CATD: A reproducible pipeline for selecting cell-type deconvolution methods across tissues

AUTHORS

Anna Vathrakokoili Pournara[1*], Zhichao Miao[1,2*], Ozgur Yilmaz Beker[1,#], Nadja Nolte[1,2],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

* To whom correspondence should be addressed. Tel: +4407522005326; Fax:- ; Email: annavp@ebi.ac.uk

Correspondence may also be addressed to Zhichao Miao(miao_zhichao@gzlab.ac.cn) and Irene Papatheodorou(irenep@ebi.ac.uk)

Present Address:

Zhichao Miao : GMU-GIBH Joint School of Life Sciences, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou, China.

Ozgur Beker: Faculty of Engineering and Natural Sciences, Sabanci University, Tuzla 34956, Istanbul, Turkey

Nadja Nolte: Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia

ABSTRACT

Cell-type deconvolution methods aim to infer cell composition from bulk transcriptomic data. The proliferation of developed methods, coupled with the inconsistent results obtained in many cases, highlights the pressing need for guidance in the selection of appropriate methods. The growing accessibility of systematic single-cell RNA sequencing datasets, often accompanied by bulk expression from related samples, makes it possible to benchmark the existing methods more objectively. Here, we propose a comprehensive assessment of 29 deconvolution methods, leveraging single-cell RNA-sequencing data from different tissues. We evaluate deconvolution across a wide range of simulation scenarios and we show that single-cell regression-based deconvolution methods perform well while their performance is highly dependent on the reference.We also study the impact of bulk-reference differences, including those associated with sample, study, and technology. We provide validation using a gold standard dataset from mononuclear cells and suggest a consensus prediction of

proportions when ground truth is not available. We f validated the consensus method on data from the stomach and studied its spillover effect. Lastly, we suggest that the Critical Assessment of Transcriptomic Deconvolution (CATD) pipeline can be employed for simultaneous deconvolution of hundreds of bulk samples and we envision it to be used for speeding up the evaluation of newly developed methods.

Keywords: deconvolution, single-cell RNA-seq, benchmarking, cell proportions, pseudo-bulk, real bulk.

Key Points:

- Thorough assessment of 29 deconvolution methods, leveraging diverse single-cell RNA-sequencing data from various tissues, alongside extensive simulations and validation against known ground truth data.
- Emphasis on the pivotal role of reference selection, tissue type, and technological nuances in determining the efficacy of deconvolution methods.
- Introduction of the user-friendly and robust Critical Assessment of Transcriptomic Deconvolution (CATD) Snakemake pipeline, enabling efficient and reproducible cell-type deconvolution in real bulk RNA-Seq datasets.

Abbreviations:

sc-RNA-seq: single-cell RNA-Sequencing
PBMC: Polymorphonuclear/Peripheral Blood Mononuclear cell
PCA: Principal Component Analysis
GEO: Gene Expression Omnibus
GTEX: Genotype-Tissue Expression Project
TCGA: The Cancer Genome Atlas Program
TARGET: Therapeutically Applicable Research to Generate Effective Treatments

INTRODUCTION

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 [1–4]. High infiltration from specific immune cells can be predictive of the disease outcome, stage and aid the treatment selection[5]. 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 true proportions of cell types observed could be masked due to technical handling, sorting of specific cell types, sub-optimal dissociation, and tissue vulnerability[6]. 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[7,8]. Additionally, there are only a handful of this type of datasets available, mostly from non-solid tissues such as PBMCs[9,10].

Nevertheless, computational deconvolution methods show a great advantage, since cell proportions from hundreds of samples can be obtained computationally very fast; in a matter of seconds to a few hours depending on the selected method. As a result, cell-type deconvolution methods - also called decomposition methods- provide a cost-effective way of measuring cell proportions when compared to experimental methods. Moreover, deconvolution methods can be very useful for inferring cell type compositions from historical bulk RNA-seq data from large databases such as GTEx[11], TCGA[12], TARGET (https://ocg.cancer.gov/programs/target) and GEO[13] in which it would be impossible to repeat experiments on. In mathematical terms, deconvolution can be understood as a linear-based technique. It takes bulk gene expression as input and aims to infer a proportions matrix that represents the relative abundance of different cell types within a tissue. Some deconvolution methods employ a reference-based approach, using as an additional input single-cell profiles, cell type signatures or marker genes, and applying linear methods to compute the mixing proportions. Alternatively, reference-free (complete) deconvolution methods use matrix factorization to uncover underlying patterns in gene expression allowing for estimation of cell-type proportions without prior reference information.

Deconvolution methods can be formally categorised into three main groups based on their input requirements: supervised, semi-supervised, and unsupervised. Supervised deconvolution methods necessitate an expression matrix (typically from bulk RNA-seq or microarray data) to be deconvolved as well as cell-type-specific information. This specific information can be derived from sources like flow cytometry, single-cell RNA-seq data, or marker gene signatures obtained from those datasets. These supervised methods can further be divided into two subcategories: bulk and single-cell, based on the type of reference data provided. Well-known "bulk" deconvolution methods in this category include CIBERSORT(absolute)[14], FARDEEP[15], NNLS (non-negative least squares) and EpiDISH[16]. Concurrently, there are contemporaneous deconvolution methods that harness available single-cell datasets to extract cell-type-specific features for deconvolution. Examples of such methods include Bseq-sc[17], DWLS[18] and MuSiC[19] and others [20–22]. Some of these approaches incorporate feature selection methods to choose informative features for deconvolution, such as DWLS-MAST, MuSIC-informative genes, and AutogeneS. All the above methods output cell proportions, enabling both intra-sample and inter-sample comparisons. In the supervised category, gene set enrichment methods like xCell[23], MCP-counter[24], and SaVant[25] can be included. However, these methods use an enrichment-based approach and output enrichment scores that can only be compared within samples, which limits their application for inter-sample comparisons. On the other hand, unsupervised methods like deconf[26] and CDSeq[27] only require a bulk sample and the number of components (representing the cell types to be identified) as input. These methods estimate both the cell-type-specific expression and cell fractions simultaneously. Semi-supervised methods, such as ssKL and ssFrobenius, can also estimate both the cell-specific expression matrix and the matrix of cell proportions simultaneously. They utilise a list of pre-determined markers, the expected number of components (cell types), and the bulk matrix as input. These semi-supervised methods employ a block-descent approach to estimate both matrices[28]. In this study, our primary focus is on comparing supervised and semi-supervised methods, and we do not extensively explore complete deconvolution methods. Complete deconvolution methods are particularly valuable when the cell composition is entirely unknown or no cell signatures are available for specific cell types. These methods have been thoroughly investigated in previous benchmarks to assess their potential and performance [29–31]. Lastly, there is a separate class of methods known as "expression deconvolution methods". These methods tackle a different problem; they aim to find the cell-type-specific expression matrix given the knowledge of cell proportions and bulk expression data. Methods falling under this category include Rodeo[31], csSAM[32], Deblender[33], cs-lsfit[28], and cs-prog[34]. Previous benchmark

