thus,

- X_{ik} represents the proportion of cell type k in sample i,
- S_{kj} represents the expression level of gene j in cell type k, and
- B_{ij} represents the observed bulk expression of gene j in sample i.

To ensure biological interpretability, two constraints are typically imposed:

- Non-negativity constraint: $X_{ik} \geq 0$ and $S_{kj} \geq 0$ for all i, j, k.
- Sum-to-one constraint: for each bulk sample i, i.e., $\sum_{k=1}^{c} X_{ik} = 1$

While many deconvolution methods explicitly assume the linear model $\boldsymbol{B}_{m\times n} = \boldsymbol{X}_{m\times c}\ \boldsymbol{S}_{c\times n} + \boldsymbol{\epsilon}_{m\times n}$, some approaches extend this framework. Bayesian methods retain the linear structure but impose priors on model components—e.g. placing Dirichlet priors on the cell-type proportions $X_{i\cdot}$ or Gaussian priors on gene-specific noise terms. These priors incorporate biological assumptions or regularization directly into the estimation procedure, yielding posterior distributions for $\boldsymbol{X}$ rather than point estimates. In contrast, deep-learning–based methods replace the explicit linear formulation with a learned nonlinear mapping $X = f_\theta(B, S) + \boldsymbol{\epsilon}_{m\times n}$, where f_θ is a neural network trained to approximate the relationship between bulk expression and cell-type proportions.

Different deconvolution algorithms estimate $\boldsymbol{X}$ and $\boldsymbol{S}$ using various optimization strategies, regularization schemes, and statistical assumptions, which could lead to differences in their robustness and sensitivity under reference mismatch conditions.

Methods evaluated

We surveyed existing methods for cell deconvolution that use scRNA-seq data as a reference. Supplementary Table 1 summarizes the identified 22 methods, their corresponding papers, websites, and package availability. A recent review by Wolfram-Schauerte *et al.* [8] also provides a comprehensive summary of deconvolution methods, and our categorization aligns closely with theirs.

The identified methods fall into three broad categories:

1) Regression-based methods: Representing the largest group, these methods model bulk RNA-seq data as a linear combination of reference gene expression profiles for individual cell types. Notable examples include CIBERSORTx (CSx), Bisque, DWLS, and MuSiC. CSx, in particular, has gained popularity for its use of support vector regression to address this problem.
2) Bayesian methods: These approaches incorporate scRNA-seq data as prior information within a probabilistic framework to estimate cell-type proportions and gene expression profiles from bulk RNA-seq data. By leveraging prior distributions derived from single-cell data, Bayesian methods can effectively model uncertainty and variability in the deconvolution process. Examples include BayesPrism and InstaPrism.
3) Deep-learning methods: These approaches leverage neural networks to capture complex, nonlinear relationships between bulk and single-cell data. Methods such as Scaden, DigitalDLsorter, and DISSECT exemplify this category.

We selected six methods, two from each methodological category, for evaluation based on popularity, recommendation from benchmarking studies [6, 7, 9–11, 25] and methodological features relevant to the specific question addressed here. Rather than conducting an exhaustive comparison of all available deconvolution approaches, our aim was to assess how reference-mixture condition mismatch affects representative methods spanning these different algorithmic classes. Each method is briefly summarized in Table 1, which lists the method category, software name, underlying algorithm, programming language, year of publication, most recent maintenance year, and

Table 1 Methods evaluated.

Category	Tool	Method	Program language	Year		Batch effect correction	
				Published	Last update	Subject-level	Cross-platform
Regression-based	CSx	Support vector regression (SVR)	R (Docker, web)	2019	2019	No	Yes
	MuSiC2	Weighted non-negative least squares (WNNLS)	R	2022	2024	Yes	No
Bayesian	BLADE	Hierarchical Bayesian model	Python	2021	2023	No	No
	InstaPrism	Topic modeling (simplified latent Dirichlet allocation)	R	2024	2024	No	No
Deep learning	DISSECT	Domain adaptation (GAN-inspired)	Python	2024	2025	Yes	No
	Scaden	Deep neural network (DNN)	Python (CML)	2020	2021	Yes	No

whether the method incorporates subject-level batch correction or cross-platform batch correction. Note that these two types of batch correction address different issues: subject-level correction accounts for variation across donors within the single-cell reference, whereas cross-platform correction addresses systematic differences between single-cell and bulk RNA-seq data. Although all methods implicitly accommodate the latter to some extent, CSx in S-mode is the only method that explicitly models cross-platform batch effects.

CSx

CIBERSORT [26] is a regression-based deconvolution method that estimates cell-type proportions in bulk RNA-seq data using a reference expression matrix derived from purified cell types. It selects informative genes for the signature matrix through a modified *t*-test, identifying those with high specificity across cell types. The method then applies ν-support vector regression (ν-SVR) to estimate cell fractions, allowing for tolerance in prediction error while maintaining robustness to multicollinearity and outliers. Since SVR does not inherently constrain outputs to be non-negative, CIBERSORT sets negative estimates to zero and renormalizes the proportions to ensure biological interpretability.

CIBERSORTx (CSx) [2] extends the original CS framework by enabling the use of single-cell RNA-seq data as a reference, offering additional features such as cross-platform batch correction to account for differences between scRNA-seq reference and bulk RNA-seq technologies. It also provides options to impute cell-type-specific gene expression profiles in the bulk data, enhancing downstream analysis. While it retains the ν-SVR model for deconvolution, CSx improves flexibility and accuracy by incorporating cross-platform alignment and noise modeling, making it more suitable for complex or heterogeneous datasets. Note that the cross-platform batch effect is different from the subject-level Batch effect in Table 1, which describes whether the algorithm is capable of adjusting for differences between batches or between different subjects present within the scRNA-seq reference. We used the CSx docker version for testing, with 100 permutations and fraction of 0, so all genes with detectable expression were included in the analysis. We tested its performance both with S-mode (S) and without S-mode (NS) batch correction.

MuSiC2

MuSiC [29] estimates cell-type proportions in bulk RNA-seq data by leveraging single-cell references from multiple donors. It uses a weighted non-negative least squares (NNLS) regression, where gene expression in the bulk sample is modeled as a linear combination of average cell-type-specific expressions from the single-cell data. To improve robustness, MuSiC incorporates gene-level weights that account for both intersubject variability and gene expression variance across cell types. This helps prioritize genes that are consistently expressed within a cell type across individuals, improving estimation accuracy.

MuSiC2 [18] builds upon MuSiC by addressing its key limitation: the assumption that all subjects share the same set of cell types. In real datasets, especially in disease contexts, certain cell types may be missing or altered in some individuals. MuSiC2 introduces an iterative subject screening algorithm to exclude genes that are differentially expressed between conditions from the scRNA-seq reference, refining the reference matrix to better match the bulk sample. This enhancement makes MuSiC2 more suitable for complex or heterogeneous tissues and improves performance when cross-subject variation in cell-type composition is present.

For MuSiC2, we used “music2_prop_t_statistics” to perform the deconvolution. We performed 20 resampling iterations using 50% of the data in each round, with significance thresholds of 0.05 for cell-type proportion differences and 0.01 for residual variance. We applied a small patch that fixes formats and gene-name alignment after subsetting; without this fix, MuSiC2 fails to execute.

BLADE

BLADE (Bayesian log-normal deconvolution) [31] is a hierarchical Bayesian model that can estimate the cell-type proportions and the cell-type-specific gene expression at the same time. It uses a log-normal probability model with normal-gamma priors for the expression level and a Dirichlet distribution for the cell-type proportions.

To define informative features, we selected the 10 000 most variable genes across the scRNA-seq reference, which were used as marker genes. All bulk samples were included in the analysis. All parameters were set to the default values recommended by the BLADE authors. Specifically, the grid search was performed over their default hyperparameter ranges (Alphas = [1, 10], Alpha0s = [0.1, 1, 5], Kappa0s = [1, 0.5, 0.1], and SYs = [1, 0.3, 0.5]), with three random restarts (Nrep = 3) and 10 repeated optimizations for the final parameter set (Nrep_final = 10).

