Patterns, Volume *3*

Supplemental information

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

Yating Lin, Haojun Li, Xu Xiao, Lei Zhang, Kejia Wang, Jingbo Zhao, Minshu Wang, Frank Zheng, Minwei Zhang, Wenxian Yang, Jiahuai Han, and Rongshan Yu

Supplementary figures

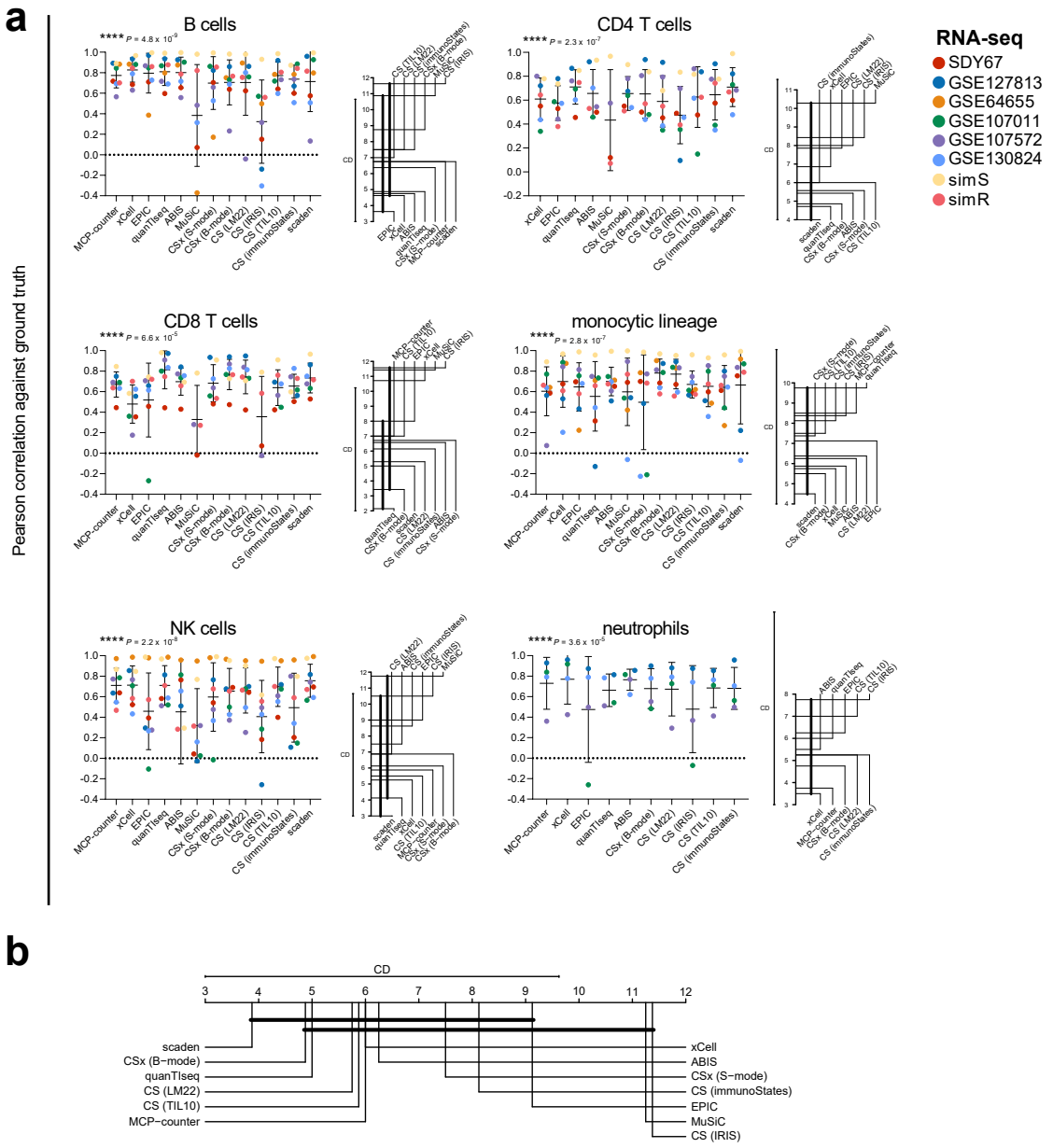

Figure S1: Evaluation of consistency in deconvolution performance across RNA-seq datasets. (a) Pearson correlations of the predicted cell type proportions of 13 deconvolution methods (including CIBERSORT using four different signature matrices) on six real-world datasets and two simulated datasets for six cell types, shown in scatter plots (left) with corresponding critical difference (CD) diagrams (right). Data are presented as means±SD. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, Friedman's rank sum test. (b) Comparison of 13 deconvolution methods on these eight datasets with post hoc two-tailed Nemenyi test. Groups of deconvolution methods that are not significantly different ($\alpha = 0.05$) are connected.

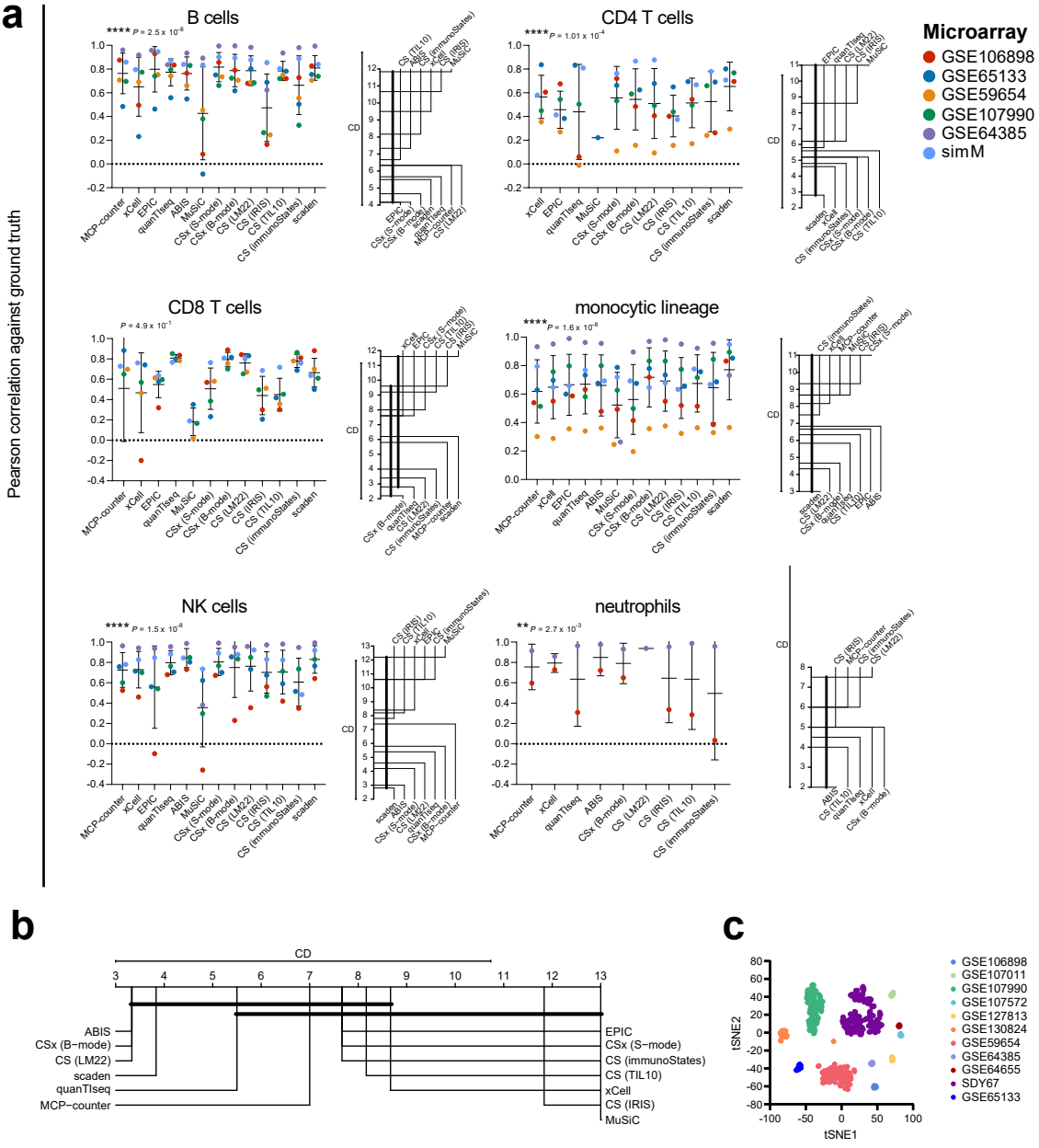

Figure S2: Evaluation of consistency in deconvolution performance across microarray datasets. (a) Pearson correlations of the predicted cell type proportions of 13 deconvolution methods (including CIBERSORT using four different signature matrices) on five real-world datasets and one simulated dataset for six cell types, shown in scatter plots (left) with corresponding critical difference (CD) diagrams (right). Data are presented as means±SD. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, Friedman's rank sum test. (b) Comparison of 13 deconvolution methods on these six datasets with post hoc two-tailed Nemenyi test. Groups of deconvolution methods that are not significantly different ($\alpha = 0.05$) are connected. (c) t-SNE projection of eleven real-world datasets. Each dot represents one single sample, coloured by datasets.

