SOFTWARE

Open Access

Deconvolution from bulk gene expression by leveraging sample-wise and gene-wise similarities and single-cell RNA-Seq data

Chenqi Wang[1†], Yifan Lin[1†], Shuchao Li[1] and Jinting Guan[1,2,3*]

Abstract

Background The widely adopted bulk RNA-seq measures the gene expression average of cells, masking cell type heterogeneity, which confounds downstream analyses. Therefore, identifying the cellular composition and cell type-specific gene expression profiles (GEPs) facilitates the study of the underlying mechanisms of various biological processes. Although single-cell RNA-seq focuses on cell type heterogeneity in gene expression, it requires specialized and expensive resources and currently is not practical for a large number of samples or a routine clinical setting. Recently, computational deconvolution methodologies have been developed, while many of them only estimate cell type composition or cell type-specific GEPs by requiring the other as input. The development of more accurate deconvolution methods to infer cell type abundance and cell type-specific GEPs is still essential.

Results We propose a new deconvolution algorithm, DSSC, which infers cell type-specific gene expression and cell type proportions of heterogeneous samples simultaneously by leveraging gene-gene and sample-sample similarities in bulk expression and single-cell RNA-seq data. Through comparisons with the other existing methods, we demonstrate that DSSC is effective in inferring both cell type proportions and cell type-specific GEPs across simulated pseudo-bulk data (including intra-dataset and inter-dataset simulations) and experimental bulk data (including mixture data and real experimental data). DSSC shows robustness to the change of marker gene number and sample size and also has cost and time efficiencies.

Conclusions DSSC provides a practical and promising alternative to the experimental techniques to characterize cellular composition and heterogeneity in the gene expression of heterogeneous samples.

Keywords Deconvolution, Cell type abundance, Cell type-specific gene expression profile, Similarity matrix, Single-cell RNA-seq data

Background

Bulk RNA-seq has been widely adopted to profile transcriptomes of samples, while for heterogeneous samples with multiple cell types, bulk RNA-seq data only represents the gene expression average across cell types contained in the bulk samples, masking the cell type heterogeneity [1, 2]. Downstream analyses based on bulk RNA-seq data, such as commonly applied differential gene expression analysis, are typically confounded by differences in cell type proportions, leading to the mask of the gene expression contribution of lowly abundant cell

[†] Chenqi Wang and Yifan Lin are co-first authors.

*Correspondence:
Jinting Guan
jtguan@xmu.edu.cn
[1] Department of Automation, Xiamen University, Xiamen, China
[2] Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai, China
[3] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China

types by that of more abundant cell types [3, 4]. The variations in gene expression between heterogeneous samples might be because of the differences in cell type composition, biological conditions, or both [5]. Therefore, identifying the cell type abundance of bulk samples facilitates the generation of profound insights into the underlying mechanisms of various biological processes [6].

Moreover, understanding the cell type composition difference of diseases is critical to the identification of cell types that could be targeted therapeutically [4]. For instance, tumors not only include malignant cells but are embedded in a complex microenvironment comprising a variable portion of immune cells [7]. Studying the composition and abundance of immune cells in cancer samples is invaluable for drug discovery, clinical treatment decisions, and cancer immunotherapy [8].

To profile the transcriptomes of individual cells for exploring cellular heterogeneity, single-cell RNA-seq (scRNA-seq) has emerged as a powerful technique, while it requires expensive and specialized resources and currently is not practical for a large number of samples or a routine clinical setting [9]. Besides, scRNA-seq contains a lot of technical noises, leading to the accuracy being lower than bulk RNA-seq. Traditional immunophenotyping techniques to measure cell type density, such as flow cytometry [10, 11] and immunohistochemistry [12], are generally dependent on the preselection of marker genes, limiting the number of cell types that can be simultaneously identified [9].

In this context, many computational deconvolution methodologies have been developed to infer cell type proportions and/or cell type-specific gene expression profiles (GEPs) from bulk transcriptomic data. Most of them either estimate cell type proportions with referenced cell type-specific GEPs as input or estimate cell type-specific GEPs with cell type compositions as input [6]. Among these methods, most were developed to infer cell type abundance by using GEPs of purified cell types, such as FARDEEP [13], CIBERSORT [14], DCQ [15], and DeconRNASeq [16]. However, due to the required input, purified cell type-specific GEPs, may be unavailable or questionable and only accessible for a few tissues, such as blood [17–19], brain [20], and pancreas [21], the use of many purified GEPs-based deconvolution methods has been limited [22]. Therefore, taking advantage of scRNA-seq data from corresponding tissue in bulk expression deconvolution is an alternative, such as MuSiC [23], SCDC [24], and DWLS [25].

Currently, there have been several deconvolution methods estimating cell type abundance and cell type-specific GEPs simultaneously, including CIBERSORTx [9], deconf [26], TOAST [22], Linseed [27], BLADE [28], BayesPrism [29], and RNA-Sieve [30]. Among them, RNA-Sieve requires scRNA-seq data, and CIBERSORTx, BLADE, and BayesPrism require cell type-specific GEPs. Others need less referenced information, even being reference-free. For instance, deconf and TOAST need the number of cell types and Linseed is a reference-free method. It has been reported that reference-based deconvolution methods are usually more accurate and robust in estimating the proportion of cell types than reference-free deconvolution methods [14, 31, 32]. Considering that there is room for improvement in deconvolution accuracy, the development of more accurate deconvolution methods to infer cell type abundance and cell type-specific GEPs is still appealing.

We hypothesize that similar samples have similar cell type compositions and similar genes have similar gene expression profiles, therefore maintaining the sample-sample similarity and gene-gene similarity in bulk expression data would benefit cell type deconvolution. To this end, we propose a deconvolution algorithm DSSC, which estimates cell type-specific GEPs and cell type density simultaneously from bulk samples by maintaining gene-gene and sample-sample similarities calculated from bulk expression and leveraging single-cell gene expression data. We compare DSSC with existing deconvolution methodologies including the ones using cell type-specific GEPs or scRNA-seq data as input to infer cell type compositions and the ones inferring both cell type abundance and cell type-specific GEPs. By deconvolution on pseudo-bulk data generated by inter-dataset and intra-dataset simulations, mixture bulk data, and real experimental bulk data, the effectiveness and accuracy of DSSC are demonstrated.

Materials and methods

DSSC algorithm

Assuming every cell type has similar expression levels across heterogeneous samples, the expression of a gene in a heterogeneous sample can be modeled as a weighted sum of the expression levels in existing cell types [5]. Let $X \in R^{g\times n}$denote the expression value of g genes in n different heterogeneous samples, i.e., the bulk gene expression matrix; $C \in R^{g\times k}$ represents the average expression levels of g genes in k cell types, i.e., cell type-specific GEPs; $P \in R^{k\times n}$ represents the proportions of k cell types in n samples, then the deconvolution problem can be formulated as:

$$X = CP \tag{1}$$

That is, based on the gene expression of heterogeneous samples, the proportion of each cell type in each sample and/or the cell type-specific GEPs are estimated [33, 34].

Different from most deconvolution methods that only infer C or P, we propose a new deconvolution method

DSSC, which leverages gene-gene similarity, sample-sample similarity, and scRNA-seq data information to infer C and P simultaneously from the bulk gene expression data. The optimization problem is:

$$\min_{C,P} \|X - CP\|_F^2 + \lambda_1 \|S_s - P^T P\|_F^2 + \lambda_2 \|S_g - CC^T\|_F^2 + \lambda_c \|C - \rho(Y)\|_F^2 \tag{2}$$

$$s.t.\ C_{il} > 0, P_{lj} > 0, \sum_l P_{lj} = 1, i = 1, 2, \cdots, g,\ l = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n$$

where $X \in R^{g \times n}$ is the bulk gene expression matrix; $C \in R^{g \times k}$ is the cell type-specific GEP matrix; $P \in R^{k \times n}$ is the cell type proportion matrix; $S_s \in R^{n \times n}$ and $S_g \in R^{g \times g}$ are the sample-sample similarity matrix and gene-gene similarity matrix calculated based on the bulk gene expression matrix X; Y represents the single-cell gene expression data of the same tissue; $\rho(Y)$ represents the averaged gene expression of each cell type for reference which can be calculated from Y or a given cell type-specific GEP matrix; $\|\cdot\|_F$ represents the Frobenius norm; λ_1,λ_2 and λ_c are regularization parameters. The goal of the optimization problem is to make $X \approx CP$, the inferred matrix P maintain the similarity between samples, and the inferred matrix C maintain the similarity between genes and approach the referenced cell type-specific GEPs.

For the calculation of the sample-sample similarity matrix S_s, we first calculated the sample-sample distance matrix D based on the bulk gene expression matrix, whose entry $d_{\alpha\beta}$ is calculated as 1 minus the Pearson's correlation coefficient between samples α and β. Then we calculated the corresponding entry of S_s as:

$$\frac{1}{1 + d_{\alpha\beta}} \tag{3}$$

Similarly, the gene-gene similarity matrix S_g is calculated.

To solve the original optimization problem of DSSC, we adopted a greedy way by solving first the same objective function but only with the non-negative constraints on matrices C and P and after obtaining the initial outputs C and P, the matrix P is processed to guarantee the sum of the proportions of all cell types in every sample is one. To solve the same objective function but only with the non-negative constraints, we derived the element-wise multiplicative updates (Supplementary Materials). Beginning with random positive initializations, we performed the following updating rules at each iteration until convergence:

$$C_{il} = \frac{(XP^T + \lambda_c \rho(Y) + 2\lambda_2 S_g C)_{il}}{(CPP^T + \lambda_c C + 2\lambda_2 CC^T C)_{il}} C_{il} \tag{4}$$

$$P_{lj} = \frac{(C^T X + 2\lambda_1 P S_s)_{lj}}{(C^T CP + 2\lambda_1 PP^T P)_{lj}} P_{lj} \tag{5}$$

The termination condition of the iterations is set to:

$$\frac{\|C_t - C_{t+1}\|_F^2}{\|C_t\|_F^2} < \theta \quad and \quad \frac{\|P_t - P_{t+1}\|_F^2}{\|P_t\|_F^2} < \theta \tag{6}$$

or the number of iterations reaches 3000. C_t and P_t are the matrices at the t-th iteration, C_{t+1} and P_{t+1} are the matrices at the $(t+1)$-th iteration, θ is a given threshold (for which the default setting is 10^{-5} while in this paper we set it as 10^{-8}). After the iterations terminate, the output matrix P is processed to guarantee the sum of the proportions of all cell types in every sample is one. Then, we determine the cell type labels for columns of C and rows of P by calculating the similarity between the output matrix C and the reference matrix $\rho(Y)$. Afterward, matrices C and P with cell type annotation are the deconvolution results of DSSC. For DSSC, we differentiated the cases using referenced GEPs (denoted as DSSC3, keeping all three regularization terms in the objective function) and not using referenced GEPs to infer initial matrices C and P (denoted as DSSC2, removing the regularization term of referenced input and only keeping the other two regularization terms on similarity matrices).

