Power analysis of cell-type deconvolution methods across tissues

Anna Vathrakokoili Pournara[1*], Zhichao Miao[1,2,¢*], Ozgur Yilmaz Beker[1,#],Alvis Brazma[1], Irene Papatheodorou[1,2*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Cambridge CB10 1SD, UK
[2]Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

*Corresponding authors

Anna Vathrakokoili Pournara
annavp@ebi.ac.uk

Zhichao Miao
miao_zhichao@gzlab.ac.cn

Irene Papatheodorou
irenep@ebi.ac.uk

Current Addresses

[¢]Z.M: Translational Research Institute of Brain and Brain-Like Intelligence and Department of Anesthesiology, Shanghai Fourth People's Hospital Affiliated to Tongji University School of Medicine, Shanghai, China.

[#]O.B: Faculty of Engineering and Natural Sciences, Sabanci University, Tuzla 34956, Istanbul, Turkey

Abstract

Cell-type deconvolution methods aim to infer cell-type composition and the cell abundances from bulk transcriptomic data. The proliferation of currently developed methods (>50), coupled with the inconsistent results obtained in many cases, highlights the pressing need for guidance in the selection of appropriate methods.. Previous proposed tests have primarily been focused on simulated data and have seen limited application to actual datasets.. The growing accessibility of systematic single-cell RNA sequencing datasets, often accompanied by bulk RNA sequencing from related or matched samples, makes it possible to benchmark the existing deconvolution methods more objectively. Here, we propose a comprehensive assessment of 29 available deconvolution methods, leveraging single-cell RNA-sequencing (scRNA-seq) data from different organs and tissues. We offer a new comprehensive framework to evaluate deconvolution across a wide range of scenarios and we provide guidelines on the preprocessing of input matrices. We then validate deconvolution results on a gold standard bulk Polymorphonuclear Peripheral/Blood Mononuclear cell (PMN/PBMC) dataset with well known cell-type proportions. We show that single-cell regression-based deconvolution methods perform well but their performance is highly dependent on the reference selection and the tissue type. Our study also explores the significant impact of various batch effects on deconvolution, including those associated with sample, study, and technology, which have been previously overlooked. Importantly, we suggest a novel methodology for consensus prediction of cell-type proportions for cases when ground truth is not available. The large-scale assessment of cell-type prediction methods is provided in a modularised pipeline for reproducibility (https://github.com/Functional-Genomics/CATD_snakemake). Lastly, we suggest/propose/recommend that the Critical Assessment of Transcriptomic Deconvolution (CATD) pipeline be employed for the efficient, simultaneous deconvolution of hundreds of real bulk samples, utilising various references. weenvision it to be used for speeding up the evaluation of newly published methods in the future.

Keywords: deconvolution, benchmarking, cell proportions, pseudo-bulk, batch effects, real bulk.

Abbreviations:

sc-RNA-seq: single-cell RNA-Sequencing
PMN/PBMCs: Polymorphonuclear/Peripheral Blood Mononuclear cell

Main

There is a growing interest in understanding the level of heterogeneity and the importance of cell-type abundances in healthy and diseased tissues. Transcriptomic heterogeneity and cell-type composition reveal critical features of tissue functionality. For example, in cancer, immune cells can either be recruited in the vicinity of the tumour to contribute to cancer progression, or they can play a tumour-suppressive role by recognising and killing abnormal cells[1–3]. High infiltration from specific immune cells can be predictive of the disease outcome, stage and aid the treatment selection[4]. However, determining the composition of a tissue sample can present significant challenges. Although techniques, such as single-cell RNA sequencing can provide us with valuable information about the heterogeneity of tissue, the absolute proportions of cell types observed could be far from the reality due to technical handling, sorting of specific cell types, sub-optimal dissociation, and tissue vulnerability[5]. Furthermore, conventional methods, such as flow cytometry and immunohistochemistry which can quantify cell abundances, rely on prior knowledge of cell markers, depend on interpretation bias and are not easily scalable[6,7]. Additionally, there are only a handful of these types of datasets available, mostly from non-solid tissues such as PBMCs[8,9]. Consequently, cell-type deconvolution methods show a great advantage, since cell proportions from hundreds of samples can be quantified computationally in minutes.

Deconvolution methods aim to derive cell type compositions and cell abundances from bulk RNA-seq data using large databases such as GTEx[10], TCGA[11], TARGET (https://ocg.cancer.gov/programs/target) and GEO[12]. Specifically, cell-type deconvolution can assist in the study interaction between cell abundances and various disease phenotypes or clinical features, coupled with expression data within large databases[13]. In addition, deconvolution can assist in identifying perturbed processes and pathways in cell-type resolution from bulk expression data[14]. Finally, the acquisition of prior knowledge on cell proportions is very crucial in Transcriptome-Association studies (TWAS) in order to remove the cell abundance bias from eQTL analysis as well as identify cell-type specific eQTLs[15].

Deconvolution methods can be classified into three main categories based on the required inputs: supervised, semi-supervised and unsupervised. Supervised deconvolution methods require an expression matrix (from bulk RNA-seq/microarray) to be deconvolved and cell-type specific information that can be derived from flow cytometry or single-cell RNA-seq data[16,17]. These methods can be classified into bulk and single-cell based on the type of reference data provided. CIBERSORT[16], FARDEEP[18], NNLS (non-negative least squares) and EpiDISH[19] are some examples of commonly used bulk deconvolution methods. Contemporaneous deconvolution methods leverage available single-cell datasets to extract cell-type specific features in order to perform deconvolution. Bseq-sc[17], DWLS[20] and MuSiC[21] and others [22–24] utilise single-cell data to deconvolve bulk RNA-seq samples. Unlike the supervised methods, unsupervised methods like deconf[25] and CDSeq[26] only require a bulk sample to be deconvolved and the number of components (number of cell types to

be found) as an input. Finally, semi-supervised methods such as ssKL and ssFrobenius, simultaneously estimate the cell-specific matrix and cell proportions, using marker information and a block-descent method that estimates both matrices[27].

Even though deconvolution remains a very promising and cost-effective approach for estimating cell-type abundances in a tissue, there are still many challenges to overcome. A plethora of methods have been developed, yet there are no clear guidelines on method selection and a systematic large-scale benchmark across tissues is missing. There are additional components such as the pre-processing of the input matrices, the selection of reference and the batch effects between data inputs that complicate deconvolution[22]. However, it remains uncertain whether the proportions of closely related cell types or cell states can be accurately resolved[26].

Previous benchmarking studies have evaluated the performance of deconvolution methods on a small number of datasets and on limited organs and tissues, with a focus on data from pancreatic islets and blood tissue[28–30]. Moreover, there is a dearth of formal guidelines for the generation of suitable simulation data to drive new method development. Most benchmarks have focused on the technical aspects of deconvolution and lack biological interpretation of real tissue data[29].

Here, we have designed a curated, large-scale, detailed assessment of 29 available deconvolution methods by utilising sc-RNA-sequencing datasets from both human and mouse tissues. We first suggest a framework for simulating realistic bulk samples from single-cell methodologies, suitable for evaluating diverse methods. Next, we provide guidelines on the use of normalisation and transformation to reduce the effect of library size and the differences in expression between bulk and single-cell references respectively. Furthermore, we study the impact of batch effects in deconvolution and provide suggestions for the optimal selection of references. We subsequently evaluate our benchmark on real PBMC data, which includes known cell proportions, to illustrate the actual capabilities and limitations of the available methods. Lastly, we propose a novel approach for the selection of suitable methods for bulk samples that lack ground truth.

Results

Computational framework for large-scale analysis of deconvolution (CATD)

We developed a computational framework that enables us to systematically investigate various factors that influence the deconvolution process, in order to provide clear guidelines for the decomposition of bulk tissue expression data. (Fig 1).

To benchmark the 29 available deconvolution methods, we collected scRNA-sequencing published datasets from different human and mouse organs (Table S1). In our pipeline, each scRNA-seq dataset is split into a training dataset, used as a "reference input" for deconvolution, and a testing dataset that was composed of pseudo-bulk mixtures to be deconvolved (Methods). The use of pseudo-bulk mixtures allows us to obtain ground truth against which we can evaluate the deconvolution methods. The collected scRNA-seq experiments were generated by numerous protocols including Smart-Seq2[31], 10X[32], CEL-seq[33], Seq-well[34], Drop-Seq[35], inDrop[36] and originate from different organs including brain, pancreas, lung, blood(PBMC) and placenta. Moreover, datasets of different granularity were included in the benchmark to account for the effect of cell-type numbers in deconvolution (Table S1). As a result, the generated pseudo-bulks encompass a wide range of real data scenarios. Additionally, different levels of noise were added to the pseudo-bulks to mimic realistic bulk samples. In addition, 11 different normalisation and 2 transformation methods (Methods, TableS2) were tested on both pseudo-bulk and reference matrices to select the most suitable combination of preprocessing to be performed prior to deconvolution. After the preprocessing step, we evaluated the performance of the 29 collected deconvolution methods (Table S3). To compare the performance of deconvolution methods we implemented various evaluation metrics in the CATD pipeline. Pearson correlation coefficient (r/pcor), root mean square error (RMSE) and average cosine similarity (avgcos) are the ones used across this study to evaluate the performance of the methods(Methods). We select the best performing methods taking into consideration the methods that have both low RMSE and high R values consistently across different simulation scenarios, datasets and tissues. All the steps of the benchmarking have been modularised in a snakemake pipeline for the reproducibility of the results (https://github.com/Functional-Genomics/CATD_snakemake). The pipeline also allows the evaluation of the scalability of each benchmarking module using additional metrics (Methods). Finally, the pipeline provides a practical, user-friendly framework to accurately deconvolve real bulk samples with or without ground truth by suggesting a new approach to obtain a fair consensus of the best-performing methods and providing visualisation of the deconvolution results.

Realistic pseudo-bulk simulations for systematic evaluation and method development

It is challenging to accurately determine the real cell-type abundances in bulk RNA-seq samples unless they are coupled with flow cytometry or immunohistochemistry experiments that can provide information about the cell proportions. The limited number of existing experiments that have captured both the expression profile and the cell-type proportions are mostly limited to non-solid

tissues like blood, focusing on immune cells[8,37]. In addition, both flow cytometry and immunohistochemistry (IHC) experiments require prior knowledge of molecular markers and, therefore, cannot detect cell types in an unbiased way[38]. Moreover, immunohistochemistry measures the cell proportions of a tissue slide (2D) that do not necessarily reflect the cell abundances of a 3D tissue. Therefore, simulating bulk samples is a very useful approach for evaluating the performance of deconvolution methods, when ground truth is not available or reliable. A common method to simulate bulk profiles is to utilise publicly available single-cell data and generate pseudo-bulk by averaging the expression of single cells. The simulation strategies can range from random sampling to more sophisticated approaches [39,40]. However, it is non-trivial to understand how the different parameters in the simulation affect the deconvolution results.

Here, we propose a formal approach with standardised parameters to simulate bulk expression profiles suitable for benchmarking and *de novo* methods development. It is of great interest to use realistic simulations in deconvolution studies in order to apply them confidently in real data. We considered two different approaches in this study. In the first approach, pseudo-bulks are generated by randomly sampling cells from three pool sizes (p1=100, p2=10K, p3=100K) from a single-cell dataset and generating 1,000 samples. In the second approach, we generate an array of pseudo-bulks by setting the following parameters: the number of cell types sampled (nCT) from single-cell, the cell proportion assignment method (variance of proportions) along with the pool size (p) and the number of samples (s) (Methods, Fig. 2a). We generated 5 pseudo-bulk datasets based on a 10X single-cell-RNA-seq from the mouse brain[41] (Table S1, Supplementary Figure 1a-b) and compared them with real bulk samples derived from the same study (Fig. 2b). Real bulk and generated pseudo-bulk samples exhibit a negative binomial distribution, as expected, for expression data[42] however, random sampling pseudo-bulk deviates the most from real bulk, with less overdispersion observed (Fig. 2b). We also compared the similarity of gene expression between samples in each pseudo-bulk strategy. A random sampling of 10K cells with replacement, generates identical samples while the other strategies produce samples with lower cosine similarity scores (Fig.2c). Our results suggest that random sampling from single cells fails to generate pseudo-bulks that mimic real bulk expression and therefore should be avoided when developing new methods, as it oversimplifies the deconvolution problem. On the other hand, sampling cell types creates pseudo-bulks that can better mimic real bulk expression variability and can challenge deconvolution methods to deal with greater noise in bulk. We also observed that in our simulations, the number of cells per sample (p) affects the overall similarity of the distributions more than the number of samples (s) (Supplementary Figure 1c). Since simulating more cells will incur a greater computational burden and in real scenarios bulk data would include thousands to millions of cells, we propose sampling at least 10k cells for pseudo-bulk simulation. A simulation with 10k cells and with at least 500 samples provides a good baseline for near-convergence conditions, which are reproducible and show good overall performance scores - depending on the distribution used (Fig.