studies have also evaluated the performance of these methods, but due to their distinct design and research focus, we have not included them in this study.

Cell-type deconvolution can contribute significantly in answering critical biological questions and understanding disease mechanisms to a greater extent. Specifically, it can unravel how cell proportions can affect certain phenotypes or specific clinical features available within genomic databases[35]. Additionally it can serve as a powerful tool for mitigating the confounding effect of cell-type proportions in bulk differential expression analysis and as a result help in the identification of reliable biomarkers for diseases and perturbed processes and pathways with cell-type resolution[36]. Moreover, in the context of Transcriptome-wide association studies(TWAS) deconvolution can rectify cell abundance bias in the eQTL analysis and facilitate the identification of cell-type specific eQTLs[37]. Finally, the abundance of bulk RNA-seq samples present a unique opportunity for training machine learning models tailored to phenotype prediction, a task that is often challenging with single-cell data due to limited sample sizes. Importantly, by integrating cell composition, we effectively enrich the information content that bulk samples are missing due to the experimental method principles.

The field of deconvolution methods has witnessed significant growth in recent years, driven by the complex nature of the deconvolution problem and the ongoing development of new and more sophisticated techniques[38]. Over the past decade, there have been notable efforts to independently evaluate deconvolution methods[29,39]. These efforts explore various aspects, such as the effectiveness of simulations, method performance, the importance of reference data, the impact of data normalisation, the influence of different technologies, and the granularity of cell-type information in the data[16,30,40–45]. However, as our knowledge in this field continues to expand, and new methods are introduced each year, there is a growing need for comprehensive, reproducible, and versatile benchmarking pipelines that can accommodate both existing and newly developed methods. This need is especially pressing because different deconvolution methods may have varying input formats, design methodologies, preprocessing steps, and output results. Therefore, a centralised and efficient platform is essential to streamline the evaluation process, facilitating a quicker and more unified assessment of deconvolution methods. Furthermore, while previous studies have primarily focused on deconvolution in peripheral blood mononuclear cells (PBMC) samples or in a limited number of tissues each time, there is a growing demand for research that covers more tissues and datasets as more and more cell type specific information becomes available.

To address these challenges, we have designed a robust and reproducible Snakemake pipeline for evaluating 29 available deconvolution methods and obtaining results on cell abundances from real data. Our evaluation involves 19 single-cell datasets from diverse sources, covering five different tissues with multiple datasets per tissue, enabling us to generate a wide range of simulated and reference data for method evaluation. Our pipeline has incorporated various simulation approaches, leveraging single-cell datasets and suggesting parameter settings to generate realistic simulated data across different tissue types. Additionally, we study the impact of preprocessing methods and the effect of distinct technologies in the reference data in different tissues. Deconvolution performance is then evaluated on real data from PBMC samples with ground truth. Moreover, we introduced a consensus deconvolution approach by

combining well-performing methods and validated the results on stomach samples using as ground truth tissue slides. Finally, we assess the spillover of the consensus method on purified samples from cerebral cortex and lung tissues.

A notable feature of our pipeline is its adaptability to newly developed deconvolution methods that require independent benchmarking. Furthermore, researchers can employ our pipeline to simultaneously deconvolve multiple samples, obtaining results from all implemented methods and benefiting from the visualisation of the results. By assessing the consistency of results across methods and references, researchers can then choose the most suitable setup to address their specific research questions.

Glossary of Key Terms

Cross-reference deconvolution: Supervised deconvolution task where the reference and the pseudo-bulk come from different samples, datasets, or technologies.

Deconvolution :
Cell type deconvolution is a computational technique that aims to infer the relative proportions or abundances of different cell types within a heterogeneous biological sample, such as a tissue or a mixture of cells. This approach leverages gene expression or other molecular data (such as methylation) from a sample to estimate the contributions of individual cell types that make up the complex mixture.

Self-reference deconvolution: Supervised deconvolution task where the reference and the pseudo-bulk come from the same dataset after splitting it into training (reference) and testing (pseudo-bulk).

Snakemake workflow manager:
Snakemake is a workflow management system and domain-specific language used that automates and manages the execution of complex, multi-step data analysis pipelines. It allows users to define the steps, inputs, outputs, and dependencies of a workflow in a human-readable format. Then it automatically manages the execution of these steps, ensuring that they are executed in the correct order while taking advantage of available computational resources efficiently. This tool is particularly valuable in bioinformatics research where reproducibility, scalability, and ease of use are essential.