Determination of regularization parameters

To determine the regularization parameters λ_1,λ_2 and λ_c, we used grid searching and five-fold cross-validation by adapting the way of Elyanow et al [35]. We divided the input bulk gene expression matrix X into five folds at random, each of which contains 20% of the entries of the input bulk matrix. Then, we ran DSSC for a range of regularization parameter combinations, masking out one-fold of entries. Next, for DSSC2, we calculated the root mean squared error (RMSE) between the masked one-fold data from X and that from CP; for DSSC3, as it can leverage reference information, we calculated the Pearson's correlation coefficient (PCC) between the inferred matrix C and the referenced GEP matrix but only for the involved genes with masked entries. The procedure was repeated for each fold, and then five RMSE or PCC values were obtained to calculate the average. Higher PCC and lower RMSE indicate better deconvolution performance. We selected the regularization parameter combination which resulted in the best average. By default, λ_1 and λ_2 can be 0, 0.001, 0.01, 0.1, 1 or 10, and λ_c can be 0, 1, 10, 100, 1000 or 10,000. The ranges can be changed in the practical usage.

Generation of simulated bulk data

We adopted a similar way as the research of Avila Cobos et al [4] to use single-cell transcriptome data to generate pseudo-bulk data but differentiated intra-dataset and inter-dataset simulations (Supplementary Materials). In the intra-dataset simulation, a single-cell gene expression matrix is divided into a training set and a testing set by stratified sampling with a ratio of 1:1, while in the inter-dataset simulation, two single-cell gene expression data are used as a training set and a testing set respectively. Based on the training set, the cell type-specific averaged gene expression GEPs for reference and cell type-specific marker genes were obtained. Based on the testing set, the pseudo-bulk data, and the ground-truths including real cell type ratio and cell type-specific GEPs were generated.

Specifically, using the training set and the testing set respectively, the average across all cells of each cell type was calculated for each gene to form a referenced cell type-specific GEP matrix and a real one. To identify cell type-specific marker genes based on the training set, we used scater package [36] for CPM normalization (if the training set was already normalized, we skipped the normalization step) and regarded the genes with $log_2FC \geq 1$ and Benjamini-Hochberg adjusted p_value < 0.1 as marker genes. For the deconvolution methods requiring referenced GEPs, we used the marker genes to extract the GEPs calculated from the training set; for the deconvolution methods that require scRNA-seq data as a reference, we used the marker genes to extract the trained scRNA-seq data, i.e., the training set, as a reference. When generating simulated bulk data based on the testing set, we generated bulk data of 1000 heterogeneous samples, where the expression of each bulk sample is the sum of the gene expression of 100 single cells randomly selected from the testing set. Note that the sampling is totally random, so the number of cells of a specific cell type could be large, small or even zero. The true cell type ratio was determined according to the actual sampling. Then, we used the marker genes obtained from the training set to extract the simulated bulk data and then supplied to the deconvolution algorithms, and also used the marker genes to extract the real GEPs calculated from the testing set for the evaluation of deconvolution results.

In the intra-dataset deconvolution experiments, we used human pancreatic cell dataset Segerstolpe [37], human brain cell datasets Camp [38], Darmanis [39], and Manno [40], and mouse hematopoietic stem cell dataset Nestorowa [41]. In the inter-dataset deconvolution, human pancreatic cell datasets Segerstolpe [37], Baron [42], and Muraro [43], and mouse retinal cell datasets Macosko [44] and Shekhar [45] were used.

Experimental bulk data

We used five mixture bulk data for testing, including BreatBlood [46] data mixing human breast and blood samples; CellLines [17] data mixing multiple human cell lines including Jurkat (T cell leukemia), THP-1 (acute monocytic leukemia), IM-9 (B lymphoblastic multiple myeloma) and Raji (Burkitt B cell lymphoma); LiverBrainLung [47] data mixing rat brain, liver, and lung tissue samples; RatBrain [3] data mixing cell culture data of rat neuronal cells, astrocytes, oligodendrocytes, and microglia; and Retina [48] data mixing retinal tissue samples from two different mouse lines. For the mixture data deconvolution, expression data of 100% purified samples was used as the training set (which can be considered as expression data of single cells) to identify marker genes and obtain referenced single-cell data or GEPs. For the deconvolution methods with GEPs as input, the referenced GEP matrix was calculated as the averaged gene expression of every cell type; for the deconvolution methods with single-cell data as input, the referenced input was the expression data of purified samples. We obtained marker genes referring to the study of Mohammadi et al [49], i.e., genes with z-score normalized Shannon entropy greater than a given threshold. Expression data of mixed samples with a known proportion was used as the testing set and the known cell type ratio is the ground-truth. The above five sets of data were directly obtained from the study of Mohammadi et al [49], among which four sets were normalized with robust multiarray average (RMA) [50] except for CellLines data.

Moreover, we used a real experiment bulk dataset, WholeBlood [9] data, for testing. The whole blood samples were collected from 12 healthy adult subjects and the expression data is TPM-normalized RNA-seq data. We used LM22 (a microarray-derived signature matrix for distinguishing 22 human hematopoietic cell subsets in bulk tissues), 3’ PBMCs and 5’ PBMCs (two publicly available PBMC datasets from healthy donors profiled by Chromium v2 (5’ and 3’ kits)) from CIBERSORT article [14] as different referenced information. LM22 is in the form of GEPs, while 3' PBMCs and 5' PBMCs data, which are CPM-normalized scRNA-seq data, were used to generate referenced GEPs and marker genes. Similarly to the CIBERSORT article, we also focused on the inference of NK.cells, Monocytes, B.cells, T.cells.CD4 and T.cells.CD8. Therefore, before deconvolution, we extracted the reference data of the cells involving the above five target cell types, and after deconvolution, we merged the sub-cell types belonging to the same cell type for comparison with the ground-truth (Supplementary Materials). The real cell type ratio was derived from the measurement result of flow cytometry and automated complete blood counts, available from the CIBETSORTx website.

Methods for benchmarking

We compared DSSC with 24 existing deconvolution methods which were tested mostly using default settings. Among the compared algorithms, 15 are deconvolution methods inferring cellular composition only, including 11 methods using GEPs as reference and 4 methods using scRNA-seq data as a reference. The GEPs-referenced methods include ordinary least squares (OLS [51]), non-negative least squares (NNLS [52]), a weighted least squares (EPIC [53]), two robust regression methods (FARDEEP [13] and RLR [54]), a support vector regression method (CIBERSORT [14]), four penalty regression methods (ridge [55], lasso [55], elastic_net [55] and DCQ [15]), and a quadratic programming method (DeconRNASeq [16]). The scRNA-seq data-referenced deconvolution methods include MuSiC [23], DWLS [25], Bisque [56], and SCDC [24]. We also evaluated 9 deconvolution methods inferring both cell type abundance and GEPs, including a support vector regression method CIBERSORTx [9] which accepts referenced GEPs, RNA-Sieve [30] which accepts scRNA-seq data, four methods contained in the CellMix [57] R package (digital sorting algorithm DSA [58], two semi-supervised non-negative matrix factorization methods ssKL [59] and ssFrobenius [59], and built-in improved deconf [26]), and three methods contained in the DeCompress [60] method (CellDistinguisher [61], Linseed [27], and TOAST + NMF [22]). For the GEPs- and scRNA-seq data-referenced methods, we provided corresponding referenced input. As to the other deconvolution methods, we also used unified ways for better comparisons. For DSA, ssKL, and ssFrobenius, we provided the names of marker genes corresponding to each cell type; for TOAST, Linseed, CellDistinguisher, deconf, and DSSC2, we provided the number of cell types to initialize the P and C matrices and associated the output matrix C with the referenced matrix to annotate cell types.

Performance indicators

When evaluating deconvolution methods inferring only cell type proportion matrix P, we calculated the PCC and RMSE between the inferred matrix P and the true value P_{true}. When evaluating deconvolution methods inferring both P and cell type-specific GEP matrix C, the PCC and RMSE between the inferred matrix C and the real value C_{true} were also calculated. For this, we transformed the two matrices into one-dimensional vectors in the same way and then calculated the PCC and RMSE between the two vectors. Then, when assessing the inferred matrix P, for each sample, we calculated the PCC between the real cell type ratio and the inferred one; when assessing the inferred matrix C, for each cell type, we calculated the PCC between the real gene expression and the inferred one.

Results

Overview of DSSC algorithm and benchmarking pipeline

Figure 1A gives a schematic representation of DSSC algorithm. DSSC uses bulk gene expression data X as input, and can also accept single-cell gene expression data Y to calculate the referenced gene expression profile (GEP) matrix $\rho\ (Y)$ or accept a given GEP matrix C_{ref} directly. DSSC calculates gene-gene similarity matrix S_g and sample-sample similarity matrix S_s based on X. Through matrix decomposition, the inferred cell type proportion matrix P maintains the sample-sample similarity, and the inferred cell type-specific GEP matrix C maintains the gene-gene similarity and approaches the averaged gene expression calculated from the referenced single-cell RNA-seq data $\rho\ (Y)$ or the given C_{ref}. After the initial matrices C and P are obtained, the PCC between C and $\rho\ (Y)/C_{ref}$ is calculated to determine the cell type label of each column in C, and then each row in P. Then each column of matrix P is processed to satisfy that the sum of the proportions of all cell types in each sample is 100%. At this time, matrices C and P are the results of DSSC. DSSC is a deconvolution method simultaneously estimating cell type-specific GEPs and cell type composition of samples by leveraging the similarities among bulk samples and genes and the information of referenced GEPs or scRNA-seq data.

