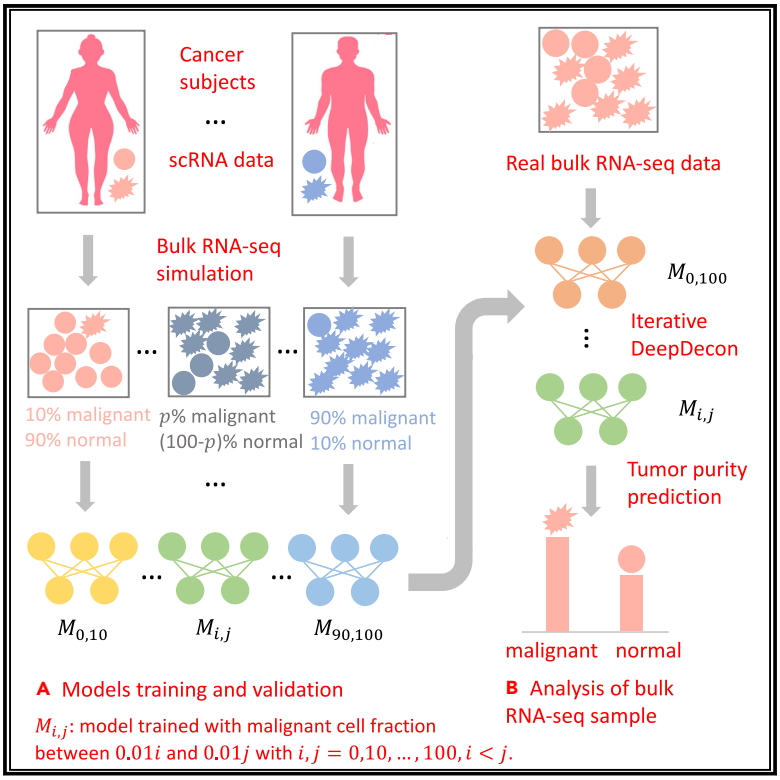
Patterns

DeepDecon accurately estimates cancer cell fractions in bulk RNA-seq data

Graphical abstract



Highlights

- DeepDecon refines estimation of malignant cell fraction via an iterative strategy
- Normalization using TF-IDF further improves the estimation accuracy
- Applications to multiple cancer samples show the wide applicability of DeepDecon

Authors

Jiawei Huang, Yuxuan Du, Andres Stucky, Kevin R. Kelly, Jiang F. Zhong, Fengzhu Sun

Correspondence

jzhong@llu.edu (J.F.Z.), fsun@usc.edu (F.S.)

In brief

The authors developed an iterative deep-learning-based method, DeepDecon, to accurately estimate the fraction of malignant cells based on bulk RNA-seq data using single malignant and normal cells as references. Extensive simulations and applications to clinical AML, neuroblastoma, and HNSCC bulk RNA-seq data show the power of DeepDecon over currently available deconvolution methods to estimate the malignant fraction of the cancer cell population accurately. DeepDecon can be used with cost-effective bulk RNA-seq data for cancer detection, prognosis, and recurrence monitoring.

Huang et al., 2024, Patterns 5, 100969
May 10, 2024
https://doi.org/10.1016/j.patter.2024.100969

CellPress

Patterns

CellPress
OPEN ACCESS

Article

DeepDecon accurately estimates cancer cell fractions in bulk RNA-seq data

Jiawei Huang,[1] Yuxuan Du,[1] Andres Stucky,[2] Kevin R. Kelly,[3] Jiang F. Zhong,[2,*] and Fengzhu Sun[1,4,*]
[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[2]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[3]Division of Hematology, University of Southern California, Los Angeles, CA 90089, USA
[4]Lead contact
*Correspondence: jzhong@llu.edu (J.F.Z.), fsun@usc.edu (F.S.)
https://doi.org/10.1016/j.patter.2024.100969

THE BIGGER PICTURE Estimating the malignant cell fraction accurately and cheaply is essential for cancer diagnosis and prognosis. Although single-cell RNA sequencing (scRNA-seq) can provide accurate information on malignant cell fraction, it is too labor intensive and expensive for clinical application. Bulk RNA-seq, on the other hand, is cost effective and widely used in clinical settings but traditionally only provides the average gene expression profiles of a cancer cell population. Using reference malignant and normal scRNA-seq data, DeepDecon provides an iterative deep-learning-based computational method for accurate estimation of the fraction of malignant cells based on the bulk-averaged gene expression profiles. This study used DeepDecon to accurately estimate acute myeloid leukemia (AML), neuroblastoma, and head-and-neck squamous cell carcinoma (HNSCC) cell fractions from bulk RNA-seq data.

SUMMARY

Understanding the cellular composition of a disease-related tissue is important in disease diagnosis, prognosis, and downstream treatment. Recent advances in single-cell RNA-sequencing (scRNA-seq) technique have allowed the measurement of gene expression profiles for individual cells. However, scRNA-seq is still too expensive to be used for large-scale population studies, and bulk RNA-seq is still widely used in such situations. An essential challenge is to deconvolve cellular composition for bulk RNA-seq data based on scRNA-seq data. Here, we present DeepDecon, a deep neural network model that leverages single-cell gene expression information to accurately predict the fraction of cancer cells in bulk tissues. It provides a refining strategy in which the cancer cell fraction is iteratively estimated by a set of trained models. When applied to simulated and real cancer data, DeepDecon exhibits superior performance compared to existing decomposition methods in terms of accuracy.

INTRODUCTION

For centuries, biologists have recognized that multicellular organisms are composed of a vast array of distinct cell types.[1] Cells and tissues play a critical role in all living organisms. Tissues are composed of cells, and cells are responsible for making up the different types of tissues in all multicellular organisms. Classifying and quantifying cells are crucial to have a detailed understanding of how tissues function and interact with one another and the microenvironment and to reveal mechanisms underlying pathological states. For example, tumor tissues are heterogeneous and consist of different fractions of cell types. Cancer identification, treatment, and clinical outcomes such as tumor growth, metastasis, recurrence, and drug resistance have a direct relation with cell-type composition and its changes.[2–4] Quantifying cell-type fractions within tumor tissues can provide insight into the role of heterogeneity in disease and how particular environments can impact tumor biology.

RNA sequencing (RNA-seq) is an alternative method to conventional microarrays for transcriptome analysis.[5,6] Bulk RNA-seq provides a view of the average gene expression profiles (GEPs) within a whole organ or tissue. It can be regarded as the sum of the product of cell-type-specific gene expressions and corresponding cell-type proportions.[7] However, information on the variations of different cell types is lost in bulk RNA-seq. Single-cell RNA-seq (scRNA-seq) instead can help solve this problem. It allows for the quantification of transcripts for each cell and the further identification of new cell types based on GEPs.[8] In addition, it enables the assessment of heterogeneity in cohorts of patient samples, providing a deeper understanding

of disease states and aiding in the development of effective treatments.[4,9–11] As a result, scRNA-seq data generated from samples with similar microenvironmental conditions can potentially help tackle the problem of bulk-tissue deconvolution.

Many methods have been developed in recent years to decompose fractions of cell types in bulk tissues, and most of them use cell-type-specific GEPs, as in Avila Cobos et al.[12] and Mohammadi et al.[13] ESTIMATE[14] uses The Cancer Genome Atlas to infer the fraction of stromal and immune cells in tumor samples, which can be further used to approximate the proportion of cancer cells in bulk RNA-seq data. Non-negative least-squares regression (NNLS)[15,16] is an optimization method to solve this deconvolution problem through matrix decomposition, but it can be easily affected by the choice of GEP. Noise, imprecision, and missing data of GEPs can lead to poor performance of NNLS. CIBERSORT/CIBERSORTx[17,18] are two widely used deconvolution methods. CIBERSORT adopts a linear support vector regression (SVR) approach, representing the gene expression of a bulk sample as a weighted sum of gene expressions from different cell types. These weights are determined based on predefined GEPs. On the other hand, CIBERSORTx is an enhanced version of CIBERSORT that enables the generation of GEPs from scRNA-seq data. Another approach, MuSiC,[7] dynamically generates reference profiles from scRNA-seq data. It assigns high weights to genes with low cross-subject variance and low weights to genes with high cross-subject variance. However, MuSiC ignores the possibility of significant variations in tumor conditions between reference data and bulk data. Bisque[19] addresses the issue of simple summation of scRNA-seq profiles by adopting a linear transformation on artificially derived bulk RNA-seq samples. The transformed data are then used for decomposition. However, the success of this transformation heavily relies on the similarity in distribution between reference single cells and actual data. An alternative method, RNA-Sieve,[20] uses a likelihood-based inference method. It assumes that the estimates of cell-type fractions are normally distributed around the true fractions. MEAD,[21] on the other hand, is a statistical inference method that introduces a gene-gene dependence structure to improve accuracy. Nonetheless, the dependence matrix used in MEAD is highly dependent on the choice of bulk samples and cannot be generated when there is only one single bulk sample to decompose. Last, Scaden[22] leverages neural networks to predict cell fractions and has demonstrated superior performance compared to traditional deconvolution methods. It generates cell fractions by averaging the outputs of three different neural networks.