MATERIAL AND METHODS

Data Collection:

Single-cell data collection for in silico experiments: We collected single-cell data from various sources in order to create references and simulated bulk for our experiments. 12 single-cell datasets were collected from human and mouse brain tissues, including dental gyrus, anterolateral motor cortex (ALM), primary visual cortex (ViSP),cerebral cortex and pons, retinas, hippocampus CA1, somatosensory cortex S1, primary motor cortex, cortex areas and glioblastoma cancer tissue. We also collected two datasets from placenta tissue from the same study but from different sequencing platforms (10X and SS2), two single-cell datasets from

pancreatic islets and two 'lung+blood' tissue datasets from the same study but different technologies(10X and SS2). More details regarding the number of samples, cells, genes and cell-types per dataset as well as the GEO accession numbers of the data can be found in Table S1.

Validation on Real bulk data using single-cell references:

Experiment 1 : We collected RNA-seq data from 9 PBMC samples[46] coupled with flow cytometry measurements [GSE107572]. To deconvolve the above PBMC samples we used a Seq-Well single-cell reference[GSE150728][47] and matched the cell-types between the two datasets.

Experiment 2 : We acquired TPM-normalised bulk expression data from GTEx v8 portal (https://storage.cloud.google.com/adult-gtex/bulk-gex/v8/rna-seq/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz) . We focused on 339 stomach bulk samples, ensuring they had associated histology slides from a pool of 939 stomach samples. For this deconvolution task we used as a reference the stomach subset of the Human Cell Landscapes sRNA-seq dataset by Han et al. in 2020, obtained from cellxgene website (https://cellxgene.cziscience.com/collections/38833785-fac5-48fd-944a-0f62a4c23ed1) For standardised reduction of cell type label granularity we employed an automated ontology-based method with the scOntoMatch (v 0.1.0) package[48]. The adapted code for this task is available at: https://github.com/nadjano/reference_preperation_for_GTEX_deconvolution/blob/main/scripts/reduceCellTypes.R. For selected stomach samples histology images were downloaded from GTEx Portal Histology Viewer (https://gtexportal.org/home/histologyPage).

Flow cytometry data and single-reference collection for validation:

We acquired bulk expression data of FACS-sorted human fetal lung cells, distinguishing between endothelial and non-endothelial populations (E-MTAB-9372 from BioStudies). As a lung specific deconvolution reference, we used a subset from the GTEx v7 single-cell dataset. Cell type labels granularity was reduced as described in the previous section.

Furthermore, we performed deconvolution on mouse brain data obtained from the cerebral cortex (E-GEOD-52564), which contains purified single cell types. As a single cell reference the cerebral cortex in the Tabula Muris dataset (E-ENAD-15) was utilised.
In our study, we downsampled the single-cell references to expedite the deconvolution computation. Specifically, for cell types that possessed more than 300 cells, we employed a random sampling approach, selecting 300 cells per cell type. The detailed code implementation for this procedure can be found in the following GitHub repository: [https://github.com/nadjano/reference_preperation_for_GTEX_deconvolution/blob/main/scripts/split_into_tissues.R].

Reanalysis of single-cell datasets:

Raw counts and metadata [minimum metadata required: cell-type annotations, cellIDs, sampleIDs] were obtained for each single-cell dataset used for the in silico experiments.

Single-cell datasets used for reference generation were re-analysed utilising the SCANPY toolkit[49]. Cells with less than 200 genes and genes expressed in less than 3 cells were filtered out. Cells with more than >5% of mitochondrial genes or too many total counts[dataset-specific threshold] were also removed. Datasets were normalised (10e4) and log-transformed, the raw count dimension was kept intact to be used as a deconvolution input later. Next, highly variable genes were identified and we performed PCA (Principal Component Analysis), computed the neighbourhood graph as well as performed leiden clustering using different resolution parameters(0.25, 0.5, 1). After clustering the data we plotted the data on UMAP space using as colour key : cellType and leiden clusters in order to evaluate the author's annotations. Cell types classified as 'others', 'unknown', 'low quality', 'unclassified' were removed from the data. Since there were differences between the cell type annotations across datasets, we re-named cell types in each tissue in order to have harmonised annotations, particularly useful for the cross-reference and real bulk deconvolution tasks.

CATD Snakemake Pipeline Overview:

The CATD pipeline is a benchmarking pipeline meant to facilitate the assessment of cell-type deconvolution methods (currently 29) across different simulation scenarios in a standardised way. It also allows the deconvolution of real bulk samples with many user-input parameters so that users can deconvolute their own in-house data following our proposed guidelines. The pipeline includes:

- Pseudo-bulk generation methods(from single-cell RNA-seq data) that allow to create diverse bulk samples and compare deconvolution methods across different scenarios.
- 17 normalisation methods implemented in the pipeline for the normalisation of the input single-cell reference and the (pseudo)bulk samples
- 4 transformation methods implemented
- 9 DE tests for the selection of marker genes from single-cell reference data(Seurat, FindMarkers)
- 29 deconvolution methods
- 7 metrics to assess the results when we test deconvolution methods on pseudo-bulks or when ground truth proportions from real data are available.
- Visualisation of performance and scalability metrics across methods, deconvolution results interpretation(cell type proportions barplots) from the consensus approach when we perform real data deconvolution
- A Consensus approach, combining three well-performing deconvolution methods
- Visualisation of deconvolution results(cell type proportions barplots) from the consensus approach when we perform real data deconvolution
- Visualisation of the correlation of the results across methods for a given dataset when ground truth is not available.

Code availability

CATD Snakemake pipeline: https://github.com/Papatheodorou-Group/CATD_snakemake

Self-reference deconvolution pipeline.