To verify the effectiveness of DSSC, we tested it on simulated bulk data (including intra-dataset and inter-dataset simulations) and experimental data (including mixture bulk data and real experimental RNA-seq data). Figure 1B gives the testing pipeline. In the intra-dataset deconvolution, a single-cell gene expression matrix is divided into training and testing sets. Based on the training set, cell type-specific averaged gene expression C_{ref} is calculated and marker genes are identified. For the deconvolution methods that do not accept the referenced input, only the marker genes or

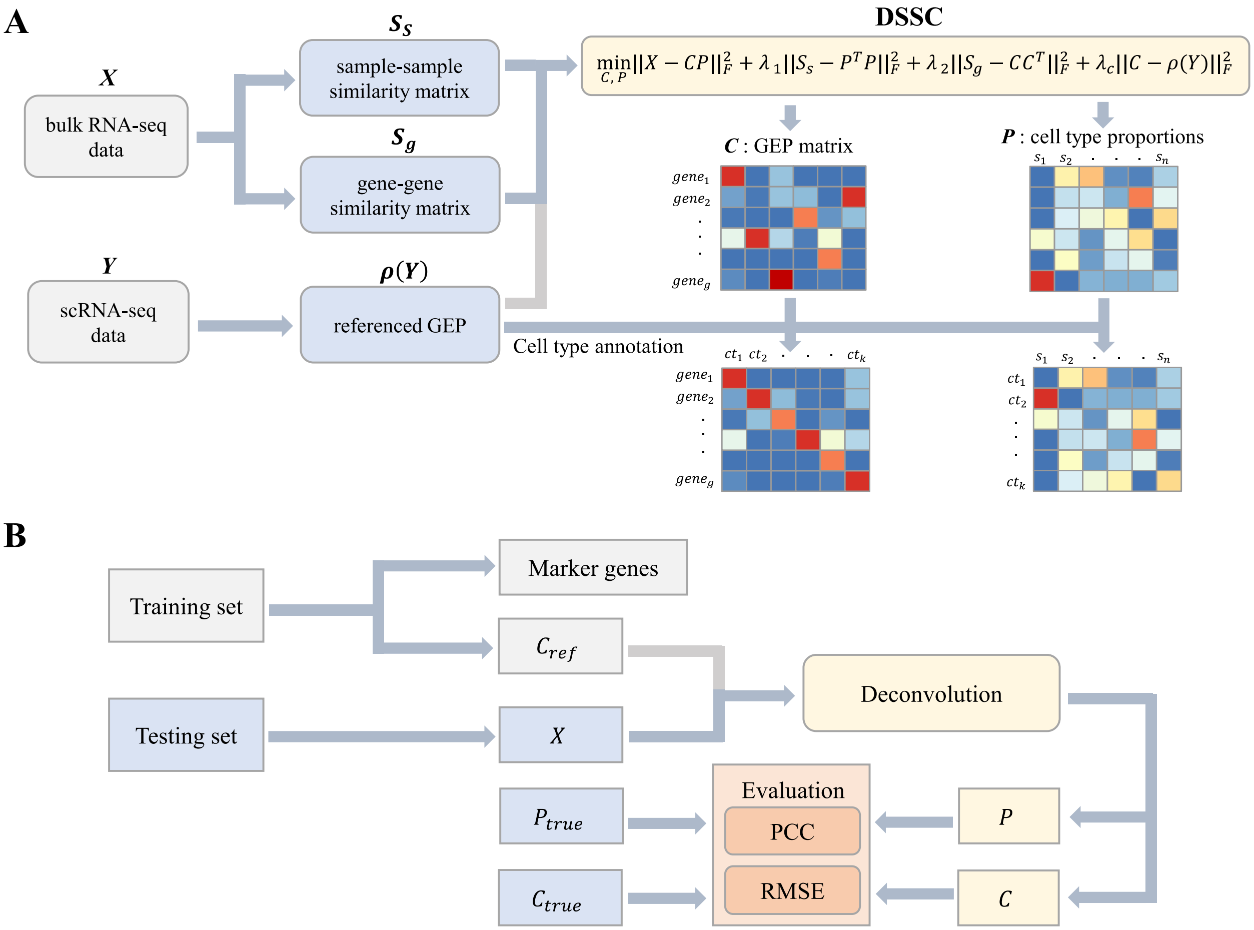


Fig. 1 Schematic representations of (**A**) DSSC algorithm and (**B**) deconvolution testing pipeline. We differentiated the cases keeping all three terms in the objective function (denoted as DSSC3) and keeping the first two terms about the sample-sample and gene-gene similarity matrices (denoted as DSSC2). DSSC2 does not need the referenced GEPs to infer the initial matrices *C* and *P*, and only needs the reference to determine the cell type label, i.e., the grey paths in the figure are not employed in DSSC2

Table 1 Details of single-cell transcriptome data used for simulation tests

Dataset	Accession	Biological sample type	Number of genes	Number of cells	Number of markers	Number of cell types	Simulation type	Reference
Nestorowa	GSE81682	Mouse hematopoietic stem	4077	1455	860	8	Intra-dataset	[41]
Manno	GSE76381	Human brain	11,507	2496	619	22	Intra-dataset	[40]
Darmanis	GSE67835	Human brain	13,488	389	1594	6	Intra-dataset	[39]
Camp	GSE75140	Human brain	11,241	551	262	5	Intra-dataset	[38]
Segerstolpe	E-MTAB-5061	Human pancreatic	13,876	898	2233	6	Intra- and inter-dataset	[37]
Baron	GSE84133	Human pancreatic	8415	7876	2235	10	Inter-dataset	[42]
Muraro	GSE85241	Human pancreatic	12,239	1930	1119	7	Inter-dataset	[43]
Macosko	GSE63473	Mouse retina	2986	34,638	429	8	Inter-dataset	[44]
Shekhar	GSE81904	Mouse retina	4707	233,811	532	4	Inter-dataset	[45]

Table 2 Details of used mixture bulk data

Dataset	Accession	Biological sample type	Number of genes	Number of samples	Number of marker genes	Number of cell types	Reference
BreastBlood	GSE29830	Human breast and blood	54,675	9	3058	2	[46]
CellLines	GSE11058	Human cell lines	54,675	12	2824	4	[17]
LiverBrain Lung	GSE19830	Rat brain, liver, and lung	31,009	33	2093	3	[47]
RatBrain	GSE19380	Rat cell lines	31,009	10	1836	4	[3]
Retina	GSE33076	Mouse retinal tissue	22,347	24	769	2	[48]

the number of cell types is required. Based on the testing set, the simulated bulk data X, the real cell type ratio P_{true} and the real GEP data C_{true} are generated. In the inter-dataset deconvolution, two single-cell transcriptome data are used as a training set and a testing set, respectively. By calculating PCC and RMSE between the inferred C/P and the real C_{true}/P_{true}, the performances of deconvolution methods are evaluated. The single-cell transcriptome data used in the simulation tests are listed in Table 1. In the deconvolution of mixture bulk data (Table 2), the training set is expression data of purified samples and is used to identify marker genes and obtain C_{ref}. The testing set is expression data of mixed samples with a known proportion P_{true}. As the real GEP data C_{true} is unknown, we used C_{ref} instead to evaluate the inferred C. In the deconvolution of real experimental bulk RNA-seq data, the testing set is WhooldBlood [9] data and the training set is LM22, 3' PBMCs or 5' PBMCs data provided in CIBERSORT article [14]. As the real GEP data C_{true} is unknown, we mainly evaluated P. The real cell type ratio P_{true} was derived from flow cytometry and automated complete blood counts.

We compared DSSC with 24 existing deconvolution methods (Table 3), including 9 deconvolution methods inferring both P and C, and 11 deconvolution methods using GEPs as reference to infer P, and 4 deconvolution methods using scRNA-seq data as reference to infer P. For the deconvolution methods inferring both cell type proportion and GEPs, we evaluated matrices P and C. For other deconvolution methods, we only evaluated matrix P. For DSSC, we differentiated the cases keeping all three terms in the objective function (denoted as DSSC3) and keeping the two terms about the sample-sample and gene-gene similarity matrices (denoted as DSSC2). DSSC2 does not need the referenced GEPs to infer the initial matrices C and P, and only needs the reference to determine the cell type label (Fig. 1).

Deconvolution on simulated bulk data

Intra-dataset deconvolution

We used five scRNA-seq data, human pancreatic cell dataset Segerstolpe [37], human brain cell datasets (Camp [38], Darmanis [39], and Manno [40]), and mouse hematopoietic stem cell dataset Nestorowa [41] to perform intra-dataset simulations. Supplementary Fig. 1A shows the PCC between the inferred cell type-specific GEPs by DSSC3 and the real GEPs. It can be seen that DSSC3 can accurately infer the gene expression of different cell types. Then we calculated PCC and RMSE

Table 3 Deconvolution methods for comparison

Deconvolution methods	Reference information	Estimate cell type proportions	Estimate GEPs	Reference
NNLS	GEPs	Yes	No	[52]
OLS	GEPs	Yes	No	[51]
FARDEEP	GEPs	Yes	No	[13]
CIBERSORT	GEPs	Yes	No	[14]
DeconRNASeq	GEPs	Yes	No	[16]
RLR	GEPs	Yes	No	[54]
DCQ	GEPs	Yes	No	[15]
elastic_net	GEPs	Yes	No	[55]
ridge	GEPs	Yes	No	[55]
lasso	GEPs	Yes	No	[55]
EPIC	GEPs	Yes	No	[53]
MuSiC	scRNA-seq data	Yes	No	[23]
BisqueRNA	scRNA-seq data	Yes	No	[56]
SCDC	scRNA-seq data	Yes	No	[24]
DWLS	scRNA-seq data	Yes	No	[25]
CIBERSORTx	GEPs	Yes	Yes	[9]
RNA-Sieve	scRNA-seq data	Yes	Yes	[30]
DSA	/	Yes	Yes	[58]
ssKL	/	Yes	Yes	[59]
ssFrobenius	/	Yes	Yes	[59]
deconf	/	Yes	Yes	[26]
TOAST	/	Yes	Yes	[22]
Linseed	/	Yes	Yes	[27]
CellDistinguisher	/	Yes	Yes	[61]