InstaPrism

BayesPrism [32] is another Bayesian framework capable of jointly estimating both cell-type proportions and cell-type-specific gene

expression. Unlike BLADE, it models bulk gene expression counts using a topic modeling approach, specifically a multinomial distribution, with gene-level probabilities constructed as weighted mixtures of cell-type-specific expression profiles. To infer these proportions and expression profiles, BayesPrism employs Gibbs sampling, providing robust posterior estimates and uncertainty quantification. While effective for capturing biological heterogeneity, Gibbs sampling is computationally intensive, motivating newer variants like InstaPrism [21], which introduce faster, deterministic inference methods. To maintain a consistent reference across all methods, we did not use InstaPrism's built-in reference. Instead, we used the custom reference we constructed from the lupus and healthy scRNA-seq datasets (see Simulation Study section for details). The hyperparameters were set to the default.

Scaden

Scaden (single-cell assisted deconvolutional network) [33] is a supervised deep-learning method that estimates cell-type proportions from bulk RNA-seq by training a neural network on synthetic mixtures generated from single-cell RNA-seq data. It uses a feedforward neural network to learn nonlinear relationships between gene expression and cell-type composition. One key strength of Scaden is its ability to capture complex patterns in the data that traditional linear models might miss. It also performs well on benchmark datasets where the reference scRNA-seq used as training data closely matches the real bulk data. However, since it is supervised, Scaden's performance can degrade if the synthetic training mixtures don't reflect the characteristics of the target bulk samples. Additionally, like many deep-learning models, Scaden requires a large, well-prepared training set.

For Scaden, we generated 1000 simulated bulk samples per sample in the scRNA-seq reference to ensure training and test data were separated at the subject level. The simulated data were processed and used to train a model for 5000 steps, after which the trained model was applied to the bulk expression data for prediction.

DISSECT

DISSECT [23] builds on the idea of training with synthetic mixtures but addresses Scaden's main limitation—domain mismatch between synthetic and real bulk data. It employs adversarial domain adaptation inspired by generative adversarial network (GAN) training, where a feature extractor learns to produce domain-invariant representations while a discriminator attempts to distinguish between synthetic and real data. Through this adversarial process, DISSECT reduces distributional differences and improves generalizability to unseen datasets. This makes it particularly suitable when the available reference is imperfect or limited. However, the adversarial component increases model complexity and computational cost, and training can be less stable. In our implementation, we accounted for donor-specific batch effects and simulated 450 bulk samples per donor from the single-cell reference dataset.

Experimental design

Figure 1 illustrates the overall workflow of this study, encompassing lupus scRNA-seq data acquisition, reference construction, bulk simulation, analysis of real peripheral blood mononuclear cell (PBMC) RNA-seq data, deconvolution, and evaluation. The following sections describe each component in detail.

Simulation study

Our simulation is grounded in scRNA-seq data from a study of systemic lupus erythematosus (SLE) in which both cell-type proportions and cell-type-specific gene expression differ between healthy controls and lupus cases [34]. In the original study, SLE samples exhibited reduced CD4+ T cell proportions and increases in classical monocytes and proliferating cells, along with differential expression of over 300 genes, including widespread changes shared across all cell types and additional cell-type-specific effects in monocytes, T cells, and B cells. This combination of compositional shifts and within-cell-type transcriptional changes creates a challenging but biologically realistic deconvolution setting. In particular, disease-associated expression changes can distort cell-type signatures learned from healthy references, reducing the accuracy of proportion estimates even when the correct cell types are present.

We compared the performance of the six deconvolution methods described above on simulated PBMC mixtures from lupus cases and healthy controls with known cell-type proportions. First, we obtained the publicly available scRNA-seq dataset [34] containing 1.2 million PBMCs from 162 SLE cases and 99 healthy controls of either Asian or European ancestry (available from NCBI under GEO accession number GSE174188). We downloaded the processed h5ad file provided by the original authors, which included donor ID, disease status, age, and cell-type annotations. For our analysis, we used the author_cell_type column in the metadata. We excluded the PB and Progenitor (Progen) cell types and retained the nine most abundant immune cell types: T4, cM, B, T8, NK, ncM, Proliferating (Prolif), cDC, and pDC. To limit the number of conditions, we selected the subset of female subjects of European ancestry ($N = 62$ managed lupus cases, $N = 72$ healthy controls). These individuals ranged in age from 23 to 75 years (median: 32) in the healthy control group and from 27 to 71 (median 47) among lupus cases.

For each run of our simulation study, we created four reference single cell datasets. For each dataset, subjects were randomly sampled a single time, and the same selected subjects were used consistently throughout all downstream analyses.

1) Managed SLE ($N = 20$): scRNA-seq from 20 randomly selected managed SLE cases.
2) Healthy Controls ($N = 20$): scRNA-seq from 20 randomly selected healthy controls.
3) Combined ($N = 20$): 10 subjects were randomly selected from each of the managed SLE reference and the healthy control references described above.
4) Combined ($N = 40$): scRNA-seq from the managed SLE and healthy control references were combined into a single reference set with data from 40 subjects.

To generate simulated bulk RNA-seq mixtures with known cell-type proportions, 40 subjects were randomly selected from the managed SLE group and 40 from the healthy group that were not chosen as part of the reference sets. For each of these subjects, we summed gene expression counts across all cells present in their sample to create a "pseudo-bulk" mixture. A summary of the demographic information and cell-type composition for the selected 40 mixture and 20 reference

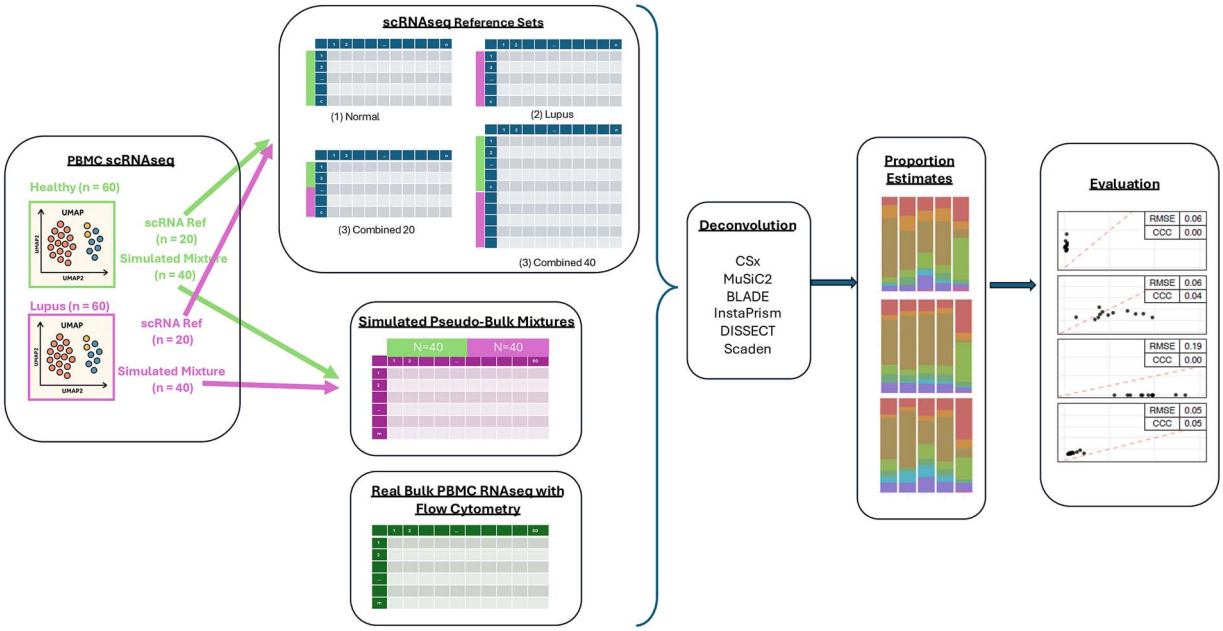

Figure 1 Overview of study design and evaluation framework. PBMC scRNA-seq data from 60 healthy donors and 60 lupus patients were used to construct reference and mixture datasets. From each condition, 20 samples were selected as scRNA-seq reference sets, and 40 were used to generate simulated pseudo-bulk mixtures. Four scRNA-seq reference sets were created: (i) healthy, (ii) lupus, (iii) combined 20 (10 healthy + 10 lupus), and (iv) combined 40 (20 healthy + 20 lupus). Both simulated pseudo-bulk mixtures and real bulk PBMC RNA-seq samples with matched flow-cytometry measurements were deconvolved using six representative methods: CSx, MuSiC2, BLADE, InstaPrism, DISSECT, and Scaden. Estimated cell-type proportions were then compared with known or experimentally measured values to evaluate deconvolution performance using metrics such as percent error, RMSE, Pearson correlation, and Lin's CCC. The simulation was repeated 10 times. For the real bulk dataset, we leveraged one split of the simulation references to mimic a real-life situation.

samples from each group, e.g. simulated dataset, for an example simulation run can be found in the Supplementary Table 2.