Description: For this task we input a single-cell dataset in Seurat or Anndata object format. The dataset is split in two, the training dataset that will be used as a reference and a test dataset that will be used to generate simulated bulk RNA-seq data. The first step of the pipeline includes converting the data from anndata format to seurat object using the sceasy package. Next, the dataset is split in half, the first half will be used for reference generation and the second half for pseudo-bulk generation. In this step the seed can be specified for reproducibility. We also make sure that all the cell types in the single-cell dataset have been included in both halves. The next step of the pipeline happens in parallel and includes the generation of suitable references -that will be later used as inputs in deconvolution methods- and simulated bulk data(see details in the methods below). Once the references and pseudo-bulk files have been generated normalisation and transformation of pseudo-bulk and reference matrices is performed where it is relevant. Next, the relevant generated inputs are fed in the deconvolution methods and cell-type predictions are obtained. The first output of the pipeline is a matrix of cell-type proportions per sample for each deconvolution method which is saved in RDS format under the directory 'Output' . Since in self-reference deconvolution ground truth proportions are known (determined while generating pseudo-bulk profiles) evaluation metrics are calculated across samples for each method. The output of the evaluation is a RDS-formatted file for each metric that includes a list of values of this metric for each deconvolution method. These files are stored under the directory 'Metrics'. Finally, an evaluation of the scalability is performed for each method, the table of values is stored under 'Benchmarks'. The values from the metrics and benchmarks file are used for visualisation of the performance and comparison of scalability across methods. Violin plots and bar charts are used to visualise the above. Resulting plots are then stored under the directory 'Plots'. (Supplementary Figure 11)

Cross-reference deconvolution pipeline.

Description: The main steps of the cross-reference pipeline are the same as described in the previous section. However, in this task we need two input single-cell datasets instead of one. The first dataset will be used for the generation of pseudo-bulk(and the ground truth) and the second for the generation of the references. For this task it is required that both datasets originate from the same tissue and hence, they include the same cell-types. In addition, the cell-type labels should be matched to allow evaluation of the results. The primary objective of the cross-reference pipeline is to facilitate the implementation of controlled in silico experiments,through which we can investigate how differences in gene expression between (pseudo)bulk and reference , stemming from sample differences, technologies, or studies, can influence the deconvolution process. Unlike self-reference deconvolution this pipeline introduces more noise and simulates a more realistic deconvolution scenario.(Supplementary Figure 12)

Real bulk deconvolution pipeline.

Description: This part of the pipeline was designed for the real bulk deconvolution as well as for future users to deconvolute their own data across the implemented methods. This pipeline is split into two smaller pipelines depending on the presence or absence of ground truth proportions. When ground truth data are present(indicated in the config.yaml file as ***realBulk: 1***) the users need to provide (a) bulk expression matrix to be deconvolved (b) a matrix of known ground truth proportions (c) single-cell data in anndata or Seurat object format. Moreover the users can decide regarding the preprocessing parameters(normalisation transformation of both the bulk and single-cell matrix), which statistical test will be used for marker selection, which deconvolution methods will run, as well as if a consensus of EpiDISH,FARDEEP and DWLS will be calculated. The outputs of the pipeline will be the same as described in the self-reference deconvolution section. In cases where there is no knowledge regarding the ground truth proportions (parameter ***realBulk-noprop : 1***) users can still run the pipeline by providing only the bulk and the single-cell reference of their choice. Since in this case evaluation metrics cannot be calculated for each method instead we perform correlation of results across methods to inform the user regarding the concordance of methods in predicting proportions.(Supplementary Figure 13)

Pseudo-bulk generation design

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 mode equals 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 equal to 2, the generation will require two additional parameters, namely 'proportional variance' (***p***) and second, 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 one.

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 references

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, this matrix is used as an input in single-cell deconvolution methods(e.g MuSiC,DWLS,bisque etc). ***C1*** is the averaged (arithmetic mean) gene expression matrix by cell types, with rows being genes and columns being cell types.The C1 matrix is used in 'bulk' deconvolution methods that require a matrix similar to FACs data. ***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. This reference is used in marker-based methods or methods that can include additional marker information (e.g EPIC, DSA, ssKL, ssFrobenius, dtangle). ***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***. refVar is used only by EPIC deconvolution methods due to its unique design.

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(transformation first or normalisation first through the configuration file.

Selection of markers genes

For the marker-based methods and the methods that can use markers as an additional input, we need to first identify the marker genes from our single-cell references. To do that we apply the `FindAllMarkers` function from the Seurat package along with the LogNormalise parameter. The `FindAllMarkers` function from Seurat has implemented different statistical tests to find significant DE genes (Differential expressed genes across cell-types). By default we use the 'wilcox' test (Wilcoxon Rank Sum test) to identify marker genes. However we have implemented in the pipeline a functionality to use two tests each time and obtain the overlap of markers from the two tests. However this could result in issues while running the pipeline when the overlap is very low.

Deconvolution methods

Here is the list of deconvolution methods currently implemented in the pipeline ordered alphabetically with their package version numbers.

1. AutogeneS:Implemented from omnideconv package, omnideconv_0.0.0.9000 , autogenes, 1.0.4
2. BisqueRNA :BisqueRNA_1.0.5

3. Bseqsc: bseqsc_1.0
4. CDSeq: CDSeq_1.0.9
5. CIBERSORT: source code available at https://cibersortx.stanford.edu/ upon request. Code was obtained in August 2022.
6. CPM: scBio_0.1.2
7. DCQ : ComICS_1.0.4
8. debCAM_marker : debCAM_1.12.0
9. Deconf : CellMix_1.6.2
10. DeconRNASeq: DeconRNASeq_1.40.0
11. DSA: CellMix_1.6.2
12. dtangle : dtangle_2.0.9
13. DWLS: DWLS_0.1
14. EpiDISH: EpiDISH_2.14.0
15. EPIC: EPIC_1.1.7
16. elasticNET: glmnet_4.1-7
17. FARDEEP: FARDEEP_1.0.1
18. lasso: glmnet_4.1-7
19. MOMF: MOMF_0.2.0, omnideconv_0.0.0.9000
20. MuSiC: MuSiC_1.0.0, TOAST_1.12.0, quadprog_1.5-8, limma_3.54.2, EpiDISH_2.14.1, ggplot2_3.4.2, nnls_1.4
21. NNLS: nnls_1.4
22. OLS: base R, R version 4.2.3 (2023-03-15) lm() function
23. proportionsInAdmixture: energy_1.7-11 WGCNA_1.72-1 fastcluster_1.2.3 dynamicTreeCut_1.63-1 ADAPTS_1.0.22
24. RLR : MASS_7.3-60 energy_1.7-11
25. ridge: glmnet_4.1-7
26. SCDC: SCDC_0.0.0.9000
27. ssFrobenius:CellMix_1.6.2
28. ssKL: CellMix_1.6.2
29. TIMER : This code is adapted from http://cistrome.org/TIMER/download.html. The method is described in Li et al. Genome Biology 2016;17(1):174., PMID 27549193. TIMER Pipeline for analysing immune cell components in the tumour microenvironment. https://github.com/Papatheodorou-Group/CATD_snakemake/tree/main/Modules/TIMER