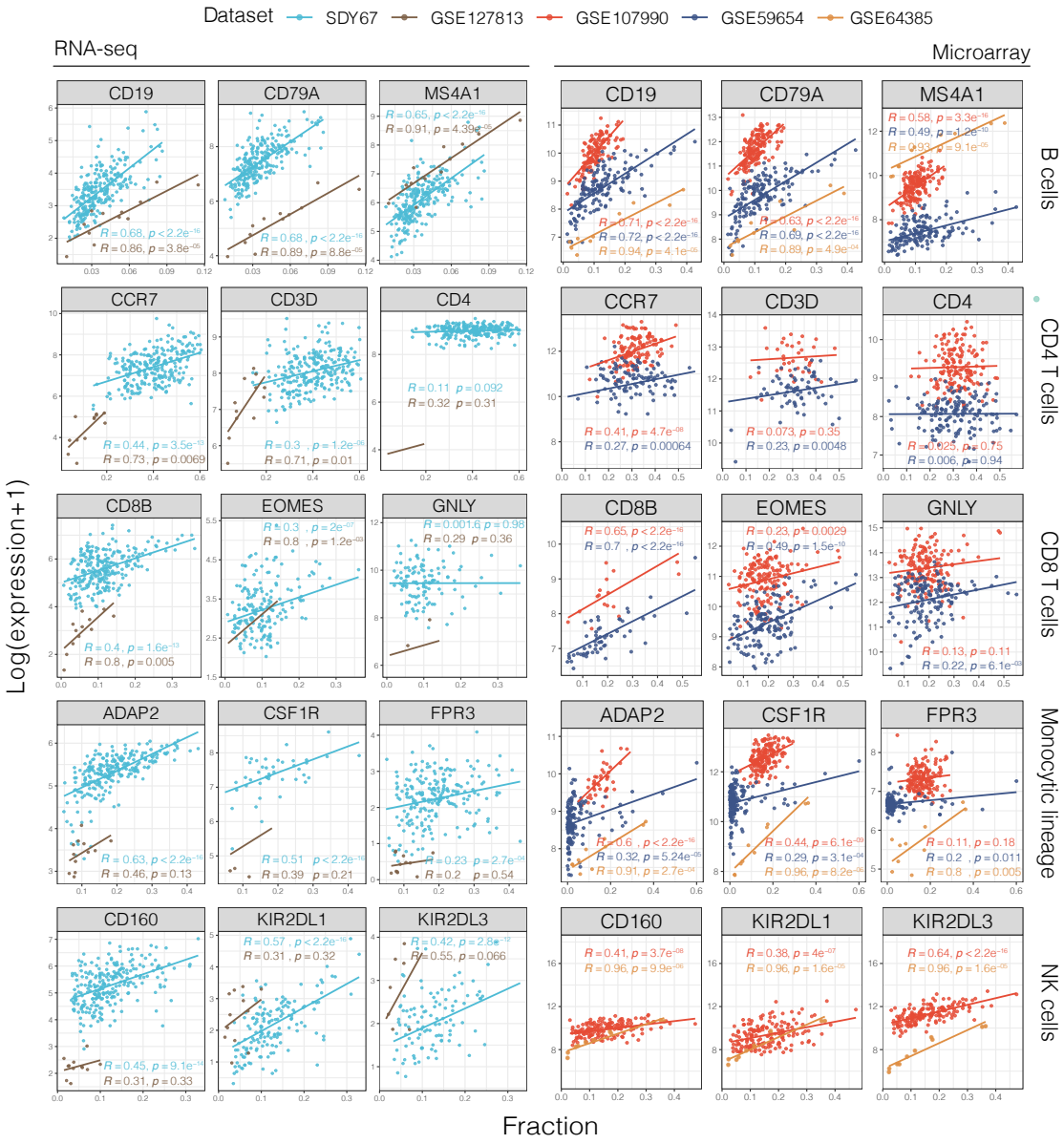

Figure S3: Relationships between the gene expression and cell type proportions on cell markers of different cell types among two RNA datasets and three microarray datasets. For each cell type, we selected three markers which were used for identifying corresdponding cell type. Numbers inside the sacatter plotting area signify Pearson correction values and p-values (Student's t-test) when performing linear regression.

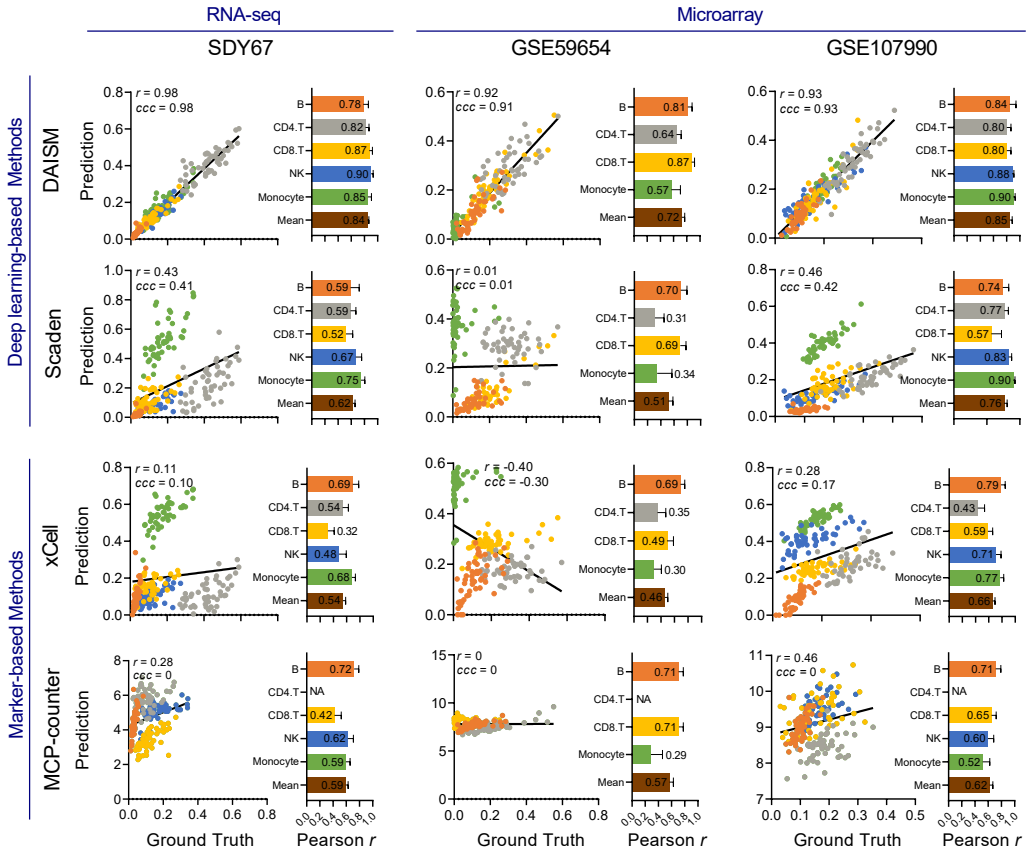

Figure S4: Performance evaluation of DAISM-DNN and other deep learning-based and marker-based deconvolution methods, tested on one RNA-seq dataset SDY67 (left) and two microarray datasets GSE59654 (middle) and GSE107990 (right). In each subplot, the scatter plot shows a global Pearson correlation (r) and concordance correlation coefficient (CCC). The bar plot displays per-cell-type Pearson correlation. The value in bar plots indicates the mean value over 30 experiments. Since different methods provide fraction of different cell types, we only validated the ones detected from flow cytometry in each dataset. NAs in the bar plots indicate cell types that cannot be predicted by this method.