2d). Using all the cell types from single-cell with a fixed variance of proportions (allCT_prop) provides a good starting point for evaluation and method development; while sampling cell-types with a fixed variance of proportions (sampleCT_prop) represents the noisiest scenarios that can challenge deconvolution methods, as it shows the highest error values (Fig.2d-e). Overall we conclude that pseudo-bulk simulation parameters affect deconvolution severely and newly developed methods should be tested for performance in a range of scenarios before applying deconvolution to real data.

The selection of pre-processing methods yields varied deconvolution results

Previous studies have reported that pre-processing of input data in deconvolution impacts the deconvolution performance [22,28], but it is yet unclear which methods should be used for each deconvolution method selected. Standard preprocessing of bulk and single-cell data includes the transformation and normalisation of the expression matrices. Transformation methods account for the mean-variance dependencies as well as the extreme count values present in a dataset. This is a crucial pre-processing step for most downstream analyses of both single-cell and bulk RNA-seq data. For instance, both the identification of DEGs and PCA analysis requires the transformation of raw count values. The same principle applies to single-cell data for performing dimensionality reduction and identification of marker genes. Gene expression normalisation is also an essential step for the above analyses. Normalisation methods correct gene expression matrices for factors that affect the number of reads that will be mapped to a gene (gene length, GC content and sequencing depth)[43]. In general, the goal of normalisation is to enable the comparison of gene counts within a sample and across samples. As mentioned before, deconvolution requires reference data that provides cell-type specific information in the form of single-cells, summed single-cells or marker genes derived from single-cell and bulk RNA-seq data to be deconvolved. Here we test how correcting data with transformation and normalisation techniques affect the deconvolution process. We implemented 2 transformation methods and 11 different normalisation methods in our pipeline (Table S3). Previous studies have applied transformation of the data first and then normalisation. Here we show that normalising the matrices first and then applying transformation yields better deconvolution results, on average in 8 brain tissues(Fig. 3a-b,Supplementary Figure 2). We then examine the effect of different normalisation and transformation methods. Results from deconvolution across self-reference data from brain tissue show that the majority of methods output better results when the input data are not transformed (raw counts). However, bseq-sc and BisqueRNA methods benefit from the logarithmic transformation of the input matrices (Fig 3c). Moreover, in many cases, sqrt-transformation seems to perform worse than no-transformation with the exception of the penalised linear regression methods lasso, elastic net and ridge where it performs the same as no transformation of the matrices. Lastly, we tested 11 different normalisation methods across 3 different brain datasets that have been produced from different single-cell technologies (Hochgerner1 and Hochegerner 3=10X, Darmanis = SS2). In all three datasets, row normalisation seems to perform

worse across all the deconvolution methods (lower Pearson Correlation values), while the other normalisation methods appear to perform differently depending on the dataset. The majority of normalisation methods yield good results for the first dataset (Darmani2017). On the other hand, for the other two 10X datasets, none, global_min_max, column-zscore and mean normalisation approaches appear to yield better results. (Fig 3d-f).

Batch effects between (pseudo)bulk and single-cell data affect deconvolution

The first methods developed for deconvolution were utilising flow cytometry data from sorted cells, marker genes or unsupervised approaches to deconstruct bulk RNA-seq as well as microarray samples [44–46]. Recently, with the advancement of single-cell protocols and the huge availability of single-cell expression profiles from different human and mouse tissues, a great opportunity has been created to leverage them in deconvolution. Newly developed deconvolution methods make use of single-cell data to deconvolve bulk profiles by selecting appropriate features (genes) to be used in deconvolution (MuSIC[47], DWLS[20]). Yet, the large differences between single-cell and bulk quantification methods of expression data can impact the outcome of deconvolution. Here, we compared the expression of each gene between single-cell data, used as references and real bulk data from the same tissue. We observed significant expression differences between summed single-cell data (pseudo-bulk) and real bulk data originating from the same tissue, here pancreas[48,49] (Supplementary Figure 3a, Supplementary Figure 3b). In contrast, when comparing real bulk mixtures with each other, or pseudo-bulks with each other we observe very similar expression profiles (Supplementary Figure 3c,d). These discrepancies, observed between bulk and single-cell, are most likely due to the batch effects deriving from the different nature of the two sequencing techniques and dissimilar quantification methods. Principal Component Analysis (PCA) of the two data types, showed that PC1 could explain 74.5% of the variance between the bulk and the single-cell (Supplementary Figure 4a,b). First, we generated pseudo-bulk profiles and single-cell references by randomly splitting the same dataset in half and then performing deconvolution afterwards. This task is a very simple case of deconvolution since both the bulk and the reference contain the same gene expression information. We name this ‘self-reference deconvolution’. While self-reference gives the opportunity to study deconvolution with simulations, it oversimplifies the problem. For this reason, we introduced cross-sample deconvolution to more systematically study the impact of batch effects in cell-type deconvolution. For this task, we split the same pancreatic single-cell dataset in half and we selected cells from two individuals to generate the reference while the other two were used to build the bulk profiles. In this way, we artificially introduced batch

effects in cellular decomposition. Deconvolution results from one “bulk” and one “single-cell” algorithm (FARDEEP, DWLS) show a significant drop in deconvolution performance in the cross-sample task compared to the self-reference task (Fig. 5a, Supplementary Figure 5a). These results suggest that cell-type deconvolution is severely affected by batch effects caused by expression differences across samples. Furthermore, since the deconvolution of real bulk samples typically entails expression data from completely different studies and the use of different technologies, we next studied how cross-studies (same tissue) and cross-technology tasks can impact the performance of deconvolution algorithms. For the cross-study deconvolution, we selected two human pancreatic single-cell datasets[17,50] (TableS1) that were generated from two different laboratories. We observed that deconvolution across studies can impact deconvolution performance severely when compared to self-reference tasks (Figure 4b, Supplementary Figure 5b). Next, we compared the performance of deconvolution when datasets from different technologies are involved. For this task, we focused on 3 independent dataset pairs from the placenta, brain and lung tissue (Supplementary Figure 6, Supplementary Figure 7). Each dataset pair was generated from the same laboratory but with two different single-cell technologies (10X, Smart-Seq2). Given that 10X and Smart-Seq2 are the most common single-cell protocols available, this test will help reveal which references are suitable in deconvolution. Results show that self-reference deconvolution with 10X sc-RNASeq data yields better results overall (higher Pearson correlation, lower RMSE values Fig. 4c, Supplementary Figure 5c,d) across most deconvolution methods compared to Smart-seq2 self-reference. Moreover, results from the top method in this task (DWLS) demonstrate that cross-technology deconvolution results in less accurate predictions, with very low Pearson Correlation values observed (Figure 4d). These results highlight the importance of the reference selection and the barrier that batch effects oppose to deconvolution.

Selecting the best performing methods in the absence of ground truth

In contrast to the simulation tasks performed above, here we decompose real bulk profiles from 9 Peripheral Blood Mononuclear Cells (PBMC)[9] samples, utilising our deconvolution pipeline to assess the results and generate an honest consensus of predictions of cell-type proportions. Apart from the publicly available RNA-seq dataset from the above study, there is also flow cytometry data available for the 9 individuals which can be used as ground truth in our deconvolution pipeline. The researchers have measured the proportions of 8 immune cell types using specific markers for each cell population (CD4+ T cells, CD8+ T cells, T-regulatory cells, B cells, NK cells, myeloid dendritic cells, monocytes and neutrophils (Supplementary figure 8a). For this deconvolution task, we selected a recent single-cell PBMC dataset[51] to use as a reference. We re-analysed the single-cell data and matched the

cell types between the single-cell and the bulk dataset so that our reference has the same cell types that had been measured in the bulk (Supplementary Figure 8c). We also summarised the proportions of CD4+positive and T-regulatory cells measured by flow cytometry since our single-cell did not contain T-regulatory cells (Supplementary Figure 8b). Before we apply deconvolution on the 9 real bulk samples we first performed a self-reference deconvolution to evaluate how the reference performs in pseudo-bulk deconvolution. Eight methods yield good results with a Pearson Correlation of more than 0.92 (Supplementary Figure 8d). Next, we applied the 29 deconvolution methods included in the pipeline on the real bulk PMN/PBMC data and obtained the cell-type predictions. DWLS and bseq-sc showed the highest Pearson Correlation values (Figure 5a). DWLS also demonstrates the lowest RMSE values (Figure 5b). Looking closer at the DWLS cell-type predictions we observe differences across samples (Figure 5c). We notice the highest cell-type prediction accuracy in two of the 9 donors (donor 1, R=0.92, p_value=0.0034 and donor 10 R=0.95,p_value=0.0009). Moreover, we examined the deconvolution predictions per cell type and noticed that CD4+ cells were consistently overestimated while CD8+ cells were underestimated in all 9 donors by the DWLS deconvolution method. The same pattern can be validated with the bseq-sc method which was ranked second in this task based on Pearson Correlation values (Figure 5d). It has been reported before[52] that it is difficult to deconvolve closely related cell types due to collinearity, a problem which some single-cell methods attempt to solve with sophisticated feature selection. Apart from collinearity, another frequent and concerning issue is the lack of ground truth data. Since there is no single method that clearly ranks first in all scenarios and it is still unclear how different methods work in data from different tissues, we believe that a fair deconvolution consensus is a reasonable approach to take. Here, we suggest a new methodology which is based on method comparison and can help obtain a fair consensus on decomposition results. For this, we perform a pairwise comparison of the deconvolution results across all samples and we estimate the Pearson Correlation coefficient. We visualise the results in a form of a heatmap where the methods that yield similar results cluster together. This indicates which methods strongly agree with each other (Supplementary Figure 8e). In the self-reference task, with the single-cell PBMC dataset, we confirm that the best-performing methods based on Pearson Correlation with ground truth also cluster together after a pairwise comparison of the results and without the use of ground truth. This suggests that when ground truth is not available, pairwise comparison of the results per cell type and per sample can help us select the methods that give more consistent results.
In parallel, we evaluated the scalability of each module of our pipeline(the pseudo-bulk generation method,normalisation and the deconvolution methods) and we report the time(s), memory, mean load and CPU time values for each one of them for both deconvolution tasks described above(Supplementary Figure 9a-b).

Discussion

In this study we developed a deconvolution framework that allows the evaluation of 29 publicly available methods across different scenarios with both synthetic and real data. We first studied the effect of different pseudo-bulk techniques by defining different ways of pseudo-bulk sampling. Our results demonstrated that pseudo-bulks generated with different methods produce highly diverse results in deconvolution, which shows that the results of deconvolution are highly sensitive to the sampling of cells and the selection of proportions. Here, we suggest the use of pseudo-bulk generation methods that can challenge deconvolution methods and reflect real bulk data, when developing new, or benchmarking existing, methods. Another important aspect of deconvolution, that we explore, is the normalisation and the transformation of the input matrices in deconvolution. Both bulk and single-cell matrices used in the deconvolution pipeline were tested across different normalisation and transformation methods. Results suggest that normalisation can be beneficial in deconvolution since it corrects for differences in the library size both in the single cell and the bulk RNA-seq data. On the other hand, logarithmic transformation of the input matrices can result in worse performance with the exception of two deconvolution methods (bisqueRNA and bseq-sc). Similar to the reports from previous studies, this suggests that keeping input data in linear scale aids in assessing the cell proportions accurately [28,29].

Another key aspect of the bulk sample deconstruction is the selection of the reference matrix, which is used to extract the key features (genes) that will be used in deconvolution. Results from Smart-seq2 and 10X single-cell datasets that have been used in this study, show that 10X data is consistently performing better as a reference, this is likely because this technology captures more cells of a specific cell type and as a result, more accurate features are extracted from these data. It should also be noted that in tasks when the bulk and the reference data come from different sources (studies or technologies) - the accuracy score of the deconvolution drops significantly. This indicates that batch effects in the expression of genes, across datasets, affect deconvolution heavily. Methods that take into account and successfully minimise these effects (DWLS, MuSiC) seem to also perform better overall in cross-reference tasks. We apply deconvolution in real PBMC samples using a suitable matching single-cell reference and conclude that two methods, DWLS and bseqsc perform best. All the above steps that should be performed before deconvolution have been implemented as part of the pipeline with a number of parameters that can be defined by the user.