In this study, we introduce DeepDecon, an iterative deep neural network model designed to accurately estimate the proportion of cancer cells in bulk RNA-seq data. DeepDecon makes use of scRNA-seq gene expression information to generate artificial bulk RNA-seq datasets with known proportions of cancer cells in each artificial bulk RNA-seq sample. The artificial bulk RNA-seq datasets can be employed to train an iterative deep neural network model, which can subsequently be employed to accurately predict the proportions of cancer cells in novel cancer tissues. Our approach utilizes an iterative process to refine predictions and enhance estimation accuracy. Through extensive benchmark evaluations using both simulated and real data, we demonstrate that DeepDecon outperforms other existing methods across different cancer tissues and is also robust to the influence of gene expression perturbations and the number of cells per bulk sample. Overall, by leveraging scRNA-seq information, employing deep neural networks, and making use of an iterative refinement process, DeepDecon achieves superior performance in cancer cell deconvolution analysis.

RESULTS

Methods overview

Figure 1 shows the graphical overview of iterative DeepDecon. DeepDecon starts with scRNA-seq datasets and assumes the cells for each subject have labeled cell types (malignant/normal) and known gene expression levels. Therefore, simulated bulk RNA-seq datasets with known cell-type fractions can be generated from these scRNA-seq datasets (Figure 1A). In addition, simulated bulk RNA-seq datasets can be generated with specific ranges of malignant cell fractions. This allows us to develop an iterative deconvolution model. During the model training process, simulated bulk samples whose malignant cell fraction $p \in [0.01i, 0.01j], i < j, i, j \in \{0, 10, 20, \cdots, 100\}$ serve as the input to train a DeepDecon model $M_{i,j}$ (Figure 1B). The whole group of DeepDecon models will be used in the iterative process. The core architecture of DeepDecon is a group of deep neural networks (DNNs) that take bulk RNA-seq data as input and output predicted malignant cell fractions. These models share the same structure, consisting of four fully connected layers with dropout layers (Figure 1C). When presented with a real bulk sample, DeepDecon first generates an initial malignant cell prediction $\widehat{P}$ using the whole range model $M_{0,100}$. Then in each iteration, DeepDecon will narrow down the prediction interval and update the prediction $\widehat{P}$ with models trained on narrow-range datasets (Figure 1D). The selection of these narrow-range models is only determined by the previous prediction value and the training datasets (see Equations 4 and 5). By incorporating datasets with all kinds of malignant cell fractions and dynamically determining fraction-specific model iterations, DeepDecon allows for estimating cell proportions of bulk RNA-seq data accurately.

Our model was constructed using artificial bulk RNA-seq samples and evaluated through leave-one-out cross-validation. Root-mean-square error (RMSE), Pearson's correlation coefficient (r), and Lin's concordance correlation coefficient (CCC) values between the predicted fractions and the true fractions of malignant cells were used to evaluate the performance of different deconvolution methods.

DeepDecon outperforms other methods for estimating malignant cell fraction

To demonstrate and evaluate the performance of DeepDecon, we first compared DeepDecon with eight other methods (Scaden [v.1.1.2],[22] CIBERSORTx [https://cibersortx.stanford.edu/],[18] Bisque [v.1.0.5],[19] ESTIMATE [v.2.0.0],[14] MuSiC [v.1.0.0],[7] MEAD [v.1.0.1],[21] RNA-Sieve [v.0.1.4],[20] and NNLS [v.1.4][7,15]) on artificial bulk RNA-seq datasets. scRNA-seq data described under "Datasets" was used as reference data for Bisque, MEAD, RNA-Sieve, MuSiC, and CIBERSORTx. MuSiC will also give the output of the NNLS method, and we used it as our NNLS result. Artificial bulk RNA-seq datasets were used to train two neural network methods, DeepDecon and Scaden. We compared all benchmark

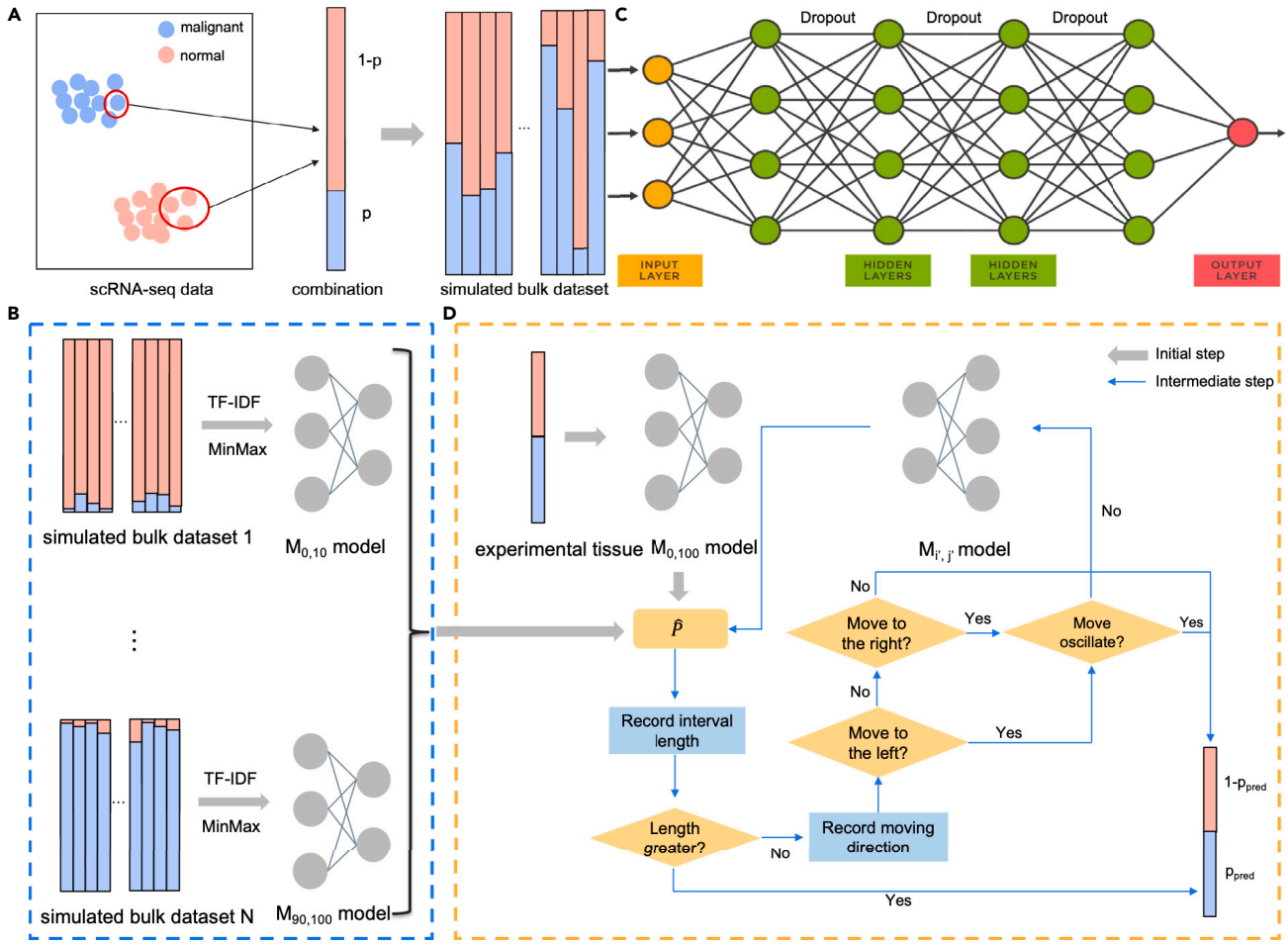


Figure 1. Overview of DeepDecon decomposition method

(A) Constructing simulated bulk RNA-seq samples with different fractions of malignant cells. p is the fraction of malignant cells in a simulated bulk sample.

(B) Training DeepDecon models using simulated bulk datasets with different malignant cell fractions. Simulated bulk samples whose malignant cell fraction $p \in [0.01i, 0.01j], i, j = 0, 10, \cdots, 100$ serve as the input to train a DeepDecon model M_{ij}.

(C) Core DeepDecon model structure. It consists of four fully connected layers with dropout layers. All DeepDecon models in the iterative process share the same structure.

(D) Predicting the fraction of malignant cells from a real bulk sample iteratively. DeepDecon designs an iterative strategy to narrow down the prediction interval of given bulk samples. When a new experimental tissue is given, DeepDecon first generates an initial malignant cell prediction $\widehat{P}$ using the whole range model $M_{0,100}$. In each iteration step, DeepDecon tries to limit the estimate to a smaller range, denoted by $[0.01i', 0.01j']$, based on the training datasets and the previous iteration prediction value. If the prediction interval can be shortened, DeepDecon will update the prediction value $\widehat{P}$ by a newly selected model $M_{i'j'}$. Ultimately, DeepDecon generates the final prediction p_{pred} when the stopping conditions are satisfied. The flowchart shows the iterative procedure and the stopping conditions.

methods with their default settings. All methods were evaluated on the same testing datasets that were separate from the training datasets used to train the above models. Details of the implementations of these compared methods are explained in supplemental experimental procedure 1.

Figure 2A shows the scatterplots of true malignant fractions with predicted malignant fractions for each method in three simulated acute myeloid leukemia (AML) datasets (Figure S1 shows the scatterplot in all 15 simulated AML datasets). Figures 2B–2D show the RMSE, correlation, and CCC metrics between the true and the estimated malignant cell fractions in all 15 simulated AML datasets. Table 1 also gives the RMSE values and average performance ranks of each method in simulated and real AML datasets. DeepDecon demonstrated exceptional performance in deconvoluting bulk RNA-seq data. It achieved the lowest RMSE values in 11 of 15 simulated datasets. Even on the three datasets where DeepDecon did not achieve the lowest RMSE values, its performance was still highly competitive, with only a marginal difference between its RMSE values and the lowest ones. Among the nine methods, we can see that the deep learning methods (i.e., DeepDecon and Scaden) performed better than traditional methods (Bisque, MEAD, RNA-Sieve, MuSiC, CIBERSORTx, ESTIMATE, and NNLS). They not only have lower RMSE values but also have higher correlations and CCC values compared to other methods (Figures 2B–2D).