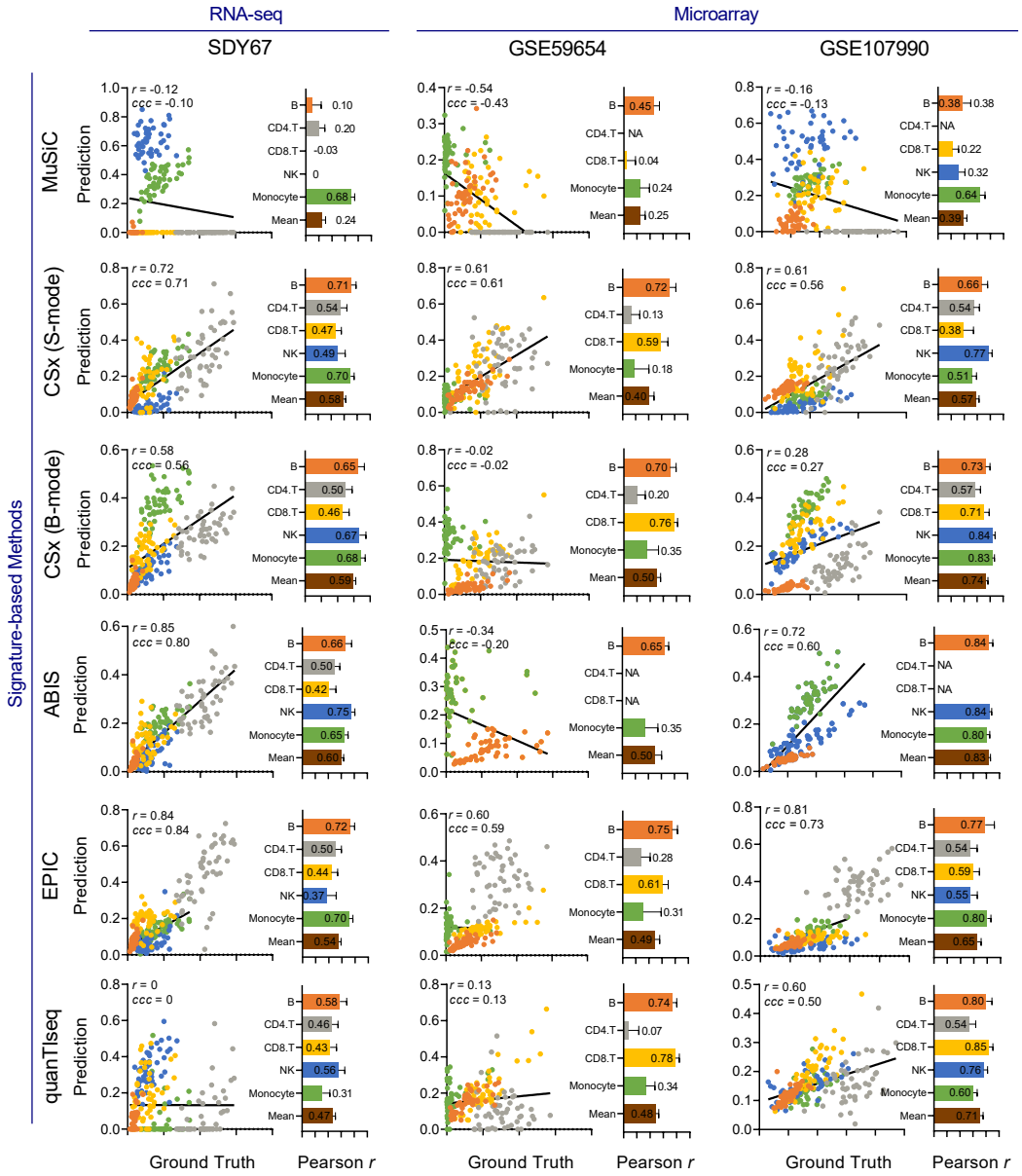

Figure S5: Performance evaluation of signature-based deconvolution methods, tested on one RNA-seq dataset SDY67 (left) and two microarray datasets GSE59654 (middle) and GSE107990 (right). In each subplot, the scatter plot shows a global Pearson correlation (r) and concordance correlation coefficient (CCC). The bar plot displays per-cell-type Pearson correlation. The value in bar plots indicates the mean value over 30 experiments. Since different methods provide fraction of different cell types, we only validated the ones detected from flow cytometry in each dataset. NAs in the bar plots indicate cell types that cannot be predicted by this method.

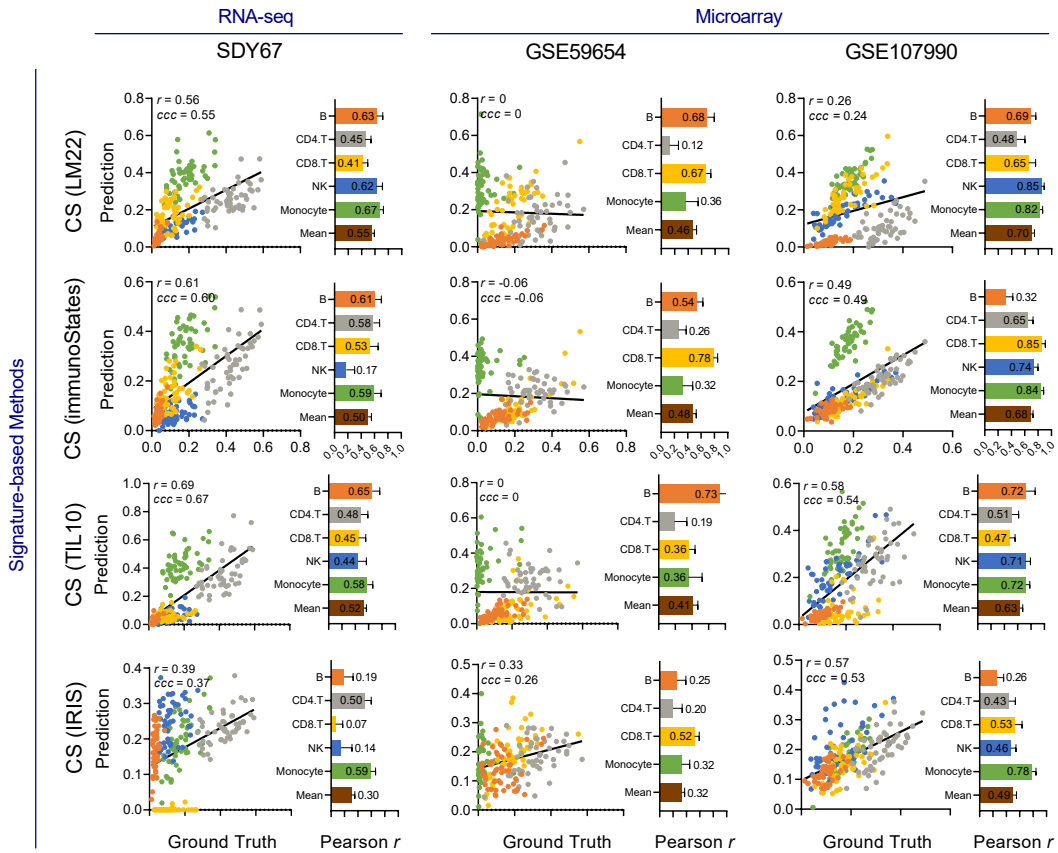

Figure S6: Performance evaluation CIBERSORT with different signature matrices, tested on one RNA-seq dataset SDY67 (left) and two microarray datasets GSE59654 (middle) and GSE107990 (right). In each subplot, the scatter plot shows a global Pearson correlation (r) and concordance correlation coefficient (CCC). The bar plot displays per-cell-type Pearson correlation. The value in bar plots indicates the mean value over 30 experiments. Since different methods provide fraction of different cell types, we only validated the ones detected from flow cytometry in each dataset. NAs in the bar plots indicate cell types that cannot be predicted by this method.

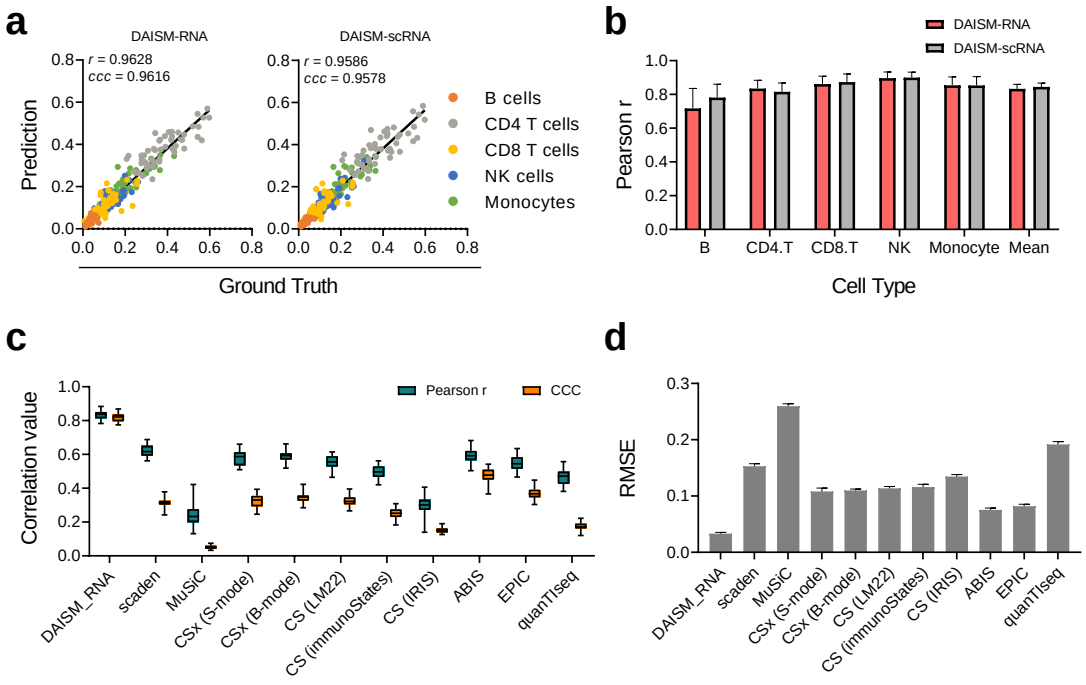

Figure S7: Performance evaluation of DAISM-RNA mode on RNA-seq dataset SDY67. (a-b) Deconvolution performance comparison between DAISM-RNA and DAISM-scRNA mode. Global Pearson correlation (r) and CCC are shown in scatter plots (a), and per-cell-type Pearson correlation are shown in the bar plot (b). (c-d) Comparison of DAISM-RNA mode with other methods by mean of per-cell type Pearson correlation and CCC (c) and RMSE (d). All data in bar plots are presented as the mean$\pm$SD.