We also propose a novel approach to select suitable deconvolution methods based on pairwise comparison of deconvolution results when ground truth is not available or reliable. The CATD pipeline, developed using snakemake, enables the reproducibility of our results and can be additionally utilised by users to deconvolve new bulk samples by providing only a bulk dataset and a single-cell dataset from the same tissue. When ground truth is provided too, the pipeline will provide evaluation metrics for each method. On the other hand when ground truth is absent the pipeline will perform pairwise comparison of the results and report a heatmap with methods to be selected. Moreover, the advent of new high-quality single-cell atlases with standardised annotation from both healthy and disease conditions will provide more opportunities to explore the contribution of cell proportions in pathological conditions. Finally, the development of robust, accurate tools for feature selection that enable clear distinction between cell-types will pave the way for use of computational deconvolution perhaps even in clinics or in settings where scRNA-Seq is not feasible.

Online Methods

Pseudo-bulk generation

For the generation of pseudo-bulk samples, at least 3 and at most 5 parameters are needed aside from the single-cell data used for generation. The 2 fundamental parameters are 'cell counts per sample' (***c***) and the 'number of samples' (***n***) to be generated. The 'mode' parameter is either 1 or 2. If 1, the sampling will randomly pick ***c*** cells with replacement from the single-cell data to generate one sample, which will be repeated ***n*** times.

If the mode parameter is 2, the generation will require two more parameters, namely 'proportional variance' (***p***) and whether or not some cell types will be left out during generation. When including all cell types, if ***p*** is negative, then random proportions ranging between 0.01 to 0.99 and summing to 1 will be generated using a uniform distribution. If ***p*** is positive, then integers between ***1*** and ***1+2p*** will be generated using a uniform distribution and then normalised to sum to 1. This enables more precise control of the noise, with the former (1-99 case) being the edge case of the latter (positive ***p***) when ***p*** tends to infinity. If cell types are sampled, a bimodal normal distribution with uniform mode probability is used to achieve the effect and the ***p*** instead becomes the standard distribution of the kept cell types. In this case, the provided ***p*** should be positive and ideally large. The absolute value of generated numbers is considered to ensure non-negative proportions and then are normalised to sum to 1.

The generated proportions in mode 2 are then scaled by ***c*** to determine how many cells will be chosen with replacement per cell type from the reference data to generate one sample, which is again repeated ***n*** times.

Building Deconvolution Inputs

Aside from the pseudo-bulk data and the corresponding proportions, deconvolution methods require different references to process the inputs. Within the pipeline, these references are encoded as ***C0***, ***C1***, ***C2***, ***phenData*** and ***refVar***.

C0 is a generic single-cell matrix, with rows being genes and columns being cells. ***C1*** is the averaged (arithmetic mean) gene expression matrix by cell types, with rows being genes and columns being cell types. ***C2*** is a list of marker genes obtained from differential gene expression tests through the R package Seurat, specifically using a combination of *NormalizeData* and *FindAllMarkers* functions. ***phenData*** is the metadata object that is associated with ***C0***. Finally, ***refVar*** is the row standard deviations matrix by cell types, similar in structure to ***C1***.

Preprocessing in deconvolution(Normalisation and Transformation)

Within the pipeline, users have a chance to apply different normalisation methods to either the pseudo-bulk matrix or the single-cell reference matrix. While most of the normalisation methods are common (*column, row, mean, column/global z-score, column/global min-max, LogNormalize, QN, TMM*), there are 4 methods which are single-cell specific (*SCTransform, scran, scater, Linnorm*). The details on the implementation of these methods are available within the pipeline. The matrices can then also be transformed in different ways if desired (log, sqrt, vst or none). Moreover, it is possible to change the order of normalisation and transformation through the configuration file.

Evaluation of Performance - Metrics

The pipeline will evaluate the predictions through various metrics, where the overall score of a method for a given metric is the mean over all samples (columns) within the pseudo-bulk proportions matrix. The metrics available in the pipeline are:

$$\text{RMSE/RMSD} = \sqrt{\frac{\sum_{i=1}^{N} x_i - \widehat{x_i}}{N}}$$

RMSE/RMSD = Root-mean-square error/deviation

i = variable i (each sample)
N= number of observations
x_i= actual observations (here ground truth cell-type proportions)
$\widehat{x_i}$= estimated proportions from the algorithm

$$\text{MAE} = \frac{\sum_{i=1}^{N} |x_i - \widehat{x_i}|}{N}$$

MAE= Mean absolute error
i = variable i (each sample)
N= number of observations
x_i= actual observations (here ground truth cell-type proportions)
$\widehat{x_i}$= estimated proportions from the algorithm

$$\text{Minkowski Distance} = \left(\sum_{i=1}^{N} |x_i - \widehat{x_i}|^p\right)^{1/p}$$

Minkowski distance (with adjustable p, p=2 by default in the pipeline)
i = variable i (each sample)
N= number of observations
x_i= actual observations (here ground truth cell-type proportions)
$\widehat{x_i}$= estimated proportions from the algorithm

$$\text{r/pcor} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})2 \ \sum \ (y_i - \bar{y})2}}$$

r/pcor = Pearson Correlation Coefficient
i = variable i (each sample)
x_i = actual observations (here ground truth cell-type proportions)
$\bar{x}$ = mean of the values of x variable
y_i = estimated proportions from the algorithm
$\bar{y}$ = mean of the values of the y variable

$$\cos(\theta) = \frac{\sum_{i=1}^{N} A_i - B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}}$$

cos(θ) = Cosine similarity (we estimate the average cosine similarity across samples)
i = variable i (each sample)
N= number of observations
A_i= actual observations (here ground truth cell-type proportions)
B_i= estimated proportions from the algorithm

$$R^2 = \left(\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})2 \sum (y_i-\bar{y})2}}\right)^2$$

R^2 = R-squared
i = variable i (each sample)
x_i = actual observations (here ground truth cell-type proportions)
$\bar{x}$ = mean of the values of x variable
y_i = estimated proportions from the algorithm
$\bar{y}$ = mean of the values of the y variable

$$dCov(X,Y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} D(x_i, x_j)\, D(y_i, y_j)$$

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dCov(X,X)\, dCov(Y,Y)}}$$

dCov & dCor = Distance covariance & correlation as described in this study[53]
*Note that dCov(X,X) = dVar(x), which is the distance variance.*n = number of observations
$D(x_i, x_j)$ = centered Euclidean distances between the i_th and j_th observations of x
$D(y_i, y_j)$ = centered Euclidean distances between the i_th and j_th observations of y

All the metrics have been estimated in R and the functions can be found here https://github.com/Functional-Genomics/CATD_snakemake/blob/main/Modules/Res_explore/Res_explore.R For each metric, the scores of the methods are saved as

named list objects throughout the pipeline. Graphs are also constructed from these lists for initial viewing of the results.

Evaluation of Scalability - Metrics

Every step of the pipeline is benchmarked automatically by Snakemake's built-in 'benchmark' directive, and the results from these benchmarks are also visualised in bar plots at the end of the pipeline. 4 of the parameters are visualised for initial scalability assessment purposes, namely the ***time*** (s)***, memory*** (maximum resident set size)***, mean load*** (CPU usage divided by total time) and ***CPU time***. All of the remaining data generated by the directive are also kept if there is a further need for computational benchmarking of certain jobs.

Code availability

CATD Snakemake pipeline:
https://github.com/Functional-Genomics/CATD_snakemake

Acknowledgements

This work was supported by: the European Molecular Biology Laboratory (A.V.P, Z.M., O.B., A.B, I.P.); the EMBL international PhD program (A.V.P.) and the OpenTargets (C.M, I.P)

Author contributions

All authors conceived the project. A.V.P. and Z.M designed and wrote the framework as well as curated the utilised datasets. A.V.P performed the data analysis and visualisation, with supervision from I.P. and inputs from Z.M. and A.B.. O.B converted the software to a reproducible snakemake pipeline with feedback from A.VP. and the supervision of I.P. A.V.P. wrote the draft of the manuscript and all authors contributed in review and editing. We would also like to thank Craig Russell for the valuable help in reviewing and editing. Ethics declaration. Competing interests: the authors declare no competing interests.

References

1. Hanahan, D. & Coussens, L. M. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell* **21**, 309–322 (2012).
2. Dumont, N. *et al.* Breast Fibroblasts Modulate Early Dissemination, Tumorigenesis, and Metastasis through Alteration of Extracellular Matrix Characteristics. *Neoplasia* **15**, 249-IN7 (2013).
3. Taube, J. M. *et al.* Implications of the tumor immune microenvironment for staging and therapeutics. *Mod. Pathol.* **31**, 214–234 (2018).
4. Zhang, Y. *et al.* High-Infiltration of Tumor-Associated Macrophages Predicts Unfavorable Clinical Outcome for Node-Negative Breast Cancer. *PLoS ONE* **8**, e76147 (2013).
5. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
6. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Comput. Biol.* **16**, e1008120 (2020).
7. Matos, L. L. de, Trufelli, D. C., de Matos, M. G. L. & da Silva Pinhal, M. A. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. *Biomark. Insights* **5**, 9–20 (2010).
8. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627-1640.e7 (2019).
9. Finotello, F. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
10. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
11. The International Cancer Genome Consortium. International network of cancer

genome projects. *Nature* **464**, 993–998 (2010).

12. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

13. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 955 (2020).

14. Inkeles, M. S. *et al.* Cell-type deconvolution with immune pathways identifies gene networks of host defense and immunopathology in leprosy. *JCI Insight* **1**, (2016).

15. Lowe, R. & Rakyan, V. K. Correcting for cell-type composition bias in epigenome-wide association studies. *Genome Med.* **6**, 23 (2014).

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

17. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).

18. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLOS Comput. Biol.* **15**, e1006976 (2019).

19. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).

20. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* **10**, 2975 (2019).

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

22. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).

23. Dong, M. *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2021).

24. Frishberg, A. *et al.* Cell composition analysis of bulk genomics using single-cell data. *Nat. Methods* **16**, 327–332 (2019).

25. Repsilber, D. *et al.* Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* **11**, 27 (2010).

26. Kang, K. *et al.* CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Comput. Biol.* **15**, e1007510 (2019).

27. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **12**, 913–921 (2012).

28. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

29. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol.* **22**, 102 (2021).

30. White, B. S. *et al. Community assessment of methods to deconvolve cellular composition from bulk gene expression*. http://biorxiv.org/lookup/doi/10.1101/2022.06.03.494221 (2022) doi:10.1101/2022.06.03.494221.

31. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).

32. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

33. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2**, 666–673 (2012).

34. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).

35. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

36. Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).

37. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **67**, 1031–1040 (2018).

38. De Matos, L. L., Trufelli, D. C., De Matos, M. G. L. & Da Silva Pinhal, M. A. Immunohistochemistry as an Important Tool in Biomarkers Detection and Clinical Practice. *Biomark. Insights* **5**, BMI.S2185 (2010).

39. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

40. Dietrich, A. *et al.* SimBu *: Bias-aware simulation of bulk RNA-seq data with variable cell type composition*. http://biorxiv.org/lookup/doi/10.1101/2022.05.06.490889 (2022) doi:10.1101/2022.05.06.490889.

41. Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).

42. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007).

43. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2017).

44. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLOS ONE* **4**, e6098 (2009).

45. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol. Clifton NJ* **1711**, 243–259 (2018).

46. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**, 2211–2212 (2013).

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

48. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13924–13929 (2014).

49. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).

50. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).

51. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).

52. Aliee, H. & Theis, F. J. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Syst.* **12**, 706-715.e4 (2021).

53. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, (2007).