We then performed deconvolution for the mixtures using all four references, using the six deconvolution methods described above. This resulted in four reference match/mismatch scenarios:

1) Matched reference: scRNA-seq from the same patient type is used to deconvolute mixtures (i.e. healthy scRNA-seq reference paired with healthy mixtures, and managed SLE reference paired with managed SLE mixtures).
2) Mismatched reference: scRNA-seq from the opposite patient type is used to deconvolute bulk mixtures (i.e. healthy scRNA-seq paired with managed SLE mixtures and managed SLE reference paired with healthy mixtures).
3) Combined 20 reference: combined scRNA-seq from both patient types, using 20 subjects. We further down-sampled 10 donors from the selected pool to match the total number of individuals in the two reference types above.
4) Combined 40 reference: same as above, but using cells from all 40 subjects.

This simulation was repeated 10 times. This design enables evaluation of overall algorithm performance, the impact of reference match or mismatch, and the extent to which reference sample size affects deconvolution accuracy.

Application to real bulk RNA-seq mixtures

In order to further evaluate the impact of reference mismatch in real bulk RNA-seq mixtures, we extended our testing to a bulk RNA-seq PBMC dataset with cell-type proportions estimated by flow cytometry. This setup mimics a real-world scenario in which researchers use an external single-cell reference dataset to deconvolve new bulk samples. This dataset was used in a previous review [30], is accessible from GSE107011 [35], and included data from healthy male and female controls, 25–35 years of age. The raw data (fastq) were downloaded and aligned to reference hg19 from 10X website to more closely align with the reference used for the scRNA-seq. The data were processed through nf-core/RNA-seq 3.19.0 with NEXTFLOW ~ version 25.04.6. Counts were extracted using salmon by rounding to the nearest integers.

This dataset was paired with the same 4 scRNA-seq references and deconvolution methods as described above. However, we applied MuSiC instead of MuSiC2, since the available bulk dataset included only one condition.

Evaluation metrics

After performing cell-type deconvolution on the real and simulated bulk RNA-seq data using each set of scRNA-seq references, we assessed and compared the accuracy of the deconvolution methods in terms of error and percent error, root-mean-squared error (RMSE), Pearson correlation, and Lin's concordance correlation coefficient (CCC). Let X_{ik} be the true proportion and $\hat{X}_{\mathrm{ik}}$ be the estimated proportion. Each metric is defined as follows:

1) Error and percent error: the deviation of the estimate from the truth. Error was calculated separately for each cell type per subject.

$$Error_{ik} = \hat{X}_{\mathrm{ik}} - \mathrm{X}_{\mathrm{ik}}$$

$$Percent\ Error_{ik} = \frac{\hat{X}_{\mathrm{ik}} - \mathrm{X}_{\mathrm{ik}}}{\mathrm{X}_{\mathrm{ik}}} \times 100\%$$

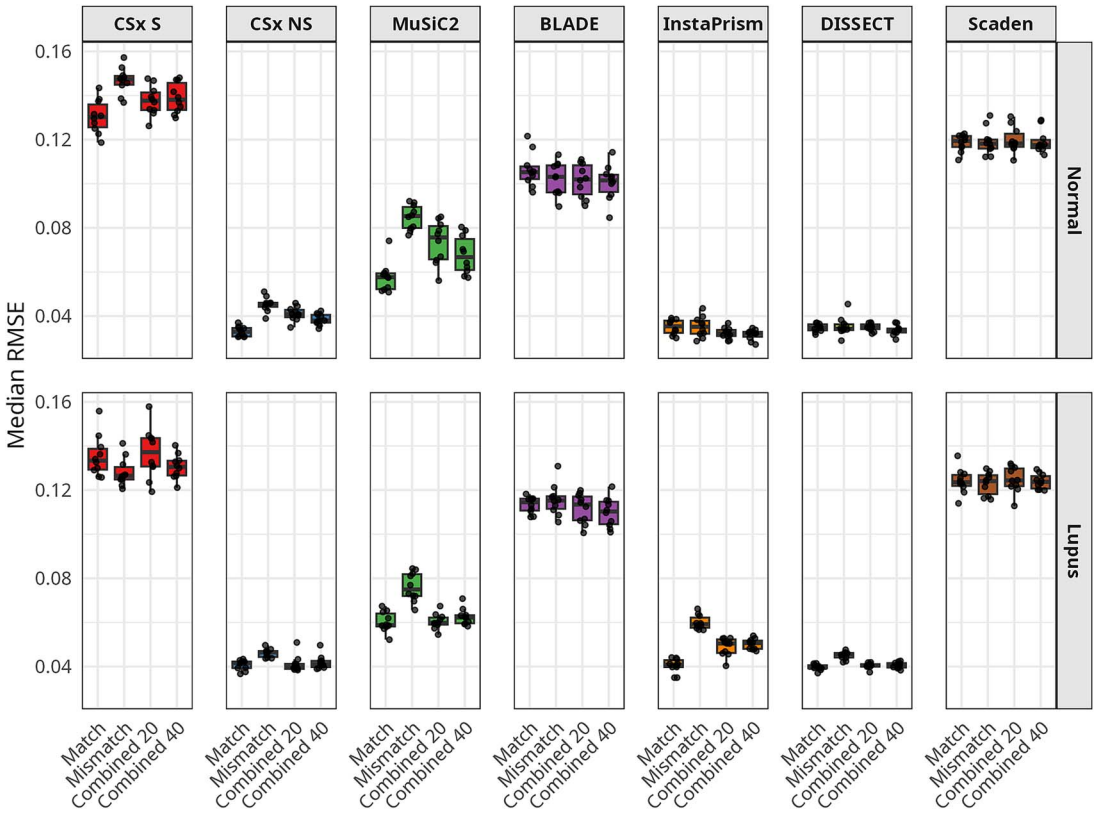

Figure 2 Distribution of median RMSE for the 10 simulation runs using different reference matching conditions for the simulated pseudo-bulk mixtures, shown on a shared *y*-axis. Box plots show the median RMSE across simulated bulk mixtures for each simulation for six deconvolution methods—CSx (S and NS), MuSiC2, BLADE, InstaPrism, DISSECT, and Scaden—under four reference configurations: matched, mismatched, combined 20, and combined 40. Results are presented separately for mixtures derived from healthy control (normal) and SLE (lupus) scRNA-seq data. Each dot represents the median RMSE across bulk samples for a single simulation. Plots share a common *y*-axis, highlighting differences between methods. Lower RMSE indicates higher deconvolution accuracy. DISSECT and InstaPrism achieved the lowest RMSE across most conditions.

2) RMSE: the square root of the expected squared deviation of the estimate from the truth. RMSE is influenced by both bias and variance of the estimates. We calculated RMSE across all cell types for each subject as an overall measure of accuracy for each subject.

$$RMSE_i = \sqrt{\frac{1}{c}\sum_{k=1}^{c}\left(\hat{X}_{ik} - X_{ik}\right)^2}$$

3) Pearson correlation (r): the strength of the linear relationship between the estimate and the truth for each cell type. We calculated the correlation across all subjects for each cell type.

$$r_k = \frac{Cov\left(\hat{X}_{\cdot k},\ X_{\cdot k}\right)}{\sigma_{\hat{X}_{\cdot k}}\sigma_{X_{\cdot k}}}$$

4) Lin's CCC: Concordance or agreement between the estimate and the truth for each cell type. CCC is influenced by both bias and correlation.

$$CCC\left(X_{\cdot k}, \hat{X}_{\cdot k}\right) = \frac{2r\,\sigma_{X_{\cdot k}}\sigma_{\hat{X}_{\cdot k}}}{\sigma^2_{X_{\cdot k}} + \sigma^2_{\hat{X}_{\cdot k}} + \left(\mu_{X_{\cdot k}} - \mu_{\hat{X}_{\cdot k}}\right)}$$

Results

Simulation study

Figure 2 summarizes the median RMSE across the 10 simulations by method and reference–mixture matching conditions, providing an overall comparison of deconvolution accuracy. Overall, CSx S-mode and Scaden consistently produced the largest RMSE values across conditions, indicating reduced accuracy in estimating cell-type proportions. DISSECT achieved the lowest RMSE, followed by InstaPrism and CSx NS, which showed comparable performance.