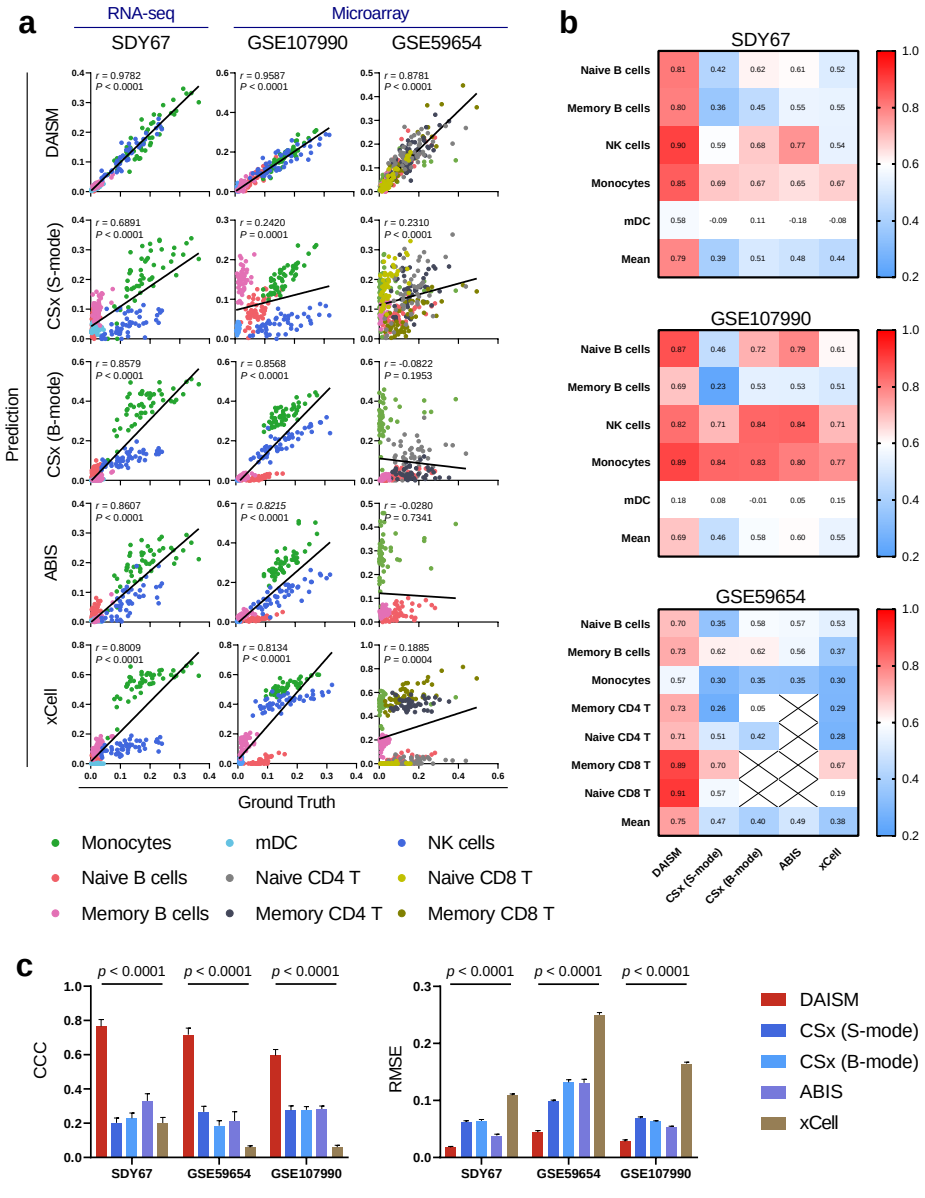

Figure S8: Performance of DAISM-DNN in fine-grained mode and benchmarking with four methods, validated using one RNA-seq dataset SDY67 and two microarray datasets GSE107990 and GSE59654. (a) Global Pearson correlation (r) between predicted and ground truth cell fraction. SDY67 and GSE107990 have five cell types: naive/memory B cells, NK cells, monocytes, myeloid dendritic cells (mDC), and GSE59654 has seven cell types: naive/memory B cells, monocytes, naive/memory CD4/CD8 T cells. (b) Heatmap summarizing performance of all methods after 30 permutation experiments by mean of Pearson correlation on each cell type and the mean of per-cell-type Pearson correlation (the last row in each heatmap). (c) Barplots of RMSE (right) and CCC (left) for five methods. Data are presented as means±SD. Two-sided paired Student's t tests were used to compare DAISM-DNN with other methods.

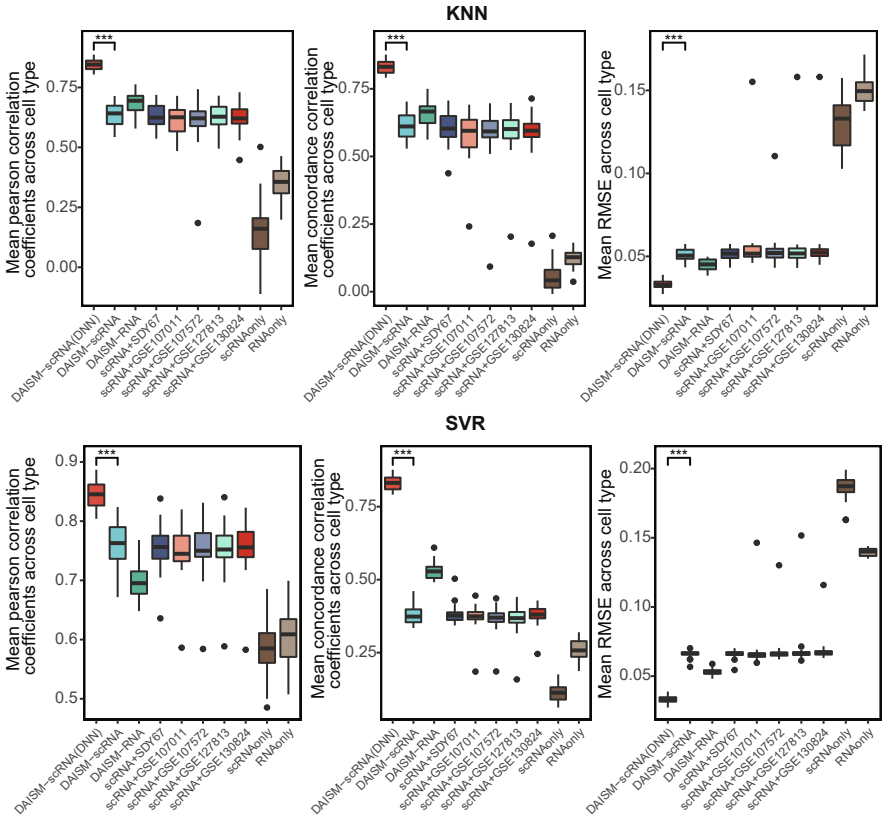

Figure S9: Performance of two machine learning models (KNN and SVR) on different training datasets generated by different mixing strategies. The first boxplot in each plotting demonstrates the performance of DNN trained by DAISM-generated training sets (paired two-sided Student's t-test).

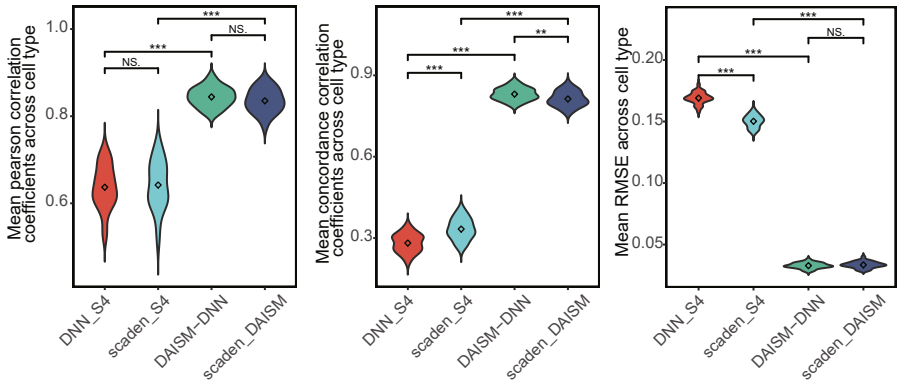

Figure S10: Evaluation of prediction accuracy of DAISM-DNN and Scaden using the same trianing sets. The performance of each model was measured from 30 permutation tests. In each permutation test, 50 samples of SDY67 dataset were randomly selected as test data, and the remaining samples were uesd as calibration data. Two training sets generated by different mixing strategies were used to train DAISM-DNN and Scaden. One is DAISM-generated training dataset. The other is the combination of four pre-generated PBMC *in silico* mixtures provided in Scaden paper (S4).