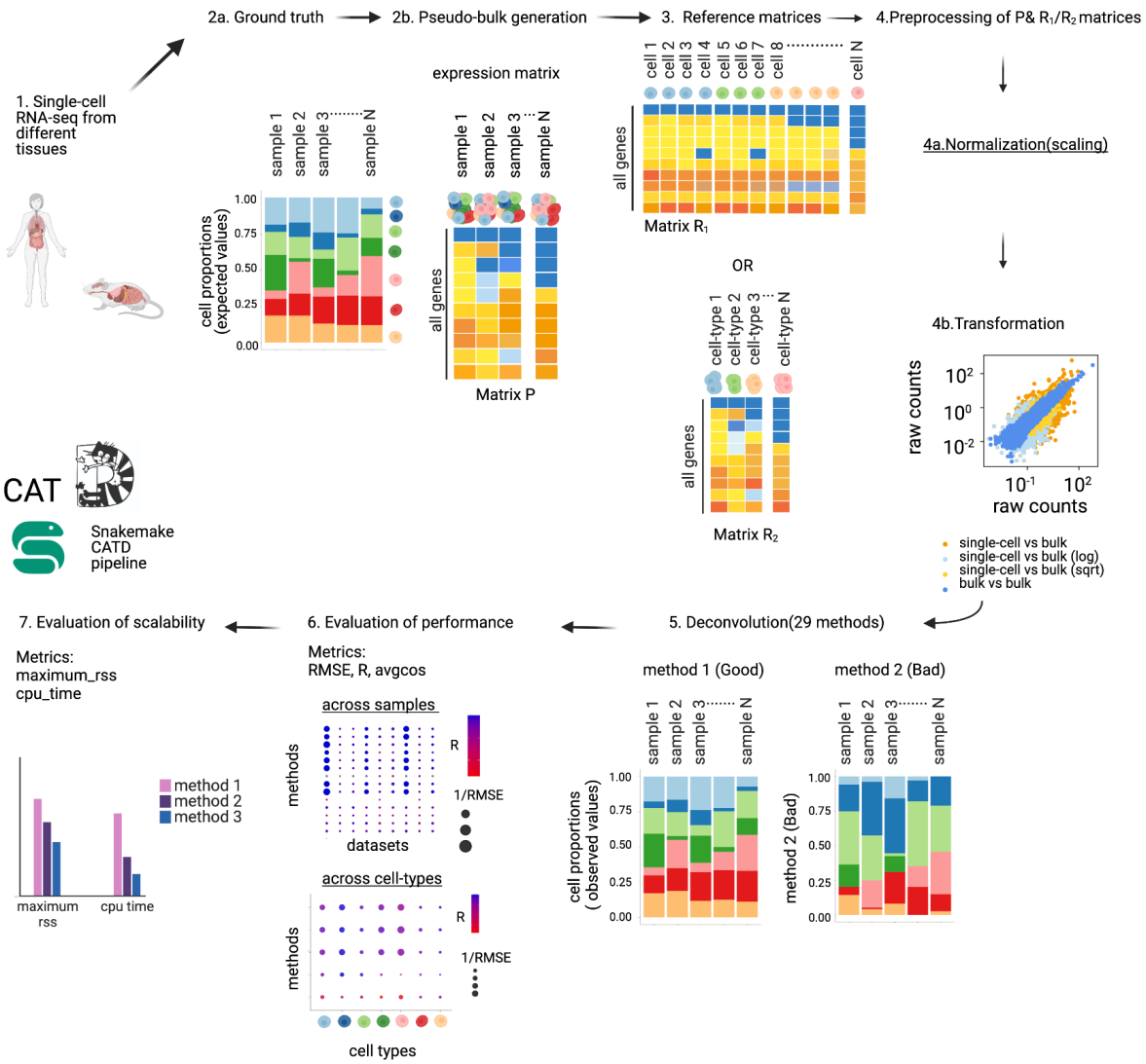


Figure 1. Critical Assessment of Transcriptomic Deconvolution(CATD)

Schematic representation of the large-scale benchmarking pipeline. 29 deconvolution methods were evaluated on simulated bulk data generated from scRNA-seq datasets and 5 separate simulation strategies. Furthermore, we test different normalisation and transformation methods on the bulk and reference matrices and we assess the results utilising various evaluation metrics (Methods). Root mean square error (RMSE) and Pearson Correlation (R) and average cosine similarity (avg cos) shown here. Finally, we evaluate the scalability of the methods. Every step of the benchmarking has been modularised in a snakemake pipeline for reproducibility.

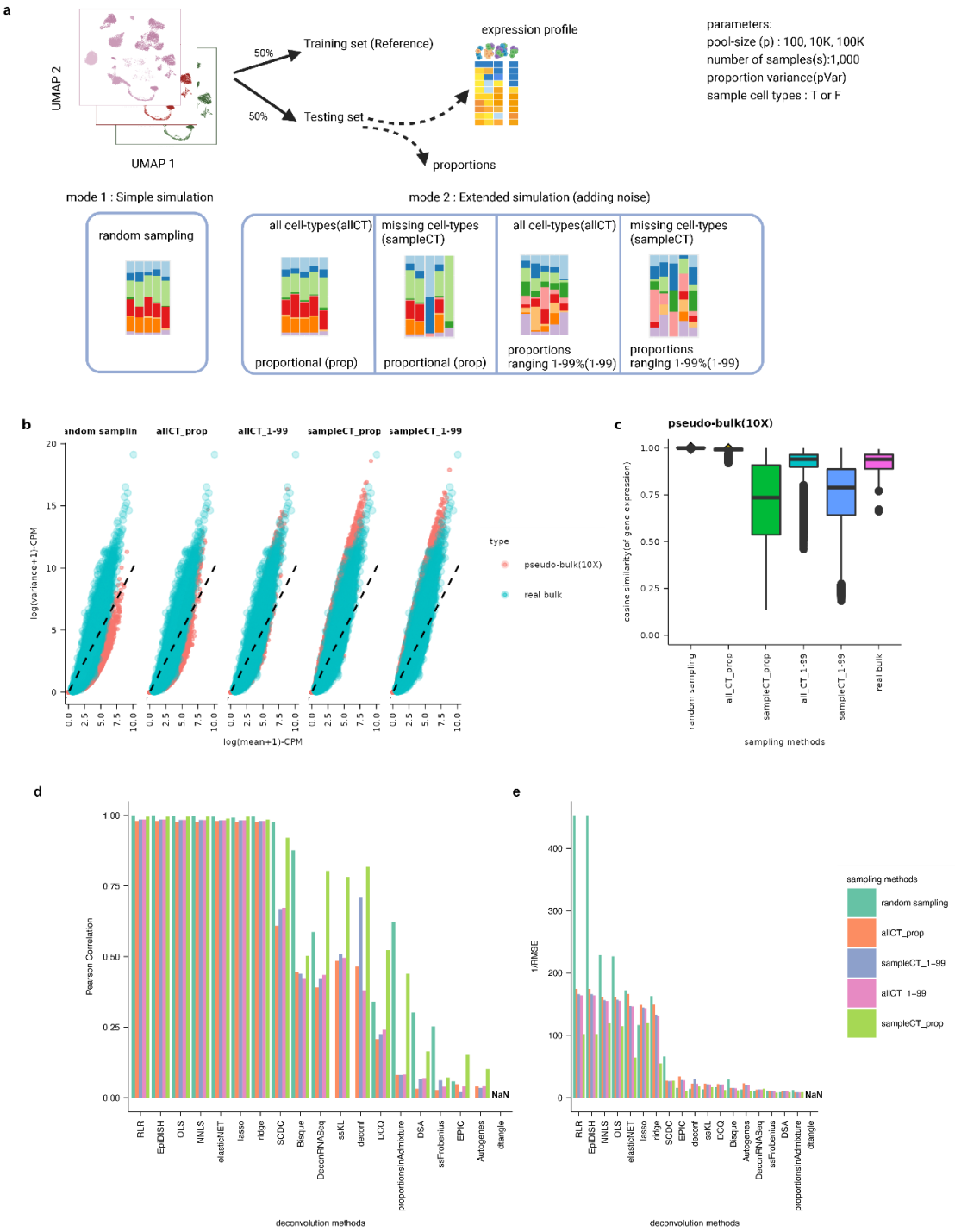

a
UMAP 2
UMAP 1
50%
Training set (Reference)
50%
Testing set
expression profile
proportions
parameters:
pool-size (p) : 100, 10K, 100K
number of samples(s):1,000
proportion variance(pVar)
sample cell types : T or F
mode 1 : Simple simulation
random sampling
mode 2 : Extended simulation (adding noise)
all cell-types(allCT)
proportional (prop)
missing cell-types (sampleCT)
proportional (prop)
all cell-types(allCT)
proportions ranging 1-99%(1-99)
missing cell-types (sampleCT)
proportions ranging 1-99%(1-99)
b
allCT_prop
allCT_1-99
sampleCT_prop
sampleCT_1-99
log(variance+1)-CPM
log(mean+1)-CPM
type
pseudo-bulk(10X)
real bulk
c
pseudo-bulk(10X)
cosine similarity(of gene expression)
random sampling
all_CT_prop
sampleCT_prop
all_CT_1-99
sampleCT_1-99
real bulk
sampling methods
d
Pearson Correlation
RLR
EpiDISH
OLS
NNLS
elasticNET
lasso
ridge
SCDC
Bisque
DeconRNASeq
ssKL
deconf
DCQ
proportionsInAdmixture
DSA
ssFrobenius
EPIC
Autogenes
dtangle
NaN
deconvolution methods
e
1/RMSE
sampling methods
random sampling
allCT_prop
sampleCT_1-99
allCT_1-99
sampleCT_prop
deconvolution methods

Figure 2. Evaluating pseudo-bulk simulations for benchmarking and new method development.

a. Simulation workflow: each single-cell matrix is split into half to generate a training set (reference) and a testing set (pseudo-bulk matrix). 5 different sampling techniques were used to generate pseudo-bulks and ground truth proportions. We set 4 different parameters for the simulation: the number of cells to be sampled from the single-cell (pool size (p)), the number of samples in each pseudo-bulk matrix (s)-the default is 1,000 samples, the variance of the proportions (pVar) for the cell-types and finally a logical vector to decide if we use all the cell-type or not (sampleCT, True or False). b. Scatter plot of mean and variance of expression data from 5 different simulation scenarios and matched real bulk data. All the simulated pseudo-bulk data derive from the same mouse 10X single-cell dataset and the real bulk RNA-seq derives from the same study as well (Hrvatin2018). c. boxplot plot showing the cosine similarity of gene expression for the samples of each simulation as well as the cosine similarity of expression across the real bulk samples. d. Pearson correlation values and e. root mean square values (RMSE) of the 5 different simulation scenarios across the tested deconvolution method that were feasible to obtain results.

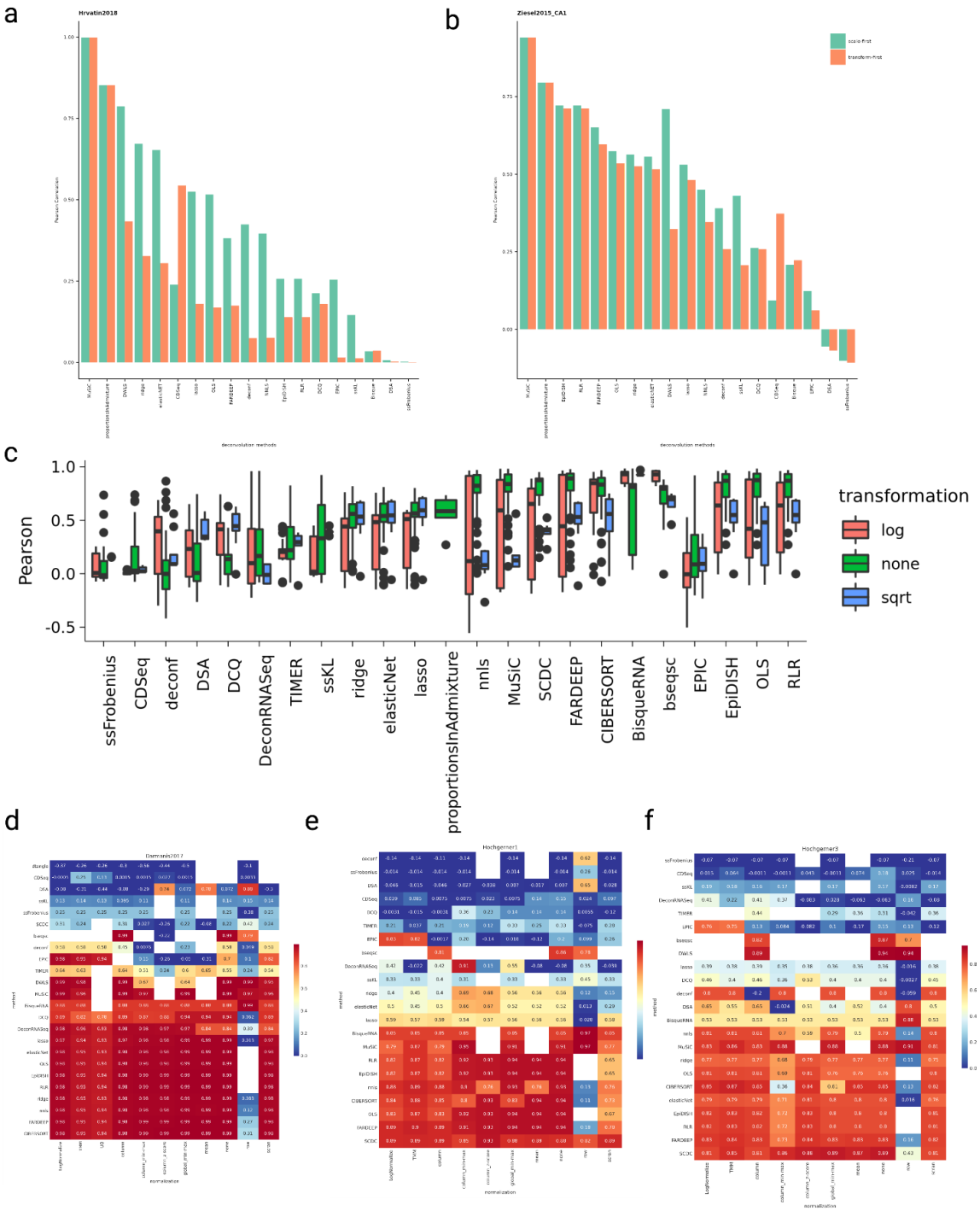


Figure 3. Pre-processing steps affect downstream decomposition of samples. Barplot showing Pearson Correlation values of a self-reference deconvolution task in which the effect of scaling first or transforming first the input matrices is tested a. Hrvatin 2018-brain tissue . b. Ziesel2015-brain tissue c. Boxplots showing Pearson correlation values estimated from the deconvolution of simulated data(from brain tissue) across 2 different transformation methods(log and sqrt) as well as no transformation(linear scale). Each data point in the boxplot represents a combination of a transformation and a normalisation method utilised in the deconvolution task.(e.g log+TMM, log+column and similar for all the feasible combinations) d-e. Results from deconvolution of 3 different simulated data from brain tissues(Darmanis2017,Hochgerner1,Hochgerner3) across 11 normalisation methods and 24 tested deconvolution methods.

