DAISM-DNN: Highly accurate cell type proportion estimation with *in silico* data augmentation and deep neural networks

Yating Lin[1,#], Haojun Li[1,#], Xu Xiao[1], Wenxian Yang[2,*], Rongshan Yu[1,*]

[#] These authors contribute equally.

[*] Corresponding author. Email: wx@aginome.com, rsyu@xmu.edu.cn

[1] *School of Informatics, Xiamen University, China.*

[2] *Aginome Scientific, Xiamen, China.*

Understanding the immune-cell abundances of cancer and other disease-related tissues has an important role in guiding cancer treatments. We propose data augmentation through *in silico* mixing with deep neural networks (DAISM-DNN), where highly accurate and unbiased immune-cell proportion estimation is achieved through DNN with dataset-specific training data created from partial samples from the same batch with ground truth cell proportions. We evaluated the performance of DAISM-DNN on three publicly available real-world datasets and results showed that DAISM-DNN is robust against platform-specific variations among different datasets and outperforms other existing methods by a significant margin on all the datasets evaluated.

Keywords: Cell type proportion estimation, deconvolution, data augmentation, data simulations, deep learning

In cancer treatments, it has been shown that the cellular composition of immune infiltrates in tumours is linked directly to tumour evolution and response to treatment[1,2]. High intratumoural infiltration of lymphocytes and dendritic cells is a favourable prognostic marker for cancer treatment[3,4], while a high stromal content of cancer-associated fibroblasts (CAFs) and M2 macrophages has been shown to associate with poor outcomes[5,6]. Particularly, the recent progress in immunotherapy has led to durable clinical benefits, but only in a subpopulation of patients with "hot" tumour immune microenvironments that are characterized by high infiltration of lymphocytes. Therefore, the knowledge of patient-specific immune cell proportion of solid tumours is invaluable in predicting disease progression or drug response as well as stratifying patients to assign the most suitable treatment options.

In the past, fluorescence activated cell sorting (FACS) and immunohistochemistry (IHC) were used as gold standards to measure the cell components in a patient sample[7]. FACS requires a large amount of cells, which limits its clinical applications. On the other hand, IHC is only able to estimate the cellular composition of a single biopsy slice and may not represent the full tumour microenvironment (TME) due to its heterogeneity. More recently, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful technique to characterize cell types and states. However, the high costs of labour, reagents and equipment of scRNA-seq at present restrain it from broad applications in routine clinical practice.

With the rapid progress of RNA quantification technologies, such as microarray, high-throughput RNA-seq, and Nanostring, large-scale expression profiling of clinical samples has become feasible in routine clinical settings[8]. These methods only measure the average expression of genes from the heterogeneous sample in their entirety, but do not provide detailed information on their cellular composition. To bridge the gap, numerous computational methods have been proposed to estimate individual cell type abundance from bulk RNA data of heterogeneous tissues (Supplementary Table S1). With the bulk gene expression values as input, the abundance of each cell type from the mixed sample can be quantified by aggregating the expressions of the marker genes into an abundance score (MCP-counter[9]), or by measuring the enrichment level of the marker genes using statistical analysis (xCell[10]), or by deconvolution algorithms that adopt computational methods, such as least squares (quanTiseq[11], EPIC[12]), support vector regression (SVR) (CIBERSORT[13], CIBERSORTx[14]), or non-negative matrix factorization (NMF)[15], to derive an optimal dissection of the original sample based on a set of pre-identified cell type-specific expression signatures. Obviously, regardless of the actual computational methods being used, the adoption of any of these methods as a reliable clinical routine for cell type proportion estimation requires that its underlying assumptions to be held over a large variety of cell types, tissues, and RNA sequencing conditions, which is challenging in practice. For example, in deconvolution-based algorithms, it is expected that the cell type-specific expression signature should truly represent the expression characteristics of the underlying immune cells from the mixture samples. Unfortunately, the signature gene expressions employed in existing methods were derived from either FACS-purified and *in vitro* differentiated or stimulated cell subsets, or single-cell experiments. The application of antibodies, culture material or physical disassociation may affect the cell status, resulting in signatures that deviate from those of the actual cells *in vivo*. Moreover, technical and biological variations between

RNA quantification experiments may introduce additional confounding factors that lead to sample or dataset-specific bias in cell type estimation. Similarly, marker gene expression aggregating methods such as MCP-counter require highly specific signature with genes that are exclusively and stably expressed on certain cell types, which may not be possible for some immune cell lineages[16].