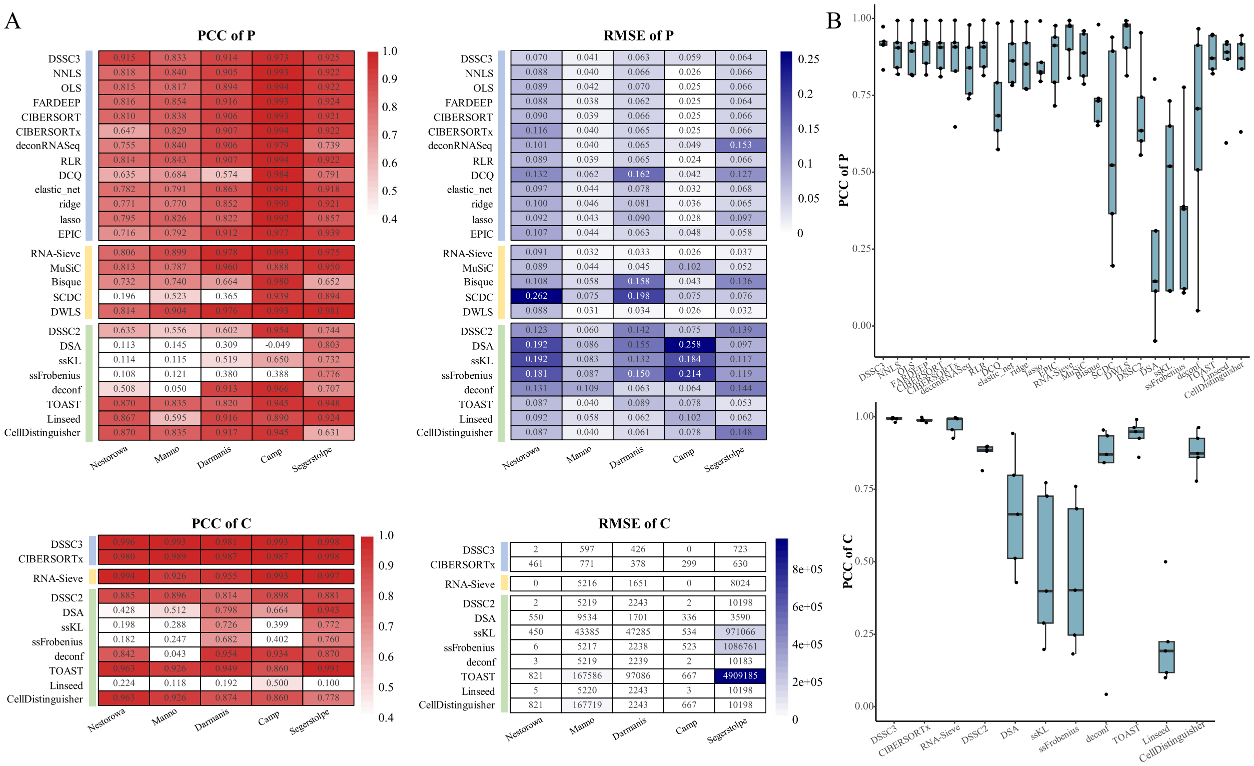


Fig. 2 Intra-dataset deconvolution results. **A** PCC and RMSE between the inferred cell type proportion matrix (or GEP matrix) and the real one. The deconvolution methods are divided into three categories: deconvolution methods with GEPs as reference (denoted as blue), deconvolution methods with scRNA-seq data as reference (denoted as yellow), and other deconvolution methods (denoted as green). **B** PCC between the inferred cell type proportion matrix (or GEP matrix) and the real matrix, each point representing each responding result in Fig. 2A

between the inferred cell type ratio by DSSC3 and the truth for each sample, and Supplementary Fig. 1B shows the means and standard deviations of PCC and RMSE across all samples, demonstrating the good performance of DSSC3 in inferring cell type proportions.

Figure 2A shows the overall indexes, PCC and RMSE between the flattened C/P estimated by all deconvolution methods and the flatten real matrix C_{true}/P_{true}. It can be noted that DSSC3 is competitive across different data, no matter in estimating cell type proportions or estimating GEPs, especially on Nestorowa dataset. For the deconvolution methods based on GEPs, most methods such as FARDEEP, CIBERSORT, CIBERSORTx are stable and effective in inferring P. For the deconvolution methods based on single-cell data, RNA-Sieve and DWLS performs well across datasets, while SCDC performs poorly on Nestorowa and Darmanis data which may be mainly because the proportion of several specific cell types in Nestorowa data could not be inferred and the performance varies widely among different samples in Darmanis data. Some deconvolution methods inferring both C and P perform well, but some do not, such as DSA, ssKL, and ssFrobenius which is perhaps because the inferred cell types are similar to a variety of real labels. When comparing DSSC3 with DSSC2 which performs well in inferring C but not in P, we can note the advantage of using referenced GEPs.

To show the overall test results in intra-dataset deconvolution, we pooled the PCC values of five datasets in Fig. 2B, each point representing the corresponding value in Fig. 2A. DSSC3 is stable and accurate, at the forefront of most methods, in the inference of GEP matrix C and cell type proportion P. DSSC2 is not prominent in the inference of P, but it performs well in inferring C relative to some other deconvolution methods. Except DSSC3, RNA-Sieve, DWLS and Linseed can also accurately infer P, but Linseed is not effective in inferring GEPs. The standard deviation of PCC of P for deconf is large, mainly due to its poor result on Manno data. We speculated that deconf may be affected by random seeds or the number of cell types, and it does not use the information of marker genes which may lead to poor performance.

Inter-dataset deconvolution

We used three human pancreas cell datasets (Baron [42], Muraro [43], and Segerstolpe [37]), and two mouse

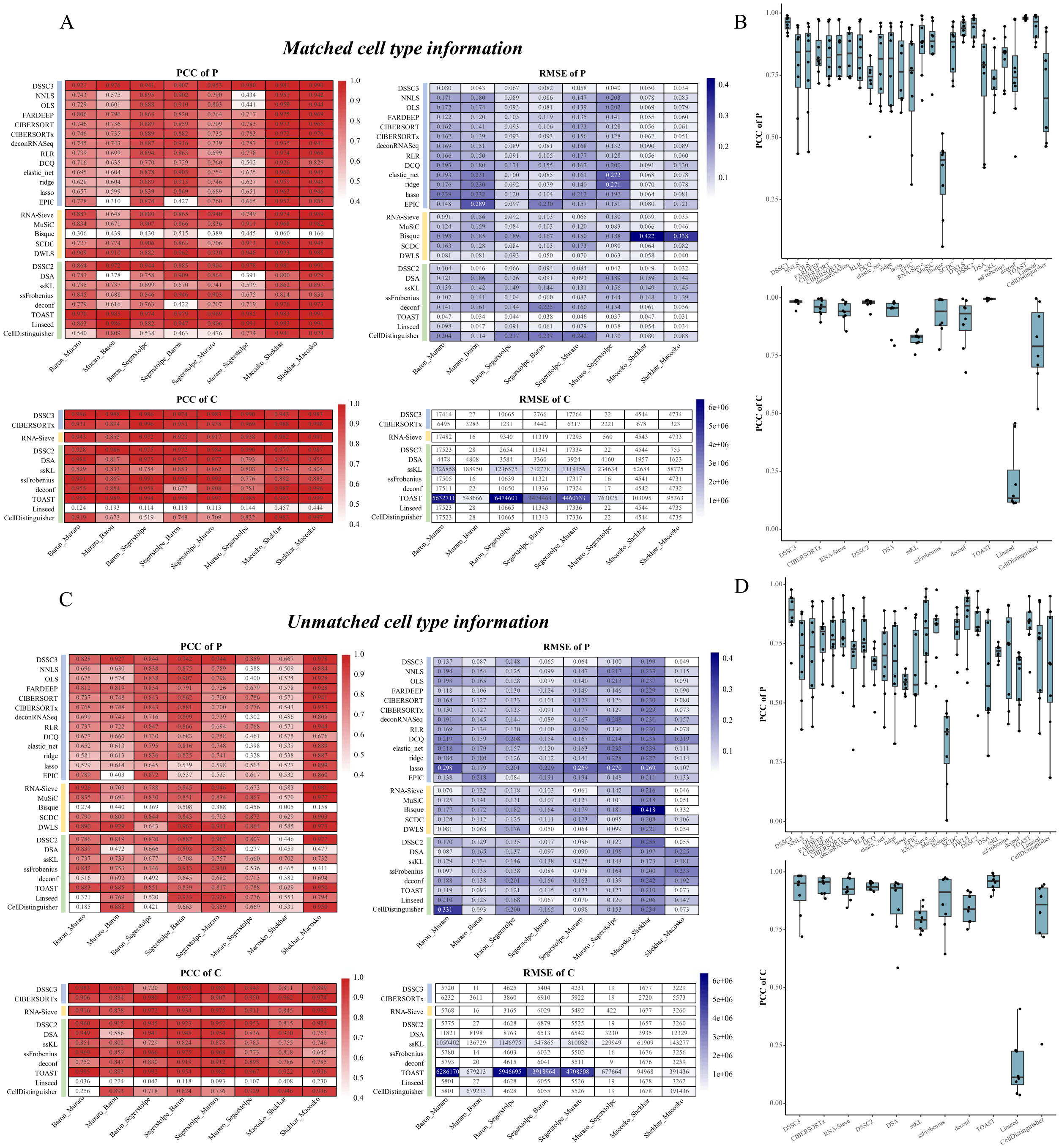


Fig. 3 Inter-dataset deconvolution results including the cases of matching and unmatching cell types between the training and testing sets before deconvolution. **A** PCC and RMSE between the inferred cell type proportion matrix (or GEP matrix) and the real one in the case of matching cell types. **B** PCC between the inferred cell type proportion matrix (or GEP) and the real matrix, each point representing each corresponding result in Fig. 3A. **C** PCC and RMSE between the inferred cell type proportion matrix (or GEP matrix) and the real one in the case of unmatching cell types. **D** PCC between the inferred cell type proportion matrix (or GEP) and the real matrix, each point representing each corresponding result in Fig. 3C. Baron_Muraro denotes Baron data as the testing set and Muraro as the training set, and others have similar meanings

retinal cell datasets (Macosko [44] and Shekhar [45]) to perform eight sets of inter-dataset simulations. Among them, the downloaded Muraro data is non-counts data, while the rest are counts data, so we could not only pay attention to the impact of batch effect on deconvolution but also the impact of different normalization methods.

Considering that the cell types are not identical between the training and the testing sets, we examined the influence of whether the cell type information is matched before simulation. First, for the case of matching cell types in advance, after quality control of single-cell data, we matched the cell types of the training set and test set (i.e., taking the intersection of cell types), and then performed data simulation. Supplementary Fig. 2A represents the PCC between the cell type-specific GEPs inferred by DSSC3 and the real ones, which indicates the excellent performance of DSSC3 in estimating GEPs. Then we calculated PCC and RMSE between the inferred cell type ratio by DSSC3 and the truth for each sample, and Supplementary Fig. 2B shows the means and standard deviations across all samples, indicating the accuracy of DSSC3 in inferring cell type proportion across different samples. Figure 3A shows the PCC and RMSE between the flattened C/P and the flattened real matrix C_{true}/P_{true}. Figure 3B summarizes the performance values of all data tests in Fig. 3A. We found that many methods are in line with expectations. DSSC3, DWLS, TOAST, DSSC2 and Linseed are effective and stable in inferring P, and DSSC3, DSSC2 and TOAST perform well in inferring C. Linseed also shows weakness in estimating GEPs, which may because it is still the method mainly inferring the proportion of cell types although it can output GEPs. For deconvolution experiments across batches, most of the reference-based deconvolution methods are effective, such as FARDEEP and CIBERSORT, but compared with the case of intra-dataset deconvolution, the performances are overall degraded, especially Bisque which is not effective on all eight tests. In addition, the inter-dataset deconvolution using two single-cell data with different normalization methods also brings some deviations. For example, in the test of Muraro_Segerstolpe (Muraro being the testing set and Segerstolpe being the training set), the deconvolution results of the least squares regression methods (NNLS and OLS) are not good, indicating the influence of referenced GEPs on the regression-based methods. Most of the reference-based deconvolution methods are interfered by different normalization methods between the training set and testing set (which can be seen from the results of tests involving Muraro data), but DSSC3, DSSC2, and DWLS still perform well.