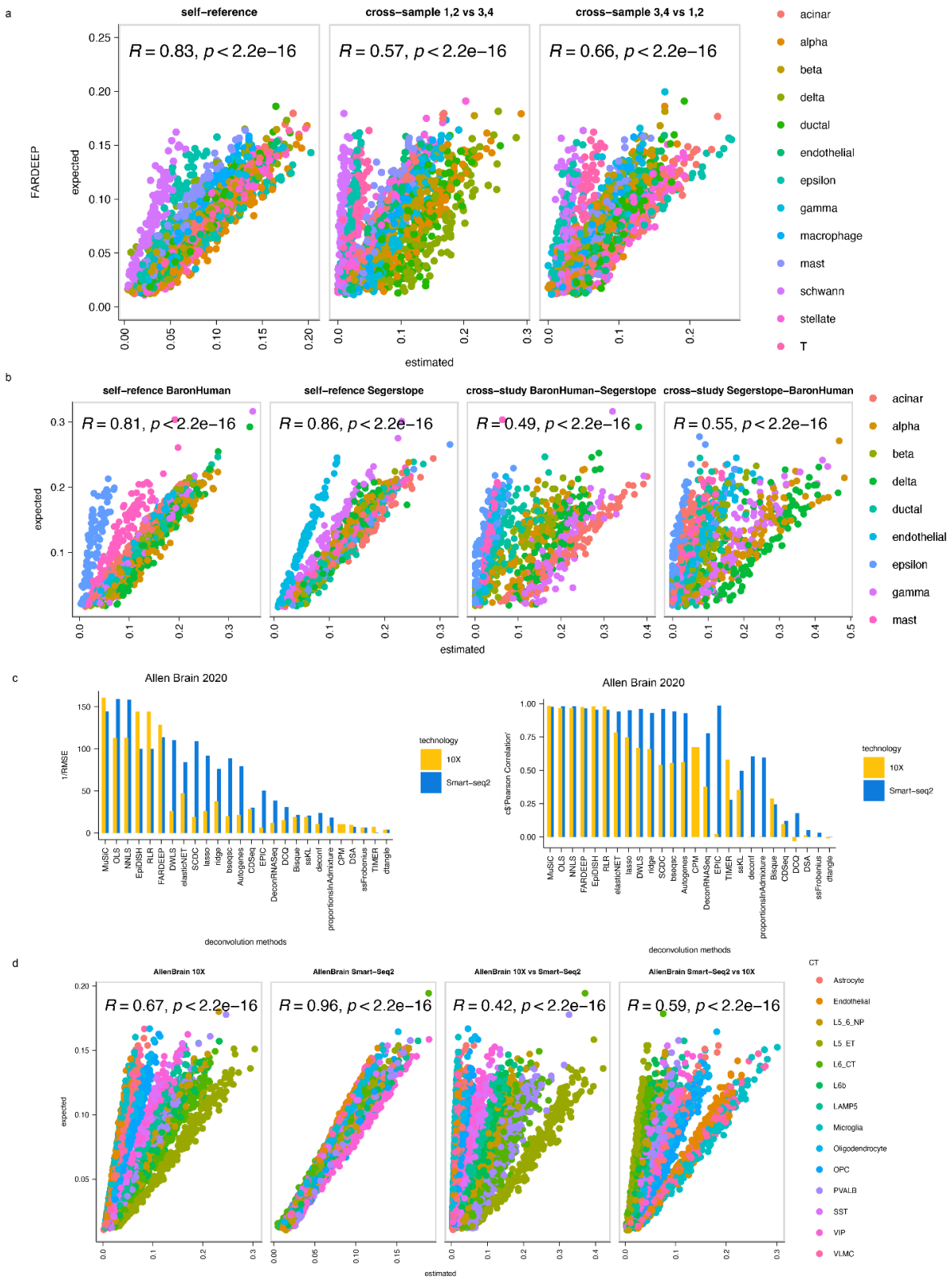

a
self−reference
cross−sample 1,2 vs 3,4
cross−sample 3,4 vs 1,2
R = 0.83, p < 2.2e−16
R = 0.57, p < 2.2e−16
R = 0.66, p < 2.2e−16
FARDEEP
expected
estimated
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
schwann
stellate
T
b
self−refence BaronHuman
self−refence Segerstope
cross−study BaronHuman−Segerstope
cross−study Segerstope−BaronHuman
R = 0.81, p < 2.2e−16
R = 0.86, p < 2.2e−16
R = 0.49, p < 2.2e−16
R = 0.55, p < 2.2e−16
expected
estimated
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
mast
c
Allen Brain 2020
1/RMSE
deconvolution methods
technology
10X
Smart−seq2
Allen Brain 2020
c$'Pearson Correlation'
deconvolution methods
technology
10X
Smart−seq2
d
AllenBrain 10X
AllenBrain Smart−Seq2
AllenBrain 10X vs Smart−Seq2
AllenBrain Smart−Seq2 vs 10X
R = 0.67, p < 2.2e−16
R = 0.96, p < 2.2e−16
R = 0.42, p < 2.2e−16
R = 0.59, p < 2.2e−16
expected
estimated
CT
Astrocyte
Endothelial
L5_6_NP
L5_ET
L6_CT
L6b
LAMP5
Microglia
Oligodendrocyte
OPC
PVALB
SST
VIP
VLMC

Figure 4. Impact of batch effects in deconvolution

a.scatter plots showing results from self-reference and 2 examples of cross-sample deconvolution from the method FARDEEP utilising a single-cell from human pancreas(13 cell-types). In the self-reference task all 4 samples are used while cross-sample tasks use two samples as reference and the other two represent the pseudo-bulk at each example. Pearson correlation(R) and p-value are reported for each task. b.scatter plots showing the results from two self-reference tasks from two pancreatic studies(9 cell-types) as well as two cross-study deconvolution tasks.c. Bar plots showing 1/RMSE and Pearson correlation values from the 26 feasible deconvolution runs from two human brain datasets that come from the same study but different technologies (10X and Smart-seq2). d. Scatter plots demonstrating results from the DWLS method from the two self-reference tasks (10 and Smart-seq2 of human brain tissues, 14 cell-types) and two cross-technology tasks. Pearson correlation(R) and p-values have been calculated for each task and shown on the plots.

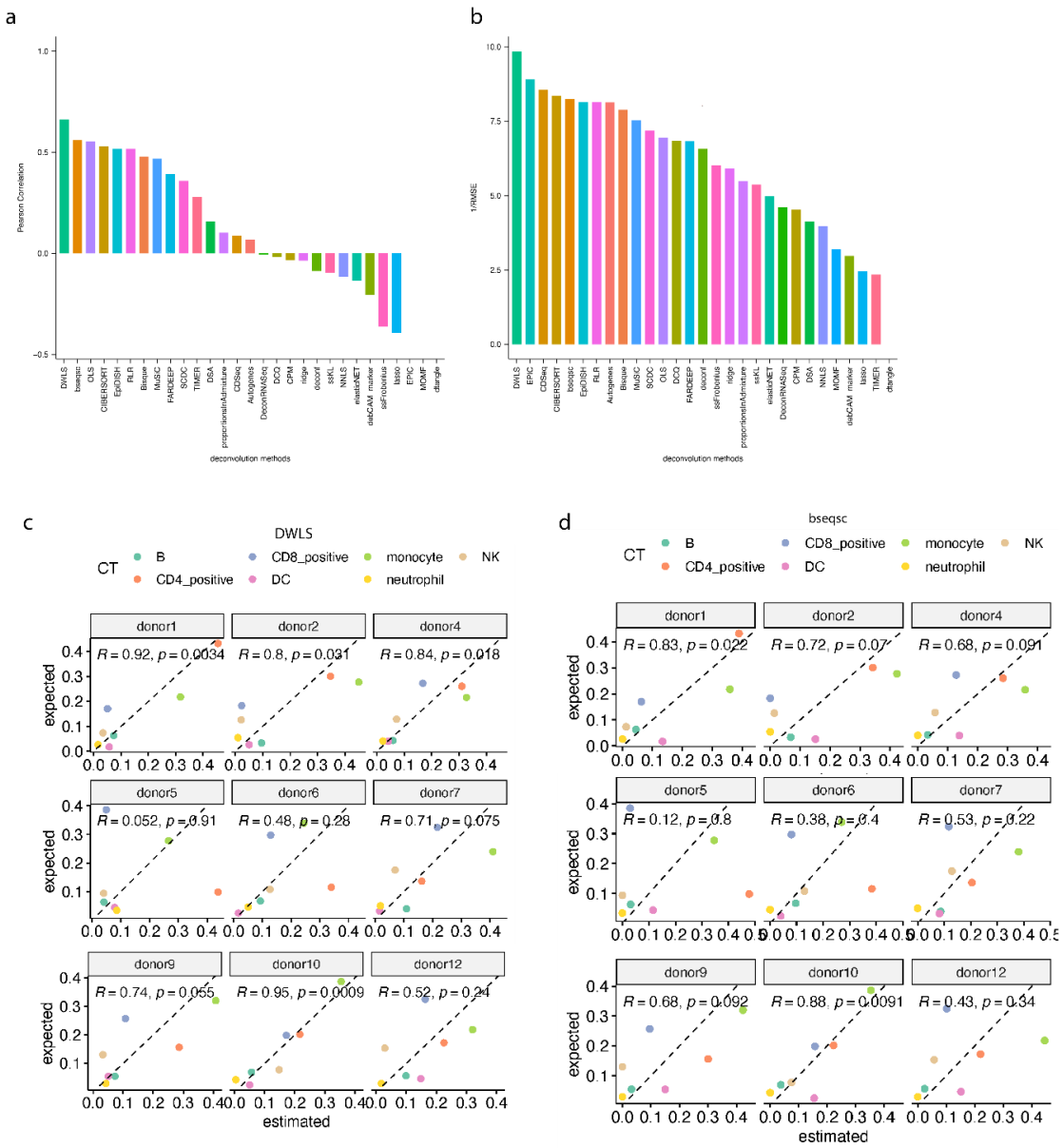
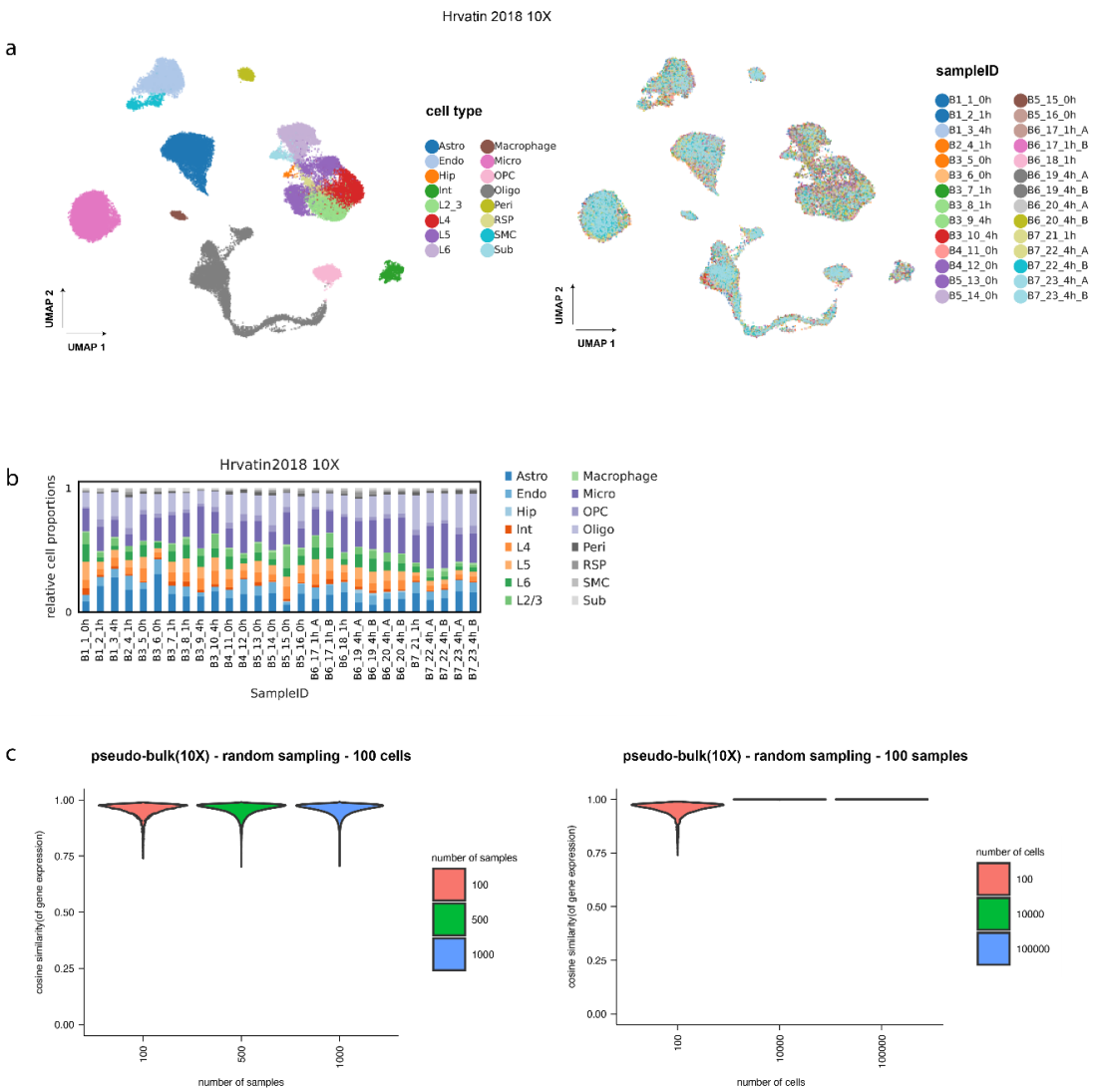