Evaluation of Performance - Metrics

The pipeline can evaluate the predictions through various metrics, where the overall score of a method for a given metric is the mean over all samples (columns). Each column in our actual observation matrix (ground truth) and estimated (predictions) proportions matrix is a sample (biological sample/pseudo-bulk) and the features are the cell-types.

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

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

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

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[50]
*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. Four 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.

Consensus deconvolution calculation :

For the consensus approach we selected three deconvolution methods that were performing well across our study. The main criteria for inclusion were: (1) well-performing methods overall (2) open-source code, (3) reasonable time of running, (5) stable and they produce results (sometimes methods won't converge and produce NAs (4) flexible (e.g they can take as input different reference every time to answer biological questions.

The displayed cell type proportions of GTEx stomach, mouse purified brain, and human fetal lung cell types are computed as the averages of deconvolution outcomes obtained through DWLS, EpiDISH, and FARDEEP, per sample and cell type. We also include the mean correlation between the results from the three methods across all samples to show the concordance of results across the methods. Standard deviation per cell type for each sample across methods is also computed.

For the barplot visualisation of deconvolution proportions in GTEx stomach samples, we employed the CiberBarFrazer() function in R[51]. Code available at https://github.com/mkrdonovan/gtex_deconvolution/blob/master/Figure_2.ipynb, R version 4.3.1). The resulting proportions are arranged in descending order, highlighting the three most abundant cell types.
The consensus barplot provides an overview of the cell type proportions for each sample. The header of the barplot includes the sample name and the mean correlation between the results from FARDEEP, DWLS, and EpiDISH across all samples.

QuPath analysis:

To validate our deconvolution results, we assessed the proportions of three stomach layers (Muscle, Mucosa, and Submucosa) on GTEx histology slides. We obtained histology files for 339 samples and used QuPath (v 0.4.1 on Mac OS X - 11.6.2) to train a pixel classifier on six slides. This classifier helped us estimate the areas of background, muscle, mucosa, and submucosa on the slides. Using the assigned areas, we calculated the relative proportions of the three tissue layers on each slide.

Mapping cell types to stomach layers:

Mucosa: endocrine, enterocyte,goblet cells, peptic cell(parietal cells),epithelial,
Submucosa: dendritic cell, endothelial,erythroid lineage cell,fibroblast,macrophages,plasma cell,stromal
Muscle: mesenchymal stem cell,neuron,smooth muscle]

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[52], 10X[53], CEL-seq[54], Seq-well[55], Drop-Seq[56], inDrop[57] 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[9,58]. 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[59]. 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 [60,61]. 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[62] (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[63] 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 [20,39], 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)[64]. 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 [65–67]. 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[68], DWLS[18]). 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[69,70] (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[17,71] (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.

Evaluation and consensus-based approach for deconvolution in real biological samples

In contrast to the simulation tasks performed above, here we decompose real bulk profiles from 9 Peripheral Blood Mononuclear Cells (PBMC)[10] 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-b). For this deconvolution task, we selected a recent single-cell PBMC dataset[47] 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 9a). 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 9b). Before we apply deconvolution on the 9 real bulk samples we first performed a self-reference deconvolution to evaluate how the reference performs in the pseudo-bulk deconvolution task. Eight methods yield good results with a Pearson Correlation of more than 0.92 (Supplementary Figure 9b). 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[72] 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 consensus method based on 3 methods DWLS,FARDEEP and EpiDISH(selection criteria on Methods) that perform well in previous pseudo-bulk scenarios and compute proportions relatively fast and reliably(Supplementary Figure 10). Bseq-sc although it performs well, because of its iterative nature, it can be very slow (> 1 week) to compute proportions so it was excluded from the consensus calculations. Next we applied the consensus on GTEx stomach data(339 samples)for which we had access to tissue slides that could inform the validation of the deconvolution results. For this task we used as a reference the stomach subset of the Human Cell Landscapes sRNA-seq dataset containing 16 cell-types(Figure 6a,b). From the deconvolution results we can observe a clear difference between the samples on the left of the barplot and the right hand of the barplot. Samples on the right seem to contain mostly smooth muscle cells, fibroblasts and endothelial cells, whereas samples on the left are mainly peptic and goblet cells (Figure 6c). Next we explored if this difference in the composition can be observed in the tissue slides of the adjacent samples. We randomly selected a small number of tissue samples (5 from the samples from the left side and 5 from the right side) . We observed that the slides that come from samples from the right side of the barplot, were in general more enriched in muscle cells compared to the samples on the left side which were more enriched in mucosa and submucosa tissue which aligns with the presence of goblet and peptic cells in the deconvolved mixtures(Figure 6e,f). We also quantified the mucosa, submucosa and muscle layers(μm2) for each stomach sample using QuPath and computed the proportions of each layer in each sample. Barplot showing the proportions computed for the 339 samples in the sample ordered the same way as in Figure 6c. Although we can observe that the samples on the right have a lot less mucosa and more muscle which aligns with the deconvolution results(absence of peptic and goblet cells on the samples on the right and presence of smooth muscle), the

measurements from the tissue slides seem to be quite different from the deconvolution results overall. This could be because of a number of reasons: (a) annotating mucosa,submucosa and muscle layers is not straightforward and is dependent on the individual who performs the annotation. (b) deconvolution predictions are objected to overfitting (c) tissue slides might have different composition compared to the bulk sample that goes to sequencing due to sampling reasons. More detailed analysis on the comparison between composition in bulk samples and tissue slides should be made to conclude if tissue slides are a good ground truth data. To further explore the differences between deconvolution results and composition calculations from tissue slides we mapped each cell-type to the respective layer based on literature knowledge(See methods) and we performed Pearson correlation analysis. We observe the highest positive correlation values between the Mucosa proportions from tissue slides and the summarised proportions of mucosa cells from deconvolution and similar Pearson correlation values between the Muscle layer and the summarised proportions from the cell types in the muscle layer. However the submucosa proportion from the tissue slide is not strongly correlated with any of the cell type proportions from deconvolution.