To illustrate the above-mentioned limitations, we performed an extensive evaluation of nine state-of-the-arts cell type proportion estimation algorithms including CIBERSORT, CIBERSORTx, EPIC, quanTIseq, MCP-counter, xCell, ABIS, MuSiC and Scaden[22] on nine independent datasets (n = 641 total samples) acquired using different techniques or platforms (Supplementary Table S2). As it was previously reported that the accuracy of deconvolution-based cell type proportion estimation approaches is strongly dependent on the basis matrix rather than the deconvolution method itself[17], we included three established basis signature matrices, namely, IRIS[18], LM22[13] and immunoStates[17], in our evaluation of CIBERSORT. Importantly, methods under evaluation included several recent developments to improve the cross-dataset robustness of cell type proportion estimation methods. ImmunoStates[17] used a basis matrix built using 6160 samples with different disease states across 42 microarray platforms to mitigate the technical bias from different platforms. In MuSiC[19], the deconvolution algorithm further included appropriate weighting of genes showing cross-subject and cross-cell consistency to enable the transfer of cell type-specific expression information from one dataset to another. CIBERSORTx implemented bulk-mode batch correction (B-mode) to reduce the potential bias from batch effects. However, none of these methods were able to fully address the estimation bias problem and deliver consistently accurate results across multiple datasets in our evaluation (Friedman test with post hoc two-tailed Nemenyi test, Supplementary Fig. S1a,b). T-distributed stochastic neighbor embedding (t-SNE) analysis of all test datasets shows significant batch effects among those datasets and the difference among the testing samples is dominated by batch effect rather than cellular composition (Supplementary Fig. S1c). This result partially explains the inconsistent performance of the existing methods on different datasets and the challenge in developing a one-size-fits-most cell type proportion estimation method that performs uniformly well under different experiment conditions.

Recently, the development of deep neural networks (DNNs) has granted the computational power to resolve complex biological problems using data-driven approaches with the vast trove of data available from the biomedical research community powered by high-throughput genomic sequencing technologies[20,21]. An application of the DNN in cell type proportion estimation was proposed in Scaden[22], where a neural network was trained on bulk RNA-seq data simulated from single-cell RNA-seq (scRNA-seq) data of different immune cell types to predict cell type proportions from bulk expression of cell mixtures. A DNN-based model could automatically create optimal features for cell fraction estimation during the training process, thus alleviating the need to generate reliable gene expression profile (GEP) matrices for different immune cell types. Moreover, it learns the potentially intricate nonlinear relationships between the gene expression composition and cell type proportions from training data, which are not possible to be captured by linear models used in other deconvolution algorithms. However, as the performance of DNN is still subject to the same statistical learning principle that test and train conditions must match, Scaden, similar to other algorithms, showed inconsistent performance on different datasets in our evaluation (Supplementary Fig. S1). On the other hand, it is possible for DNN-based algorithms to deliver consistent performance under different experiment conditions, provided that sufficient ground truth data are available to train a specific predictive model for each distinct experiment condition. As a DNN model usually requires tens of thousands of training samples, the cost of implementing such a method would be prohibitive in practice.

To address these challenges, we developed an *in silico* data augmentation method to work together with a DNN model (DAISM-DNN) for robust and highly accurate cell type proportion estimation. The DAISM-DNN method performs model training on a dataset augmented from a calibration dataset, which refers to a small portion of the actual data from the same batch of RNA-seq experiment of which the ground truth cell type proportions are available for calibration. In order to create sufficient training data for DNN, the calibration dataset is augmented using purified cell RNA-seq data or scRNA-seq data that are publicly available, or derived from the same batch of samples (Fig. 1; all procedures are described in details in Methods). The DAISM-DNN method is also highly customizable and can be tailored to estimate the proportions of a large variety of cell types including those which are difficult for existing methods to estimate due to the lack of marker genes or GEP signature matrices. Moreover, it is able to handle complicated tasks on immune cells with overlapping marker or signature genes that are challenging, if not impossible, for existing methods[23,24].

We evaluated the performance of DAISM-DNN on RNA-seq dataset SDY67 containing 250 samples with ground truth proportions of five cell types. The training data included 200 randomly selected real samples from SDY67 augmented with scRNA-seq data of the five cell types. The performance of DAISM-DNN was then tested on the remaining 50 samples that were not used in training. For comparison, we performed cell type proportion estimation for the same 50 samples using other algorithms. Overall, by using the augmentation of a portion from the real-data as training data, DAISM-DNN outperformed all other algorithms by a significant margin in all the cell types under evaluation (Fig. 2 and Supplementary Fig. S2). When evaluated by the average per-cell-type Pearson correlation between the predicted and ground truth cell proportions, DAISM-DNN achieved the highest correlation, followed by Scaden (Box plot in Fig. 2b; results after 30 bootstrapping permutations). In addition, DAISM-DNN had the lowest root mean square error (RMSE) and the highest Lin's concordance correlation coefficient (CCC) followed by ABIS (Bar plots in Fig. 2b). Results show that DAISM-DNN significantly outperformed other methods by using pairwise Student's t-tests between the methods evaluated. We further tested DAISM-DNN on two microarray datasets, i.e., GSE59654 and GSE107990, with 153 samples and 164 samples respectively. Similarly, we randomly selected 50 samples as the test dataset and augmented the remaining samples with scRNA-seq data of the respective cell types to generate the training dataset for the DAISM-DNN model (Methods). As expected, DAISM-DNN outperformed other algorithms by a significant margin over all the cell types under evaluation (Fig. 2 and Supplementary Fig. S2), except for the ABIS method on dataset GSE107990.