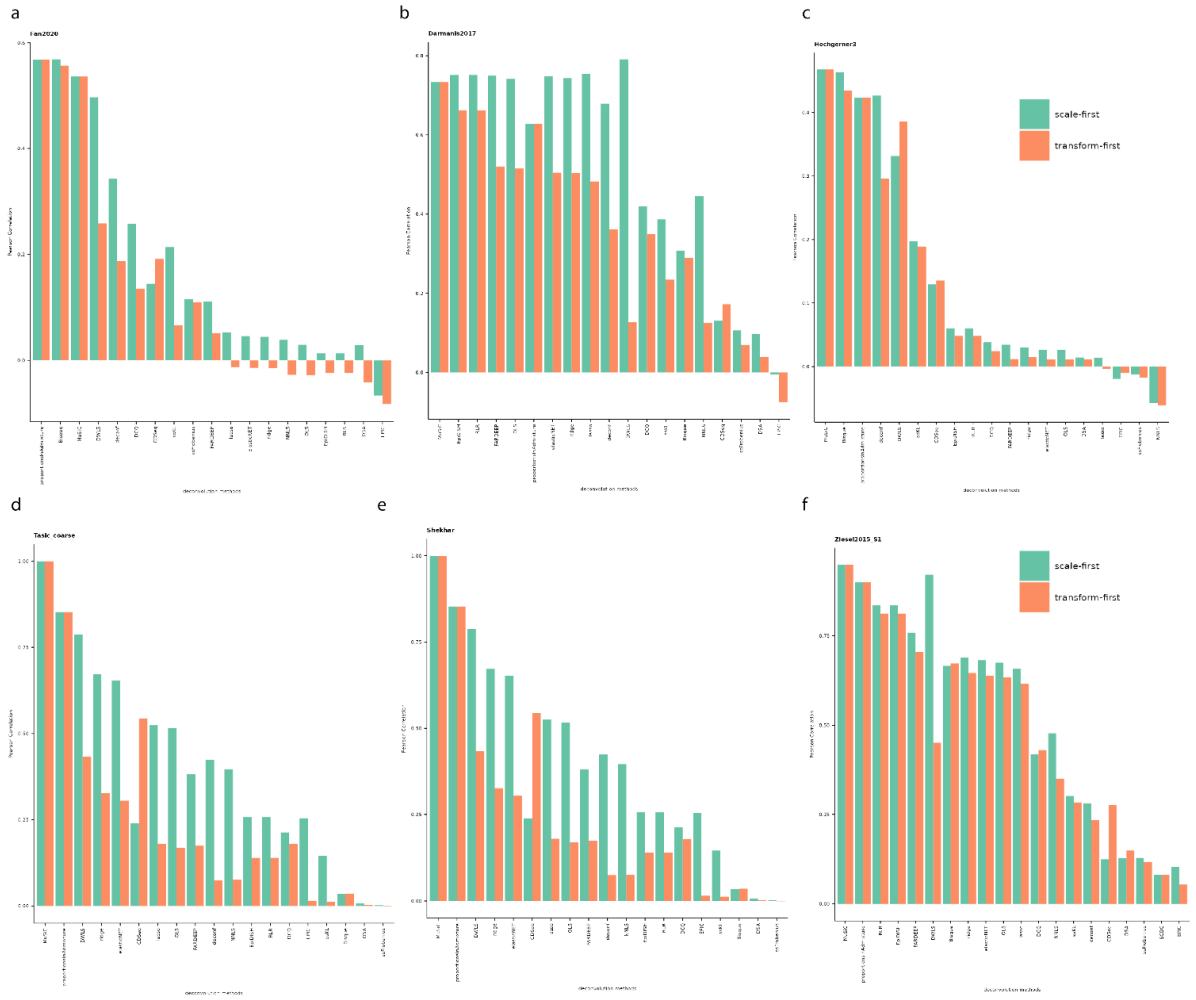
Figure 5. Deconvolution performance of real bulk RNA-seq samples

a.Pearson correlation values from 29 deconvolution methods on real bulk PBMC data with ground truth using a single-cell PBMC dataset as reference(Wilk2020). b. RMSE values from the same task. For methods that reported NA values we show them in the bar plot without existing bars corresponding to them. c. scatter plots showing the 'per sample/donor' deconvolution results(DWLS method) of the real PBMC dataset. Different R values are observed across samples. d. Same results for the bseq-sc deconvolution method.



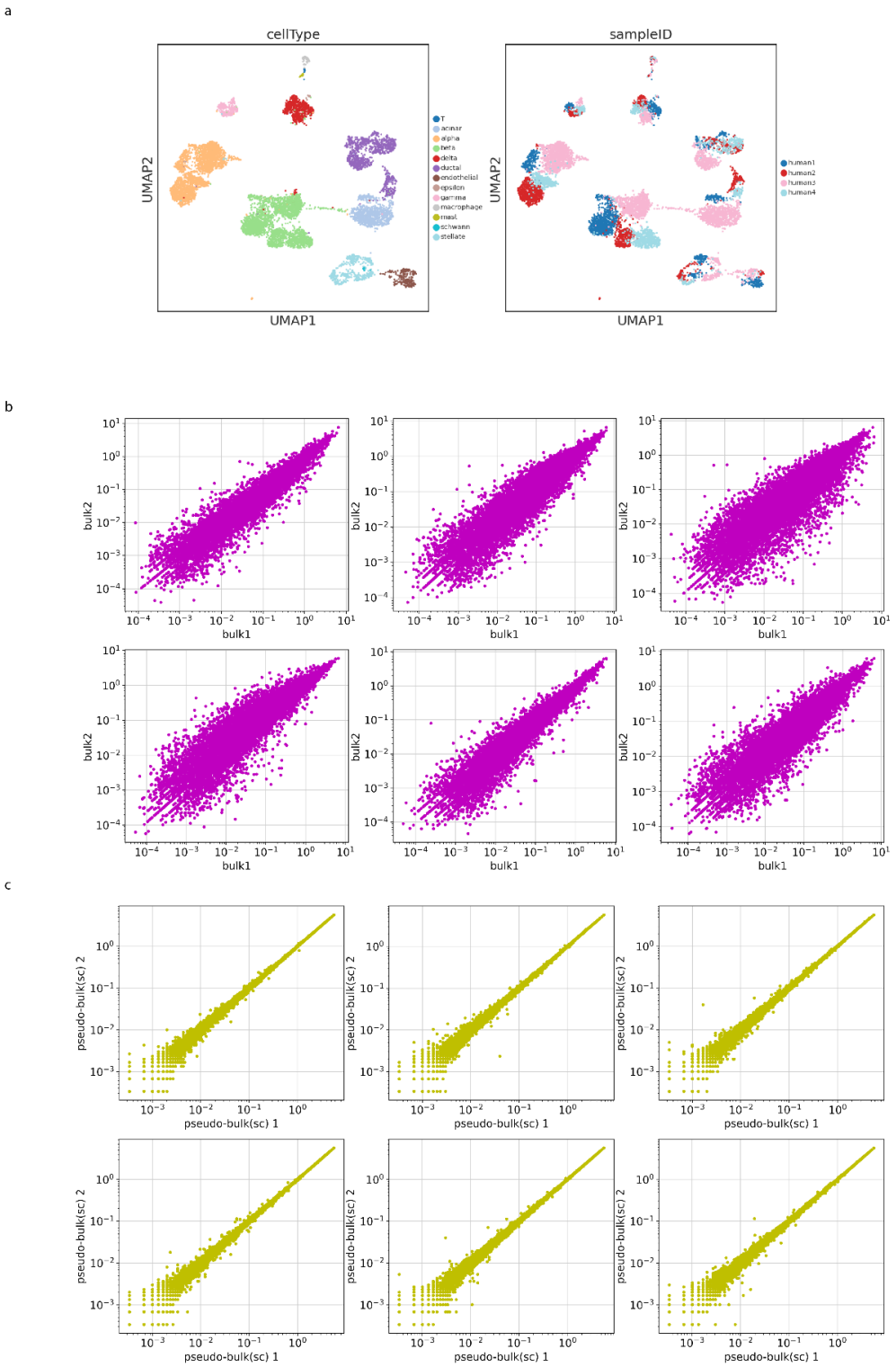
Supplementary Figure 1. Multi-Parametric simulation for pseudo-bulk RNA-seq data deriving from single-cell RNA-Seq data

a.UMAP plot of 10x mouse brain single-cell dataset after filtering and reanalysis. The dataset contains 16 cell-types as shown with different colours (left) and 28 samples. b. bar plot of cell-type proportions for each sample based on the single-cell dataset. c. violin plot of the cosine similarity of gene expression in randomly sampled pseudo-bulks with 100,500 and 1000 samples (left) as well as across different numbers of sampled cells 100, 10K and 100K (right).



Supplementary Figure 2: Preprocessing of simulated data from 6 brain tissue datasets.

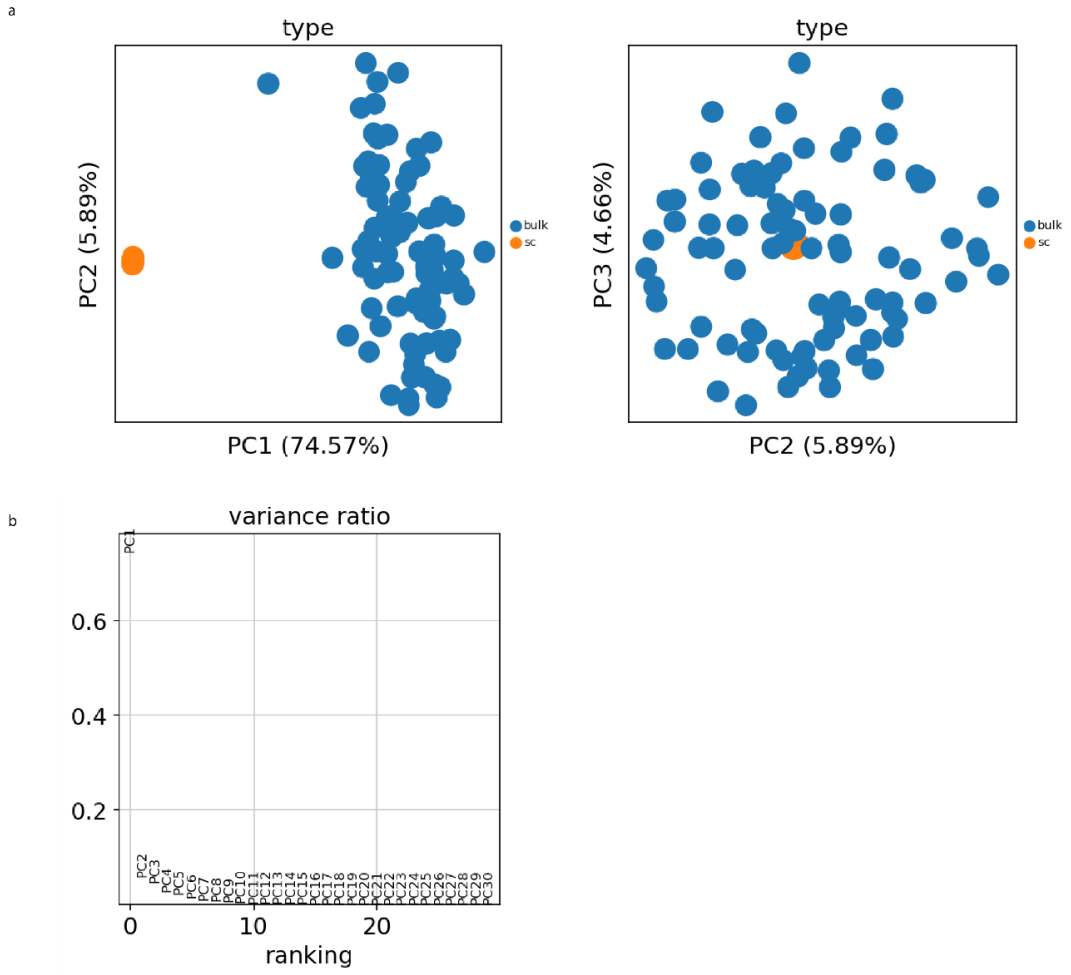
a-e. Barplots showing Pearson Correlation values of 6 self-reference deconvolution tasks in which the effect of scaling or transforming first the input matrices is tested.