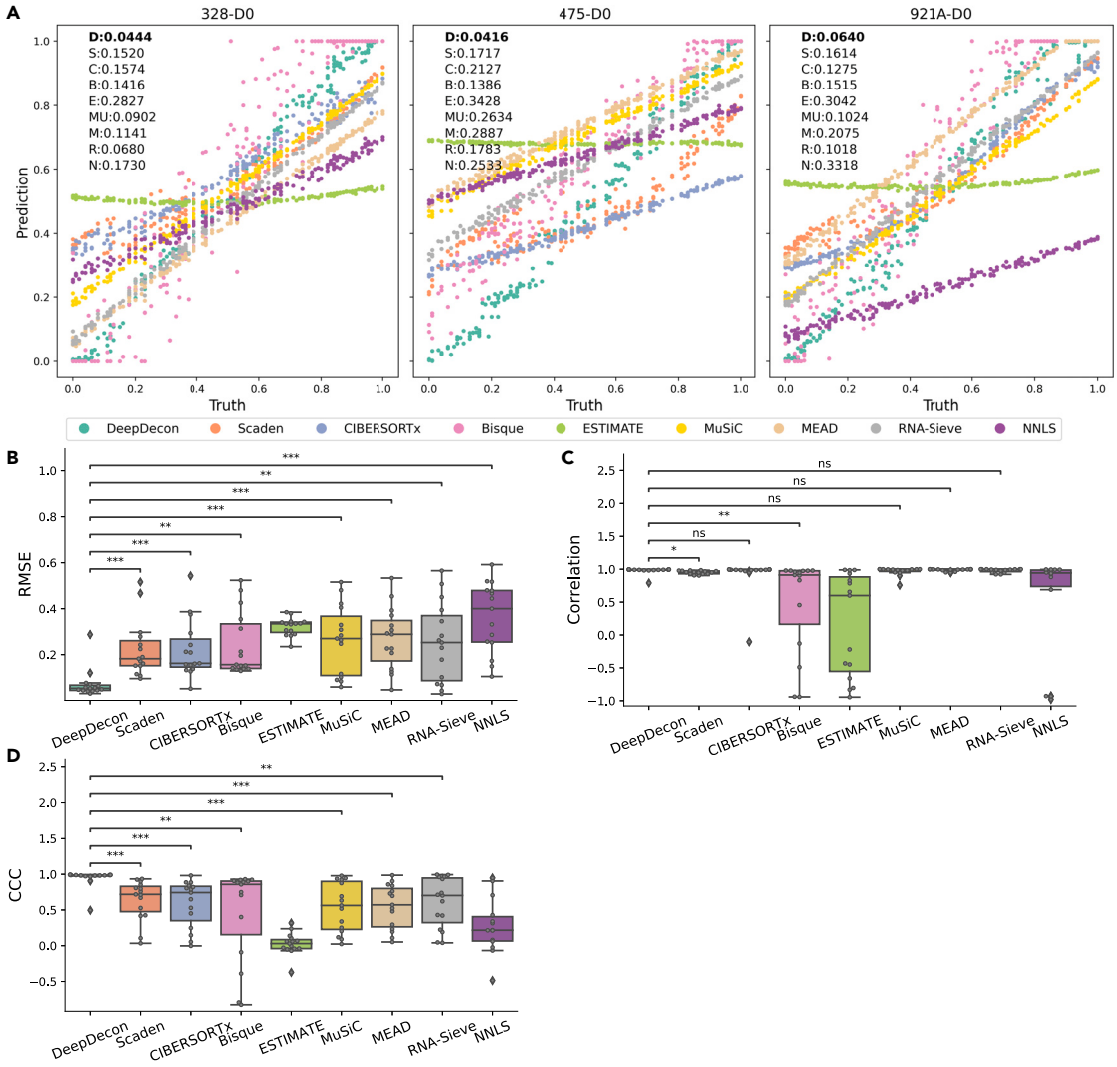


Figure 2. DeepDecon outperforms other methods in predicting malignant cell-type fractions on AML simulated bulk RNA-seq datasets

(A) Scatterplots of true versus predicted malignant cell fractions based on DeepDecon (D), Scaden (S), CIBERSORTx (C), Bisque (B), ESTIMATE (E), MuSiC (MU), MEAD (M), RNA-Sieve (R), and NNLS (N) on three selected AML simulated datasets. The x axis is the true fraction and the y axis is the predicted fraction. The numbers on each subplot are the root-mean-square error (RMSE) values between the true and the predicted fraction of each method.

(B) Boxplots of RMSE values between the predicted and the true fractions of malignant cells on 15 AML simulated bulk RNA-seq datasets.

(C) Boxplots of Pearson's correlation coefficient (*r*) values between the predicted and the true fractions of malignant cells on 15 AML simulated bulk RNA-seq datasets.

(D) Lin's concordance correlation coefficient (CCC) values between the predicted and the true fractions of malignant cells on 15 AML simulated bulk RNA-seq datasets. The correlation and CCC values of NNLS contain not-available (NA) values. Therefore, paired tests of correlation and CCC values between DeepDecon and NNLS are not available. *$0.01 < p$value ≤ 0.05, **$0.001 < p$ value ≤ 0.01, ***p value ≤ 0.001.

Figure 3A demonstrates the effectiveness of term frequency-inverse document frequency (TF-IDF) transformation on DeepDecon. Among all 15 simulated AML datasets, DeepDecon with TF-IDF transformation exhibited lower RMSE values in 14 datasets compared to DeepDecon without TF-IDF transformation. We used the paired Wilcoxon signed-rank test to compare the RMSE values of DeepDecon with versus without TF-IDF normalization by combining all 15 simulated datasets, and the resulting p value was 0.00099. This suggests that the use of TF-IDF can enhance the predictive power of DeepDecon. We also compared TF-IDF transformation with other existing normalization methods (fragments per kilobase per million mapped fragments [FPKM] and transcripts per kilobase million [TPM] normalization), and the corresponding results are given in Figure S2. The figure shows that DeepDecon with TF-IDF normalization outperforms DeepDecon with FPKM and TPM normalization methods. Figure 3B shows the effects of iterations on DeepDecon. Non-iterative DeepDecon is only one neural network trained on datasets with malignant cell fractions ranging from 0.0 to 1.0. The RMSE values of iterative DeepDecon are lower than those of non-iterative DeepDecon in all 15 simulated AML datasets (p value = 0.00065). Figure S3 also gives the scatterplot of iterative DeepDecon and

Table 1. DeepDecon outperforms other methods in simulated and real AML datasets

Method	Subject ID															Real AML data			
	210AD0	328 D0	328 D113	328 D171	328 D29	329 D0	329 D20	419AD0	420BD0	475 D0	556 D0	707B D0	916 D0	921AD0	1012D0	Mean	Median	Mean rank	Primary	Recurrent	Beat AML
DeepDecon	5	4*	7*	6*	5	6	8*	4*	4*	4*	3*	5*	12*	6*	29	7.2	5	1.4	13*	19*	17*
Scaden	19	15	22	12	11	10	24	28	15	17	18	30	47	16	52	22.4	18	4.4	17	21	20
Bisque	14	14	16	13	15	35	48	31	43	14	14	52	21	15	20*	24.33	16	5.07	27	25	28
MEAD	5	11	23	33	14	13	29	39	23	29	31	37	45	21	53	27.07	29	5.6	27	28	28
RNA-Sieve	4*	7	28	25	3*	7	26	45	23	18	35	39	57	10	51	25.2	25	4.73	27	31	32
CIBERSORTx	15	16	16	13	14	5*	29	24	21	21	16	38	39	13	54	22.27	16	4.6	29	28	23
MuSiC	6	9	25	28	8	12	31	52	27	26	41	33	42	10	48	26.53	27	5.33	27	29	23
ESTIMATE	34	28	33	33	31	29	24	29	34	34	34	38	38	30	34	32.2	33	6.67	22	23	29
NNLS	11	17	15	47	52	26	40	59	52	25	48	44	48	33	29	36.4	40	7.2	30	30	32

The root-mean-square errors (RMSEs) (%) for the estimated fraction of malignant cells in leave-one-subject-out cross-validation for DeepDecon, Scaden, Bisque, MEAD, RNA-Sieve, CIBERSORTx, MuSiC, ESTIMATE, and NNLS on AML datasets. Asterisks indicate the best RMSE value.

non-iterative DeepDecon on simulated AML datasets. It shows non-iterative DeepDecon's poor prediction accuracy when the malignant cell fraction is close to 0 or 1.

Figure S4 shows this heterogeneity by presenting the uniform manifold approximation projection (UMAP)[23] of all 15 subject datasets based on their scRNA-seq gene expression levels. Each subject has a specific clinical outcome, leading to gene expression variations and model performance differences. The projection indicated the heterogeneity across different subjects and further proved that it is necessary to simulate artificial bulk samples separately across different single-cell subjects.