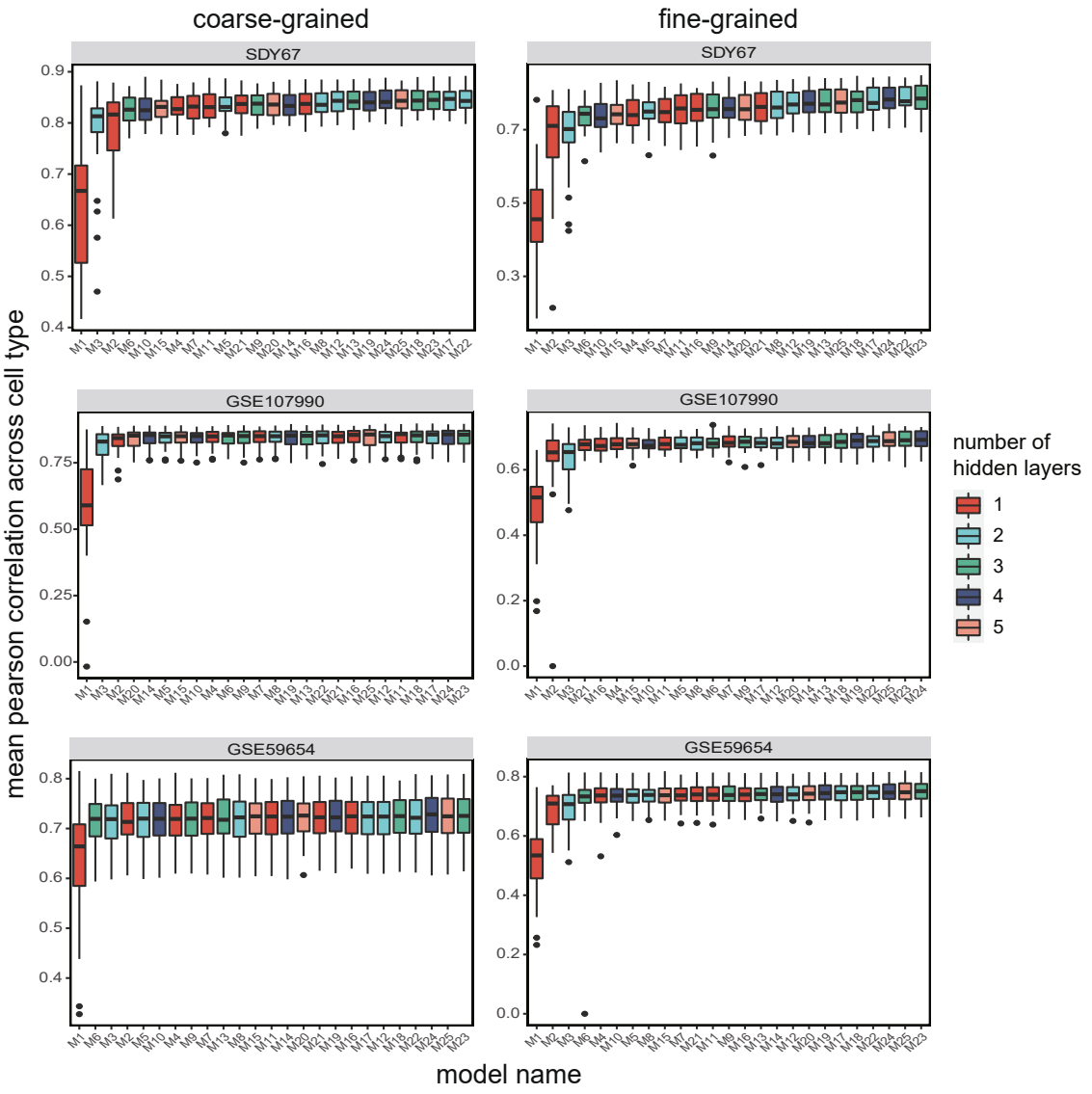

Figure S11: Performance of DAISM-DNN with deep neural networks of different hidden layers and number of neurons. The detailed architectures of each model were listed in Table S3. ReLU activation was used for all hidden layers. We considered 25 combinations of different numbers of neurons for the layers in the neural network. We tested DAISM-DNN with different hyperparameters on three PBMC datasets (SDY67, GSE107990 and GSE59654). For each hyperparameter setting, the model was built and trained on the same training set (16,000 samples). The performance of each model was measured from 30 permutation tests. In each permutation test, 50 randomly selected samples were held out as the test samples and the remaining samples from the same dataset were used as calibration samples.

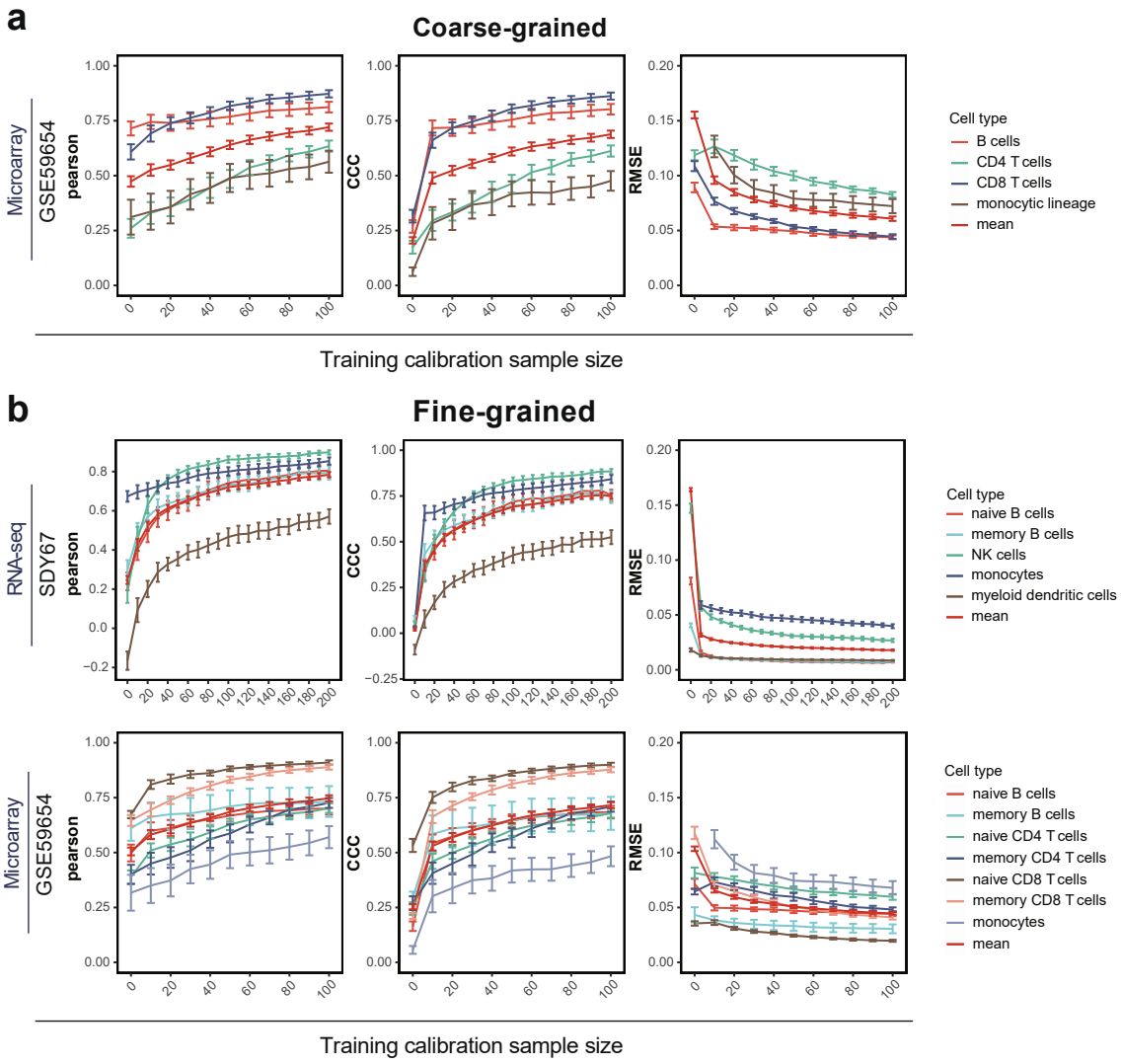

Figure S12: The effect of calibration sample size on the DAISM-DNN pipeline, assessed by the Pearson correlation, CCC and RMSE of the cell type proportion estimation results. (a) Results for microarray dataset GSE59654 for coarse-grained cell types deconvolution. Data are presented as means±SEM, coloured by different cell types in each dataset. (b) Results for RNA-seq dataset SDY67 and microarray dataset GSE59654 for fine-grained cell types deconvolution.

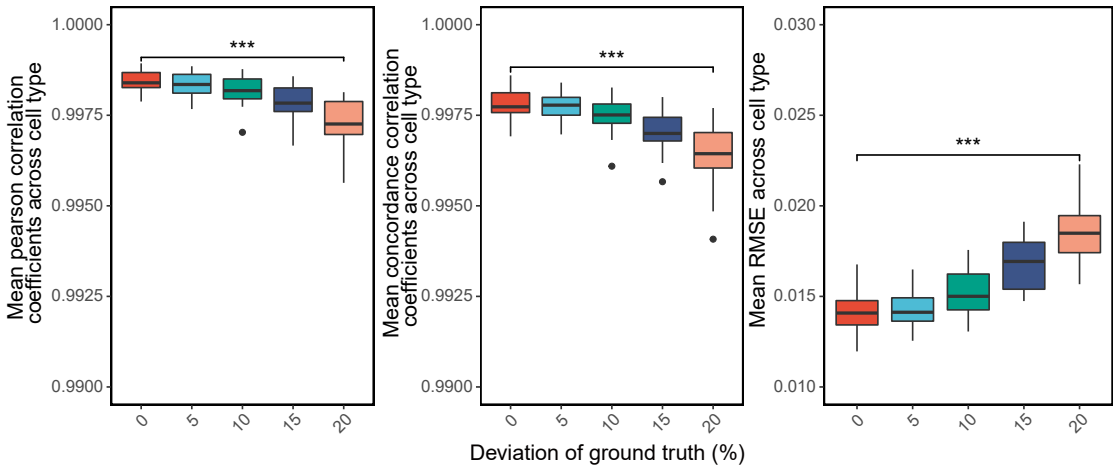

Figure S13: The performance of DAISM-DNN when ground truth of calibration samples (n = 200) has different degrees of deviation (5, 10, 15, 20%).***$P < 0.001$, paired one-way ANOVA test.