Next, we designed an experiment in order to evaluate the background prediction of the consensus method in two datasets with purified cell types and/or mixtures with known cell types. Our data collection included mouse cerebral cortex and human lung datasets, for which we carefully selected relevant single-cell references tailored to the task (Figure 7a,e). Initially, we applied the consensus deconvolution method to 17 samples obtained from the mouse cerebral cortex (Figure 7b). Each sample in this task represented a single cell type, serving as the ground truth, and aimed to evaluate the method's ability to accurately predict the presence of that specific cell type while quantifying background prediction. To quantify background prediction, we employed logarithmic loss, a widely used metric for evaluating classification model performance. In the context of deconvolution, this metric measures the accuracy of estimated cell type proportions within a mixture, taking into account uncertainties and penalising incorrect estimates arising from low-level gene expression that may lead to background predictions. Across the 17 samples, we observed consistently low log-loss values, ranging from approximately 0.004 to 0.104 (Figure 7c). Notably, among all samples, two endothelial flow cytometry samples and one myelinating oligodendrocyte sample exhibited the lowest log-loss values, signifying precise predictions of the purified cell type within each sample. Moving forward, we extended our analysis to encompass 26 samples from human fetal lung. Of these, the initial 16 samples were predominantly composed of endothelial cells, while the subsequent 20 samples contained non-endothelial cells. Our objective was to observe whether the consensus approach could effectively differentiate between these two sample categories. Notably, the endothelial samples predominantly exhibited high proportions of endothelial cells in the predictions, albeit with predictions of fibroblasts, lymphocytes, and myeloid cells. In contrast, the non-endothelial samples were primarily predicted to consist of fibroblasts, with minor proportions of endothelial cells detected. These low-level endothelial cell predictions may be attributed to background predictions generated by the consensus algorithm (Figure 7d).

In situations where ground truth data is unavailable, our pipeline incorporates a unique feature to address this. We perform a pairwise comparison of the deconvolution results across all samples, enabling us to estimate Pearson Correlation coefficients. These coefficients serve as a quantitative measure of the agreement between different methods. The results of this comparison are visually presented in the form of a heatmap, where methods that yield similar results cluster together (Supplementary Figure 9c). This clustering provides valuable insights into which methods exhibit strong agreement, both in the presence and absence of ground truth data.

It is important to mention that while the clustering approach successfully identifies method agreement, it does not inherently imply that the clustered methods consistently perform well. Finally, to comprehensively assess the reliability and scalability of our pipeline, we conducted parallel evaluations of each module, encompassing pseudo-bulk generation, normalisation, and deconvolution methods. These assessments provide valuable insights into the computational efficiency of our pipeline, reporting essential metrics such as processing time, memory usage, mean load, and CPU time for each module(Supplementary Figure 10a-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 [29,39].

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 differences 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.
Notably, in the absence of ground truth data, our pipeline introduces a valuable feature for pairwise comparisons of deconvolution results across all samples, allowing the estimation of Pearson Correlation coefficients. This innovative approach offers a quantitative measure of method agreement, even without the presence of ground truth data, providing researchers with insights into method performance. It is imperative to understand that while this clustering technique effectively identifies method agreement, it doesn't inherently imply consistent

method performance. We also suggest a consensus method based on three robust techniques: DWLS, FARDEEP, and EpiDISH and test it across 3 real bulk datasets with available ground truth testing its accuracy and background prediction.
Notably, 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 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 as well as calculating proportions based on the consensus method.
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.

DATA AVAILABILITY

The source code for the CATD pipeline can be downloaded from https://github.com/Papatheodorou-Group/CATD_snakemake . All data used in this study are publicly available. Accession numbers and urls can be found in the Supplementary Table 1 and Supplementary Table 4.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

AUTHOR CONTRIBUTIONS

Anna Vathrakokoili Pournara: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing-original draft.

Zhichao Miao: Conceptualization, Data Curation, Methodology, Software, Supervision, Writing-review & editing.

Ozgur Beker: Methodology, Formal Analysis, Software

Nadja Nolte: Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualisation

Alvis Brazma: Conceptualization, Supervision, Writing -review & editing

Irene Papatheodorou: Conceptualization, Project Administration, Supervision, Writing-review & editing

ACKNOWLEDGEMENTS

We would also like to thank Craig Russell for the valuable help in reviewing and editing.

FUNDING

This work was supported by: the European Molecular Biology Laboratory (A.V.P, Z.M., O.B.,N.N, A.B, I.P.); the EMBL international PhD program (A.V.P.) and the OpenTargets (N.N,C.M, I.P,) [OTAR2067]

CONFLICT OF INTEREST

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. Jorge, N. A. N. *et al.* Poor clinical outcome in metastatic melanoma is associated with a microRNA-modulated immunosuppressive tumor microenvironment. *J. Transl. Med.* **18**, 56 (2020).
5. Zhang, Y. *et al.* High-Infiltration of Tumor-Associated Macrophages Predicts Unfavorable Clinical Outcome for Node-Negative Breast Cancer. *PLoS ONE* **8**, e76147 (2013).
6. 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).
7. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Comput. Biol.* **16**, e1008120 (2020).
8. 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).

9. 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).
10. 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).
11. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
12. The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
13. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
14. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
15. 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).
16. 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).
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. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* **10**, 2975 (2019).
19. 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).
20. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through

robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).