Supplementary Figure 3. Comparison of gene expression in pseudo-bulk (sc) and real bulk data.

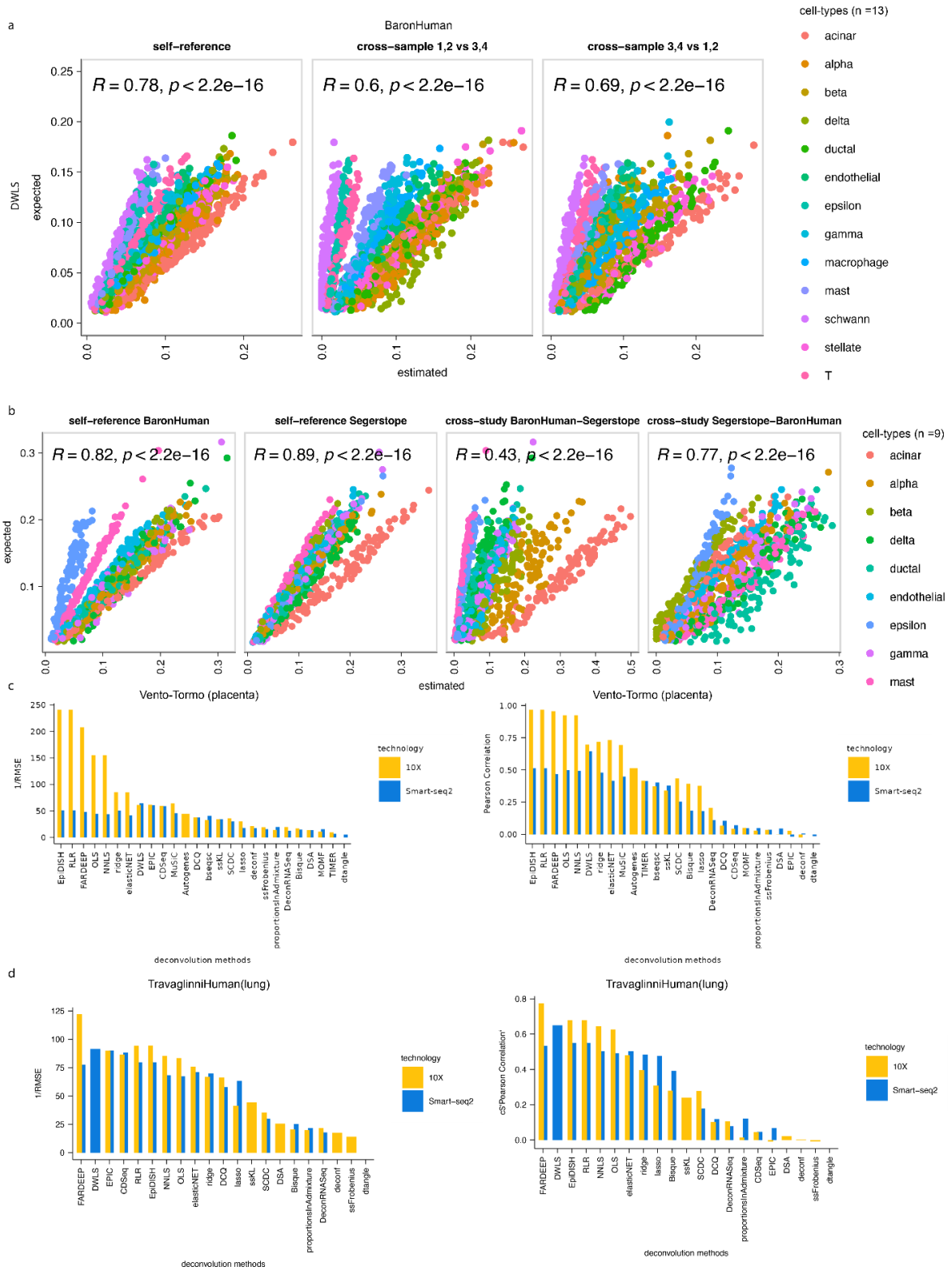
a.UMAP plot of a single-cell dataset (BaronHuman2018) from 4 pancreatic islets after re-analysis, different colours indicate different cell-types based on author's annotations (left) or the different samples (right). b. Scatter plot demonstrating gene expression between 12 different bulk samples

from Fadista2015 pancreatic RNA-seq dataset. c. scatter plot of gene expression deriving from 12 pseudo-bulk profiles generated from the pancreatic single-cell dataset



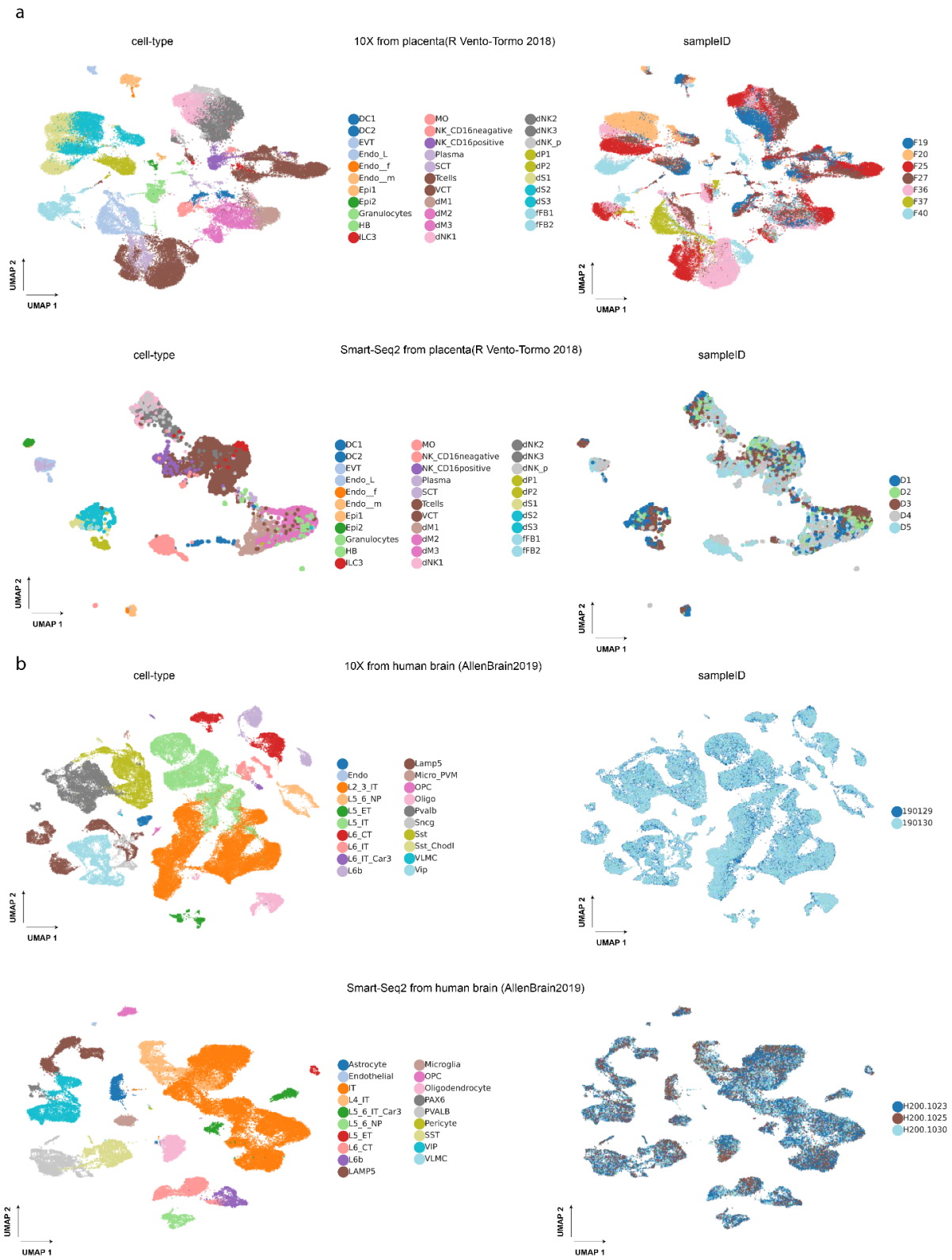
Supplementary Figure 4 : Combined Principal Component Analysis of real bulk and pseudo-bulk samples.

a.PCA results from 78 real bulk samples (Fadista 2014) and 78 pseudo-bulk samples (BaronHuman2018). PC1 and PC2 can explain 80.46% of the variance (left) , PC2 and PC3 shown on the right scatter plot. b. Variance ratio plot, "Explained" variance ratio as a function of the number of principal components. The inclusion of 30 PCs accounts for 98.08% of the variance.


a
BaronHuman
self−reference
cross−sample 1,2 vs 3,4
cross−sample 3,4 vs 1,2
R = 0.78, p < 2.2e−16
R = 0.6, p < 2.2e−16
R = 0.69, p < 2.2e−16
DWLS
expected
estimated
cell-types (n =13)
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
schwann
stellate
T
b
self−reference BaronHuman
self−reference Segerstope
cross−study BaronHuman−Segerstope
cross−study Segerstope−BaronHuman
R = 0.82, p < 2.2e−16
R = 0.89, p < 2.2e−16
R = 0.43, p < 2.2e−16
R = 0.77, p < 2.2e−16
expected
estimated
cell-types (n =9)
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
mast
c
Vento-Tormo (placenta)
1/RMSE
Pearson Correlation
deconvolution methods
technology
10X
Smart-seq2
d
TravaglinniHuman(lung)
1/RMSE
cS*Pearson Correlation
deconvolution methods
technology
10X
Smart−seq2

Supplementary Figure 5: Cross-reference deconvolution of pseudo-bulk samples across different tissues.

a.scatter plots showing results from self-reference and 2 examples of cross-sample deconvolution from the method DWLS utilising a single-cell from human pancreas(13 cell-types). Pearson correlation(R) and p-value are reported for each task. b.scatter plots showing the results from two self-reference tasks from two pancreatic studies(9 cell-types) as well as two cross-study deconvolution tasks. c. Bar plots showing 1/RMSE and Pearson correlation values from the 26 feasible deconvolution runs from two placenta tissue datasets that come from the same study but different technologies (10X and Smart-seq2). d. Bar plots showing 1/RMSE and Pearson correlation values from the 26 feasible deconvolution runs from two lung tissue datasets that come from the same study but different technologies (10X and Smart-seq2)


a
cell-type
10X from placenta(R Vento-Tormo 2018)
sampleID
DC1
DC2
EVT
Endo_L
Endo__f
Endo__m
Epi1
Epi2
Granulocytes
HB
ILC3
MO
NK_CD16neagative
NK_CD16positive
Plasma
SCT
Tcells
VCT
dM1
dM2
dM3
dNK1
dNK2
dNK3
dNK_p
dP1
dP2
dS1
dS2
dS3
fFB1
fFB2
F19
F20
F25
F27
F36
F37
F40
UMAP 2
UMAP 1
cell-type
Smart-Seq2 from placenta(R Vento-Tormo 2018)
sampleID
D1
D2
D3
D4
D5
b
cell-type
10X from human brain (AllenBrain2019)
sampleID
Endo
L2_3_IT
L5_6_NP
L5_ET
L5_IT
L6_CT
L6_IT
L6_IT_Car3
L6b
Lamp5
Micro_PVM
OPC
Oligo
Pvalb
Sncg
Sst
Sst_Chodl
VLMC
Vip
190129
190130
Smart-Seq2 from human brain (AllenBrain2019)
Astrocyte
Endothelial
IT
L4_IT
L5_6_IT_Car3
L5_6_NP
L5_ET
L6_CT
L6b
LAMP5
Microglia
OPC
Oligodendrocyte
PAX6
PVALB
Pericyte
SST
VIP
VLMC
H200.1023
H200.1025
H200.1030

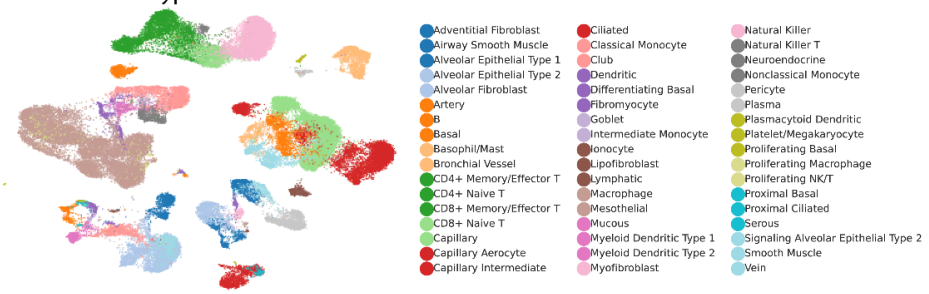
Supplementary Figure 6 : Single-cell references from human placenta and brain for cross-technology deconvolution

a.UMAP plots of two single-cell dataset(10X and Smart-Seq2) from human placenta after re-analysis, different colours indicate different cell-types based on author's annotations (left) or the different samples (right). b. UMAP plots of two single-cell dataset(10X and Smart-Seq2) from human brain after re-analysis, different colours indicate different cell-types based on author's annotations (left) or the different samples (right).