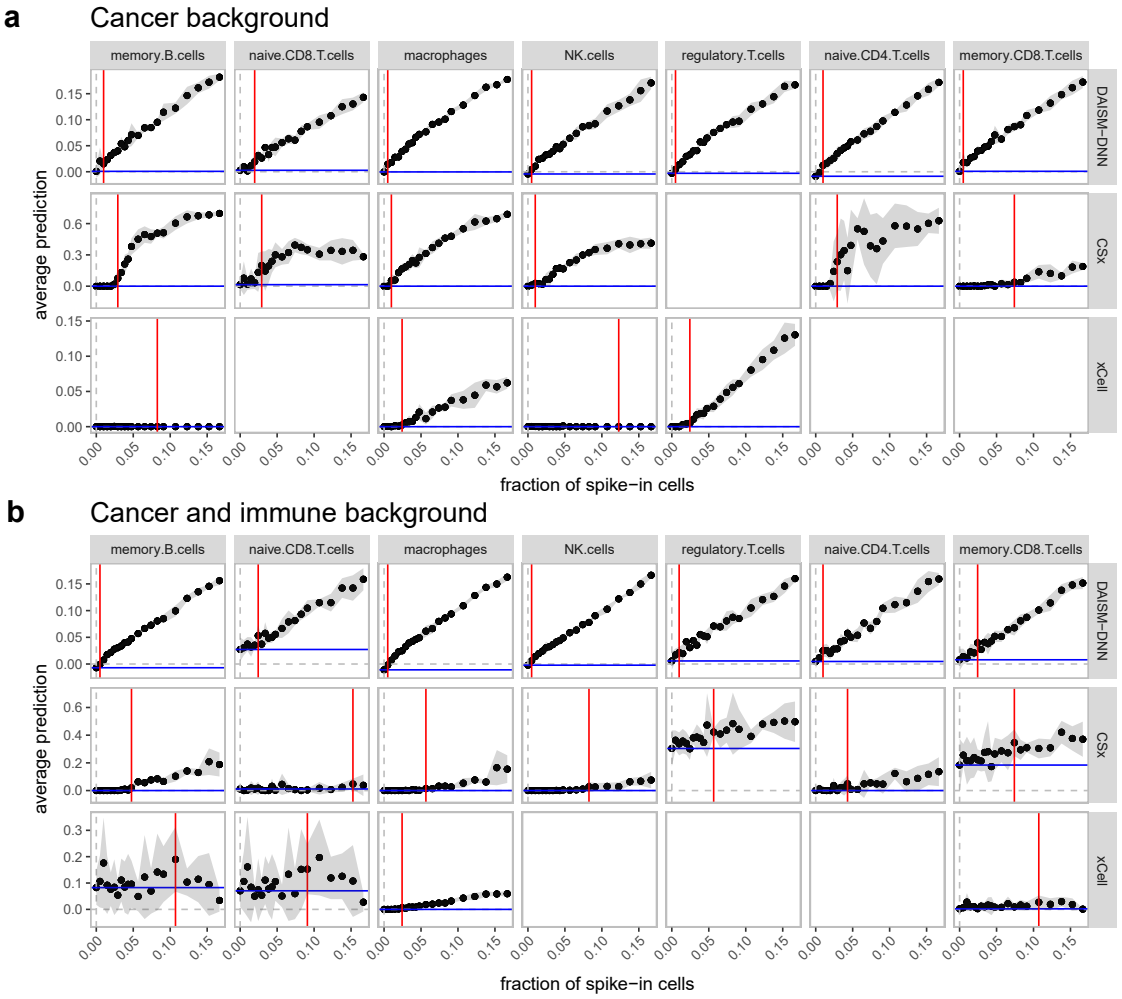

Figure S14: Minimal detection fraction of DAISM-DNN compared with CIBERSORTx and xCell. For each panel, we created simulated bulk RNA-seq samples by spiking-in an increasing amount of the cell type of interest and a background of 1000 cells randomly sampled from the other cell types. The dots show the mean predicted score across five independently simulated samples for each fraction of spike-in cells. The grey ribbon indicates the 95% confidence interval. The red line refers to the minimal detection fraction, i.e., the minimal fraction of an immune cell type needed for a method to reliably detect its abundance as significantly different from the background (P-value < 0.05, one-sided Student's t-test). The blue line refers to the background prediction level, i.e., the average estimate of a method while the cell type of interest is absent. (a) A background of 1000 cells randomly sampled from cancer cells. (b) A background of 1000 cells randomly sampled from cancer and all other immune cells.

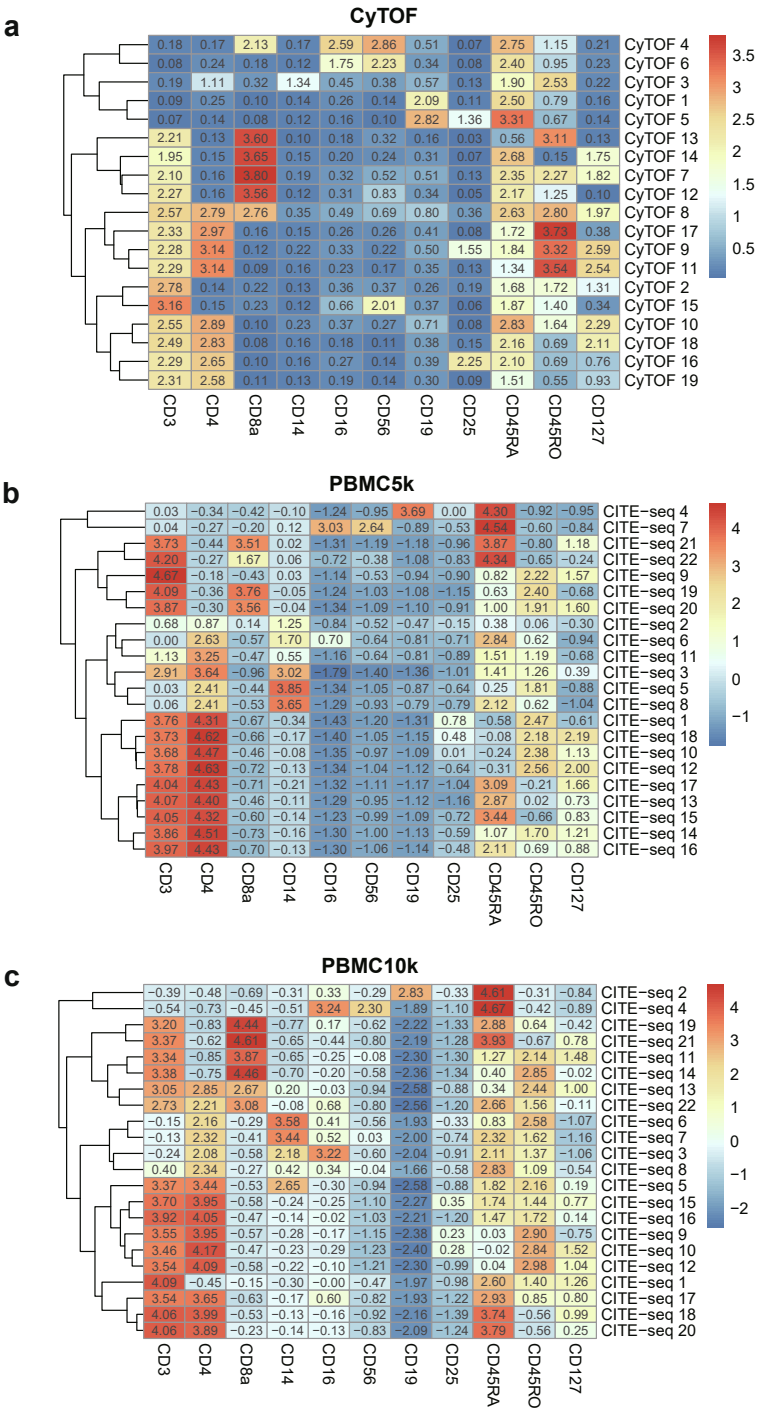

Figure S15: Heatmap showing mean values of normalized markers expression in each Phenograph clusters. CyTOF in-house dataset (a) and two public CITE-seq datasets (b) PBMC5k and (c) PBMC10k.