Figure 3 displays the RMSE across simulations but uses a method-specific *y*-axis to highlight differences in performance across the reference conditions. For the regression-based methods (CSx S, CSx NS, and MuSiC2), RMSE was generally higher under mismatched than matched reference conditions. An exception occurred for lupus mixtures analyzed with CSx S-mode, where mismatched references yielded slightly lower RMSE than matched, though the distributions of RMSE for matched and mismatched references are overlapping. For the combined references, median RMSE values were intermediate between match and mismatch cases. These results suggest that a matched reference is optimal for regression-based methods, whereas mixed-condition references are acceptable and opposite-condition references are least favorable. Increasing the reference size

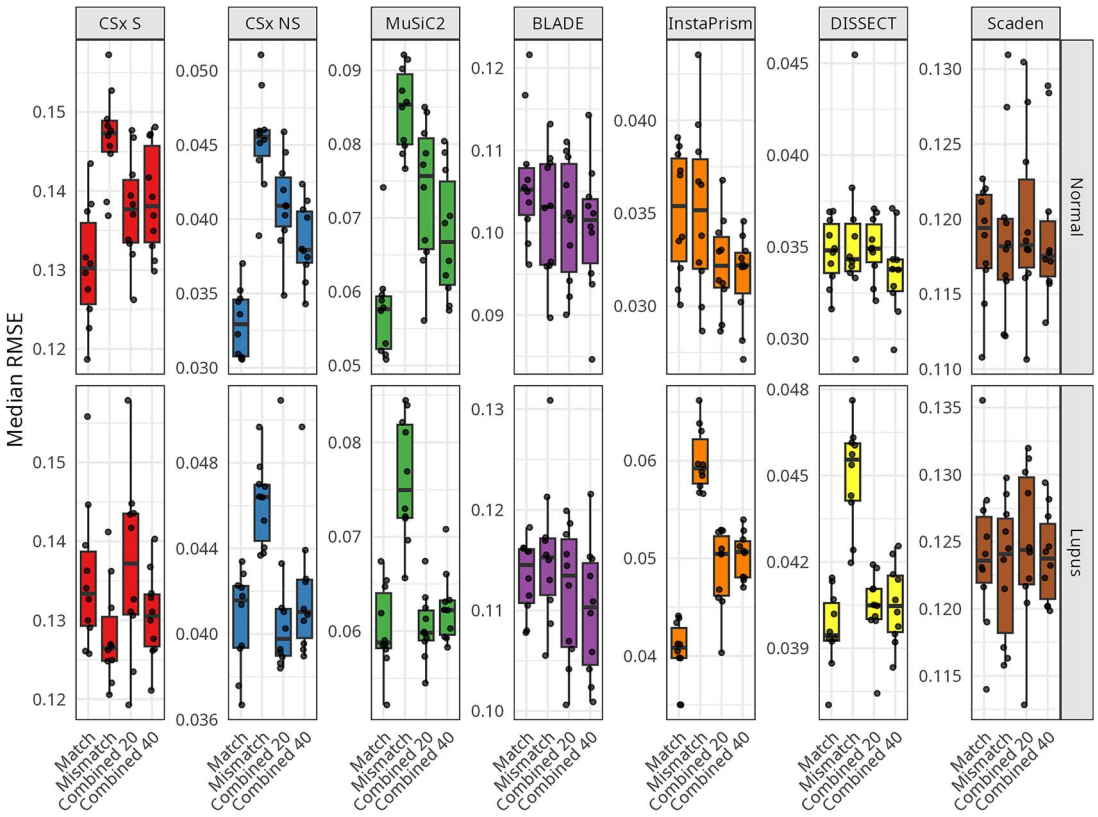

Figure 3 Distribution of median RMSE for the 10 simulation runs using different reference matching conditions for simulated pseudo-bulk mixtures, shown on a method-specific *y*-axis. Box plots show the median RMSE across simulated bulk mixtures for each simulation for six deconvolution methods—CSx (S and NS), MuSiC2, BLADE, InstaPrism, DISSECT, and Scaden—under four reference configurations: matched, mismatched, combined 20, and combined 40. Results are presented separately for mixtures derived from healthy control (normal) and SLE (lupus) scRNA-seq data. Each dot represents the median RMSE across bulk samples for a single simulation. Plots have a method-specific *y*-axis, highlighting differences between references within a method. Lower RMSE indicates higher deconvolution accuracy. Regression-based methods such as MuSiC2 and CSx (NS) showed greater sensitivity to reference mismatch.

from 20 to 40 samples did not consistently improve performance, indicating that larger reference panels alone do not guarantee higher accuracy.

For the Bayesian and deep-learning methods (BLADE, InstaPrism, DISSECT, and Scaden), reference–mixture matching status had less impact on RMSE; however, there were still some notable differences. RMSE was similar between matched and mismatched references for the healthy mixtures, but for the lupus mixtures, DISSECT and InstaPrism had higher RMSE using the mismatched reference. This may be due to lupus samples exhibiting disease-specific gene expression changes that are not captured by healthy scRNA-seq references. The healthy mixtures may be less affected by reference mismatch since the lupus reference retains the core baseline transcriptional programs of shared cell types. Additionally, increasing the reference size from 20 to 40 samples modestly reduced RMSE for BLADE and InstaPrism, suggesting a benefit from larger reference panels.

Figure 4 presents heatmaps of Lin's CCC for the simulated mixtures by cell type. While RMSE captures squared deviations, Lin's CCC jointly assesses bias and correlation. However, both measures led to broadly similar conclusions. Heatmaps of Pearson correlation can be found in Supplementary Fig. 1. Across most cell types, Scaden, BLADE, and CSx S-mode showed poor concordance, with CCC values near zero, whereas DISSECT achieved the highest CCC and InstaPrism produced comparable results. For the regression-based methods (CSx NS and MuSiC2), mismatched references generally reduced CCC relative to matched ones, consistent with their higher RMSE under mismatched conditions. DISSECT and Scaden also exhibited reduced CCC under mismatch in certain scenarios. The level of impact of reference-mixture mismatch on CCC was cell type-specific, with the most pronounced effect for immune populations known to exhibit disease-associated transcriptional changes in lupus. In particular, lower concordance under reference mismatch was observed for B cells in lupus mixtures (InstaPrism), for CD8+ T cells in both normal and lupus mixtures (DISSECT, InstaPrism, MuSiC, and CSx NS), and for nonclassical monocytes in lupus mixtures across multiple methods. Increasing the reference size again had minimal effect.

Finally, certain cell types such as cDCs displayed near-zero CCC regardless of method or matching status. Interestingly, MuSiC2 and CSx NS achieved relatively higher CCC for cDCs than DISSECT. Examining Pearson correlations, DISSECT exhibited stronger correlations between true and estimated cDC proportions (Supplementary Fig. 1) than either MuSiC2 or CSx NS, indicating that DISSECT more accurately captured the relative ordering of samples from low to high cDC abundance. The lower CCC observed for DISSECT reflects systematic bias in the estimated proportions (i.e. consistent overestimation), rather than a failure to track biological variation across samples. In contrast,