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

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

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

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

25. Lopez, D. *et al.* SaVanT: a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. *BMC Genomics* **18**, 824 (2017).

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

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

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

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

30. Sutton, G. J. *et al.* Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat. Commun.* **13**, 1358 (2022).

31. Jaakkola, M. K. & Elo, L. L. Computational deconvolution to estimate cell type-specific gene expression from bulk data. *NAR Genomics Bioinforma.* **3**, lqaa110 (2021).

32. Shen-Orr, S. S. *et al.* Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).

33. Dimitrakopoulou, K., Wik, E., Akslen, L. A. & Jonassen, I. Deblender: a

semi−/unsupervised multi-operational computational method for complete deconvolution of expression data from heterogeneous samples. *BMC Bioinformatics* **19**, 408 (2018).

34. Gong, T. *et al.* Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLoS ONE* **6**, e27156 (2011).

35. 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).

36. 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).

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

38. Menden, K. *et al.* Deep learning–based cell composition analysis from tissue expression profiles. *Sci. Adv.* **6**, eaba2619 (2020).

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. Nadel, B. B. *et al.* Systematic evaluation of transcriptomics-based deconvolution methods and references using thousands of clinical samples. *Brief. Bioinform.* **22**, bbab265 (2021).

41. Cobos, F. A. *et al.* Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes. *Genome Biol.* **24**, 177 (2023).

42. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

43. Jiménez-Sánchez, A., Cast, O. & Miller, M. L. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res.* **79**, 6238–6246 (2019).

44. 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.

45. Alonso-Moreda, N. *et al.* Comparative Analysis of Cell Mixtures Deconvolution and Gene Signatures Generated for Blood, Immune and Cancer Cells. *Int. J. Mol. Sci.* **24**, 10765 (2023).

46. 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).

47. 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).

48. Song, Y., Miao, Z., Brazma, A. & Papatheodorou, I. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nat. Commun.* **14**, 6495 (2023).

49. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

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

51. 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).

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

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

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

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

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

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

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

59. 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).

60. 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).

61. 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.

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

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

64. 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).

65. 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).

66. 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).

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

68. 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).

69. 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).

70. 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).