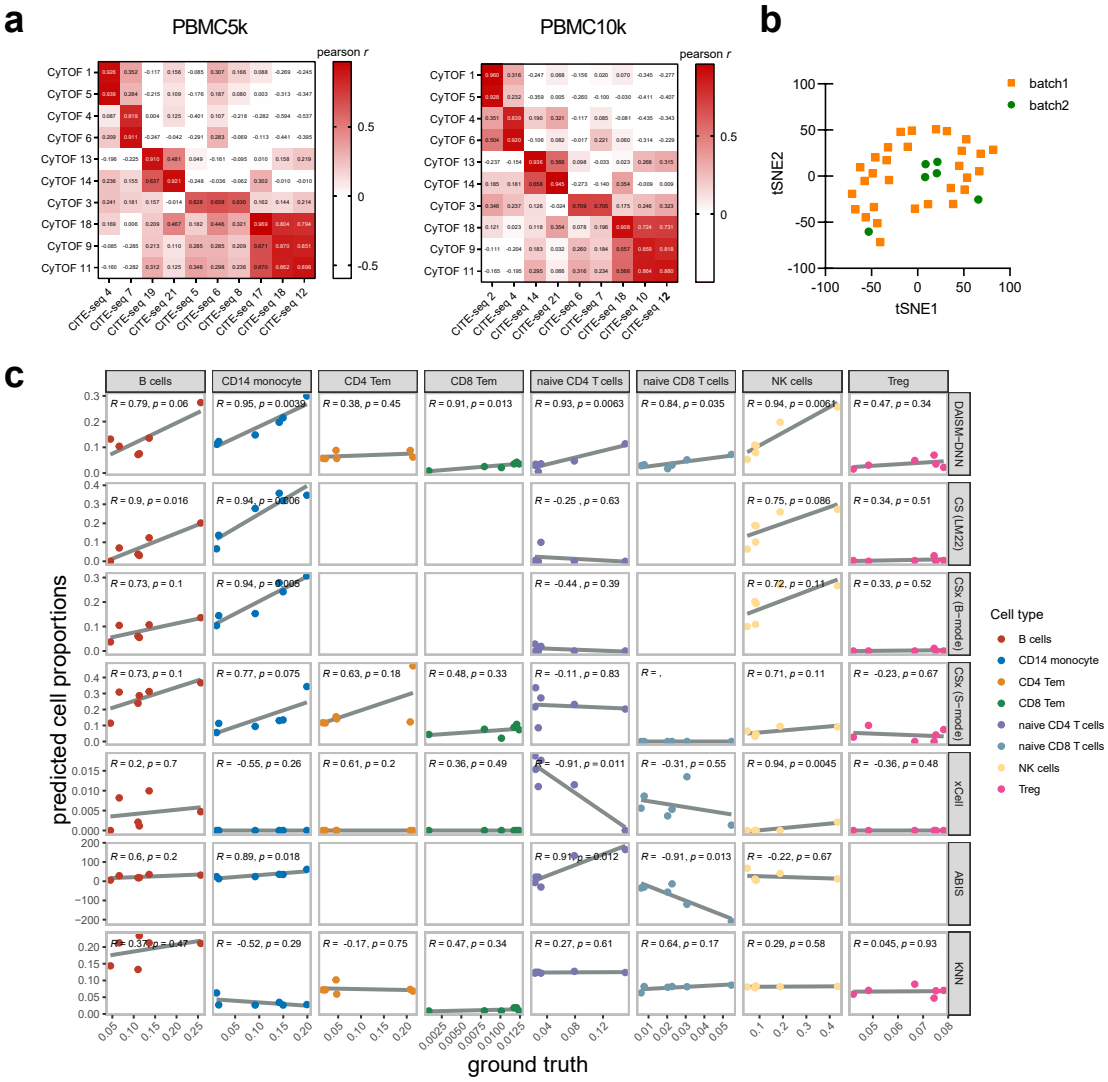

Figure S16: DAISM-DNN enables cross-batch accurate cell type proportion estimation. (a) Heatmap showing pearson correlation between CyTOF data clusters and CITE-seq data clusters used for augmentation. (b) The t-SNE projection of in-house dataset colored in different batches. (c) Scatterplots for eight immune cell types of ground truth (x axis) and predicted values (y axis) for DAISM-DNN, CSx (including two batch-correction modes), xCell, ABIS and KNN on in-house data. Numbers inside the plotting area signify Pearson correction values and p-values (Student's t-test).

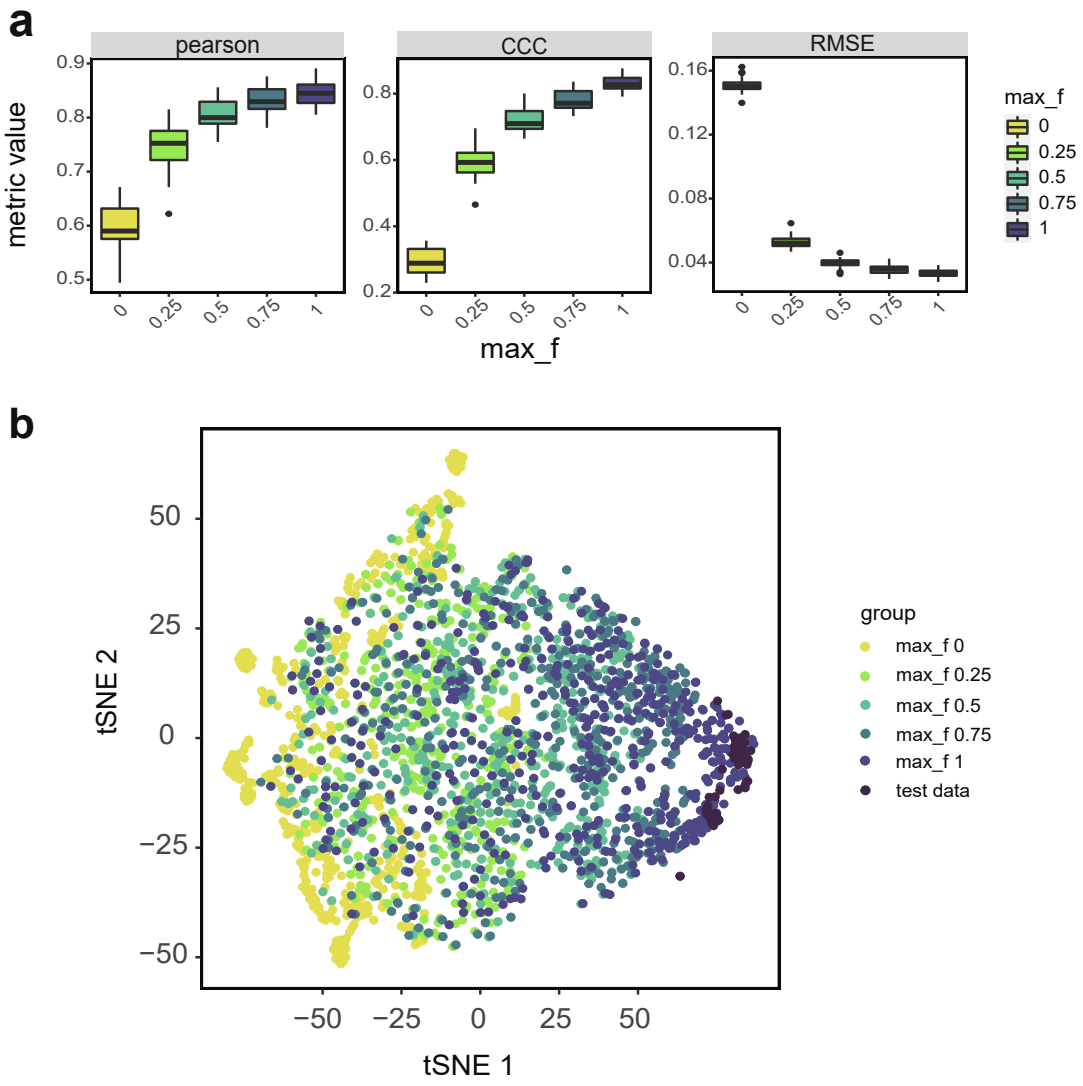

Figure S17: Assessment of the effect of max fraction of calibration samples on DAISM-DNN. (a) Performance of DAISM-DNN on different levels of max fraction of calibration samples when creating *in silico* mixing samples. The fraction of calibration samples was a random variable with uniform distribution between 0 and max_f. We performed 30 permutation experiments. In each permutation tests, 50 randomly selected samples of SDY67 dataset were used as testing and the remaining samples were used as calibration samples for creating DAISM-generated training set. (b) The t-SNE projection of the SDY67 dataset (n = 50 samples) and training datasets (n = 500 samples per set) constructed by DAISM mixing strategies, coloured according to different max fraction of calibration samples in mixed samples.