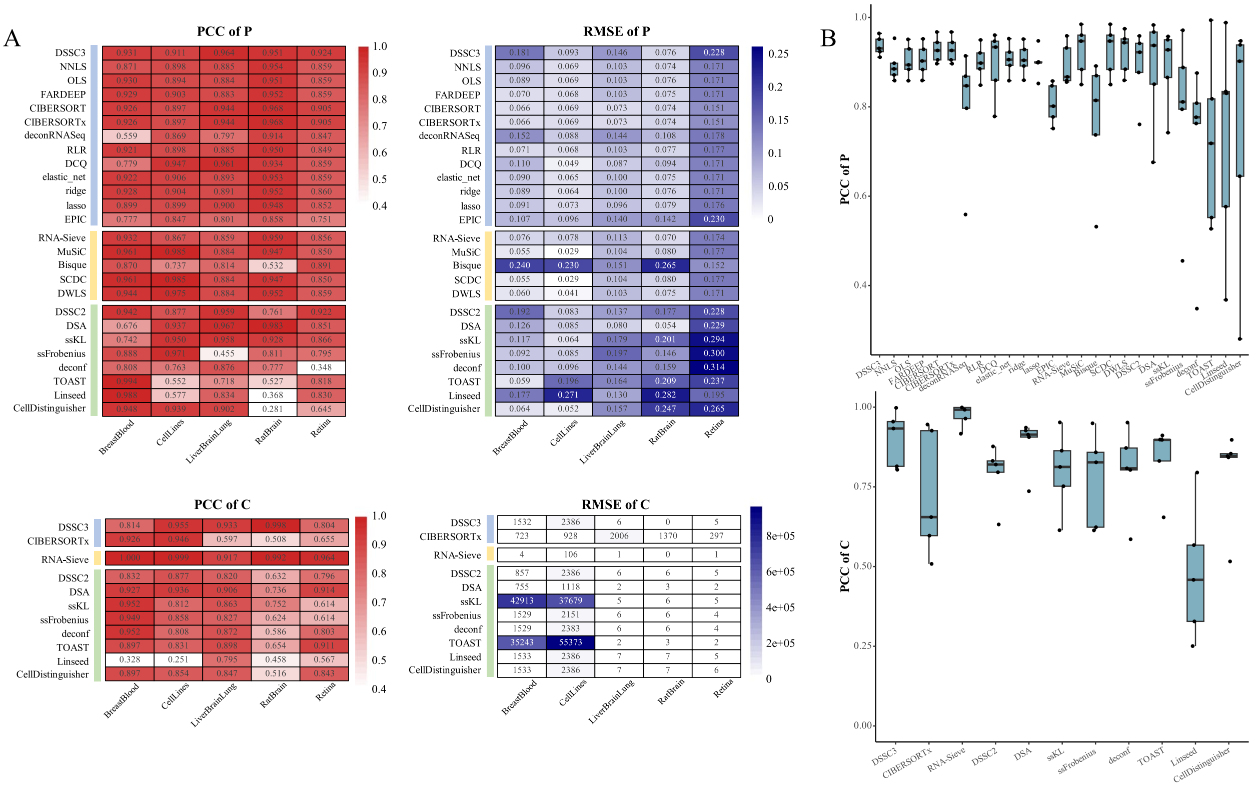
Next, we tested the case of unmatching cell types between the training and testing sets in advance. After deconvolution, the similarity between C/P and the real C_{true}/P_{true} is calculated only for the cell types of the testing set. Supplementary Fig. 3A shows the results of DSSC3 for the same eight groups of inter-data deconvolution but the cell type information is not matched, and Supplementary Fig. 3B shows the PCC and RMSE of P for every sample. Compared with the results of matching cell types, DSSC3 can still infer cell type-specific GEPs and cell type ratio. The PCC of P at the sample level in Macosko_Shekhar test is relatively low, but the mean can still maintain 0.7. Conversely, the performance of DSSC3 in Shekhar_Macosko test is more stable, perhaps because training on Macosko data can learn more general information for prediction on the testing set. Figure 3C and D show the overall results. Compared with Fig. 3A and B, it can be found that the performances of all deconvolution methods have a little decrease under the case of unmatched cell types no matter in the estimation of GEPs or cell type proportions, which is in line with the law of reality, but DSSC3 and DSSC2 have relatively excellent performances. Moreover, checking Fig. 3C, the results of all deconvolution methods are significantly reduced in Macosko_Shekhar test, but DSSC3 remains at the forefront.

In these heterogeneous simulation experiments, DSSC algorithm especially DSSC3 can accurately and stably estimate cell type proportion and GEPs. Although the performance in the case of unmatching cell types would be reduced relative to the case of matching cell types, it is more realistic and DSSC still performs well. Compared with intra-dataset deconvolution, the results of most deconvolution methods can still be believed in inter-dataset simulation experiments, among which DSSC performs more prominently, indicating its robustness to the influence of batch effect to a certain extent.

Deconvolution on mixture data

We used mixture data with known proportions as testing data, including human BreatBlood [46] and CellLines [17] data, Rat LiverBrainLung [47] and RatBrain [3] data, and mouse Retina [48] data. The expression data of 100% purified samples was used as the training set. For deconvolution methods that require GEPs as reference, we supplied the GEPs obtained from the purified data. For the deconvolution methods that require scRNA-seq data as a reference, we supplied the data of purified samples by regarding the purified samples as cells. It is noted that as the real GEP matrix C_{true} is unknown, we used C_{ref} instead to evaluate as the reference is from purified samples which makes the two GEP matrices similar.

The PCC between the cell type-specific GEPs inferred by DSSC3 and the reference is shown in Supplementary Fig. 4A. Supplementary Fig. 4B shows the mean and standard deviation of PCC and RMSE calculated for every sample for datasets with more than two cell types. It indicates the stability of DSSC3 in inferring the cell type proportions across different samples and also confirms that the homology between reference data and tested bulk data could improve the deconvolution performance. Figure 4A shows the overall indexes, from which



PCC of P

	BreastBlood	CellLines	LiverBrainLung	RatBrain	Retina
DSSC3	0.931	0.911	0.964	0.951	0.924
NNLS	0.871	0.898	0.885	0.954	0.859
OLS	0.930	0.894	0.884	0.951	0.859
FARDEEP	0.929	0.903	0.883	0.952	0.859
CIBERSORT	0.926	0.897	0.944	0.968	0.905
CIBERSORTx	0.926	0.897	0.944	0.968	0.905
deconRNASeq	0.559	0.869	0.797	0.914	0.847
RLR	0.921	0.898	0.885	0.950	0.849
DCQ	0.779	0.947	0.961	0.934	0.859
elastic_net	0.922	0.906	0.893	0.953	0.859
ridge	0.928	0.904	0.891	0.952	0.860
lasso	0.899	0.899	0.900	0.948	0.852
EPIC	0.777	0.847	0.801	0.858	0.751
RNA-Sieve	0.932	0.867	0.859	0.959	0.856
MuSiC	0.961	0.985	0.884	0.947	0.850
Bisque	0.870	0.737	0.814	0.532	0.891
SCDC	0.961	0.985	0.884	0.947	0.850
DWLS	0.944	0.975	0.884	0.952	0.859
DSSC2	0.942	0.877	0.959	0.761	0.922
DSA	0.676	0.937	0.967	0.983	0.851
ssKL	0.742	0.950	0.958	0.928	0.866
ssFrobenius	0.888	0.971	0.455	0.811	0.795
deconf	0.808	0.763	0.876	0.777	0.348
TOAST	0.994	0.552	0.718	0.527	0.818
Linseed	0.988	0.577	0.834	0.368	0.830
CellDistinguisher	0.948	0.939	0.902	0.281	0.645

RMSE of P

	BreastBlood	CellLines	LiverBrainLung	RatBrain	Retina
DSSC3	0.181	0.093	0.146	0.076	0.228
NNLS	0.096	0.069	0.103	0.074	0.171
OLS	0.089	0.069	0.103	0.076	0.171
FARDEEP	0.070	0.068	0.103	0.075	0.171
CIBERSORT	0.066	0.069	0.073	0.074	0.151
CIBERSORTx	0.066	0.069	0.073	0.074	0.151
deconRNASeq	0.152	0.088	0.144	0.108	0.178
RLR	0.071	0.068	0.103	0.077	0.177
DCQ	0.110	0.049	0.087	0.094	0.171
elastic_net	0.090	0.065	0.100	0.075	0.171
ridge	0.089	0.064	0.100	0.076	0.171
lasso	0.091	0.073	0.096	0.079	0.176
EPIC	0.107	0.096	0.140	0.142	0.230
RNA-Sieve	0.076	0.078	0.113	0.070	0.174
MuSiC	0.055	0.029	0.104	0.080	0.177
Bisque	0.240	0.230	0.151	0.265	0.152
SCDC	0.055	0.029	0.104	0.080	0.177
DWLS	0.060	0.041	0.103	0.075	0.171
DSSC2	0.192	0.083	0.137	0.177	0.228
DSA	0.126	0.085	0.080	0.054	0.229
ssKL	0.117	0.064	0.179	0.201	0.294
ssFrobenius	0.092	0.085	0.197	0.146	0.300
deconf	0.100	0.096	0.144	0.159	0.314
TOAST	0.059	0.196	0.164	0.209	0.237
Linseed	0.177	0.271	0.130	0.282	0.195
CellDistinguisher	0.064	0.052	0.157	0.247	0.265

PCC of C

	BreastBlood	CellLines	LiverBrainLung	RatBrain	Retina
DSSC3	0.814	0.955	0.933	0.998	0.804
CIBERSORTx	0.926	0.946	0.597	0.508	0.655
RNA-Sieve	1.000	0.999	0.917	0.992	0.964
DSSC2	0.832	0.877	0.820	0.632	0.796
DSA	0.927	0.936	0.906	0.736	0.914
ssKL	0.952	0.812	0.863	0.752	0.614
ssFrobenius	0.949	0.858	0.827	0.624	0.614
deconf	0.952	0.808	0.872	0.586	0.803
TOAST	0.897	0.831	0.898	0.654	0.911
Linseed	0.328	0.251	0.795	0.458	0.567
CellDistinguisher	0.897	0.854	0.847	0.516	0.843

RMSE of C

	BreastBlood	CellLines	LiverBrainLung	RatBrain	Retina
DSSC3	1532	2386	6	0	5
CIBERSORTx	723	928	2006	1370	297
RNA-Sieve	4	106	1	0	1
DSSC2	857	2386	6	6	5
DSA	755	1118	2	3	2
ssKL	42913	37679	5	6	5
ssFrobenius	1529	2151	6	6	4
deconf	1529	2383	6	6	4
TOAST	35243	55373	2	3	2
Linseed	1533	2386	7	7	5
CellDistinguisher	1533	2386	7	7	6

Fig. 4 Deconvolution results on mixture data. **A** PCC and RMSE between the inferred cell type proportion matrix and the real one and those between the inferred GEP matrix and the referenced one. The real GEP matrix is unknown, we used the referenced matrix instead to evaluate as the reference is from purified samples which makes the two GEP matrices similar. **B** The boxplot of PCC values, each point representing each corresponding result in Fig. 4A

can be seen that many methods perform well in the inference of *P* and *C*. DSSC3, CIBERSORT, and CIBERSORTx are more prominent when inferring cell type proportions. RNA-Sieve performs well in inferring GEPs on all datasets and keeping the similarity between the inferred GEPs and the referenced GEPs is a main feature of RNA-Sieve. Compared with other deconvolution methods inferring both *C* and *P*, DSSC3 is outstanding in inferring *C* on RatBrain data. Linseed is good on LiverBrainLung in inferring GEPs compared with its performance on other data. As to Linseed, it is good at inferring *P* in the simulation experiments but underperforms on the mixture data. ssFrobenius has low performance in inferring the proportion of cell types on the LiverBrainLung data, but ssKL, also a semi-supervised non-negative matrix factorization method, works well. As seen in Fig. 4B, the task of deconvolution on mixture data would be more difficult than that of simulation experiments, but DSSC3 is still at the forefront of all deconvolution methods. Overall, the stability and accuracy of the reference-based deconvolution methods are higher than those of the reference-free complete deconvolution methods, because the reference is from purified data which is homologous with the tested bulk data.