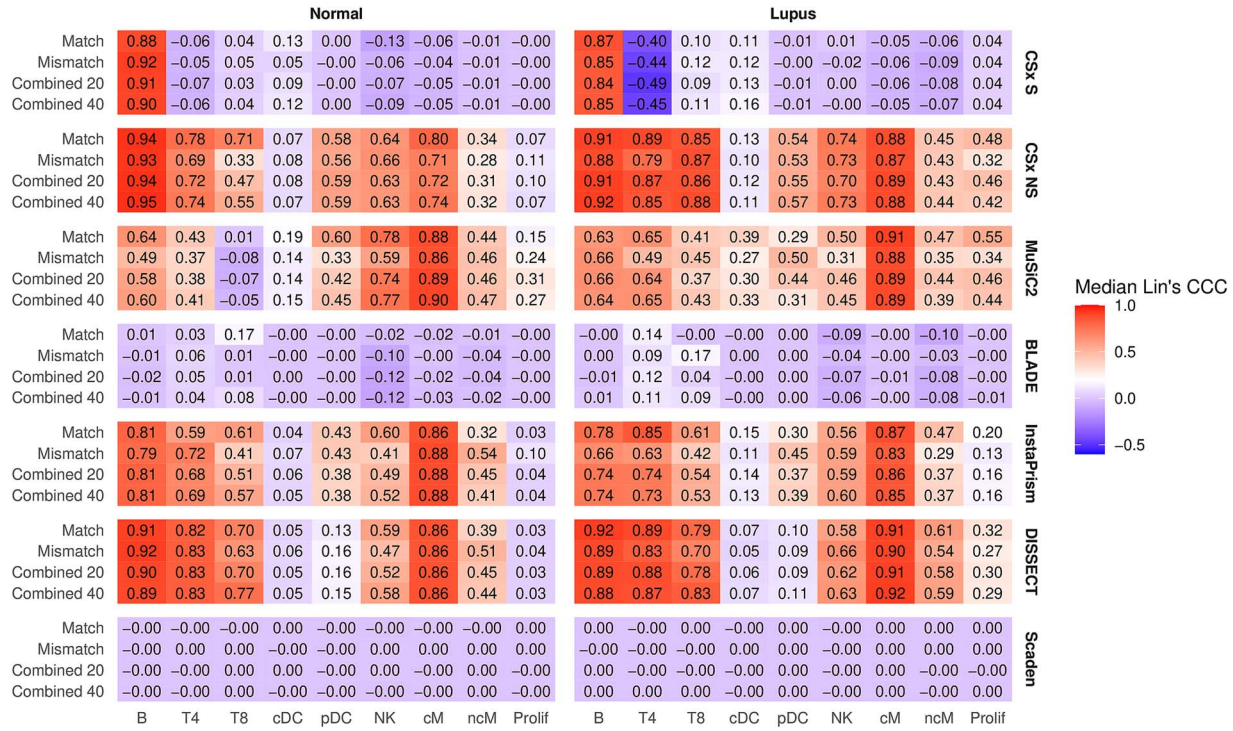

Figure 4 Median Lin's CCC across the 10 simulation runs using different reference matching conditions for simulated pseudo-bulk mixtures. Heatmaps show the median Lin's CCC across the 10 simulation runs for six deconvolution methods—CSx (S), CSx (NS), MuSiC2, BLADE, InstaPrism, DISSECT, and Scaden—across four reference configurations (rows: matched, mismatched, combined 20, combined 40) and eight major PBMC cell types (columns). Results are presented separately for simulated healthy control (normal) and SLE (lupus) mixtures. Higher values indicate higher concordance between estimated and true cell-type proportions. Regression-based methods such as MuSiC2 and CSx were typically more sensitive to reference mismatch.

MuSiC2 and CSx NS produced estimates with less bias for cDCs in this dataset, which improved CCC despite weaker correlation and poorer overall performance. Additional boxplots of the percent error for each cell type (Supplementary File 1) and scatter plots of true versus estimated proportions under each reference-mixture condition (Supplementary Figs 3–11) can be found in the Supplementary Materials.

Run time and computational resources

Median runtimes (minutes) across 10 data splits varied substantially among methods (Table 2). Across the 20-subject reference settings (Managed, Normal, and Combined 20), InstaPrism was consistently the fastest (2.2–2.5 minutes), followed by DISSECT and CSx NS (<50 minutes). Scaden required moderate computational time (137.5–153.9 minutes). In contrast, CSx S, MuSiC2, and BLADE were substantially more computationally intensive, with MuSiC2 showing the longest runtimes among 20-subject references (436.2–541.8 minutes).

Increasing the reference size from 20 to 40 subjects (Combined40) affected methods differently. BLADE and DISSECT showed minimal change in runtime. CSx NS, InstaPrism, and Scaden exhibited approximately two-fold increases, consistent with near-linear scaling. In contrast, CSx S and MuSiC2 demonstrated more than two-fold increases, indicating superlinear growth in computational cost with reference expansion.

Reported runtimes exclude method-specific input data management. Analyses were conducted on a SLURM-based high-performance computing cluster using 6 CPU cores and 256 GB RAM per job unless otherwise specified (BLADE: 20 CPU cores; CSx S Combined 40: 512 GB RAM).

Real bulk dataset

We next evaluated the performance of the deconvolution methods on real bulk PBMC RNA-seq data from healthy adults with known cell-type proportions measured by flow cytometry. This evaluation reflects a realistic application in which investigators deconvolve bulk RNA-seq samples using an external single-cell reference. Figures 5 and 6 show the distribution of sample-level RMSE across reference conditions. Because the real bulk dataset included only a single condition, MuSiC (rather than MuSiC2) was used for this analysis, as MuSiC2 requires at least two bulk conditions for iterative adjustment. RMSE values were generally higher in the real bulk RNA-seq mixtures than in the simulated mixtures, likely reflecting additional technical and biological variability between the scRNA-seq reference and bulk RNA-seq mixtures. Consistent with the simulation results, DISSECT achieved the lowest RMSE across all reference conditions, with median values consistently below 0.1. Interestingly, Scaden, which performed poorly on the simulated data, produced the second-lowest RMSE in the real bulk setting. However, inspection of scatterplots between flow cytometry and Scaden-estimated proportions revealed that Scaden predicted nearly identical cell-type compositions across samples (Supplementary Figs 12–15), explaining the deceptively low RMSE. CSx S, CSx NS, BLADE, and InstaPrism showed intermediate performance, while MuSiC consistently produced the highest RMSE (0.1–0.15). Additional boxplots of the percent error for each cell type can be found in Supplementary File 2.

Table 2 Median runtime (IQR) in minutes across 10 data splits for each method and reference setting in the simulation study. Reported runtimes exclude method-specific input data management. Simulations were performed on a SLURM-based high-performance computing cluster using six CPU cores.

Method	Managed	Normal	Combined 20	Combined 40
CSx (NS)	43.2 (40.9–46.9)	48.8 (44.3–51.4)	44.2 (41.1–51.1)	97.0 (90.0–98.9)
CSx (S)	211.2 (196.1–217.2)	248.1 (213.4–263.7)	223.2 (207.7–250.6)	611.5 (564.5–624.8)
MuSiC2	436.2 (415.8–453.4)	541.8 (519.3–574.0)	473.5 (433.7–537.3)	2164.4 (2019.9–2303.8)
BLADE	932.1 (872.8–1001.4)	832.5 (806.3–875.9)	895.3 (847.4–948.3)	900.3 (844.0–948.0)
InstaPrism	2.2 (2.1–2.3)	2.5 (2.2–2.8)	2.3 (2.1–2.5)	4.7 (4.5–4.8)
DISSECT	39.4 (37.9–40.2)	40.4 (39.8–41.5)	40.3 (39.9–40.8)	47.4 (47.1–48.5)
Scaden	137.5 (135.7–150.0)	153.9 (142.1–166.0)	146.5 (136.4–157.1)	261.3 (256.0–268.6)

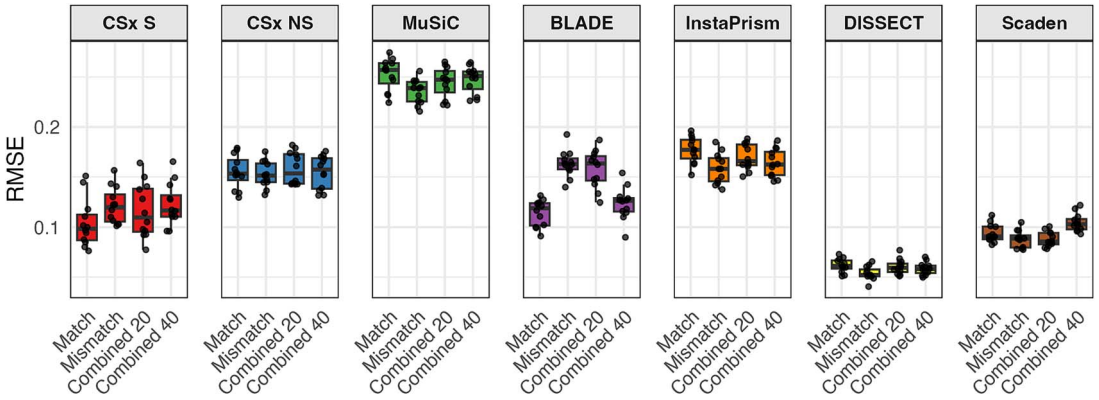

Figure 5 Distribution of sample-level RMSE in the real bulk PBMC RNA-seq data under different reference matching conditions, shown on a shared *y*-axis. Box plots show the distribution of RMSE for six deconvolution methods—CSx (S), CSx (NS), MuSiC, BLADE, InstaPrism, DISSECT, and Scaden—applied to real PBMC bulk RNA-seq data benchmarked against flow-cytometry-derived cell-type proportions. Four reference configurations (normal, managed, combined 20, and combined 40) were evaluated. DISSECT and InstaPrism maintained the lowest RMSE across most reference conditions, while regression-based methods such as MuSiC and CSx showed greater variability in performance. Each dot represents the RMSE for a single sample. Plots have a shared *y*-axis, highlighting differences between methods.