For RNA-seq datasets, we provided an alternative mode (DAISM-RNA) using purified RNA-seq data to augment the data with ground truth cell type porportions (Methods). Compared with DAISM-scRNA mode and other algorithms, DAISM-RNA mode performed equally well as DAISM-scRNA mode and was significantly better than other algorithms (Supplementary Fig. S3). Results show that the global Pearson correlation and concordance correlation are very close in DAISM-RNA and DAISM-scRNA modes (Supplementary Fig. S3a), as well as per-cell-type Pearson correlation (Supplementary Fig. S3b), suggesting that purified RNA-seq data can be used for data augmentation on RNA-seq datasets. To better elucidate this, we further compared DAISM-RNA with other methods, showing that DAISM-RNA outperformed these methods and achieved the highest mean of per-cell-type Pearson correlation and concordance correlation (Supplementary Fig. S3c) and the lowest RMSE (Supplementary Fig. S3d) among all methods evaluated.

We performed t-SNE analysis to compare *in silico* mixed training data created using DAISM versus direct mixing of RNA-seq or microarray data from sorted cells or scRNA-seq data of selected cell types (Methods). All the *in silico* mixed training datasets followed the same cell proportions and only the mixing strategies differ. The t-SNE plot reveals highly distinct clusters of these datasets. Importantly, only the clusters of the DAISM-mixed dataset strongly overlapped with SDY67 while the clusters from the remaining datasets showed a clear gap with SDY67, demonstrating strong batch effects between them and the real samples (Supplementary Fig. S4). We further compared the deconvolution performance of the DAISM-DNN model trained using different *in silico* training datasets and found highly different cell fraction estimation performance when different training datasets were used (Supplementary Fig. S4b), indicating that the choice of the training data indeed plays a critical role in determining the performance of DNN-based models for cell type proportion estimation. Among all these training datasets, DAISM-DNN achieved significantly better performance on SDY67 when the DAISM-mixed training datasets were used (Supplementary Fig. S4b), suggesting the effectiveness of the proposed method in creating a training dataset that matches the real-life data to enable highly accurate cell type proportion estimation.

To identify the minimum number of samples needed to be sorted to achieve satisfactory performance, we further tested the cell fraction estimation performance of DAISM-DNN with respect to the number of real samples used for creating the augmented training data. We found that in general, the estimation accuracy improved with an increasing number of real samples used in creating the *in silico* mixed training data (Supplementary Fig. S5). When evaluate by CCC or RMSE, which requires that the predicted cell fractions should follow the real fractions in terms of absolute numbers, the estimation performance improved dramatically when the number of real samples used in *in silico* mixing increased from zero to around 20∼40. Beyond that, the rate of improvement slowed down significantly even if more real samples were used. Therefore, a calibration data consisting of 20∼40 samples with ground truth cell type proportions would be appropriate for DAISM to generate a suitable dataset to train a deep network with satisfactory performance.

We developed DAISM-DNN to meet the challenge of accurate cell type proportion estimation from bulk RNA expression quantification data. The DAISM method is able to populate a large quantity of *in silico* mixed training data that match the distribution of the target dataset from a small amount of tagged real samples (20∼40) with the aid from the large quantities of publicly available expression data of purified immune cells from microarray, RNA-seq or single-cell sequencing platforms. When using this *in silico* mixing strategy in combination with a DNN-based prediction model, DIASM-DNN delivers consistent cellular fraction estimation performance across different experiment conditions and breaks the performance barrier of existing algorithms by a significant margin. As such, DAISM-DNN would enable wide adoption of the digital sorting framework in biomedical research and clinical applications for exploring cellular composition in complex tissues given its accuracy and robustness.

Methods

Microarray datasets of purified cells. In total, 686 microarray expression profiles of eight purified immune cell types (B cells, CD4+ T cell, CD8+ T cell, monocytes, NK cells, neutrophils, endothelium and fibroblast) from Affymetrix Human Genome 133A platform were downloaded from Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo/; Supplementary Table S3). Expression profiles obtained as CEL files were RMA-normalized using R package "affy" (vesion 1.56.0) in Bioconductor. Probes were converted to HUGO gene symbols using chipset definition files available from the NCBI GEO. The maximum expression values at gene level were chosen when multiple probes mapped to the same gene. Finally, the *normalizeBetweenArrays* function from R package "limma" was used to normalize the expression intensities of all samples.