Deconvolution on real experimental data

We tested the bulk RNA-seq data of human peripheral blood collected from 12 healthy adult subjects (denoted

(See figure on next page.)

Fig. 5 Deconvolution results on real experimental WholeBlood data for three different references. The single cell data-based methods were only tested using the references with the form of single cell data, i.e., 3' PBMCs and 5' PBMCs. (A) PCC and RMSE between the inferred cell type proportion matrix and the real one. (B) PCC between the inferred cell type proportion and the real one for each sample, each point representing a sample. (C) The averages of PCC of each sample across all three references and two references (3' PBMCs and 5' PBMCs), with each point representing a sample. (D) PCC values of each sample calculated based on the averaged cell type proportion matrix across three references and two references

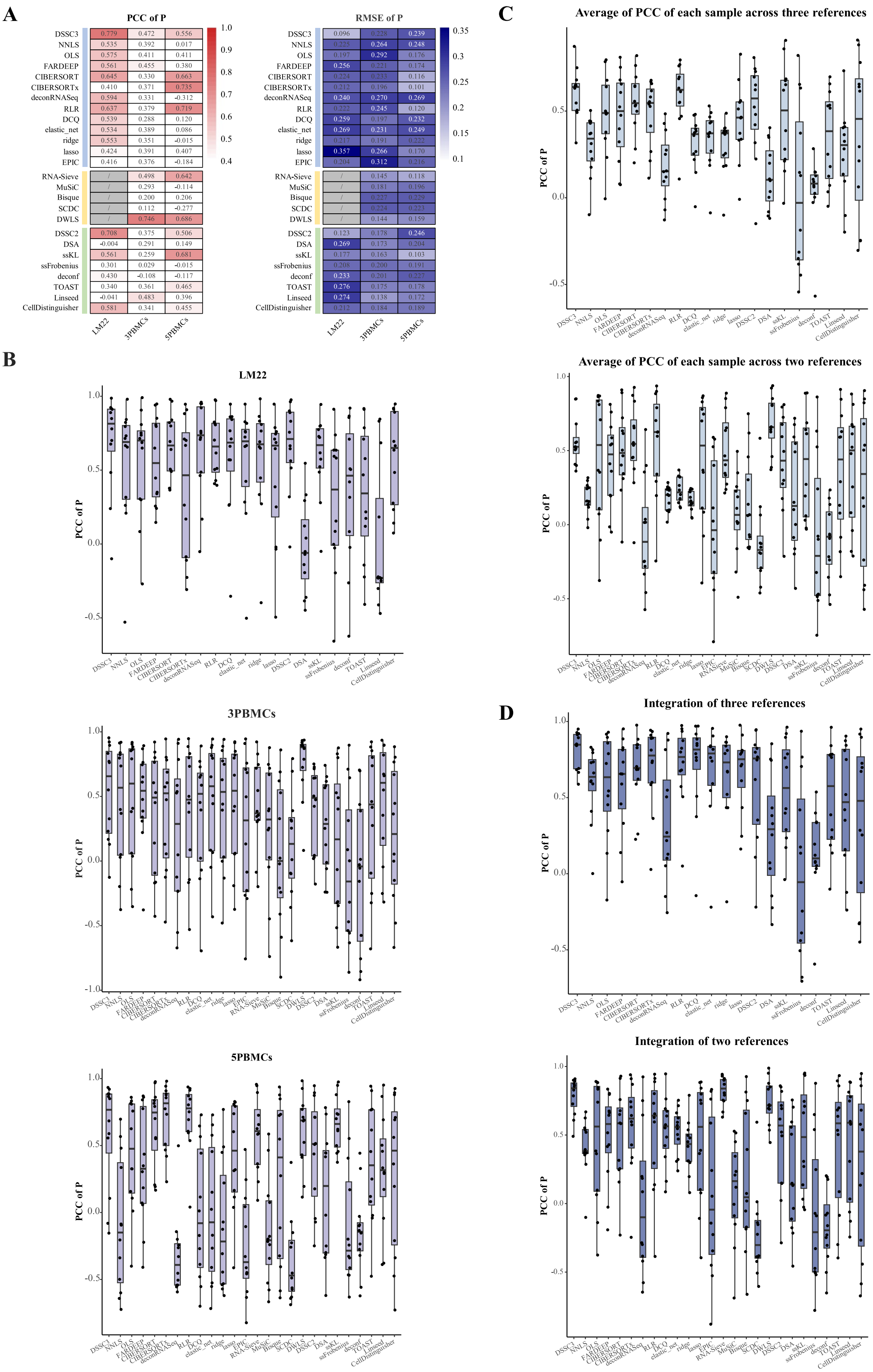


Fig. 5 (See legend on previous page.)

as WholeBlood [9]) and used LM22 and two PBMC datasets of healthy donors (3' PBMCs and 5' PBMCs) as different references. Referring to the CIBERSORT article, we also focused on the inference of five cell types, NK.cells, Monocytes, B.cells, T.cells.CD4, and T.cells.CD8. Therefore, before deconvolution, we extracted the reference data of the sub-cell types belonging to the above five target cell types, and after deconvolution, we merged the sub-cell types for comparison with the ground-truth. The real cell type ratio comes from the measurement result of flow cytometry and automated complete blood counts. As the real GEPs information is unknown, we used the referenced GEPs to just check the proximity between the referenced GEPs and the estimated one and mainly focused on the result of *P*. We compared DSSC with other deconvolution methods and because LM22 itself is in the form of GEPs, the deconvolution methods using single-cell data as input were only tested using references of 3' PBMCs and 5' PBMCs.

Supplementary Fig. 5A shows the PCC between the inferred cell type-specific gene expression by DSSC3 and the referenced gene expression for each different reference. It can be found that T.cells.CD4 and T.cells.CD8 are similar when 5' PBMCs data was used as a reference, which may affect the annotation of cell type labels to a certain extent. Figure 5A shows the overall performance of each deconvolution method when using different reference data. It can be noted that DSSC3 has a certain degree of advantages in estimating cell type proportions, while like all other methods, there is still room for improvement. When using LM22 as a reference, most deconvolution methods would be better in inferring *P* relative to using other references. Some methods perform better in estimating *P* when using 3' PBMCs as a reference than using 5' PBMCs, such as NNLS, DeconRNASeq, ridge, and EPIC. As to the methods using scRNA-seq data as input, DWLS performs outstandingly. To further check each sample, Fig. 5B shows the PCC of *P* calculated for every sample based on the cell type proportions estimated by all deconvolution methods. We found that for most methods, the PCC varies widely among the samples under a given reference data and was also influenced by the choice of reference, such as NNLS. Among all methods, DSSC3 overall is more stable and performs well, especially for LM22 and 5' PBMCs, and DWLS is outstanding when using 3' PBMCs. We also checked the result at the cell type level. Supplementary Fig. 5B shows the PCC between the inferred cell type-specific gene expression and the referenced one calculated for every cell type based on the cell type-specific GEPs estimated by the deconvolution methods inferring *C* and *P*. To summarize the results under different references, we calculated the average of PCC of each cell type across different references (Supplementary Fig. 5C). It can be seen that DSSC3 and RNA-Sieve can keep the similarity between the referenced GEPs and the inferred one. Same as the case in mixture data, RNA-Sieve also demonstrates this characteristic. However, we do not know the proximity of the reference data to the real ground-truth, so we mainly focused on the result of *P*.

To summarize the performance indicators under different references, we calculated the average of PCC of each sample across different references (Fig. 5C), showing the relatively better performance of DSSC3 and DWLS. Moreover, in practical use, we may not know the proximity of the reference data to the real ground-truth, therefore we would like to integrate the deconvolution results using different reference data. We first calculated the average of the cell type proportion matrices obtained using different references and then calculated the performance index of each sample (Fig. 5D). It can be found that the performances of many deconvolution methods improve to varying degrees, and DSSC3 is still advantageous. Since for real experimental data deconvolution, different reference data will generate different marker genes, we did not integrate the cell type-specific GEPs obtained from different reference data. It can be seen from this that integrating multiple reference data would improve the deconvolution performance. Here, we used the simplest way to integrate the results from different references, while a weighted average of different references may be better.

Impact of marker gene and sample size

To explore the robustness of deconvolution methods to the number of marker genes and the number of bulk samples, we took the Nestorowa data in the intra-dataset deconvolution and the Baron_Muraro data in the inter-dataset deconvolution as examples and used the PCC between the flattened cell type proportion matrix and the flattened matrix of ground-truth as an indicator. Figure 6A and Supplementary Fig. 6A show the change of PCC with the increase of the number of marker genes (from using 25–50%, 75%, and 100% of all marker genes) when the number of simulated bulk samples is 10, 50, 100, 500, 1000, and 2000, respectively. It can be found that relative to some deconvolution methods, DSSC3 is generally stable and does not require a high number of marker genes to achieve a good inference. Figure 6B and Supplementary Fig. 6B shows the change of PCC with the increase of sample size in the case of using 25%, 50%, 75%, and 100% of all marker genes, respectively. Overall reference-based deconvolution methods are more stable, while the reference-free deconvolution methods fluctuate more greatly.