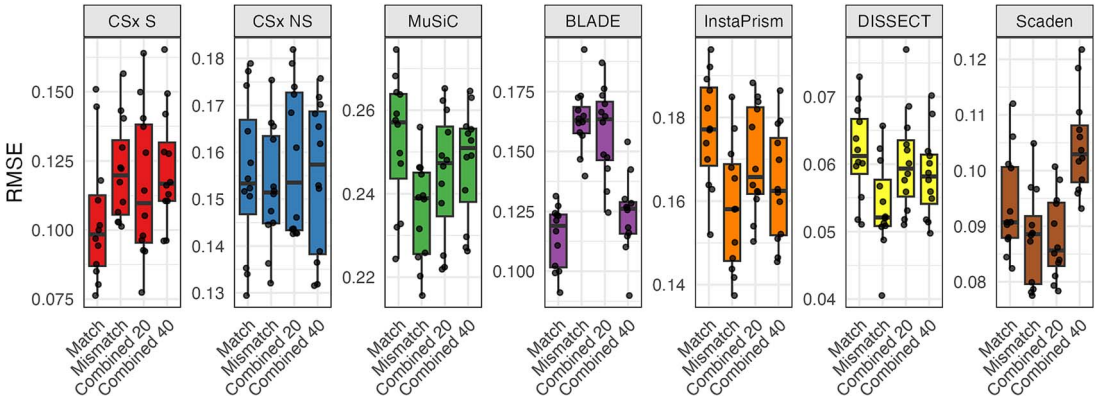

Figure 6 Distribution of sample-level RMSE in the real for real bulk PBMC RNA-seq data under different reference matching conditions, shown on a method-specific *y*-axis. Boxplots show the distribution of RMSE values for six deconvolution methods—CSx (S), CSx (NS), MuSiC, BLADE, InstaPrism, DISSECT, and Scaden—across four reference configurations (normal, managed, combined 20, and combined 40). Lower RMSE values indicate higher deconvolution accuracy. Consistent with Fig. 5, DISSECT and InstaPrism achieved the best overall performance, while regression-based methods such as MuSiC and CSx exhibited greater variability across reference conditions. Each dot represents the RMSE for a single sample. Plots have a method-specific *y*-axis, highlighting differences between references within a method.

Counterintuitively, several methods, including MuSiC, InstaPrism, DISSECT, and Scaden, showed lower RMSE under the mismatched reference. There are several factors that may contribute to these results. First, in the real bulk dataset, the reported flow cytometry proportions do not sum to one (with an average total of approximately 0.9). In contrast, all deconvolution methods evaluated enforce proportions that sum to one. This mismatch introduces an unavoidable source of error that directly affects RMSE and Lin's CCC, making these metrics less informative in this setting and potentially favoring methods that distribute error differently across cell types. In addition, differences between the real bulk dataset and the scRNA-seq reference extend beyond disease status, including age, sex, and race distributions. Thus,

any modest improvement observed under mismatch should be interpreted cautiously.

For concordance with flow cytometry proportions (Fig. 7), we focus on Pearson correlation instead of Lin's CCC (Supplementary Fig. 2), since the flow cytometry proportions did not sum to one. Similar to the simulated mixtures, DISSECT again outperformed all other methods, achieving the highest correlations across most cell types (often exceeding 0.7 for common populations). InstaPrism followed, with moderate correlations (0.5–0.7) for several cell types including B, T4, cDC, and cM. CSx S exhibited variable performance, with correlations ranging from positive to negative depending on the cell type, whereas BLADE and Scaden performed poorly, showing near-zero or negative correlations across most cell types. MuSiC returned zeros for several abundant cell types (B, T4, T8) as well as pDCs; interestingly, Xu *et al.* [30] also found a similar result, with zeros estimated for T4, T8, and B cells across all samples, when using MuSiC to perform deconvolution on the same publicly available bulk RNA-seq dataset. CSx NS also yielded 0 values for cM, possibly due to its low representation in the dataset. Overall, correlations were weaker in the real bulk analysis than in simulations, consistent with increased measurement noise and reference mismatch. Some cell-type-specific effects of reference-mixture mismatch were observed but are more difficult to interpret than in the simulation study, likely due to multiple differences between reference and bulk cohorts, including disease status, age, sex, and race. For example, while correlation for CD8+ T cells was generally lower under mismatch for InstaPrism, CIBERSORTx S-mode, and BLADE, as expected, it was higher for DISSECT. In addition, correlation for B cells was unexpectedly higher under mismatch for InstaPrism, CSx S, and BLADE.

Because predicted proportions from all deconvolution methods inherently sum to one, whereas the original ground-truth proportions from flow cytometry do not, direct comparison can introduce bias. As a sensitivity analysis, we rescaled the ground-truth proportions to sum to one and observed similar results (Supplementary Figs 16–19). However, forcing the ground truth to sum to one may also introduce new distortions by inflating or deflating specific cell-type proportions. For this reason, the unscaled ground-truth proportions remain our primary evaluation metric, and the rescaled analysis is presented only as a sensitivity check.

In general, the effect of the reference condition was small compared to that of the deconvolution algorithm. Although higher correlations might be expected when reference and target conditions matched, this trend was not consistently observed. For example, InstaPrism achieved a higher correlation for B cells when using a mismatched reference (0.85) than when matched (0.74), and DISSECT showed higher correlation for T8 cells under mismatched conditions (0.82 versus 0.72). No consistent pattern emerged when comparing matched, mismatched, and combined reference sets, indicating that algorithmic design played a more dominant role than reference condition in determining deconvolution accuracy.

Discussion

In this study, we examined whether the matching disease conditions between the scRNA-seq reference and target bulk RNA-seq mixtures influences the performance of cell-type deconvolution methods. Using single-cell PBMC data, we evaluated six methods spanning three algorithmic categories across four reference–mixture matching scenarios, applied to both simulated and real bulk datasets.

Overall, we found that the choice of deconvolution algorithm had a much greater impact on performance than reference–mixture matching. Among all methods, DISSECT, representing the deep-learning category, consistently performed best across nearly all scenarios, achieving the lowest error and RMSE and the highest correlation and CCC across most cell types. InstaPrism, a representative Bayesian method, achieved comparable performance in the simulated bulk setting but performed less well in the real bulk data. The effect of matching status was modest overall, though, as expected, RMSE values were smaller under matched conditions in our simulation study and tended to be more pronounced when using healthy references to perform deconvolution on lupus mixtures. The impact of reference mismatch was most evident among regression-based methods. The cell types most affected by reference–mixture mismatch in the simulations, B cells, CD8 T cells, and classical and nonclassical monocytes, are also those reported in the underlying scRNA-seq study to exhibit substantial disease-associated transcriptional changes in SLE. Reduced concordance for these populations under reference mismatch is therefore consistent with biological shifts in cell-type-specific expression profiles, which can distort reference-derived signatures and disproportionately impact deconvolution accuracy for affected cell types.

Beyond disease status, demographic differences, particularly age, between scRNAseq references and RNAseq mixtures represent a biologically meaningful and often overlooked source of reference mismatch. In our simulation study, healthy controls and lupus cases differed not only in disease status but also in age distribution (median 32 versus 47 years), and these factors cannot be disentangled in our simulation framework. Aging is known to substantially reshape peripheral immune composition and these changes could alter cell-type-specific expression signatures used to construct deconvolution references [36, 37]. The compounded effects of disease, age, and additional demographic mismatches observed in our real bulk RNA-seq validation, where performance was more variable and sometimes unexpected compared to the simulation, are consistent with this view. These findings underscore a broader challenge in reference-based deconvolution: real-world reference cohorts will rarely match the target mixtures along all important axes of biological variation, and performance degradation likely reflects the cumulative impact of multiple demographic and disease-related differences.