RNA-seq datasets of purified cells. For RNA-seq data of purified immune cells, we used 1533 purified cell samples of the same eight immune cell types as above in this study (Supplementary Table S4). Briefly, raw FASTQ reads were downloaded from the NCBI website, and transcription and gene-level expression quantification were performed using Salmon[25] (version 0.11.3) with Gencode v29 after quality control of FASTQ reads using fastp[26]. All tools were used with default parameters. Transcription per million (TPM) normalization was performed on all samples before further analysis.

scRNA-seq datasets. For scRNA-seq data of different immune cell types, we downloaded two scRNA-seq datasets of PBMCs from patient blood samples from the 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets, "8k PBMCs from a Healthy Donor" and "Frozen PBMCs (Donor B)"), which are denoted as PBMC8k and DonorB, respectively. PBMC8k was sequenced on Illumina Hiseq4000 with approximately 92,000 reads per cell, detecting 8,381 cells in total. There are 7,783 cells detected in DonorB, which was sequenced on Illumina Hiseq2500 with around 14,000 reads per cell. Briefly, raw scRNA-seq reads from both datasets were aligned to the GRCh38 reference genome and quantified by Cell Ranger (10X Genomics version 2.1.0). The resulting expression matrix was then processed using Seurat (v. 3.1.1)[27]. First, cells with less than 500 genes or greater than 10% mitochondiral RNA content, and genes expressed in less than 5 cells were excluded from analysis. Then, cells with abnormal high number of gene counts were considered as cell doublet and were excluded. Raw UMI counts were log-normalized and the top 2000 highly variable genes were called based on the average expression (between 0.0125 and 3) and average dispersion (> 0.5). Principal component analysis (PCA) was performed on the highly variable genes to further reduce the dimensionality of the data. Finally, clusters were identified using the shared nearest neighbor (SNN)-based clustering algorithm on the basis of the first 20 principle components with an appropriate resolution (between 0.4 and 1.2 based on the total number of cells).

Clusters were annotated on the basis of marker genes' expressions. The marker genes were obtained from the CellMarker database[28] based on the target cell types in peripheral blood, specifically, CD4 for CD4+ T cells, CD8A and CD8B for CD8+ T cells, MS4A1 and CD79A for B cells, CD14 and FCGR3A for monocytes, GNLY for NK cells, FLT3 and FCER1A for dendritic cells. Cell types were identified manually by checking if the respective marker genes were highly differentially expressed in each cluster. Clusters without high expression on the selected marker genes or with high expression on the marker genes of other cell types were grouped into the "unknown" type.

Datasets for benchmarking. We used nine public microarray and bulk RNA-seq datasets to evaluate the performance of different cell type proportion estimation methods (Supplementary Table S2), of which the expression data and the corresponding ground truth cell fractions were publicly available with reference to the original publications and used accordingly in our benchmarking tests.

Data augmentation through *in silico* mixing (DAISM). Deep learning-based approaches require a large amount of training data, but real tissue samples with known fractions of cell types and gene expressions are clearly insufficient for use as a training set. In this regard, we extracted a small number of real-life samples with ground truth cell type proportions to use as a calibration dataset, and used the DAISM strategy to create a large number of *in silico* mixing samples from this calibration dataset.

The expression profile of an *in silico* mixed sample of DAISM is calculated as follows. First, we generated a random variable r with uniform distribution between 0 and 1 to determine the fractions of the real sample in the mixed sample, and C random variables with Dirichlet distribution p_k, $k = 1, \ldots, C$ such that $\sum_{k=1}^{C} p_k = 1$ to determine the fractions of the immune cells in the mixed sample, where C is the number of cell types. The expression profile of the final mixed sample $\boldsymbol{e}$ is then calculated as:

$$\boldsymbol{e} = r\boldsymbol{\theta} + (1-r)\boldsymbol{\phi},$$

where $\boldsymbol{\theta}$ is the expression a real-life sample randomly selected from the calibration dataset as a seed sample for this *in silico* mixed sample, and $\boldsymbol{\phi}$ is the aggregated expression of single cell samples or purified samples used for data augmentation. When scRNA-seq dataset was used for data augmentation (DAISM-scRNA), we have

$$\boldsymbol{\phi} = \sum_{k=1}^{C} \sum_{j=1}^{n_k} \boldsymbol{\epsilon}_{kj},$$

where $n_k = 500 \cdot p_k$ is the number of cells of type k extracted randomly from scRNA-seq datasets for mixing, and $\boldsymbol{\epsilon}_{kj}$ denote their expression profiles. Note that $\boldsymbol{\phi}$ were further TPM-normalized before mixing. When RNA-seq or microarray data from purified cells were used (DAISM-RNA), we have

$$\boldsymbol{\phi} = \sum_{k=1}^{C} p_k \boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\epsilon}_k$ is the expression profile of a randomly selected purified sample of cell type k from the respective RNA-seq dataset. Once the expression profile of the *in silico* sample was created, its "ground truth" cell fractions can be calculated following the same concept as follows:

$$\rho_k = r\lambda_k + (1-r)p_k,$$

where ρ_k is the fraction of cell type k in the *in silico* mixed sample, λ_k is the ground truth fraction of cell type k in the selected seed sample which is known *a priori* through experiments, e.g., flow cytometry analysis.