We then investigated the decomposition performance of the nine methods using real bulk RNA-seq data. We utilized all 15 artificial bulk RNA-seq datasets to train DeepDecon and Scaden. To obtain the single-cell reference data for Bisque, MEAD, RNA-Sieve, MuSiC, and CIBERSORTx, we selected single cells from all 15 scRNA-seq datasets and combined them together. Figure 4 shows the decomposition performance of the nine methods on real AML RNA-seq datasets ("TARGET-AML" [primary and recurrent] and "BeatAML"). The RMSE values for each method on real datasets are also given in Table 1. Figures S5, S6, and S7 show the scatterplots of true malignant fractions with predicted malignant fractions for each method in real "primary," "recurrent," and "BeatAML" AML datasets, respectively. DeepDecon outperforms Scaden, Bisque, MEAD, RNA-Sieve, MuSiC, CIBERSORTx, ESTIMATE, and NNLS in deconvolving the malignant cell fraction on real AML datasets.

DeepDecon outperforms other deconvolution methods for other cancer types

To test DeepDecon's performance on other cancer types, we also applied DeepDecon to other cancer types.[24,25] Specifically, we constructed artificial bulk RNA-seq samples for each subject separately. Then, we trained each DeepDecon model using the generated artificial bulk RNA-seq datasets and evaluated the performance of DeepDecon and other methods using leave-one-out cross-validation. Figures 5 and 6 show the boxplots of the RMSE, correlation, and CCC values between the true and the estimated cancer cell fractions among all methods on the simulated neuroblastoma and head-and-neck squamous cell carcinoma (HNSCC) datasets. Tables S1 and S2 also give the RMSE values and average performance ranks of each method on simulated and real neuroblastoma and HNSCC datasets. They show that DeepDecon still achieves the lowest RMSE values, the highest correlations, and CCC values in neuroblastoma and HNSCC cancers, indicating that DeepDecon is robust and applicable to other cancer types.

We also compared DeepDecon with regression-based methods such as CIBERSORTx, RNA-Sieve, and NNLS that do not use subject-specific information in their original publications. We designed a way to incorporate subject information in these methods and showed that DeepDecon outperforms them in both ways. Details are given in supplemental experimental procedure 2 and Figure S8.

The impacts of gene expression perturbations and cell number per bulk sample on the performance of DeepDecon

Under "Methods," we discuss that bulk RNA-seq gene expression perturbations and the number of cells N in a bulk sample

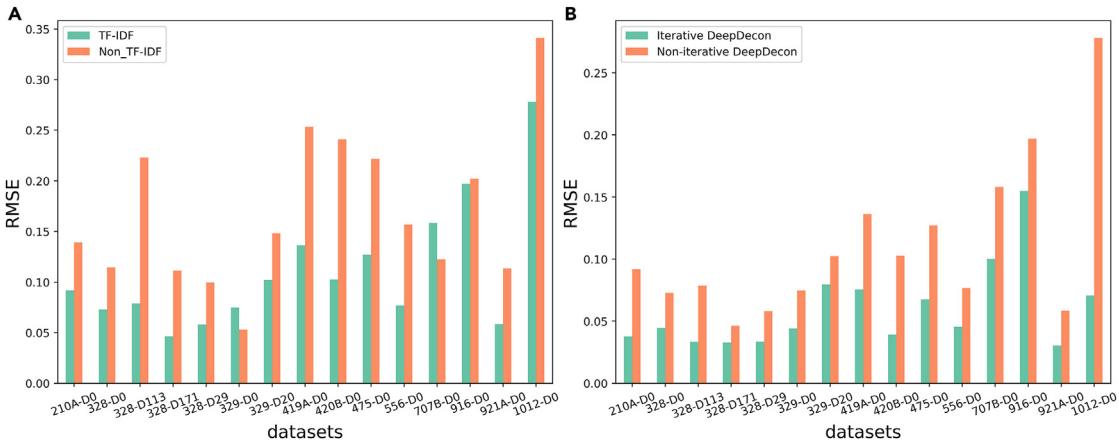


Figure 3. The TF-IDF transformation and the iterative strategy improve the performance of DeepDecon

(A) Bar plots of RMSE values of DeepDecon models with and without TF-IDF transformation. DeepDecon with TF-IDF transformation achieves the lowest RMSE values in 14 of 15 simulated AML datasets.

(B) Bar plots of RMSE values on DeepDecon models with and without the iterative strategy. Iterative DeepDecon achieves the lowest RMSE values in all 15 simulated AML datasets. The x axis is the simulated AML dataset. The y axis is the RMSE value.

can influence the accuracy of the decomposition algorithms. Figures 7, S9, and S10 show the influence of different levels of perturbations on the performance of various decomposition methods. The RMSE values for most methods except Bisque slightly increase with the noise level. DeepDecon consistently achieves the lowest RMSE among all methods under different noise levels, showing its robustness.

We also investigated the influence of the number of cells in a bulk sample on the prediction accuracy of DeepDecon using AML datasets. Figure 8 shows the RMSE values between true and predicted malignant cell fractions under different combinations of cell numbers in a bulk sample. More specifically, when the training model is fixed, the RMSE value decreases with the cell number in bulk samples in the testing data. This shows that a better prediction performance can be achieved when the testing bulk sample contains more single cells. If the number of cells per bulk sample exceeds a certain threshold ($> 3,000$), the performance of the DeepDecon model becomes stable. On the other hand, when the number of cells per bulk sample in testing datasets is above 3,000, the number of cells in the training dataset does not have a strong influence on DeepDecon. The RMSE values are stable, showing the robustness of DeepDecon to the number of cells in the training datasets when the number of cells in the testing data is above 3,000.

DeepDecon was trained on a high-performance cluster (HPC) with a Xeon-2640 6-core CPU node and it took $\sim$ 20 min to train a model and took $\sim$ 3 s to predict on one bulk tissue.

DISCUSSION

DeepDecon is an innovative deep-learning-based algorithm that leverages scRNA-seq information to accurately predict cancer cell fractions. Due to the latent feature engineering capabilities of neural networks, which can automatically extract non-linear features in the hidden layers, DeepDecon can achieve superior performance by incorporating all input genes ($\sim 10^4$). We showed that DeepDecon is applicable to multiple cancer datasets. DeepDecon can iteratively predict malignant cell fractions with lower RMSE compared to other methods, making it a powerful tool for accurate and reliable prediction of cancer cell fractions.

DeepDecon adopts a TF-IDF approach to weigh the expression of different genes, which addresses the issue of imbalanced expression levels across genes. In addition, our algorithm employs an iterative approach to refine the prediction, as opposed to the three deep neural network outputs average used by Scaden. These two steps have significantly improved the estimation accuracy of malignant cell fractions in bulk RNA-seq samples. By iteratively using small-range models $M_{i,j}$ to predict the same bulk samples, where these models share similar structures but work in different malignant fraction ranges, we have achieved better prediction accuracy compared to using only one initial model $M_{0,100}$. We have also evaluated DeepDecon's performance with respect to gene expression perturbations and varying numbers of cells per bulk sample. We showed that DeepDecon is robust to gene expression perturbations and the number of cells in the training set, if the number of cells in the testing data is at least 3,000. These findings make DeepDecon a valuable tool for the accurate and reliable prediction of malignant cells.

DeepDecon accurately estimates the malignant cell fraction in a tissue based on its transcriptomic features from bulk RNA-seq data. In particular, for AML, this novel approach could be used to accurately detect malignant clones in patients who appear to be in complete remission by standard morphology and flow cytometric analysis. DeepDecon can also be used to measure residual disease in AML patients with morphological remission and classify patients into different phases, such as accelerated or blast phase crisis, depending on malignant cell fractions.

While DeepDecon can achieve good performance on different cancer samples and tissues, we note that there are still limitations to this deep-learning-based method. First, the quality of training data is very important. If the number of subjects is small or the single-cell data are dominated by one specific cell type,

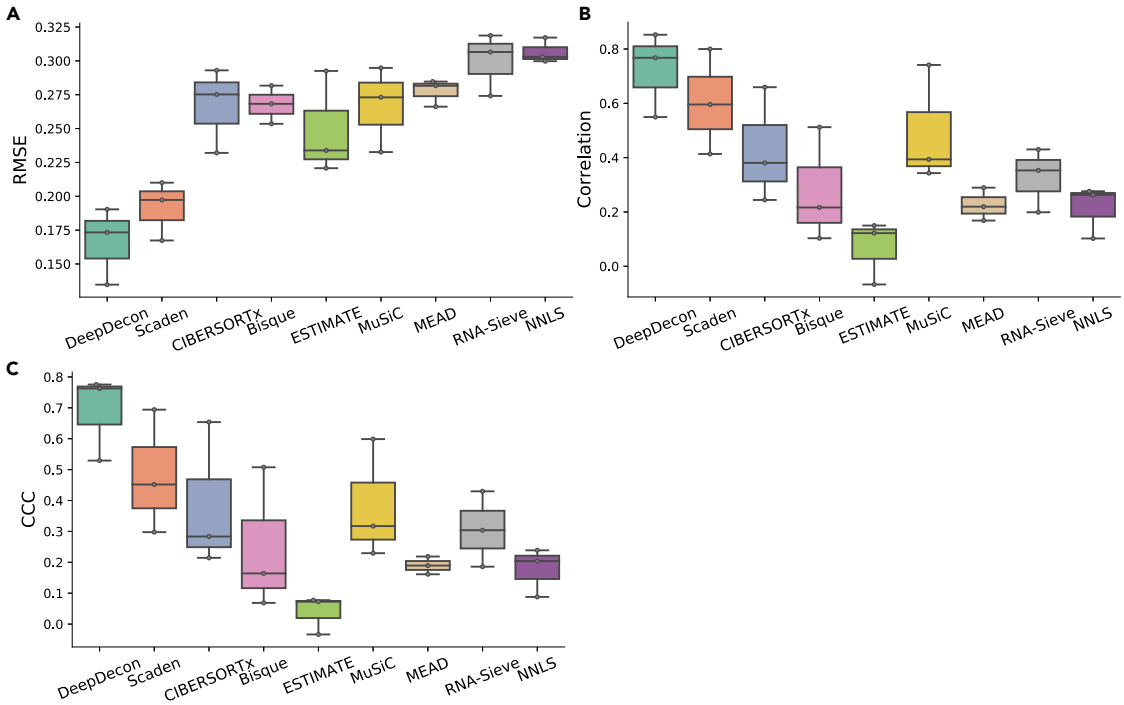


Figure 4. DeepDecon outperforms other deconvolution methods on real AML RNA-seq datasets

Boxplots of root-mean-square error (RMSE) (A), Pearson's correlation coefficient (PCC) (B), and Lin's concordance correlation coefficient (CCC) (C) values between the predicted and the true fractions of malignant cells. Each bar in the boxplots contains three points corresponding to three real AML bulk RNA-seq datasets, namely "primary," "recurrent," and "BeatAML" datasets.