Results from deconvolution of the real bulk RNA-seq dataset sometimes produced counterintuitive results for some methods, such as lower RMSE or higher Pearson correlation for mismatched references for certain cell types, such as B cells. As the flow cytometry proportions from the real bulk dataset did not sum to one, whereas all deconvolution methods enforced a sum-to-one constraint, RMSE may have been influenced by systematic differences rather than true accuracy. The unexpected pattern for B cells for Pearson correlation likely reflects complex interactions among disease-specific expression shifts, differences in age, sex, and race distributions between the scRNA-seq and real bulk RNA-seq cohorts, and method-specific biases. Together, these observations highlight the challenge of interpreting reference mismatch effects in real-world datasets where the contributions of individual biological and demographic variables cannot be isolated, a limitation shared broadly across reference-based deconvolution benchmarking studies.

DISSECT likely had the strongest performance because deep-learning methods are able to capture nonlinear relationships between the scRNA-seq reference and the target bulk mixtures and account for between-subject differences in cell-type-specific gene expression

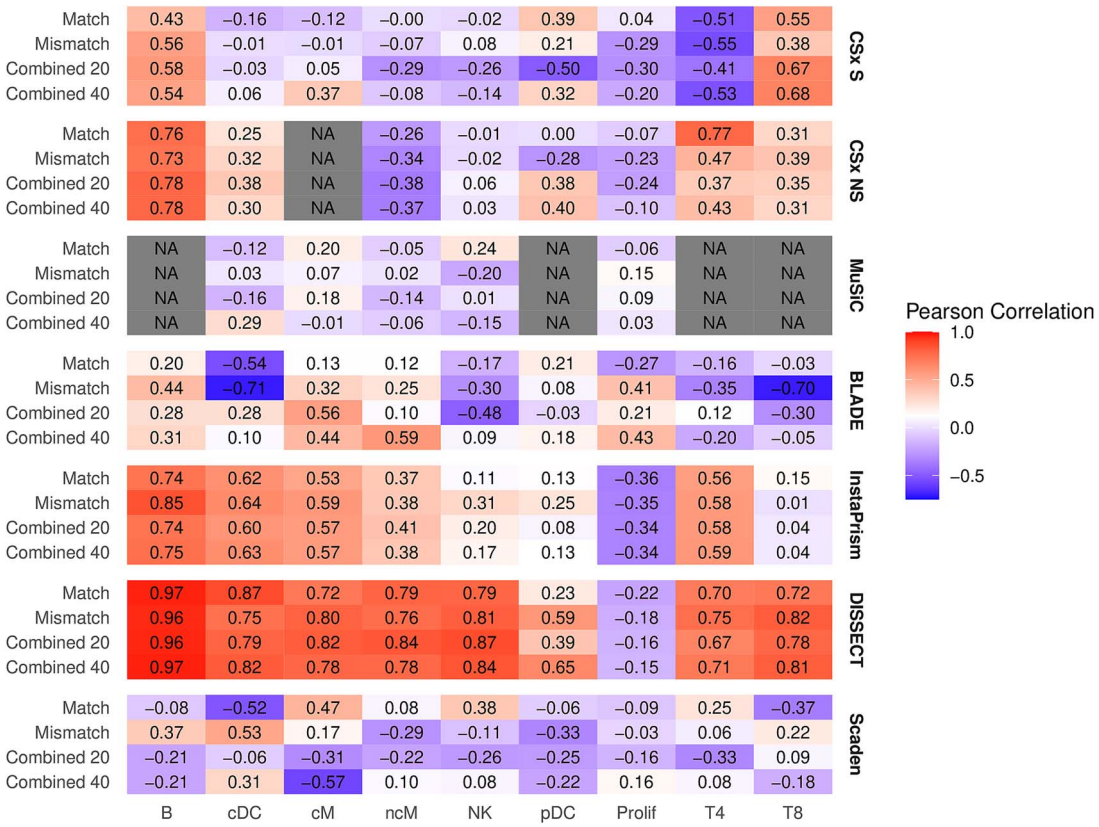

Figure 7 Pearson correlation across reference conditions for real bulk PBMC RNA-seq samples. Heatmaps show Pearson correlation coefficients (r) between estimated and flow-cytometry-derived cell-type proportions for six deconvolution methods—CSx (S), CSx (NS), MuSiC, BLADE, InstaPrism, DISSECT, and Scaden—across four reference configurations (normal, managed, combined 20, and combined 40). Higher values indicate stronger positive correlations. DISSECT and InstaPrism achieved consistently high correlations across most cell types, while regression-based methods such as MuSiC and CSx displayed reduced concordance in certain references and cell types.

in their simulated training sets. Although both Scaden and DISSECT are deep-learning–based methods trained using subject-level data, their performance differed markedly in our analyses. This likely reflects differences in model architecture and learning objectives: Scaden aggregates predictions across an ensemble of neural networks, whereas DISSECT leverages simultaneous consistency regularization between the target and training domains. Additionally, deep-learning methods often require careful hyperparameter tuning to achieve optimal performance. For the fairness of comparison, we evaluated both methods using their default settings. It is possible that Scaden's performance could improve with further hyperparameter optimization, which was beyond the scope of this study.

Rare cell populations, such as cDC, pDC, and proliferating cells, were consistently more difficult to estimate, and deconvolution performance was consistently lower for these rare cell types, likely due to limited representation in the scRNA-seq reference and a restricted range of pseudo-bulk proportions, both of which complicate accurate estimation and evaluation using correlation-based metrics. As shown in Figs 4 and 7 and Supplementary Figs 1–2, these cell types exhibited lower correlations with the ground truth and higher bias, highlighting the challenges of accurately deconvolving low-abundance populations.

Runtime considerations are also important for practical implementation. InstaPrism and MuSiC2 offer near-immediate results, whereas CSx and BLADE require substantially longer compute time. Although the deep-learning approaches (Scaden and DISSECT) incur additional upfront cost during training, their prediction steps are fast, making them efficient once models have been trained.

Our study has several important limitations. Simulated mixtures, which are the sums of single cells, do not reflect technical differences between bulk and scRNA-seq technologies. This could lead to overly optimistic estimates of performance in simulations. In the real bulk PBMC dataset, the accuracy of all methods decreased. In addition to differences between real and simulated bulk RNA-seq mixtures, this decline in performance was likely influenced by additional factors. For example, the cell-type proportions estimated from flow cytometry summed to only ∼90% per subject on average, leaving an inherent 10% error that affected the calculation of RMSE and Lin's CCC, both of which depend on bias terms. In addition, cell-type definitions may have differed slightly between the flow cytometry data and the scRNA-seq reference. For example, "proliferative cells" may represent distinct populations in the two datasets. Furthermore, flow cytometry measures surface protein abundance, which may correlate with—but does not directly reflect—mRNA expression levels. Another notable observation was that several methods underestimated the proportions of abundant cell types, a pattern warranting further investigation.

Another limitation of this study is that we evaluated a limited subset of available deconvolution methods rather than performing an exhaustive comparison. While the selected methods span multiple algorithmic approaches and include widely used methods, our findings may not generalize to all existing or future methods, particularly

those with modeling assumptions or implementations that differ substantially from the evaluated methods.

Our study is also limited by the lack of a comprehensive benchmark dataset that includes matched bulk RNA-seq, scRNA-seq, and ground-truth abundance measures under multiple conditions. As noted by Wolfram-Schauerte *et al.* [8], very few such datasets exist, and none fully met the needs of our study, which aimed to evaluate condition-specific effects. In addition, our evaluation was restricted to PBMCs, which were selected due to the availability of large scRNA-seq references and bulk RNA-seq datasets with experimentally measured cell-type proportions. Although results from other tissue types may differ, we believe our findings remain informative, as PBMCs are a commonly analyzed tissue in translational research. Many disease studies aim to compare cellular composition between diseased and healthy individuals using scRNA-seq of PBMCs, given that blood is easily accessible and cost-effective for potential diagnostic applications. Accurate estimation of PBMC cell-type composition thus remains an important goal [8]. Whether the effects of reference–mixture condition mismatch observed here generalize to other complex tissues, such as the brain or pancreas, remains an important question for future studies as suitable reference and validation datasets become available.

Beyond benchmarking, an open question is how to translate deconvolution results into clinically meaningful inference. We saw wide variation in the accuracy of deconvolution methods at the individual sample level; therefore, interpreting the composition for any given sample must be done with caution. Though most methods do not provide estimates of uncertainty, InstaPrism does provide credible intervals for estimated proportions, which may aid in the interpretation of deconvolution results. In addition, best practices for testing for differences in deconvolution proportions between disease groups when using differing references that are tailored to each group require additional research.