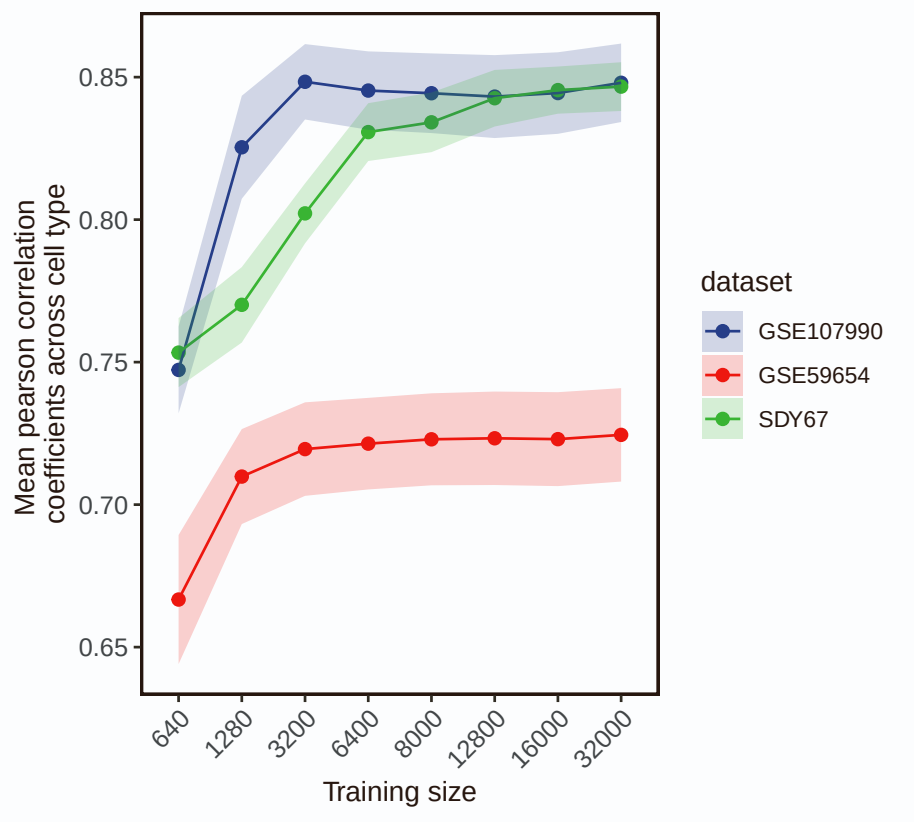

Figure S18: Assessment of the effect of training data size on mean pearson correlation coefficients across cell types. We used the pbmc8k single cell dataset as augmentation dataset to generate training data of different sizes between 640 samples and 32,000 samples for training the DNN. For each real world dataset, 50 samples were randomly selected for testing and the remaining samples of the corresponding dataset were used as calibration samples to perform DAISM mixing strategy to generate training dataset. The ribbon indicates the 95% confidence interval.

Supplementary Tables

Table S1: Overview of related methods for immune cells abundance prediction.

Methods	Type	Comparisons	Data Algorithm	Model Algorithm	Reference
DAISM-DNN	DNN-based	intra, inter	data augmentation	deep learning	this paper
Scaden	DNN-based	intra, inter	bulk RNA-seq mixtures simulated with scRNA-seq data	deep learning	Menden et al., 2019
MuSiC	signature-based	intra, inter	weighting of genes showing cross-subject and cross-cell consistency	weighted non-negative least squares regression	Wang et al., 2019
CIBERSORTx（CSx）	signature-based	intra, inter		v-support vector regression	Newman et al., 2019
CIBERSORT (CS)	signature-based	inter		v-support vector regression	Newman et al., 2015
ABIS	signature-based	intra, inter	TMM normalization, mRNA scaling factor	robust linear modeling	Monaco et al., 2019
EPIC	signature-based	intra, inter	renormalization based on cell-type-specific mRNA content	constrained least square regression	Racle et al., 2017
quanTIseq	signature-based	intra, inter		constrained least square regression	Finotello et al., 2017
MCP-counter	marker-based	inter	transcriptomic markers selection	geometric mean of expression of marker genes	Becht et al., 2016
xCell	marker-based	inter	spillover compensation	ssGSEA	Aran et al., 2017

Table S2: Inventory of expression datasets analyzed in this work.

Name	Species	Accession number or URL	Tissues	No. total samples/ cells	No. analyzed samples/cells	Data type	Platform	Reference
GSE64385	Human	GEO: GSE64385	PBMC and Polymorphonuclear Cells (PMN)	12	10	Microarray	Affymetrix Human Genome U133 Plus 2.0 Array	Becht et al.
GSE65133	Human	GEO: GSE65133	PBMC	20	20	Microarray	Illumina HumanHT-12 V4.0 expression beadchip	Newman et al.
GSE106898	Human	GEO: GSE106898	PBMC	12	12	Microarray	Illumina HumanHT-12 V4.0 expression beadchip	Monaco et al.
GSE64655	Human	GEO: GSE64655	PBMC	8	8	RNA-seq	Illumina HiSeq 2000 (Homo sapiens)	Hoek et al.
GSE127813	Human	GEO: GSE127813	whole blood	12	12	RNA-seq	Illumina HiSeq 4000 (Homo sapiens)	Newman et al.
GSE107011	Human	GEO: GSE107011	PBMC	12	12	RNA-seq	Illumina HiSeq 2000 (Homo sapiens)	Monaco et al.
GSE59654	Human	GEO: GSE59654	PBMC	156	153	Microarray	Illumina HumanHT-12 V4.0 expression beadchip	Thakar et al.
GSE107990	Human	GEO: GSE107990	PBMC	671	164	Microarray	Illumina HumanHT-12 V4.0 expression beadchip	Narang et al.
SDY67	Human	https://www.immport.org/shared/study/SDY67	PBMC	477	250	RNA-seq	Illumina HiSeq 2000 (Homo sapiens)	Zimmermann et al.
GSE107572	Human	GEO: GSE107572	PBMC/PMN cell mixture	9	9	RNA-seq	Illumina HiSeq 2500 (Homo sapiens)	Finotello et al.
GSE130824	Human	GEO: GSE130824	whole blood	35	35	RNA-seq	Illumina HiSeq 2500 (Homo sapiens)	Mao et al.

PBMC8k	Human	https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k	PBMCs	8381	8038	scRNA-seq	10x Chromium v2 (3' kit)	10x Genomics
PBMC6k	Human	https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k	PBMCs	5419	4745	scRNA-seq	10x Chromium v1	10x Genomics
GSE115978	Human	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115978	PBMCs	7186	7186	scRNA-seq	Illumina NextSeq 500 (Homo sapiens)	Arnon et al.
PBMC10k	Human	https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3	PBMCs	7865	6630	CITE-seq	Illumina NovaSeq	10x Genomics
PBMC5K	Human	https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_protein_v3	PBMCs	5247	3902	CITE-seq	Illumina NovaSeq	10x Genomics

Table S3: Hyperparameter settings of different numbers of neurons and layers of DAISM-DNN.

Model	N neurons layer1	N neurons layer2	N neurons layer3	N neurons layer4	N neurons layer5
M1	16				
M2	32				
M3	32	16			
M4	64				
M5	64	32			
M6	64	32	16		
M7	128				
M8	128	64			
M9	128	64	32		
M10	128	64	32	16	
M11	256				
M12	256	128			
M13	256	128	64		
M14	256	128	64	32	
M15	256	128	64	32	16
M16	512				
M17	512	256			
M18	512	256	128		
M19	512	256	128	64	
M20	512	256	128	64	32
M21	1024				
M22	1024	512			
M23	1024	512	256		
M24	1024	512	256	128	
M25	1024	512	256	128	64

Table S5: List of Antibodies.

List	Label	Antibody Target	clone	Category
1	89Y	CD45	HI30	Surface
2	115In	CD3	UCHT1	Surface
3	141Pr	CD56	NCAM16.2	Surface
4	142Nd	CD123	6H6	Surface
5	143Nd	CD274/PDL1	29E.2A3	Surface
6	144Nd	CD14	M5E2	Surface
7	145Nd	CD27	O323	Surface
8	146Nd	CD11a	HI111	Surface
9	147Sm	CD197/CCR7	G043H7	Surface
10	148Nd	CD19	HIB19	Surface
11	149Sm	CD33	WM53	Surface
12	150Nd	CD223/LAG3	874501	Surface
13	151Eu	CD45RO	UCHL1	Surface
14	152Sm	CD195/CCR5	J418F1	Surface
15	153Eu	CD366/Tim_3	F38-2E2	Surface
16	154Sm	Tbet	4B10	intra
17	155Gd	CD45RA	HI100	Surface
18	156Gd	CD194/CCR4	L291H4	Surface
19	157Gd	CD206	44242	Surface
20	158Gd	CD127	A019D5	Surface
21	159Tb	CD11c	BU15	Surface
22	160Gd	CD25	24212	Surface
23	161Dy	CD152/CTLA4	14D3	Surface
24	162Dy	Foxp3	PCH101	intra
25	163Dy	CD163	GHI/61	Surface
26	164Dy	CD38	HIT2	Surface
27	165Ho	CD44	BJ18	Surface
28	166Er	CD69	FN50	Surface
29	167Er	CD36	5-271	Surface
30	168Er	CD137/4-1BB	4B4-1	Surface
31	169Tm	Ki67	SolA15	intra
32	170Er	CD7	CD7-6B7	Surface
33	171Yb	CD279/PD1	EH12.2H7	Surface
34	172Yb	CD273/PDL2	24F.10C12	Surface

35	173Yb	CD278/ICOS	C398.4A	Surface
36	174Yb	CD134/OX40	BER-ACT35	Surface
37	175Lu	CD16	3G8	Surface
38	176Yb	HLA-DR	L243	Surface
39	197Au	CD4	RPA-T4	Surface
40	198Pt	CD8a	RPA-T8	Surface
41	209Bi	CD11b	M1/70	Surface