DeepDecon can learn less information about real cell fraction distribution and cannot generalize and represent the latent features well. Second, experimental bias and noise can greatly limit decomposition accuracy. These limitations can potentially be alleviated by including more training subjects to increase the training set size and by reducing noise in the expression data. However, more computational resources will be needed to train the DeepDecon models. How to efficiently train the model with large training data is a topic for further research. Third, DeepDecon constructed simulated bulk RNA-seq datasets by assuming random sampling of single cells from the tissue. However, it should be noted that simulated bulk RNA-seq is not necessarily the same as real bulk RNA-seq samples. A potential limitation of DeepDecon is that the exact cell-type information may not be available. Preparation methods for generating single-cell suspensions may result in the underrepresentation of certain cell types, particularly those that are rare or do not survive disassociation. Therefore, the resulting cell composition may differ from that in the real tissues. However, particular cell types that are consistently missing from all single-cell suspensions are less likely when using multiple training datasets. Since we analyze both malignant and normal cell types in this study, this is less an issue than a general cell-type deconvolution study where a large number of cell types are considered in solid tissues. We alleviate this issue further by only selecting subjects with more than 100 malignant cells and 100 normal cells. In future studies involving multiple cell types, we could adopt similar requirements and add the cell type "unknown" to cover potential missing cell types.

We plan to further improve the performance and applicability of DeepDecon by implementing several key modifications to the existing methodology. First, we want to extend DeepDecon's capacity to include multiple cell types or subtypes. For instance, we considered two main cell types in this study: malignant and normal cells. However, it has been reported that both cell types consist of molecular subtypes.[26] Thus, it is important to extend DeepDecon to multiple cell types. Second, it is essential to consider both known and unknown cell types in deconvolution. Cell composition derived from biological experiments can contain cells that do not belong to any of the existing cell types. These cells are labeled as unknown cell types and have more complex gene expression patterns. Third, the current DeepDecon model takes all genes into account. Selecting genes that are only relevant to the cell types of interest may further increase prediction accuracy.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for data should be directed to and will be fulfilled by the lead contact, Prof. Fengzhu Sun (fsun@usc.edu).

Materials availability

The study did not generate new unique reagents.

Data and code availability

Single-cell AML data were obtained from the Gene Expression Omnibus (GEO) under accession no. GSE116256.[27] Single-cell neuroblastoma data were downloaded from the GEO with accession no. GSE137804.[24] HNSCC cancer data were collected from the database TISCH.[25] All real bulk RNA-seq data sources are listed in Table S3. All original code has been deposited at Github

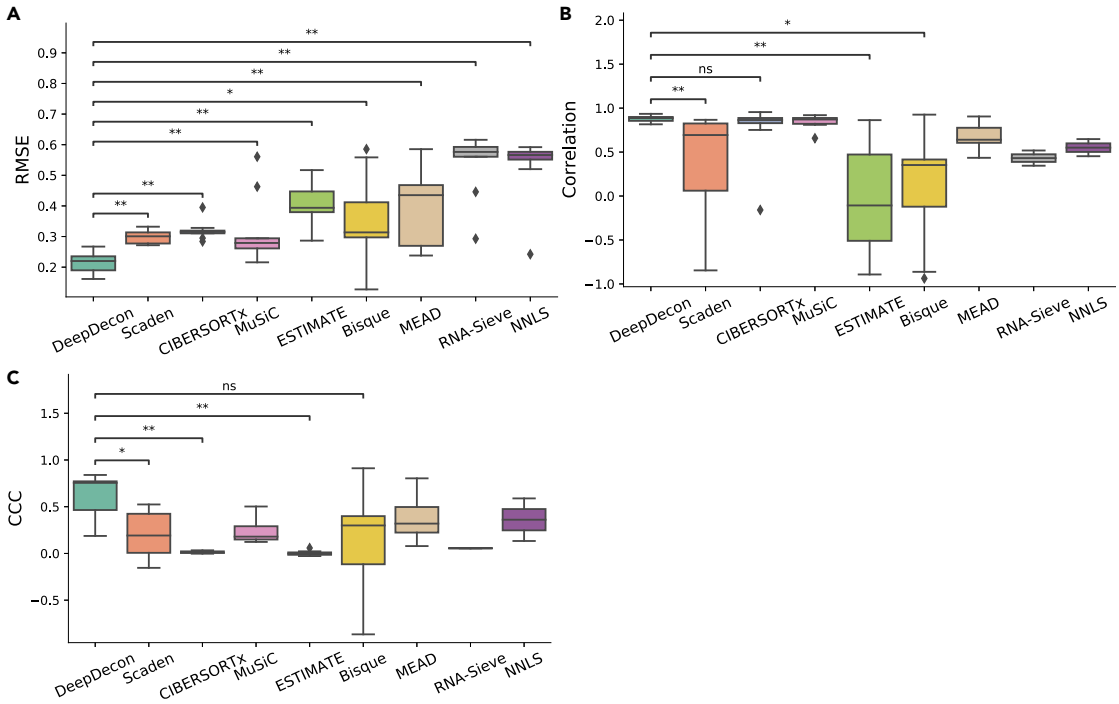


Figure 5. DeepDecon outperforms other deconvolution methods on the simulated neuroblastoma datasets

Boxplots of root-mean-square error (RMSE) (A), Pearson's correlation coefficient (PCC) (B), and Lin's concordance correlation coefficient (CCC) (C) values between the predicted and the true fractions of malignant cells on nine simulated neuroblastoma bulk RNA-seq datasets. Each bar in the boxplot contains nine points corresponding to nine simulated neuroblastoma bulk RNA-seq datasets. Each simulated neuroblastoma dataset contains bulk samples constructed from only one subject. The correlation and CCC values of methods MEAD, RNA-Sieve, and NNLS contain not-available (NA) values. Therefore, paired tests of correlation and CCC values between DeepDecon and MEAD, RNA-Sieve, and NNLS are not available. *$0.01 < p$ value ≤ 0.05, **$0.001 < p$ value ≤ 0.01.

(https://github.com/Jiawei-Huang/DeepDecon) and has been archived at Zenodo.[28]

Methods

Datasets

AML is a heterogeneous disease in which hemopoietic progenitor cells (blasts) lose the ability of normal differentiation and proliferation.[29] The diagnosis of AML has a direct relation with the malignant cell percentage in bone marrow (BM) tissues.[30,31] Therefore, we chose AML as our primary disease in this study. The single-cell AML datasets were downloaded from the GEO with accession no. GSE116256.[27] This dataset contains scRNA-seq gene expression sequenced from subjects who have different degrees of AML disease. Each cell in the dataset has labeled cell type (malignant or normal). A total of 15 subjects were selected to simulate artificial bulk RNA-seq datasets. The scRNA-seq data were processed following the preprocessing workflow of the widely used single-cell gene expression python package, Scanpy (v.1.7.2).[32] Initially, cell-gene matrices were filtered to exclude cells with fewer than 500 detected genes and genes expressed in fewer than five cells. Subsequently, the count matrix for each subject was filtered to remove extreme outliers in gene expression values (Table S4). Then, gene expression was normalized by Scanpy's "normalize_total" function so that every cell had the same total count after normalization. This will counteract the effect of different library sizes. Finally, the resulting normalized matrix of all filtered cells and genes was saved for subsequent simulated bulk data generation. The details of data selection and preprocessing are given in supplemental experimental procedure 3 and Figure S11.

DeepDecon was tested on real AML bulk RNA-seq datasets. We first downloaded AML data from the GDC data portal (https://portal.gdc.cancer.gov/) with the project name "TARGET-AML." The AML samples were further divided into primary AML and recurrent AML categories according to different cancer stages. As a result, there were a total of 117 primary AML samples and 38 recurrent AML samples. Ground-truth cancer cell fractions from flow cytometry are available in these bulk RNA-seq data. Moreover, an additional real AML dataset, BeatAML,[33] was collected from cBioportal.[34] BeatAML contains a total of 451 bulk RNA-seq samples, and 300 of them have corresponding ground-truth cancer cell fractions. The study used the "SureSelect" sequencing platform, which is different from the sequencing platform used to generate the single-cell data on the TARGET-AML dataset (Table S3). These datasets enable us to evaluate DeepDecon's performance on data from different sources.