Conclusion

In summary, our results suggest that while reference–mixture matching can modestly affect deconvolution performance, the choice of algorithm remains the dominant factor. For users seeking accurate cell-type proportion estimates, we recommend consulting recent benchmark studies and selecting the best-performing methods accordingly. When feasible, matched references should be used, though the degree of impact may depend on the chosen algorithm. For those exploring diagnostic applications, methods that provide uncertainty quantification, such as InstaPrism, may be preferable. Finally, for method developers, creating well-curated benchmark datasets and maintaining accessible, reproducible software will be crucial for advancing the field.

Key Points

- We evaluated how reference–mixture matching affects performance for six bulk RNA-seq deconvolution methods using single-cell RNA-seq references across simulated and real datasets.
- Method choice had a much larger impact on deconvolution accuracy than reference–mixture matching. Regression-based methods were more sensitive to mismatched references.
- DISSECT, a deep-learning–based algorithm, consistently achieved the best RMSE and correlation performance. We recommend using robust methods in practice and interpreting estimates for rare cell types with caution.

Acknowledgements

The authors would like to thank Dr. Katerina Kechris for her service on Ms. Zhao's thesis committee and for her valuable insights and feedback on this work.

Author contributions

Yifan Zhao (Data curation, Formal analysis, Visualization, Writing—original draft, Writing—review & editing), Brian E. Vestal (Supervision, Writing—review & editing), and Camille M. Moore (Conceptualization, Methodology, Supervision, Writing—review & editing).

Supplementary data

Supplementary data are available at Briefings in Bioinformatics online.

Conflicts of interest

The authors declare no competing interests.

Funding

This study was supported in part by the National Institutes of Health (5U01AI173502-02 to C.M.M.).

Data availability

All datasets used in this study are publicly available. The single-cell RNA-seq dataset used for reference construction and bulk simulation is available from NCBI GEO under accession GSE174188. The bulk RNA-seq dataset used for real-data evaluation is available under accession GSE107011.

References

1. Jew B, Alvarez M, Rahmani E. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun* 2020;**11**:1971. https://doi.org/10.1038/s41467-020-15816-6.
2. 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.
3. Torroja C, Sanchez-Cabo F. Digitaldlsorter: deep-learning on scRNA-Seq to deconvolute gene expression data. *Front Genet* 2019;**10**:978. https://doi.org/10.3389/fgene.2019.00978.
4. Wang C, Lin Y, Li S. *et al.* Deconvolution from bulk gene expression by leveraging sample-wise and gene-wise similarities and single-cell RNA-Seq data. *BMC Genomics* 2024;**25**:875. https://doi.org/10.1186/s12864-024-10728-x.
5. Xie D, Gui L, Wang J. Statistical Inference for Cell Type Deconvolution. arXiv 2025;2202.06420. https://doi.org/10.48550/arXiv.2202.06420.

6. Le T. *et al.* A review of digital cytometry methods: estimating the relative abundance of cell types in a bulk of cells. *Brief Bioinform* 2021;**22**:bbaa219. https://doi.org/10.1093/bib/bbaa219.
7. Sturm G, Finotello F, Petitprez F. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* 2019;**35**:i436–45. https://doi.org/10.1093/bioinformatics/btz363.
8. Wolfram-Schauerte M, Vogel T, Tuoken H. *et al.* Approaching the holistic transcriptome—convolution and deconvolution in transcriptomics. *Brief Bioinform* 2025;**26**:bbaf388. https://doi.org/10.1093/bib/bbaf388.
9. Im Y, Kim Y. A comprehensive overview of RNA deconvolution methods and their application. *Mol Cells* 2023;**46**:99–105. https://doi.org/10.14348/molcells.2023.2178.
10. 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.
11. 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.
12. Cai B, Zhang J, Li H. *et al.* Statistical inference of cell-type proportions estimated from bulk expression data. *J Am Stat Assoc* 2024;**119**:2521–32. https://doi.org/10.1080/01621459.2024.2382435.
13. Chen Y, Wang Y, Chen Y. *et al.* Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nat Commun* 2022;**13**:6735. https://doi.org/10.1038/s41467-022-34550-9.
14. Cobos FA, Panah MJN, Epps J. *et al.* Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes. *Genome Biol* 2023;**24**:177. https://doi.org/10.1186/s13059-023-03016-6.
15. 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.
16. Du R, Carey V, Weiss ST. deconvSeq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics* 2019;**35**:5095–102. https://doi.org/10.1093/bioinformatics/btz444.
17. Erdmann-Pham DD, Fischer J, Hong J. *et al.* Likelihood-based deconvolution of bulk gene expression data using single-cell references. *Genome Res* 2021;**31**:1794–806. https://doi.org/10.1101/gr.272344.120.
18. Fan J, Lyu Y, Zhang Q. *et al.* MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief Bioinform* 2022;**23**:bbac430. https://doi.org/10.1093/bib/bbac430.
19. Frishberg A, Peshes-Yaloz N, Cohn O. *et al.* Cell composition analysis of bulk genomics using single-cell data. *Nat Methods* 2019;**16**:327–32. https://doi.org/10.1038/s41592-019-0355-5.
20. Goldfarbmuren KC, Jackson ND, Sajuthi SP. *et al.* Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat Commun* 2020;**11**:2485. https://doi.org/10.1038/s41467-020-16239-z.
21. Hu M, Chikina M. InstaPrism: an R package for fast implementation of BayesPrism. *Bioinformatics* 2024;**40**:btae440. https://doi.org/10.1093/bioinformatics/btae440.
22. Karimnezhad A. More accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Bioinform Adv* 2022;**2**:vbac049. https://doi.org/10.1093/bioadv/vbac049.
23. Khatri R, Machart P, Bonn S. DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. *Genome Biol* 2024;**25**:112. https://doi.org/10.1186/s13059-024-03251-5.
24. Luo Y, Fan R. Deconvolution analysis of cell-type expression from bulk tissues by integrating with single-cell expression reference. *Genet Epidemiol* 2022;**46**:615–28. https://doi.org/10.1002/gepi.22494.
25. Nava A, Alves da Quinta D, Prato L. *et al.* Novel evaluation approach for molecular signature-based deconvolution methods. *J Biomed Inform* 2023;**142**:104387. https://doi.org/10.1016/j.jbi.2023.104387.
26. 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.
27. Tsoucas D, Dong R, Chen H. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nat Commun* 2019;**10**:2975. https://doi.org/10.1038/s41467-019-10802-z.
28. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res* 2021;**31**:1807–18. https://doi.org/10.1101/gr.268722.120.
29. 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.
30. Xu X, Li R, Mo O. *et al.* Cell-type deconvolution for bulk RNA-seq data using single-cell reference: a comparative analysis and recommendation guideline. *Brief Bioinform* 2025;**26**:bbaf031. https://doi.org/10.1093/bib/bbaf031.
31. Andrade, Barbosa B, van Asten SD, Oh JW. *et al.* Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. *Nat Commun* 2021;**12**:6106. https://doi.org/10.1038/s41467-021-26328-2.
32. Chu T, Wang Z, Pe'er D. *et al.* Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* 2022;**3**:505–17. https://doi.org/10.1038/s43018-022-00356-3.
33. 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.
34. Perez RK, Gordon MG, Subramaniam M. *et al.* Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* 2022;**376**:eabf1970. https://doi.org/10.1126/science.abf1970.
35. Xu W, Monaco G, Wong EH. *et al.* Mapping of γ/δ T cells reveals Vδ2+ T cells resistance to senescence. *EBioMedicine* 2019;**39**:44–58. https://doi.org/10.1016/j.ebiom.2018.11.053.
36. Gong Q, Sharma M, Glass MC. *et al.* Multi-omic profiling reveals age-related immune dynamics in healthy adults. *Nature* 2025;**648**:696–706. https://doi.org/10.1038/s41586-025-09686-5.
37. Filippov I, Schauser L, Peterson P. An integrated single-cell atlas of blood immune cells in aging. *npj Aging* 2024;**10**:59. https://doi.org/10.1038/s41514-024-00185-x.

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Briefings in Bioinformatics, 2026, **27**, bbag184
https://doi.org/10.1093/bib/bbag184
Problem Solving Protocol