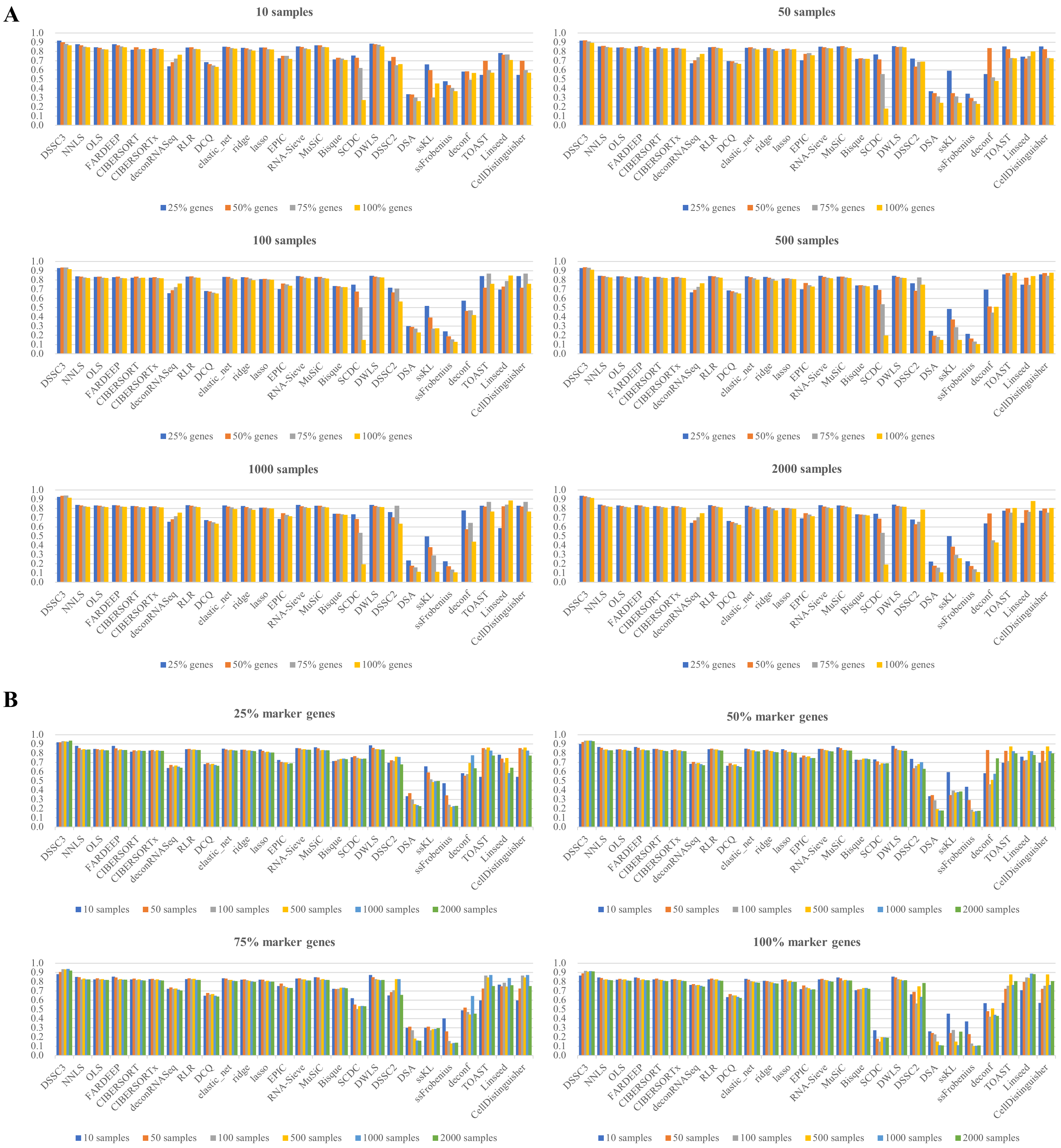


Fig. 6 The influence of marker gene number and sample size on intra-dataset deconvolution results using Nestorowa data as an example. **A** Under the same number of samples, with the increase of the number of marker genes, the change of PCC of cell type proportion matrix. **B** Under the same number of marker genes, with the increase of sample size, the change of PCC of cell type proportion matrix

Until now, all the simulated bulk samples we generated consist of 100 randomly selected single cells, which is a fixed number. To investigate the influence of variable number of cells contained in every bulk sample, we used the Nestorowa data in the intra-dataset deconvolution and the Baron_Muraro data in the inter-dataset deconvolution as examples to generate simulated bulk data, for which each bulk sample consists of a random number of cells, maybe 50, 100, 150, 200, 250, or 300. Then we checked the performance of all deconvolution algorithms (Supplementary Fig. 7) and found our method can still be better than many other methods.

To better compare the performance in inferring cell type-specific GEPs and cell type proportions, we used a unified way to make the dimensions of inputs and outputs are corresponding across different methods, making the calculation of *P* and *C* performance measures be based on the same genes and the same cell types. Specifically, we used a unified way to provide cell type-specific averaged gene expression of marker genes for the methods whose inputs are cell type-specific GEPs, or provide gene expression of marker genes for the methods whose inputs are single-cell expression data. For the methods requiring cell type-specific GEPs, the referenced inputs are the same across these methods as the training sets are the same. We calculated the PCC and RMSE between the unified referenced input and the ground-truth for these methods and also for showing the actual similarity between the reference and the truth (Supplementary Fig. 8A). For the methods which accept single-cell RNA-seq data, indeed they may filter genes or identify marker genes in their ways and then perform deconvolution using the genes. Therefore, for these methods, we re-ran all related tests only providing the same original single-cell gene expression. As the number of genes even the genes may be different across different methods after internal normalization, filtering or even marker gene identification, for each method we calculated the performance indexes for the genes actually outputted. Supplementary Fig. 8B shows the PCC and RMSE values for these methods. Compared with the results based on using the unified marker genes, the performances decrease in some datasets and are close to the ones based on the unified makers in the other datasets. This may because that our way of identifying marker genes may be stringent to provide important genes for deconvolution.

Impact of heterogeneity in cell type proportions

To check if our method can be applicable to the deconvolution of cell types with heterogeneous proportions across different bulk samples compared with the existing methods, we used the Nestorowa dataset with 50 simulated samples and all marker genes (shown in Fig. 6A) as a visualization example and plotted the true cell composition of each sample and the predicted one by the top methods on this dataset, DSSC3, FARDEEP, DWLS, and Linseed (Supplementary Fig. 9). The results demonstrate that the simulation process is random where for some samples the cell type proportion is relatively uniform while the proportion is heterogeneous for some other samples. For bulk samples containing few kinds of cell types, DSSC3 still exhibits certain advantages compared with the other methods. Moreover, from these 50 simulated bulk samples, we selected the ones containing only one or two major cell types whose proportions are larger than 95% (Supplementary Fig. 10A). Then we also plotted the predicted results by DSSC3, FARDEEP, DWLS, and Linseed and calculated the PCC and RMSE for each sample (Supplementary Fig. 10B). Although there is still room for improvement, DSSC3 shows a better overall performance.

Runtime and memory requirements

We used the Segerstolpe data in the intra-dataset experiment as an example to test the time and memory requirements of deconvolution algorithms in the case of different numbers of marker genes (25%, 50%, 75%, and 100% of all marker genes) and different numbers of simulated bulk samples (10, 50, 100, 500, 1000 and 2000). Since CIBERSORTx conducts experiments on the website, we could not record its time and memory requirements. Supplementary Fig. 11 shows the effect of the number of marker genes on time and memory requirements under the same sample size. Supplementary Fig. 12 shows the effect of sample size on time and memory requirements under the same number of marker genes. DWLS, TOAST, CIBERSORT, and RNA-Sieve need more runtime relative to other methods, and as the number of marker genes or sample size increases, the runtime will increase. Overall, deconvolution methods need more memory with the increase of sample sizes or marker genes. DSSC, like some deconvolution methods, does not require a lot of runtime and memory.

Convergence and parameter tuning of DSSC

DSSC is an algorithm based on non-negative matrix factorization. To examine its convergence, we recorded the change of the objective function with iterations for one example test of intra-dataset deconvolution (Segerstolpe data), inter-dataset deconvolution with matching cell types (Baron_Muraro), inter-dataset deconvolution with unmatching cell types, mixture data deconvolution (LiverBrainLung), and real experimental data deconvolution (WholeBlood + LM22), respectively. Supplementary Fig. 13A shows that DSSC algorithm can usually converge quickly.

As to the parameters of DSSC, the parameters of the sample-sample similarity matrix (λ_1), gene-gene similarity matrix (λ_2), and referenced data (λ_c) affect the inferred cell type proportion matrix and cell type-specific GEPs. They can balance the information obtained from bulk data and reference data to improve deconvolution accuracy. Using the representative tests, we recorded the PCC between the flattened inferred cell type proportion matrix and the flattened matrix of ground-truth for each parameter combination of λ_1,λ_2, andλ_c. Supplementary Fig. 13B indicates that λ_1 and λ_2 are usually small values, and if the referenced information is reliable, parameter λ_c could be set larger to generate a positive impact (seen as the cases of intra-dataset and mixture data deconvolution), otherwise, one can increase the value of λ_1 and λ_2. Supplementary Fig. 13C lists the hyperparameters used in DSSC3 and DSSC2 for each dataset. In practical usage, users can apply the cross-validation function we provide to select the optimal parameters of DSSC from the ranges set by users.

Discussion

Single-cell RNA-seq profiles the transcriptomic variations among individual cells to characterize cellular heterogeneity, while it is now impractical for a clinical setting or a large number of samples. Moreover, the noises contained in the process of single-cell RNA-seq make the accuracy of single-cell gene expression data lower than bulk gene expression. In recent years, there have been many computational methods developed to perform cell deconvolution from bulk RNA-seq data, while many of them only estimate the proportions of cell types. Unlike these methods, we proposed a new deconvolution method DSSC to estimate cell type-specific GEPs and cell type proportions at the same time, by matrix factorization and leveraging the information of sample-sample similarity, gene-gene similarity, and single-cell gene expression.

To test the performance of DSSC, we conducted simulation experiments and experimental data tests and compared it with the existing deconvolution methods. In the simulation experiments, we differentiated the situations of the referenced data and the simulated bulk data coming from the same source and different sources, i.e., intra-dataset and inter-dataset deconvolution. Moreover, in the case of inter-dataset simulations, we explored if matching cell types between the training set and testing set before deconvolution or not would influence the performance and also examined the impact of different normalization methods between the training set and testing set. The results demonstrated that DSSC has strong stability and accuracy, and is still reliable even in the interferences of batch effect and unmatched cell types. We also conducted deconvolution on five groups of mixture bulk data. Compared with simulation experiments, the sample sizes and the numbers of cell types contained in mixture data are smaller, while these have not affected the effectiveness of DSSC. Considering the practical usage scenario, we tested on real experimental bulk data with a variety of different references and also integrated the results obtained from different references. There is room for improvement for all deconvolution methods, while the overall performance of DSSC is still at the forefront. When different references are available, it is recommended to average or weighted average the cell type proportion matrices obtained from different references to obtain the final inferred cell type proportions. Moreover, the performance of DSSC is robust to the change of marker gene and sample size and does not require a lot of runtime and memory.