To test DeepDecon's performance on other cancer tissues, we also collected 19,173 single cells from nine neuroblastoma cancer patients[24] and 184,868 single cells from 27 HNSCC patients.[25] They were used to simulate artificial RNA-seq bulk samples to build and evaluate DeepDecon. A real neuroblastoma bulk RNA-seq dataset consisting of 99 bulk RNA-seq samples with known cancer cell fractions was collected from cBioportal,[34] and another real HNSCC bulk RNA-seq dataset, "TCGA-HNSC," consisting of 518 bulk RNA-seq samples with known cancer cell fractions, was collected from LinkedOmics.[35] These two real datasets were used for testing. The details of data selection and preprocessing are given in supplemental experimental procedure 3.

Generating artificial bulk RNA-seq datasets

We used scRNA-seq datasets described under "Datasets" to construct artificial bulk RNA-seq samples. The generated samples were designed to have predetermined malignant cell fractions, which were then employed as training data for the DeepDecon model. Specifically, we first fixed the total number of cells in an artificial bulk sample to be N, and a malignant cell number n_m was randomly generated from a uniform distribution between 0 and N. Subsequently, n_m malignant cells and $N - n_m$ normal cells were randomly sampled from the same scRNA-seq dataset. If the total number of malignant or normal cells in the scRNA-seq dataset was smaller than n_m or $N - n_m$, respectively, the cells were chosen with replacement, that is, each cell was chosen uniformly from all the single cells available; otherwise, the cells were chosen without replacement, that is, each cell was chosen from the remaining cells.

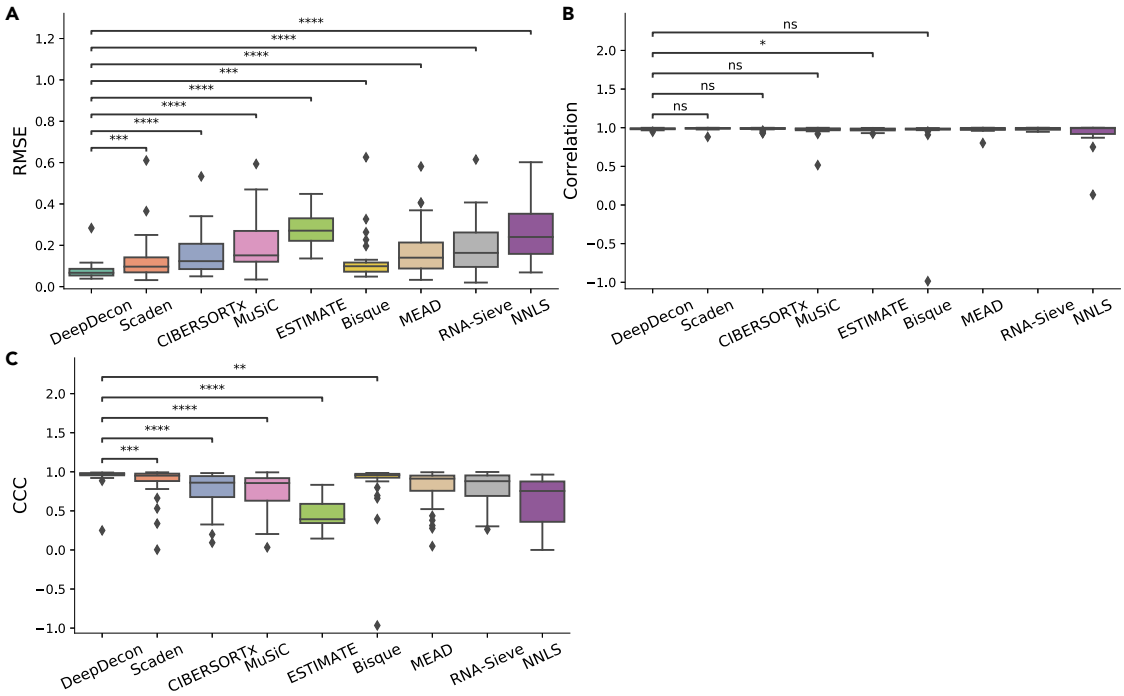


Figure 6. DeepDecon outperforms other deconvolution methods on the simulated HNSCC dataset

Boxplots of root-mean-square error (RMSE) (A), Pearson's correlation coefficient (PCC) (B), and Lin's concordance correlation coefficient (CCC) (C) values between the predicted and the true fractions of malignant cells on 27 simulated HNSCC bulk RNA-seq datasets. Each bar in the boxplot contains 27 points corresponding to 27 simulated HNSCC bulk RNA-seq datasets. Each simulated HNSCC dataset contains bulk samples constructed from only one subject. The correlation and CCC values of method MEAD, RNA-Sieve, and NNLS contain not-available (NA) values. Therefore, paired tests of correlation and CCC values between DeepDecon and MEAD, RNA-Sieve, and NNLS are not available. *$0.01 < p$ value ≤ 0.05, **$0.001 < p$ value ≤ 0.01, ***$1.00e-04 < p$ value $\leq 1.00e-03$, ****p value $\leq 1.00e-04$.

Importantly, cells from different subjects (i.e., individuals) were not merged into an aggregated sample. This decision was motivated by two primary motivations. First, the aim was to safeguard within-subject relationships among genes by preserving the unique gene expression patterns inherent to each subject. Second, the intention was to capture the variability between subjects, commonly referred to as cross-subject heterogeneity.[22] The single cells were merged into one bulk sample by summing their expression values, and the resulting artificial bulk sample was labeled with the fraction of malignant cells n_m/N. This process was repeated for each scRNA-seq dataset, generating a corresponding artificial bulk RNA-seq dataset. Each bulk dataset contained T samples with known malignant cell-type proportions (supplemental experimental procedure 4). We set $N = 3,000$ and $T = 200$ here for model training. We also investigated the impacts of N on DeepDecon. This procedure provides a valuable resource for training and evaluating the DeepDecon algorithm.

Data processing

To ensure consistency between the data used for training and prediction, the artificial bulk RNA-seq samples underwent a preprocessing procedure before model training. Specifically, only genes that were present in both training and testing datasets were retained, and genes with low expression variances (below 0.1) were removed. Next, a TF-IDF transformation was applied to the raw RNA-seq count matrix. This transformation, commonly used in information retrieval and text mining,[36,37] involves calculating the "term frequency" (TF) for

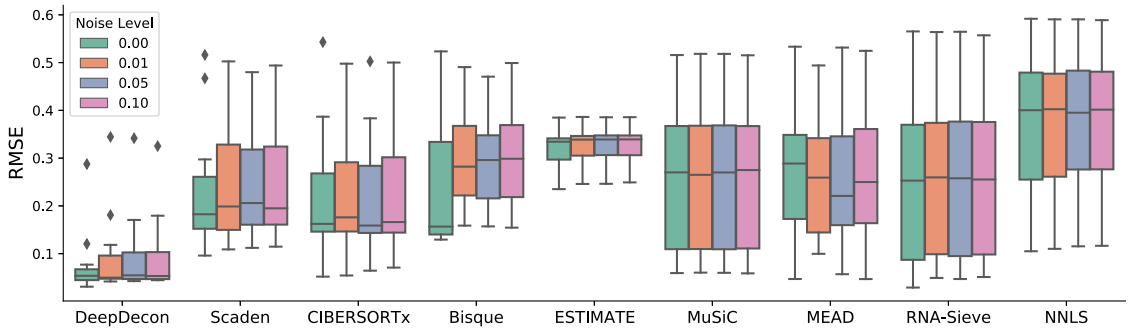


Figure 7. DeepDecon is robust to gene expression perturbations

Boxplots of RMSE values between the true and the estimated malignant cell fractions on simulated AML datasets under different noise levels. We added random noise generated from a Gaussian distribution with zero mean and variance that equals $\alpha(\alpha = 0.01, 0.05, 0.1)$ times gene expression level for each gene in each sample. We also randomly selected 10% of the genes for each sample and masked its gene expression values into 0. Each bar contains a total of 15 points, representing 15 separate AML datasets. The color represents different levels of noise level α.

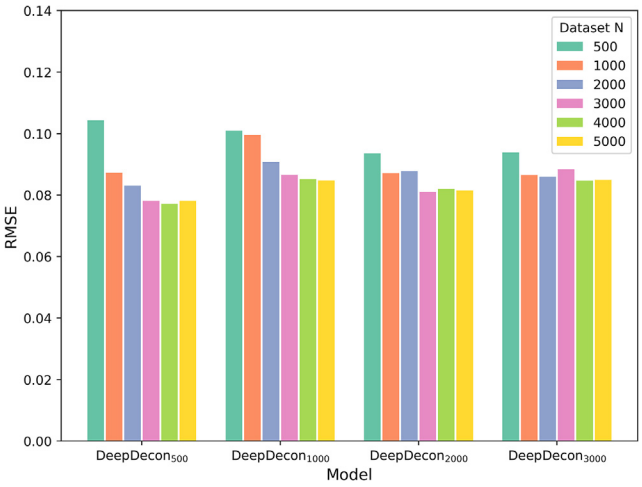


Figure 8. DeepDecon is robust to the number of cells per bulk sample when the number of cells in testing data is above 3,000

The x axis is the trained DeepDecon model. The subscript is the number of cells per bulk sample. DeepDecon_N means a DeepDecon model trained on a dataset in which one bulk sample consists of N single cells. The y axis is the RMSE value between the true and the estimated malignant cell fractions. The color represents the number of single cells per bulk sample in the testing data.

each gene in each sample by normalizing the GEP (see Equation 1). The "inverse document frequency" (IDF) was then calculated by dividing the total number of bulk samples by the total gene expression values of the gene across all samples (see Equation 2), followed by log transformation and multiplication by the TF value. The TF-IDF transformation weights genes with lower expression levels more heavily, which helps to adjust for the imbalanced expression levels across genes.[38] This preprocessing procedure is an important step in ensuring the quality and consistency of the data used for training the deep learning models:

$$\mathrm{TF}(X_{i,j}) = \frac{X_{i,j}}{\sum_j X_{i,j}}, \qquad \text{(Equation 1)}$$

$$\mathrm{IDF}(G_j) = \log\left(\frac{T}{\sum_i X_{i,j}} + 1\right), \qquad \text{(Equation 2)}$$

where $X_{i,j}$ is the expression level of the j-th gene in the i-th sample, G_j indicates the j-th gene, and T is the number of bulk samples.