DAISM-DNN algorithm. We trained deep feed-forward, fully connected neural networks (multi-layer perceptron networks) on *in silico* mixed training data by DAISM to predict the cell fractions from bulk expression data. The network consists of one input layer, five fully-connected hidden layers and one output layer and was implemented with the PyTorch framework (v1.0.1) in Python (v3.7.3). As DNN can fit a large feature space with its large number of parameters (connection weights), we did not perform any feature selection in advance. Instead, we used all the genes that were present in both the training and testing datasets as input to the neural network. Moreover, the expression profile of each sample was log2-transformed, and scaled to the range of [0,1] through min-max scaling before using as inputs to DNN:

$$\hat{e}_i = \frac{e_i - \min(\boldsymbol{e})}{\max(\boldsymbol{e}) - min(\boldsymbol{e})}.$$

Here, e_i is the log2-transformed expression of gene i, $\boldsymbol{e}$ is the vector of the log2-transformed expressions of all genes of a sample, and $\hat{e}_i$ is the min-max scaled expression.

The network was trained using the back-propagation algorithm with randomly initialized network parameters. Mean square error (MSE) between the actual and predicted absolute cell fractions was used as the loss function. The optimization function is set as Adam with an initial learning rate of 1e-4. During the training process, the training set was randomly divided into a number of mini-batches with a batch size of 32, which were then fed into the input layer of the network to obtain the predicted fraction values. The neural network then continuously adjusts its network parameters to minimize the difference between the predicted and actual cell fractions throught stochastic gradient descent. When the average MSE of all mini-batches in the current epoch is higher than that of the last epoch, the learning rate was multiplied by an attenuation coefficient until a minimum of 1e-5 is reached to avoid training noise from excessivly large learning rates when the network converge to steady state. Furthermore, we randomly split the training set and the validation set at a ratio of 9:1. Early-stopping strategy was adopted in the training process where training is stopped when the validation error does not decrease in 10 epochs, and the model producing the best results on the validation set during the training process was used as the final model for prediction.

Performance comparison. Since cell types abundances were resolved at different granularities in different deconvolution methods, regularizing the cell types of all methods to the same granularities has to be performed to facilitate a fair comparison. In this study, for comparison of the methods, we only test the performance of all methods on six specific coarse-grain cell types (B cells, CD4+ T cells, CD8+ T cells, NK cells, monocytes, neutrophils). The fine-grain cell type results of some methods were mapped to coarse-grain cell types according to a hierarchy of cell types as defined in Strum et al.[24].

We ran quanTIseq, MCP-counter, EPIC, xCell through an R package, *immunedeconv*[24], which provided an integrated inference to benchmark on six deconvolution methods. The parameter *tumor* was set FALSE on performing deconvolution on all test PBMC datasets. As recommended in the original paper, EPIC was run with BRef as the signature set on PBMC samples. More details on other methods tested in this paper are listed as follows.

- CIBERSORT (CS)

  CIBERSORT is a signature-based deconvolution algorithm which uses ν-support vector regression to estimate cell abundance. Predicted cell proportions were obtained using R source code requested from the CIBERSORT website (https://cibersort.stanford.edu/). In this study, to compare the effect of different signature matrices on the deconvolution performance, we used CIBERSORT with three signature matrices (LM22, IRIS and immunoStates) and denoted as three methods. The input data for CIBERSORT was in linear domain and all parameters were used as in their default settings.

- CIBERSORTx (CSx)

  CIBERSORTx is an extended version of CIBERSORT. CIBERSORTx enables generating the signature matrix from scRNA-seq data and provides two distinct batch correlation strategies (B-mode and S-mode) for cross-platform deconvolution. The B-mode was designed to remove technical differences between bulk profiles and signature matrices derived from bulk sorted reference profiles while S-mode was used for signature matrices derived from dropket-based or UMI-based scRNA-seq data. In this study, we tested both B-mode and S-mode batch correction methods to perform deconvolution. For B-mode, LM22 were used as the signature matrix. For S-mode, scRNA-seq dataset PBMC8k was chosen to generate the signature matrix using CIBERSORTx, which was then applied in further deconvolution. We ran CIBERSORTx on its website (https://cibersortx.stanford.edu/). Quantile normalization was disabled when input was RNA-seq or scRNA-seq simulated mixtures data.

- MuSiC

  MuSiC predictions were obtained from R source package *MuSiC* (https://github.com/xuranw/MuSiC). MuSiC is a deconvolution method which takes scRNA-seq data with cell type labels as reference. Deconvolution using MuSiC was performed following the tutorial provided by the authors with five coarse-grained cell types (B cells, CD4+ T cells, CD8+ T cells, NK cells and monocytes) from three single cell PBMC datasets download from the 10X Genomics website (PBMC6k, PBMC8k and Donor C) used as reference data for the deconvolution process.