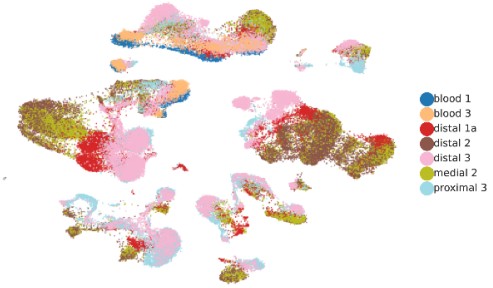
a

KJ Travaglini 2020 (10X)

cell type



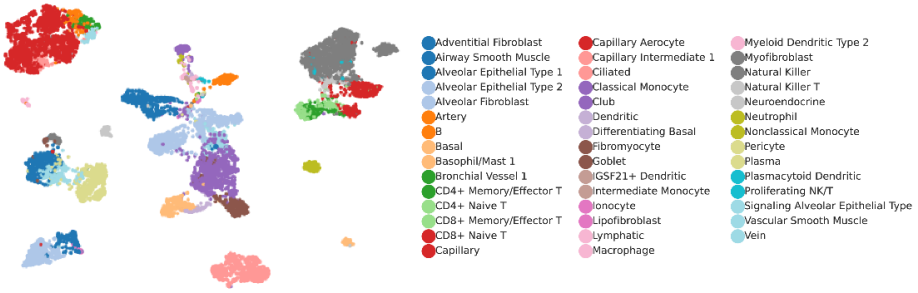
sampleID



b

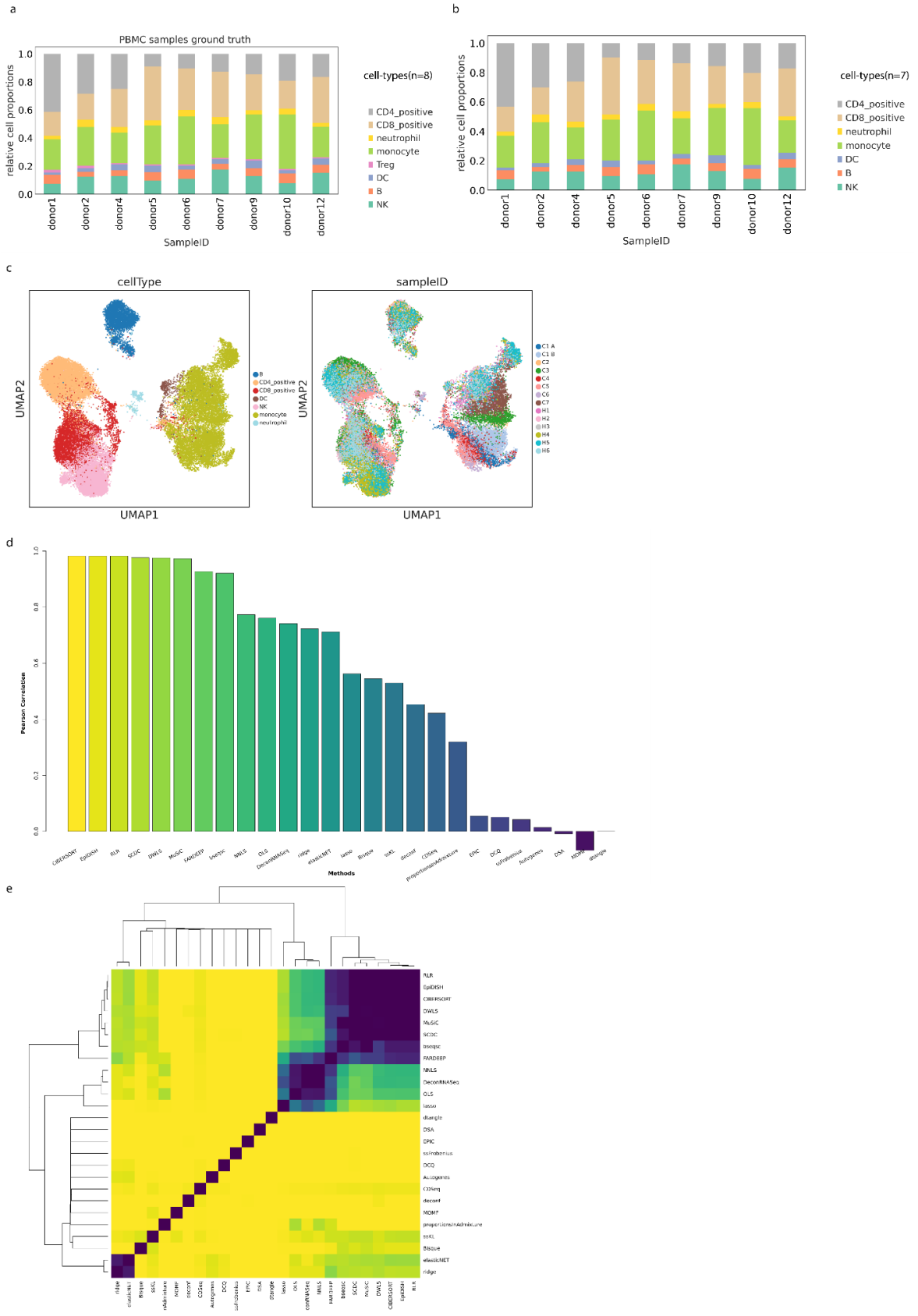
KJ Travaglini 2020 - SS2 analysis of plate-sorted cells

cell_label



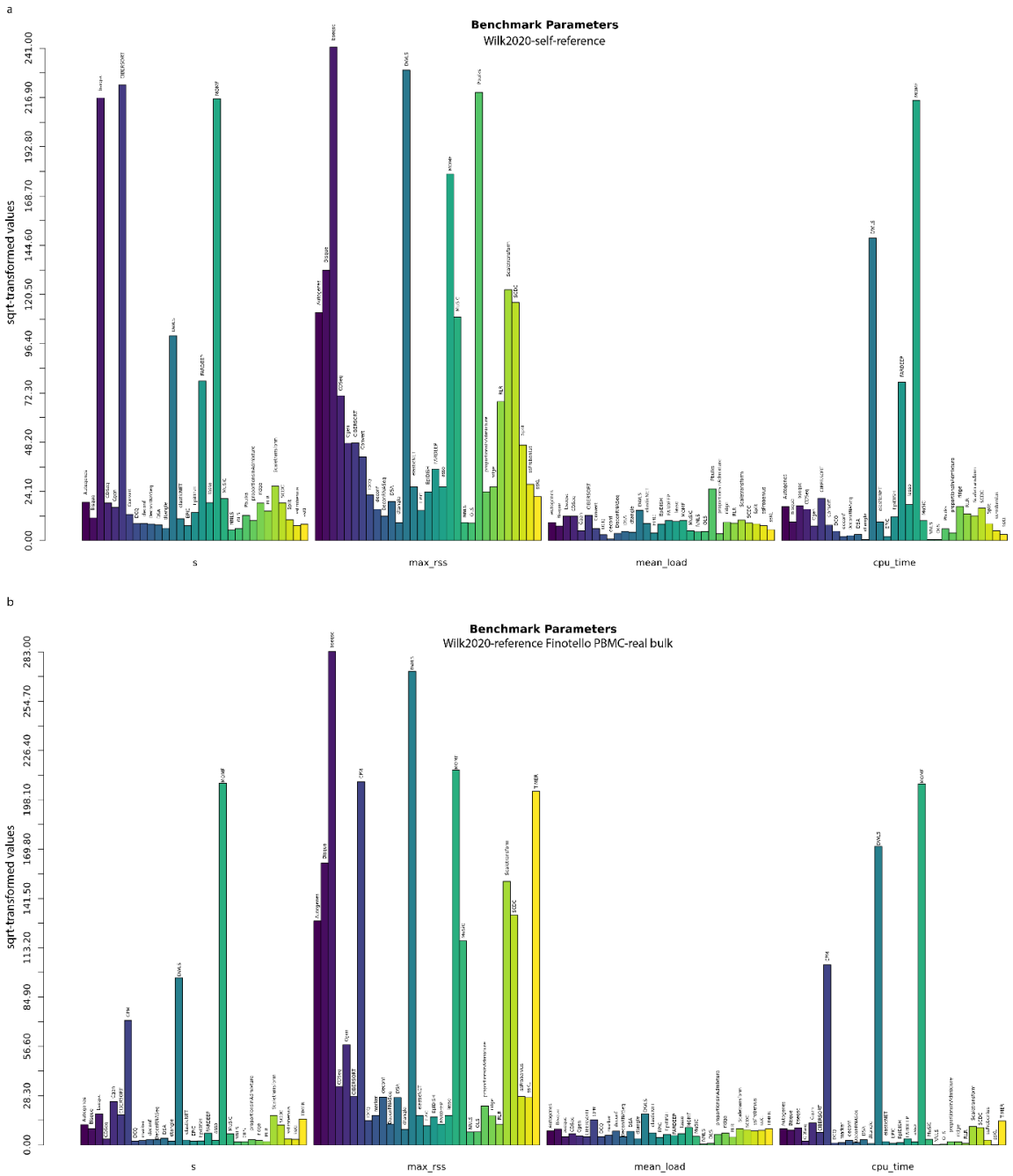
sampleID

Supplementary Figure 7 : Single-cell references from human lung tissue.
a.UMAP plots of a single-cell 10X dataset from human lung after re-analysis, different colours indicate different cell-types based on author's annotations (top) or the different samples (bottom). b. UMAP plots of a single-cell Smart-seq2 dataset from human lung after re-analysis, different colours indicate different cell-types based on author's annotations (top) or the different samples (bottom).


a
PBMC samples ground truth
relative cell proportions
SampleID
cell-types(n=8)
CD4_positive
CD8_positive
neutrophil
monocyte
Treg
DC
B
NK
b
relative cell proportions
SampleID
cell-types(n=7)
CD4_positive
CD8_positive
neutrophil
monocyte
DC
B
NK
c
cellType
sampleID
UMAP1
UMAP2
d
Pearson Correlation
Methods
e

Supplementary Figure 8. Bulk and single-cell PBMC data for real data deconvolution.

a. Bar plot showing initial ground truth proportions of 9 PMBC samples measured by flow cytometry. b. Bar plot with ground truth proportions after summing up cell-type proportions of CD4+ T cells and T-regs. c. Umap plots of a single-cell PBMC dataset from Wilk2020 after re-analysis. Cell-type and sampleIDs are shown in the two UMAPs respectively. d. Results from self-reference deconvolution task using Wilk2020 single-cell RNA-Seq dataset. e. Pairwise correlation approach to compare results across methods.



Supplementary Figure 9. Evaluation of scalability on simulated and real PBMC data.

a-b. for scalability assessment purposes, time (s), memory (maximum resident set size), mean load (CPU usage divided by total time) and CPU time are visualised in barplot for Wilk2020-self-reference task as well as for the deconvolution of Finotello PBMC real bulk samples.