Let X' denote the gene expression matrix after TF-IDF transformation. A MinMax transformation was applied to the resulting expression matrix X' to scale the expression values to the (0, 1) range (see Equation 3). This is a common practice in deep learning models that use gradient-based optimization algorithms[22,39]:

$$X_i^{\mathrm{norm}} = \frac{X_i' - \min(X_i')}{\max(X_i') - \min(X_i')}, \qquad \text{(Equation 3)}$$

where X_i' is the i-th row of X' and X_i^{norm} is the i-th row of the resulting expression matrix after the MinMax transformation.

There are also several existing normalization methods, including FPKM and TPM. These methods are mainly used for different sequencing methods.[40] For example, gene expression data from the unique molecular identifier (UMI) counting can represent the real expression value, while gene expression data from the Smart-Seq protocol need to be further normalized using methods like TPM or FPKM.[41] We compared these normalization methods with TF-IDF normalization, and the details can be accessed in supplemental experimental procedure 5.

The DeepDecon model

The deep learning model used in this study consisted of two main components. The first component consisted of four fully connected layers with a dropout regularization between each two layers, and the rectified linear unit (ReLU) was used as the activation function in every internal layer. The second component was a softmax function used to predict the malignant and normal cell fractions. All model parameters were optimized using the Adam optimization algorithm[42] with a learning rate of 0.0001 and a batch size of 128. The output of the DeepDecon model is the estimated fraction of malignant (tumor) cells of given bulk RNA-seq samples. The model was trained as a regression task, with the RMSE as the loss function. Various combinations of learning rates, batch sizes, and dropout rates in the deep learning model were tested, and the results are shown in supplemental experimental procedure 6 and Table S5. The Keras (v.1.0.8) library (https://keras.io/) was used to implement the deep learning model.

To address the issue of poor prediction accuracy when the malignant cell fraction is close to 0 or 1 (Figure S3), an iterative deep-learning model was developed. This model involves iteratively narrowing down the prediction interval of giving samples. More specifically, let $d_{i,j}$ denote the set of artificial bulk RNA-seq samples whose malignant cell fractions $p \in [0.01 \cdot i, 0.01 \cdot j]$, $i < j, i, j \in \{0, 10, 20, \cdots, 100\}$ and $M_{i,j}$ denote a DeepDecon model trained on $d_{i,j}$; that is, $M_{i,j}$ was trained on artificial bulk samples with a particular range of cell fraction. A total of 55 models were trained in this experiment. DeepDecon model $M_{i,j}$ was trained to minimize the error between the predicted cell fraction and the true cell fraction. After training, the difference between the predicted and the true malignant fractions was calculated for each artificial sample in $d_{i,j}$, and the set of differences was defined as $\mathrm{diff}(i,j)$.

To predict the malignant fraction for a given real bulk sample X, the full-range model $M_{0,100}$ (with $i = 0, j = 100$) is used to provide an initial estimate $\widehat{P}$. DeepDecon tries to limit the estimate to a smaller range, denoted as $[0.01i', 0.01j']$, based on the previous prediction value $\widehat{P}$ and training datasets difference $\mathrm{diff}(i,j)$ (see Formulas 4 and 5). Model $M_{i',j'}$ is then used to predict the malignant cell fraction of bulk sample X again, and the process continues to refine the estimation. During each iteration, DeepDecon either shortens the intervals or moves them to the left or right. Direction flags f_l and f_r are used to indicate the directions in which DeepDecon moves. The number of intervals is finite, and DeepDecon cannot shrink the intervals indefinitely. The intervals are also not allowed to oscillate between left and right. Therefore, the algorithm is finally forced to stop:

$$\begin{aligned} L(i,j) &= \widehat{P} + \mathrm{diff}(i,j)_{\lambda/2}, \\ i' &= \max(0, \lfloor 100 * L(i,j) \rfloor), \end{aligned} \qquad \text{(Equation 4)}$$

$$\begin{aligned} U(i,j) &= \widehat{P} + \mathrm{diff}(i,j)_{1-\lambda/2}, \\ j' &= \min(100, \lceil 100 * U(i,j) \rceil), \end{aligned} \qquad \text{(Equation 5)}$$

where λ is a hyperparameter we use to select the lower and upper percentile of $\mathrm{diff}(i,j)$ and help define the new model interval. The default value of λ is set at 10%. $\mathrm{diff}(i,j)_{\lambda/2}$ and $\mathrm{diff}(i,j)_{1-\lambda/2}$ indicate the $\lambda/2$ and $1 - \lambda/2$ percentiles of the set $\mathrm{diff}(i,j)$, respectively. $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ indicate the floor and ceiling of a number, respectively. The specific steps of iterative DeepDecon are given in Algorithm 1.

The impact of gene expression perturbations and the number of cells per bulk sample on DeepDecon

To test the model's robustness to gene expression perturbations, we added different levels of Gaussian noise to the expression levels of the simulated datasets. Specifically, we added random noise generated from a Gaussian distribution with zero mean and variance that equals $\alpha(\alpha = 0.01, 0.05, 0.1)$ times gene expression level for each gene in each sample (see Equation 6). Moreover, for each simulated bulk sample, we randomly selected 10% of the genes and masked their gene expression values into 0 to simulate data missing issues in practice:

$$X_{ij}^{\mathrm{noise}} = \max(0, X_{ij} + \mathrm{N}(0, X_{ij})), \qquad \text{(Equation 6)}$$

where X_{ij} is the gene expression value of gene j in simulated bulk sample i and α is the noise level.

For each subject, we generated the simulated bulk datasets with different noise levels separately. Leave-one-out cross-validation was used to evaluate model performance across subjects. Specifically, we selected one of the k artificial bulk RNA-seq datasets as the testing dataset, while the remaining $k - 1$

Algorithm 1. Iterative DeepDecon

Require: Trained DeepDecon models, $M = \{M_{i,j}, i<j;\ i,j \in \{0,10,20,\cdots,100\}\}$; Difference sets, which are differences between the prediction and the true malignant fractions from training datasets, $DIFF = \{diff(i,j), i<j;\ i,j \in \{0,10,20,\cdots,100\}\}$; Testing bulk sample X

Ensure: Malignant cell fraction estimate, $\widehat{P}$

1: Record the direction of the interval that DeepDecon moves compared to the last iteration, denoted by f_l and f_r

2: Initialization: model start interval index $i = 0$, end interval index $j = 100$, left direction $f_l = 0$, right direction $f_r = 0$, iteration end flag $flag = 0$, and percentile hyperparameter $\lambda = 10\%$

3: $\widehat{P} = M_{0,100}(X)$

4: $L(i,j) = \widehat{P} + diff(i,j)_{\lambda/2}$; $i' = \max(0, \lfloor 100 * L(i,j) \rfloor)$

5: $U(i,j) = \widehat{P} + diff(i,j)_{1-\lambda/2}$; $j' = \min(100, \lceil 100 * U(i,j) \rceil)$

6: **while** $flag = 0$ **do**

7: **if** $i' \geq i$ and $j' \leq j$ then

8: $flag = 0$

9: **else if** $i' \leq i$ and $j' \leq j$ then

10: $flag = 0$; $f_l = 1$

11: **else if** $i' \geq i$ and $j' \geq j$ then

12: $flag = 0$; $f_r = 1$

13: **end if**

14:

15: **if** $i' \geq j'$ or $min(f_l, f_r) > 0$ or $(i' \leq i$ and $j' \geq j)$ **then**

16: $flag = 1$

17: **end if**

18:

19: **if** $flag = 0$ then

20: $i = i'$; $j = j'$

21: $\widehat{P} = M_{i,j}(X)$

22: $L(i,j) = \widehat{P} + diff(i,j)_{\lambda/2}$; $i' = \max(0, \lfloor 100 * L(i,j) \rfloor)$

23: $U(i,j) = \widehat{P} + diff(i,j)_{1-\lambda/2}$; $j' = \min(100, \lceil 100 * U(i,j) \rceil)$

24: **end if**

25: **end** while

26: **return** $\widehat{P}$

datasets served as the training set. This process was repeated k times to fully evaluate the performance of our model.

The total number of cells N in bulk RNA-seq samples can vary from sample to sample, and it can be challenging to accurately estimate the number of cells in a given sample. In addition, bulk RNA-seq samples in practice could also be influenced by factors such as cell isolation, cell size, and clustering, which can further complicate the estimation of cell numbers. In order to evaluate the performance of DeepDecon under different numbers of single cells, we first fixed our DeepDecon model and generated a set of testing datasets $Q = \{q_{i,n}, |i = 0, 10,\cdots,80,90,100; n = 500,1,000,2,000,3,000,4,000,5,000\}$. Each bulk sample in the dataset $q_{i,n}$ contains n single cells, and the number of malignant cells follows a binomial distribution $\text{Binomial}\left(n, \frac{i}{100}\right)$. This simulates the variation of a random sampling of malignant cells. In practice, the number of single cells in tissue samples can vary widely among different patients and even among different sampling periods for the same patient.[43] Finally, we used DeepDecon to estimate the fraction of malignant cells for each sample in the testing datasets. By evaluating the performance of DeepDecon under different numbers of single cells, we can assess the robustness and accuracy of the model in real-world scenarios.