- ABIS

  ABIS is a deconvolution method that enables absolute estimation of cell abundance from both bulk RNA-seq and microarray data. Deconvolution was performed through an R/Shiny app downloaded from https://github.com/giannimonaco/ABIS. The results provided were absolute cell frequencies and were divided by 100 in this study for comparison with other methods.

- Scaden

  Scaden is a DNN-based deconvolution algorithm. For Scaden, we used the same training datasets provided by Scaden (https://github.com/KevinMenden/scaden), which contains 32,000 artificial mixtures from four scRNA-seq datasets. The training was performed for 5000 steps per model on each dataset as recommended in the original paper.

Performances of DNN on different training datasets. We trained DAISM-DNN on training datasets generated from expression data of purified cells to compare the performance of DAISM-DNN with and without using real-life calibration samples. To this end, we generated three training datasets that used RNA-seq and microarray expression profiles of sorted cells and scRNA-seq data respectively. The generation of these three *in silico* bulk data followed the same linear mathematical operation as previously defined, with the only exception that real-life samples with ground truth cell fractions were not used in the mixing process. Briefly, for RNA-seq or microarray datasets, the expression of a simulated sample $\boldsymbol{e}$ was calculated as

$$\boldsymbol{e} = \sum_{k=1}^{C} p_k \boldsymbol{\epsilon}_k,$$

where C is the number of cell types involved in mixing, p_k, $k = 1, \ldots, C$ are random variables with Dirichlet distribution that determined the fractions of different cells in the *in silico* mixed sample, and $\boldsymbol{\epsilon}_k$ is the expression profile of a randomly selected purified sample of cell type k from the respective RNA-seq or microarray dataset. For scRNA-seq dataset, $\boldsymbol{e}$ is given by

$$\boldsymbol{e} = \sum_{k=1}^{C} \sum_{j=1}^{n_k} \boldsymbol{\epsilon}_{kj},$$

where $n_k = 500 \cdot p_k$ is the number of cells of type k extracted randomly from scRNA-seq datasets for mixing, and $\boldsymbol{\epsilon}_{kj}$ denote their expression profiles. Note that in this case $\boldsymbol{e}$ were further TPM-normalized before used as training data.

For comparison, we also generated two training datasets by the DAISM strategy with 200 randomly real-life samples from the SDY67 dataset as the calibration samples, which were further augmented using scRNA-seq data or RNA-seq data of purified samples to train the DAISM-DNN model. The remaining 50 samples from SDY67 not used in creating the training set were used for testing. For fair comparison, the same number of training samples (16,000) and the same cell fractions were used in the five training sets. For t-SNE analysis, all the 50 test samples of SDY67 and 500 randomly selected artificial mixtures from each training datasets were plotted based on the common genes from SDY67 and the five training datasets. The perplexity was set to 30 in this experiment and the other parameters were set as default.

Statistical analysis. We used three evaluation criteria to compare the performance of DAISM-DNN and other methods. Pearson corrlation r was used to measure the linear concordance between predicted cell proportions and the FACS ground truth. Lin's CCC and the RMSE were further used to evaluate the performance for methods which enable absolute cell type proportion estimation.

Differences in continuous measurements were tested using the two-tailed Student's t-test. Two-sided p-values were used unless otherwise specified, and a p-value less than 0.05 was considered significant. Ranking of the algorithms over multiple testing sets was determined using Friedman test with post hoc two-tailed Nemenyi test[29]. PRISM was used for basic statistical analysis and plotting (http://www.graphpad.com), and the Python or R language and programming environment were used for the remainder of the statistical analysis.

Data availability

All expression datasets analyzed in this work, including accession codes and web links (if available), are listed in Supplementary Table S2.

Code availability

The source code for DAISM-DNN will be available once the paper is accepted for publication.

Authors' contributions

Y.L., H.L. and X.X. performed the computational analysis, analyzed data and generated figures. W.Y. and R.Y. conceived the project and designed the methodology. All authors assisted to write the manuscript.

Acknowledgements

Li Chenxin helped to perform some simulation tests in this paper.

List of abbreviations

RNA-seq: Next Generation RNA Sequencing

scRNA-seq: Single cell RNA-seq

GEP: Gene expression profile matrix

SVR: Support Vector Regression

DNN: Deep Neural Network

PBMC: Peripheral Blood Mononuclear Cells

CCC: Lin's Concordance Correlation Coefficient

r: Pearson's correlation coefficient

RMSE: Root Mean Square Error

CS: CIBERSORT

CSx: CIBERSORTx

1. Fridman, W. H., Pagès, F., Sautes-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer* **12**, 298–306 (2012).

2. Palucka, A. K. & Coussens, L. M. The basis of oncoimmunology. *Cell* **164**, 1233–1247 (2016).

3. Huang, A. C. *et al.* T-cell invigoration to tumour burden ratio associated with anti-pd-1 response. *Nature* **545**, 60–65 (2017).