Compared with the other deconvolution methods estimating cell type abundance and cell type-specific GEPs simultaneously by leveraging single-cell RNA-seq data, such as CIBERSORTx, BLADE, and BayesPrism, we leverage the information of sample-sample similarity and gene-gene similarity contained in bulk gene expression. We hypothesize that similar samples have similar cell type compositions and similar genes have similar gene expression profiles, therefore we would like to maintain the sample-sample similarity and gene-gene similarity in the process of deconvolution. Based on non-negative matrix factorization, the gene-gene similarity and sample-sample similarity matrices calculated from the bulk gene expression were introduced to constrain the inferred cell type-specific GEPs and cell type proportion matrix, respectively. As to the information of single-cell RNA-seq data, it can be obtained from single-cell expression data or directly supplied a given cell type-specific gene signature, to further constrain cell type-specific GEPs. By implementing the objection function, the inferred cell type-specific GEPs maintain the gene-gene similarity and approach the referenced signature and the inferred cell type proportion matrix maintain the sample-sample similarity, thus improving the accuracy of deconvolution. The corresponding parameters can be tuned to balance the information calculated from bulk data and that from reference for achieving a good deconvolution performance. DSSC algorithm can also be tuned between reference-based DSSC3 and reference-free DSSC2. It has been demonstrated that DSSC can accurately infer cell type-specific gene expression and cell type abundance and provides a practical and promising alternative to the techniques that require expensive experimental equipment and a lot of labor to separate single cells from heterogeneous samples.

For future improvements or applications of DSSC, several aspects can be considered. Currently, DSSC algorithm requires the number of cell types which is determined by the dimension of referenced GEPs, singular value decomposition can be applied to determine the possible number of cell types. Considering the room for improvement on estimating the proportion of not included cell types, other calculation methods can be designed to update sample-sample and gene-gene similarity matrices. Other methods can also be used to implement DSSC algorithm, such as the gradient descent algorithm. Additionally, DSSC can be adapted to spatial transcriptomic data of bulk samples, by incorporating sample location information to update the calculation of sample-sample similarity for enhancing the deconvolution accuracy of the algorithm.

Conclusions

We proposed a new deconvolution algorithm DSSC to simultaneously estimate cell type-specific gene expression profiles and cell type abundance from bulk gene expression by leveraging gene-gene and sample-sample similarities and using single-cell RNA-seq data or cell type-specific gene signature as a reference. Using pseudo-bulk data generated by intra- and inter-dataset simulations, mixture bulk data, and real experimental data, we compared DSSC with various deconvolution methods to demonstrate its reliability, effectiveness, and robustness. DSSC facilitates the study of cell heterogeneity in gene expression based on bulk RNA-seq data and the deconvolution of cell types in heterogeneous tissues.

Availability and requirements

- Project name: DSSC.
- Project home page: https://github.com/JGuan-lab/DSSC.
- Operating system(s): Platform independent.
- Programming language: R.
- Other requirements: R 4.1.0 or higher.
- License: GNU GPL.
- Any restrictions to use by non-academics: None.

Abbreviations

GEP	Gene expression profile
scRNA-seq	Single-cell RNA-seq
PCC	Pearson's correlation coefficient
RMSE	Root mean square error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10728-x.

Supplementary Materials 1.

Acknowledgements

Not applicable.

Authors' contributions

JG conceived and designed the study. JG, CW, YL, and SL conducted the analyses. JG, CW, and YL wrote the manuscript. All authors approved the final manuscript.

Funding

This work was supported by National Science and Technology Major Project (No. 2021ZD0112600), National Natural Science Foundation of China (No. 61803320), Natural Science Foundation of Fujian Province of China (No. 2022J05012), and the Foundation of Key Laboratory of System Control and Information Processing, Ministry of Education, China (No. Scip20240104).

Availability of data and materials

DSSC can be accessed at GitHub: https://github.com/JGuan-lab/DSSC. The single-cell expression datasets used for intra- and inter-dataset simulation were originally downloaded from: https://github.com/hemberg-lab/scRNA.seq.datasets. The five sets of mixture data and corresponding ground-truth were obtained from: https://github.com/shmohammadi86/DeconvolutionReview. The WhooldBlood data, the references LM22, 3' PBMCs, and 5' PBMCs data, and the ground-truth were obtained from: https://cibersortx.stanford.edu. All analyzed inputs and corresponding ground-truths are deposited at: https://zenodo.org/record/8020767.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have declared no competing interests.

Received: 30 January 2024 Accepted: 20 August 2024
Published online: 18 September 2024

References

1. Bennett DA, Schneider JA, Buchman AS, Mendes de Leon C, Bienias JL, Wilson RS. The rush memory and aging roject: study design and baseline characteristics of the study cohort. Neuroepidemiology. 2005;25(4):163–75.
2. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
3. Kuhn A, Kumar A, Beilina A, Dillman A, Cookson MR, Singleton AB. Cell population-specific expression analysis of human cerebellum. BMC Genomics. 2012;13(1): 610.
4. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11(1):5650.
5. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34(11):1969–79.
6. Kang K, Meng Q, Shats I, Umbach DM, Li M, Li Y, Li X, Li L. CDSeq: a novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. PLoS Comput Biol. 2019;15(12): e1007510.
7. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12(4):298–306.
8. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, List M, Aneichyk T. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35(14):i436-45.
9. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82.
10. Basu S, Campbell HM, Dittel BN, Ray A. Purification of specific cell population by fluorescence activated cell sorting (FACS). J Vis Exp. 2010;41:e1546.

11. Schmitz B, Radbruch A, Kümmel T, Wickenhauser C, Korb H, Hansmann ML, Thiele J, Fischer R. Magnetic activated cell sorting (MACS)--a new immunomagnetic method for megakaryocytic cell isolation: comparison of different separation techniques. Eur J Haematol. 1994;52(5):267–75.
12. Coons AH, Creech HJ, Jones RN. Immunological properties of an antibody containing a fluorescent group. Proc Soc Exp Biol Med. 1941;47(2):200–2.
13. Hao Y, Yan M, Heath BR, Lei YL, Xie Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLoS Comput Biol. 2019;15(5): e1006976.
14. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
15. Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, Meningher T, Mendelson E, Mandelboim M, Gat-Viks I, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol. 2014;10(2):720.
16. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29(8):1083–5.
17. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS ONE. 2009;4(7): e6098.
18. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE. 2012;7(7):e41361.
19. Vallania F, Tam A, Lofgren S, Schaffert S, Azad TD, Bongen E, Haynes W, Alsup M, Alonso M, Davis M, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. Nat Commun. 2018;9(1):4735.
20. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics. 2013;8(3):290–302.
21. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, Samet Y, Maoz M, Druid H, Arner P, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9(1):5068.
22. Li Z, Wu H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome Biol. 2019;20(1):190.
23. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):380.
24. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, Jiang Y. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2021;22(1):416–27.
25. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. Accurate estimation of cell-type composition from gene expression data. Nat Commun. 2019;10(1):2975.
26. Repsilber D, Kern S, Telaar A, Walzl G, Black G, Selbig J, Parida S, Kaufmann S, Jacobsen M. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics. 2010;11:27.
27. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat Commun. 2019;10(1):2209.
28. Andrade Barbosa B, van Asten SD, Oh JW, Farina-Sarasqueta A, Verheij J, Dijk F, van Laarhoven HWM, Ylstra B, Garcia Vallejo JJ, van de Wiel MA, et al. Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. Nat Commun. 2021;12(1):6106.
29. Chu T, Wang Z, Pe'er D, Danko CG. 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(4):505–17.
30. Erdmann-Pham DD, Fischer J, Hong J, Song YS. Likelihood-based deconvolution of bulk gene expression data using single-cell references. Genome Res. 2021;31(10):1794–806.
31. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC Bioinformatics. 2017;18(1):105.
32. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, Irizarry RA, Teschendorff AE. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. Nat Methods. 2017;14(3):216–7.
33. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25(5):571–8.
34. Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour–immune cell interactions. Nat Rev Genet. 2016;17(8):441–58.
35. Elyanow R, Dumitrascu B, Engelhardt BE, Raphael BJ. netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. Genome Res. 2020;30(2):195–204.
36. McCarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179–86.
37. Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24(4):593–607.
38. Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Bräuninger M, Lewitus E, Sykes A, Hevers W, Lancaster M, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proc Natl Acad Sci U S A. 2015;112(51):15672–7.
39. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, Quake SR. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A. 2015;112(23):7285–90.
40. La Manno G, Gyllborg D, Codeluppi S, Nishimura K, Salto C, Zeisel A, Borm LE, Stott SRW, Toledo EM, Villaescusa JC, et al. Molecular diversity of midbrain development in mouse, human, and stem cells. Cell. 2016;167(2):566-e580519.
41. Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, Wilson NK, Kent DG, Göttgens B. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood. 2016;128(8):e20-31.
42. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals Inter- and Intra-cell population structure. Cell Syst. 2016;3(4):346-e360344.
43. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, van Gurp L, Engelse MA, Carlotti F, de Koning EJ, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst. 2016;3(4):385-e394383.
44. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using Nanoliter droplets. Cell. 2015;161(5):1202–14.
45. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166(5):1308-e13231330.
46. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS ONE. 2011;6(11): e27156.
47. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type–specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9.
48. Siegert S, Cabuy E, Scherf BG, Kohler H, Panda S, Le Y-Z, Fehling HJ, Gaidatzis D, Stadler MB, Roska B. Transcriptional code and disease map for adult retinal cell types. Nat Neurosci. 2012;15(3):487–95.
49. Mohammadi S, Zuckerman N, Goldsmith A, Grama A. A critical survey of deconvolution methods for separating cell types in omplex tissues. Proc IEEE. 2017;105(2):340–66.
50. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
51. Chambers JM, Hastie TJ. Statistical models in S. Technometrics. 1993;35(2):227–8.

52. Mullen KM. nnls: The Lawson-Hanson NNLS algorithm for non-negative least squares. 2007.
53. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017;6: 6.
54. Ripley BD. Support functions and datasets for venables and ripley's MASS. 2015.
55. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
56. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, Sul JH, Pietiläinen KH, Pajukanta P, Halperin E. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun. 2020;11(1):1971.
57. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics. 2013;29(17):2211–2.
58. Zhong Y, Wan Y-W, Pang K, Chow LML, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. 2013;14(1):89.
59. Gaujoux R, Seoighe C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. Infect Genet Evol. 2012;12(5):913–21.
60. Bhattacharya A, Hamilton AM, Troester MA, Love MI. DeCompress: tissue compartment deconvolution of targeted mRNA expression panels using compressed sensing. Nucleic Acids Res. 2021;49(8):e48.
61. Newberg LA, Chen X, Kodira CD, Zavodszky MI. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. PLoS ONE. 2018;13(3): e0193067.

Publisher's note