In addition to testing the scenario in which the DeepDecon model is fixed and the testing datasets are varied, we also conducted additional experiments examining the impact of varying the number of cells per sample during the training process. Specifically, we fixed the testing datasets and trained different DeepDecon models using datasets where each bulk sample consisted of a different number of single cells, ranging from 500 to 3,000. These models are denoted as $\text{DeepDecon}_N, N = 500, 1,000, 2,000, 3,000$, where N represents the number of cells per bulk sample (i.e., $N = 500, 1,000, 2,000, 3,000$). These DeepDecon models were then used to predict the fraction of malignant cells on the same testing bulk RNA-seq dataset $q_{i,n}$, which was generated as described earlier. This analysis provides insights into the performance of DeepDecon under different training scenarios and can help with the optimal selection of training cell numbers for a given experimental setup.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100969.

ACKNOWLEDGMENTS

This research was supported in part by NIH/NCI R01CA197903 (to J.F.Z.) and NIH/NCI R01CA251848 (to J.F.Z.).

AUTHOR CONTRIBUTIONS

J.H., J.F.Z., and F.S. conceived the study. J.H. designed the DeepDecon method conceptually, designed the neural network architecture, and developed code for implementing, training, and evaluating models. A.S. helped with data management and applications to the real data. Y.D. helped with the finalization of the manuscript and applications of the model. J.F.Z. and K.R.K. helped with the applications to real datasets, explanations of the implications of the computational results, and the finalization of the manuscript. F.S. supervised the project and helped design the DeepDecon method, data analysis, and finalization of the manuscript. All authors contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 26, 2023
Revised: January 15, 2024
Accepted: March 21, 2024
Published: April 15, 2024

REFERENCES

1. Corchete, L.A., Rojas, E.A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N.C., and Burguillo, F.J. (2020). Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis. Sci. Rep. *10*, 19737. https://doi.org/10.1038/s41598-020-76881-x.
2. Xiao, H., Zhang, J., Wang, K., Song, K., Zheng, H., Yang, J., Li, K., Yuan, R., Zhao, W., and Hui, Y. (2021). A Cancer-Specific Qualitative Method for Estimating the Proportion of Tumor-Infiltrating Immune Cells. Front. Immunol. *12*, 672031. https://doi.org/10.3389/fimmu.2021.672031.
3. Li, X., and Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. Int. J. Oral Sci. *13*, 36. https://doi.org/10.1038/s41368-021-00146-0.
4. Qin, Y., Zhang, W., Sun, X., Nan, S., Wei, N., Wu, H.-J., and Zheng, X. (2020). Deconvolution of heterogeneous tumor samples using partial reference signals. PLoS Comput. Biol. *16*, e1008452. https://doi.org/10.1371/journal.pcbi.1008452.
5. Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using rna-seq. Nat. Methods *8*, 469–477. https://doi.org/10.1038/nmeth.1613.
6. Finotello, F., and Di Camillo, B. (2015). Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. Brief. Funct. Genomics *14*, 130–142. https://doi.org/10.1093/bfgp/elu035.
7. Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. *10*, 380. https://doi.org/10.1038/s41467-018-08023-x.
8. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.
9. Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P.S., Povinelli, B.J., Booth, C.A.G., Sopp, P., Norfo, R., Rodriguez-Meira, A., Ashley, N., et al. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat. Med. *23*, 692–702. https://doi.org/10.1038/nm.4336.
10. Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611–1624.e24. https://doi.org/10.1016/j.cell.2017.10.044.
11. Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. *9*, 75. https://doi.org/10.1186/s13073-017-0467-4.
12. Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics *34*, 1969–1979. https://doi.org/10.1093/bioinformatics/bty019.
13. Mohammadi, S., Zuckerman, N., Goldsmith, A., and Grama, A. (2017). A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proc. IEEE *105*, 340–366. https://doi.org/10.1109/JPROC.2016.2607121.
14. Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612. https://doi.org/10.1038/ncomms3612.
15. Sugino, K., Clark, E., Schulmann, A., Shima, Y., Wang, L., Hunt, D.L., Hooks, B.M., Tränkner, D., Chandrashekar, J., Picard, S., et al. (2019). Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain. Elife *8*, e38619. https://doi.org/10.7554/eLife.38619.
16. Chen, D., and Plemmons, R.J. (2009). Nonnegativity Constraints in Numerical Analysis (WORLD SCIENTIFIC). https://doi.org/10.1142/9789812836267_0008.
17. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.
18. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782. https://doi.org/10.1038/s41587-019-0114-2.
19. Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat. Commun. *11*, 1971. https://doi.org/10.1038/s41467-020-15816-6.
20. Erdmann-Pham, D.D., Fischer, J., Hong, J., and Song, Y.S. (2021). A likelihood-based deconvolution of bulk gene expression data using single-cell references. Genome Res. *31*, 1794–1806. https://doi.org/10.1101/gr.272344.120.
21. Xie, D., and Wang, J. (2022). Robust statistical inference for cell type deconvolution. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.06420.
22. Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., and Bonn, S. (2020). Deep learning–based cell composition analysis from tissue expression profiles. Sci. Adv. *6*, eaba2619. https://doi.org/10.1126/sciadv.aba2619.
23. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.
24. Dong, R., Yang, R., Zhan, Y., Lai, H.-D., Ye, C.-J., Yao, X.-Y., Luo, W.-Q., Cheng, X.-M., Miao, J.-J., Wang, J.-F., et al. (2020). Single-cell characterization of malignant phenotypes and developmental trajectories of adrenal neuroblastoma. Cancer Cell *38*, 716–733.e6. https://doi.org/10.1016/j.ccell.2020.08.014.
25. Sun, D., Wang, J., Han, Y., Dong, X., Ge, J., Zheng, R., Shi, X., Wang, B., Li, Z., Ren, P., et al. (2021). Tisch: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Res. *49*, D1420–D1430. https://doi.org/10.1093/nar/gkaa1020.
26. Wang, L., Sebra, R.P., Sfakianos, J.P., Allette, K., Wang, W., Yoo, S., Bhardwaj, N., Schadt, E.E., Yao, X., Galsky, M.D., and Zhu, J. (2020). A reference profile-free deconvolution method to infer cancer cell-intrinsic subtypes and tumor-type-specific stromal profiles. Genome Med. *12*, 24. https://doi.org/10.1186/s13073-020-0720-0.
27. van Galen, P., Hovestadt, V., Wadsworth Ii, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. Cell *176*, 1265–1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.
28. Huang, J., Du, Y., Stucky, A., Kelly, K.R., Zhong, J.F., and Sun, F. (2024). Codes for the Paper "DeepDecon Accurately Estimates Cancer Cell Fractions in Bulk RNA-Seq Data (Zenodo). https://doi.org/10.5281/zenodo.10798618.
29. Estey, E., and Döhner, H. (2006). Acute myeloid leukaemia. Lancet *368*, 1894–1907. https://doi.org/10.1016/S0140-6736(06)69780-8.

30. Bennett, J.M., Catovsky, D., Daniel, M.T., Flandrin, G., Galton, D.A., Gralnick, H.R., and Sultan, C. (1985). Proposed Revised Criteria for the Classification of Acute Myeloid Leukemia. Ann. Intern. Med. *103*, 620–625. https://doi.org/10.7326/0003-4819-103-4-620.

31. Vardiman, J.W., Harris, N.L., and Brunning, R.D. (2002). The World Health Organization (WHO) classification of the myeloid neoplasms. Blood *100*, 2292–2302. https://doi.org/10.1182/blood-2002-04-1199.

32. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

33. Tyner, J.W., Tognon, C.E., Bottomly, D., Wilmot, B., Kurtz, S.E., Savage, S.L., Long, N., Schultz, A.R., Traer, E., Abel, M., et al. (2018). Functional genomic landscape of acute myeloid leukaemia. Nature *562*, 526–531. https://doi.org/10.1038/s41586-018-0623-z.

34. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.

35. Vasaikar, S.V., Straub, P., Wang, J., and Zhang, B. (2018). Linkedomics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. *46*, D956–D963. https://doi.org/10.1093/nar/gkx1090.

36. Teller, V. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Comput. Ling. *26*, 638–641. https://doi.org/10.1162/089120100750105975.

37. Chowdhury, G.G. (2010). Introduction to Modern Information Retrieval (Facet publishing).

38. Moussa, M., and Măndoiu, I.I. (2018). Single cell RNA-seq data clustering using TF-IDF based methods. BMC Genom. *19*, 569. https://doi.org/10.1186/s12864-018-4922-4.

39. Chen, Y., Wang, Y., Chen, Y., Cheng, Y., Wei, Y., Li, Y., Wang, J., Wei, Y., Chan, T.-F., and Li, Y. (2022). Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nat. Commun. *13*, 6735. https://doi.org/10.1038/s41467-022-34550-9.

40. Zhao, Y., Li, M.-C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H., and McShane, L.M. (2021). Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. J. Transl. Med. *19*, 269. https://doi.org/10.1186/s12967-021-02936-w.

41. Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell rna sequencing methods. Mol. Cell *65*, 631–643.e4. https://doi.org/10.1016/j.molcel.2017.01.023.

42. Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

43. Herzenberg, L.A., Tung, J., Moore, W.A., Herzenberg, L.A., and Parks, D.R. (2006). Interpreting flow cytometry data: a guide for the perplexed. Nat. Immunol. *7*, 681–685. https://doi.org/10.1038/ni0706-681.