4. Fridman, W. H., Zitvogel, L., Sautès-Fridman, C. & Kroemer, G. The immune contexture in cancer prognosis and treatment. *Nature reviews Clinical oncology* **14**, 717 (2017).

5. Kalluri, R. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer* **16**, 582 (2016).

6. DeNardo, D. G. & Ruffell, B. Macrophages as regulators of tumour immunity and immunotherapy. *Nature Reviews Immunology* **19**, 369–382 (2019).

7. Petitprez, F., Sun, C., Lacroix, L. *et al.* Quantitative analyses of the tumor microenvironment composition and orientation in the era of precision medicine. *Frontiers in oncology* (2018).

8. Hrdlickova, R., Toloue, M. & Tian, B. Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA* **8**, e1364 (2017).

9. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17**, 218 (2016).

10. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 220 (2017).

11. Finotello, F. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data. *Genome Medicine* **11**, 34–34 (2019).

12. Racle, J., De Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, 1–25 (2017).

13. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453–457 (2015).

14. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773–782 (2019).

15. Chang, W. *et al.* Ictd: A semi-supervised cell type identification and deconvolution method for multi-omics data. *bioRxiv* 426593 (2019).

16. Danaher, P. *et al.* Gene expression markers of tumor infiltrating leukocytes. *Journal for immunotherapy of cancer* **5**, 18 (2017).

17. Vallania, F. *et al.* Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature Communications* **9**, 4735 (2018).

18. Abbas, A. R. *et al.* Immune response in silico (iris): immune-specific genes identified from a compendium of microarray expression data. *Genes and Immunity* **6**, 319–331 (2005).

19. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications* **10**, 380 (2019).

20. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20**, 389–403 (2019).

21. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **15**, 20170387 (2018).

22. Menden, K., Marouf, M., Dalmia, A., Heutink, P. & Bonn, S. Deep-learning-based cell composition analysis from tissue expression profiles. *bioRxiv* 659227 (2019).

23. Monaco, G. *et al.* Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell Reports* **26**, 1627 (2019).

24. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35** (2019).

25. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).

26. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics* **34** (2018).

27. Butler, A., Hoffman, P. J., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018).

28. Zhang, X. *et al.* Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research* **47** (2019).

29. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* **7**, 1–30 (2006).

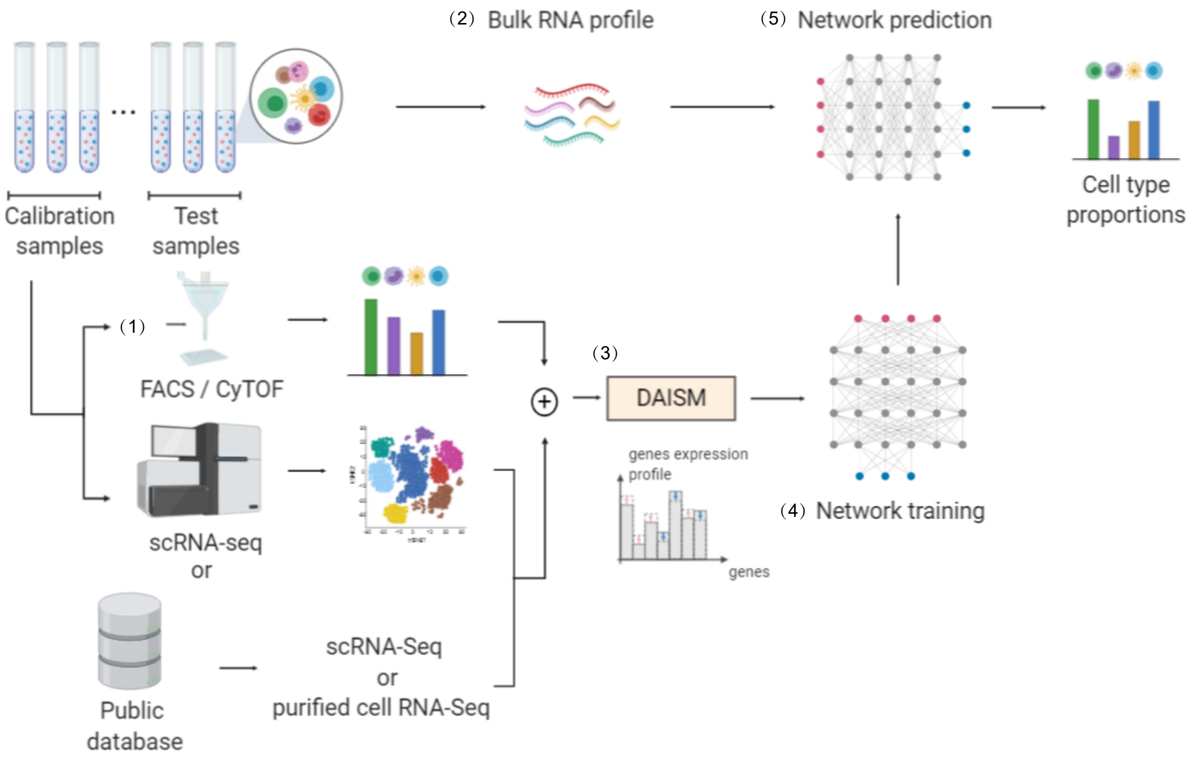

Figure 1: The framework of DAISM-DNN. A typical DAISM-DNN workflow involves several steps when performing cell type proportion estimation: (1) measuring the ground truth proportions of cell types of interest on a small portion of calibration samples (e.g., 20∼40 samples) from the batch of samples to be evaluated; (2) performing bulk RNA-seq on the calibration samples to obtain their expression profiles; (3) performing data augmentation on the expression profiles of the calibration samples through *in silico* mixing with RNA-seq data of purified cells or scRNA-seq data (DAISM); (4) performing DNN training on the augmented data; (5) using the trained DNN model for cell type proportion estimation for the remaining samples with their bulk expression profiles. Note that steps (1)-(4) could be optional if the DNN model has already been trained for the given RNA-seq experiment conditions.

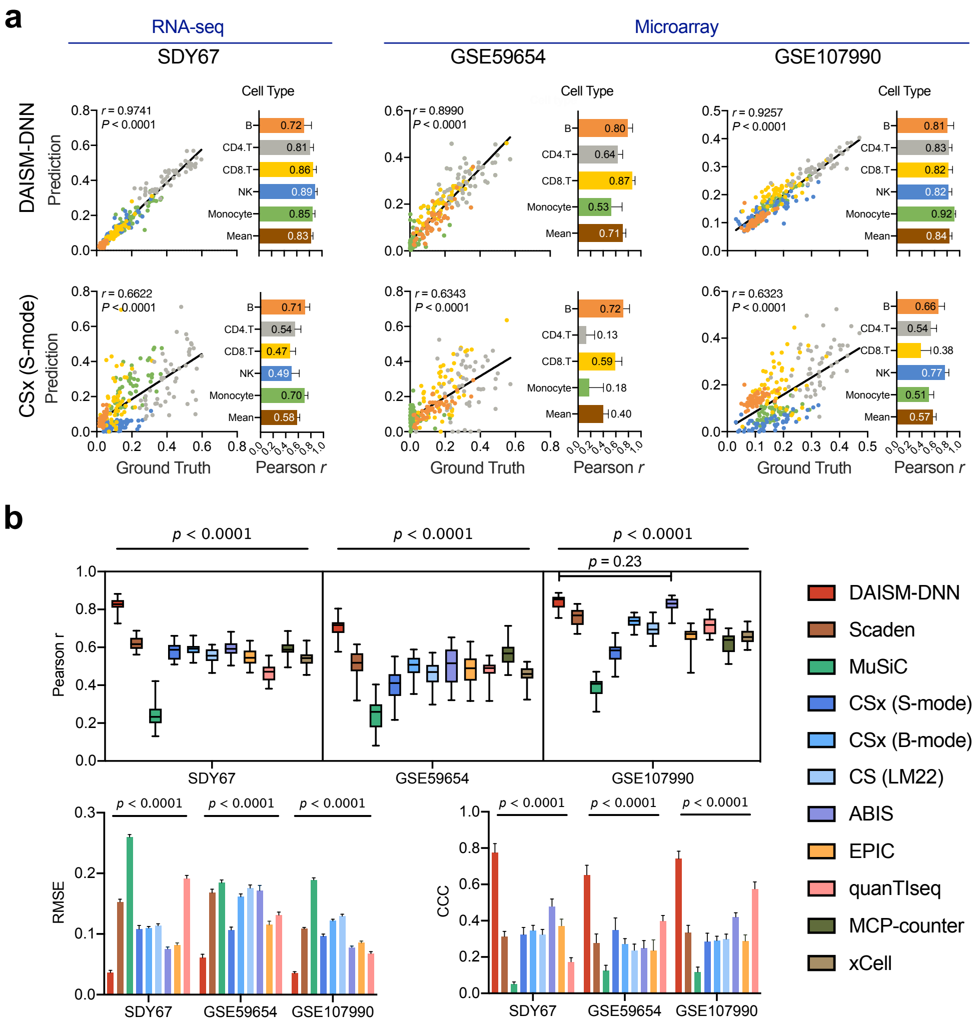

Figure 2: Performance of different algorithms on datasets SDY67, GSE59654, and GSE107990. (a) Scatter plots of ground truth fractions (x-axis) and predicted cell fractions (y-axis) for DAISM-DNN, CIBERSORTx (S-mode). Bar plots aside show the Pearson correlation on each cell type. (b) Box plot of average per-cell-type Pearson correlation on 11 methods. Bar plots of RMSE (right) and CCC (left) for 9 methods. Note that RSEM and CCC are not suitable to evaluate the two marker-based methods, MCP-counter and xCell.