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

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

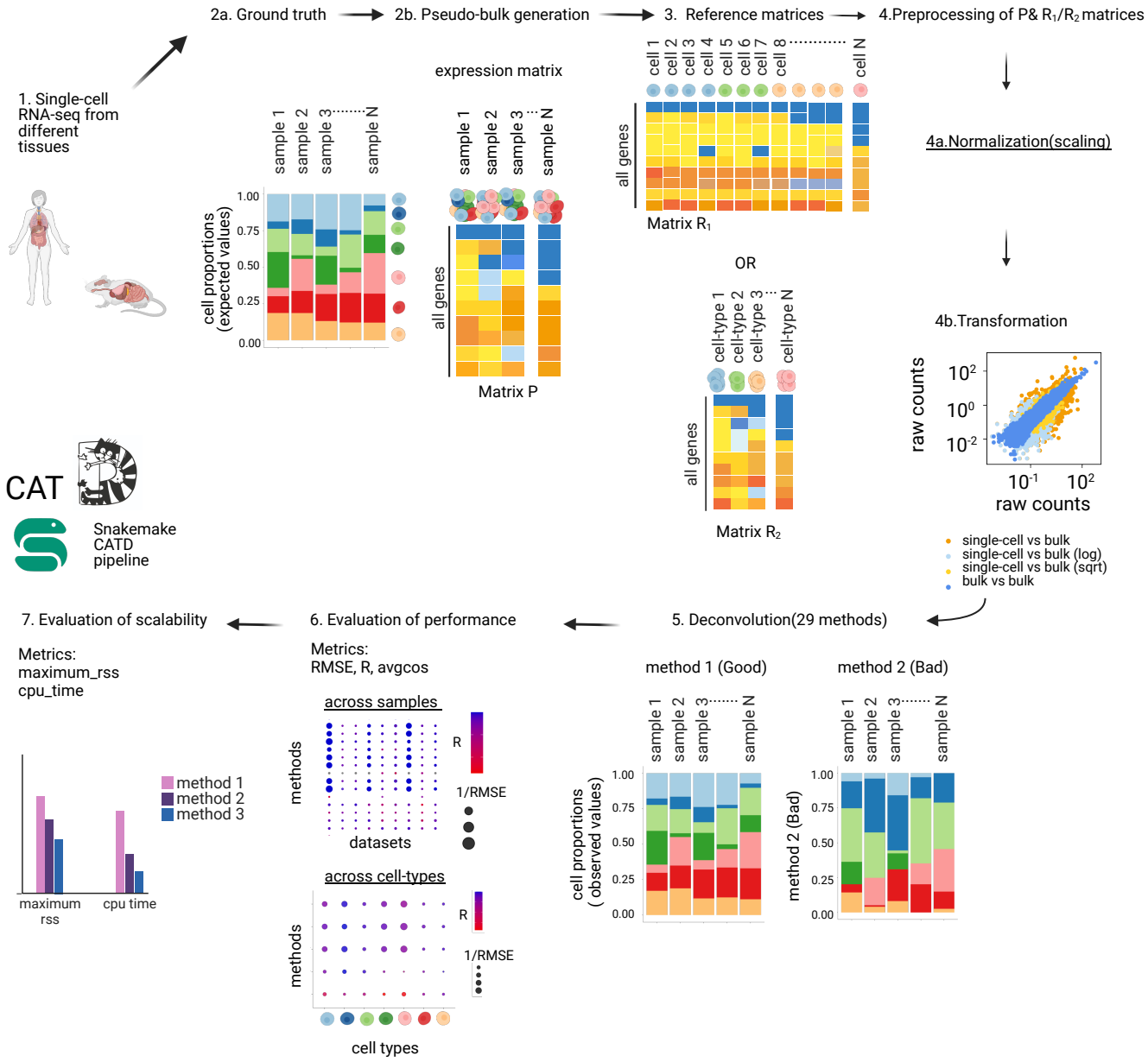

2a. Ground truth
2b. Pseudo-bulk generation
3. Reference matrices
4.Preprocessing of P& R1/R2 matrices
1. Single-cell RNA-seq from different tissues
expression matrix
sample 1
sample 2
sample 3
sample N
cell proportions (expected values)
1.00
0.75
0.50
0.25
0.00
all genes
Matrix P
cell 1
cell 2
cell 3
cell 4
cell 5
cell 6
cell 7
cell 8
cell N
Matrix R1
OR
cell-type 1
cell-type 2
cell-type 3
cell-type N
Matrix R2
4a.Normalization(scaling)
4b.Transformation
raw counts
single-cell vs bulk
single-cell vs bulk (log)
single-cell vs bulk (sqrt)
bulk vs bulk
CAT
Snakemake
CATD
pipeline
5. Deconvolution(29 methods)
method 1 (Good)
method 2 (Bad)
cell proportions (observed values)
6. Evaluation of performance
Metrics:
RMSE, R, avgcos
across samples
methods
datasets
R
1/RMSE
across cell-types
cell types
7. Evaluation of scalability
Metrics:
maximum_rss
cpu_time
method 1
method 2
method 3
maximum rss
cpu time

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.

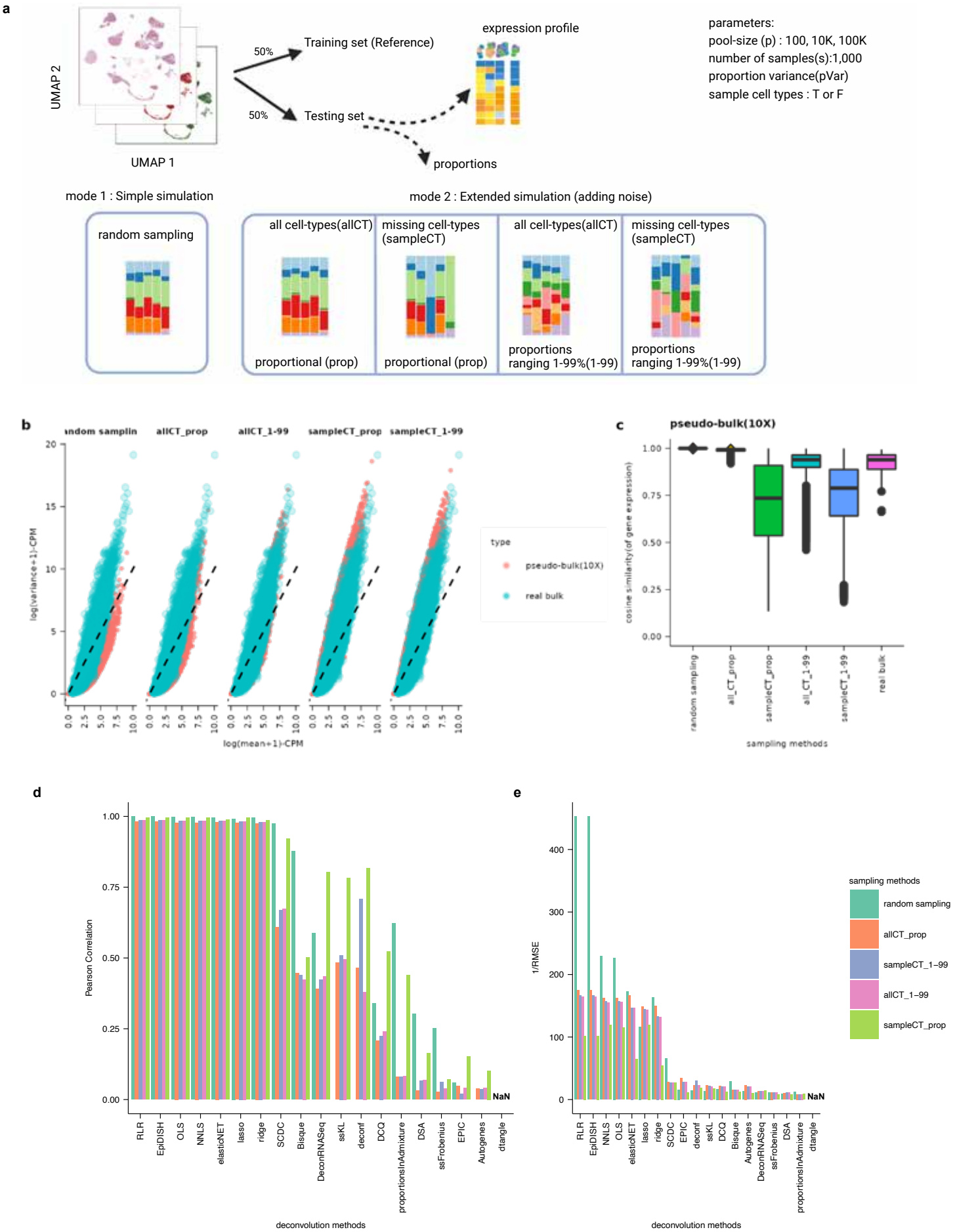

a
UMAP 2
UMAP 1
50%
Training set (Reference)
50%
Testing set
expression profile
proportions
parameters:
pool-size (p) : 100, 10K, 100K
number of samples(s):1,000
proportion variance(pVar)
sample cell types : T or F
mode 1 : Simple simulation
random sampling
mode 2 : Extended simulation (adding noise)
all cell-types(allCT)
missing cell-types (sampleCT)
all cell-types(allCT)
missing cell-types (sampleCT)
proportional (prop)
proportional (prop)
proportions ranging 1-99%(1-99)
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
NaN
deconvolution methods
sampling methods
random sampling
allCT_prop
sampleCT_1–99
allCT_1–99
sampleCT_prop

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.

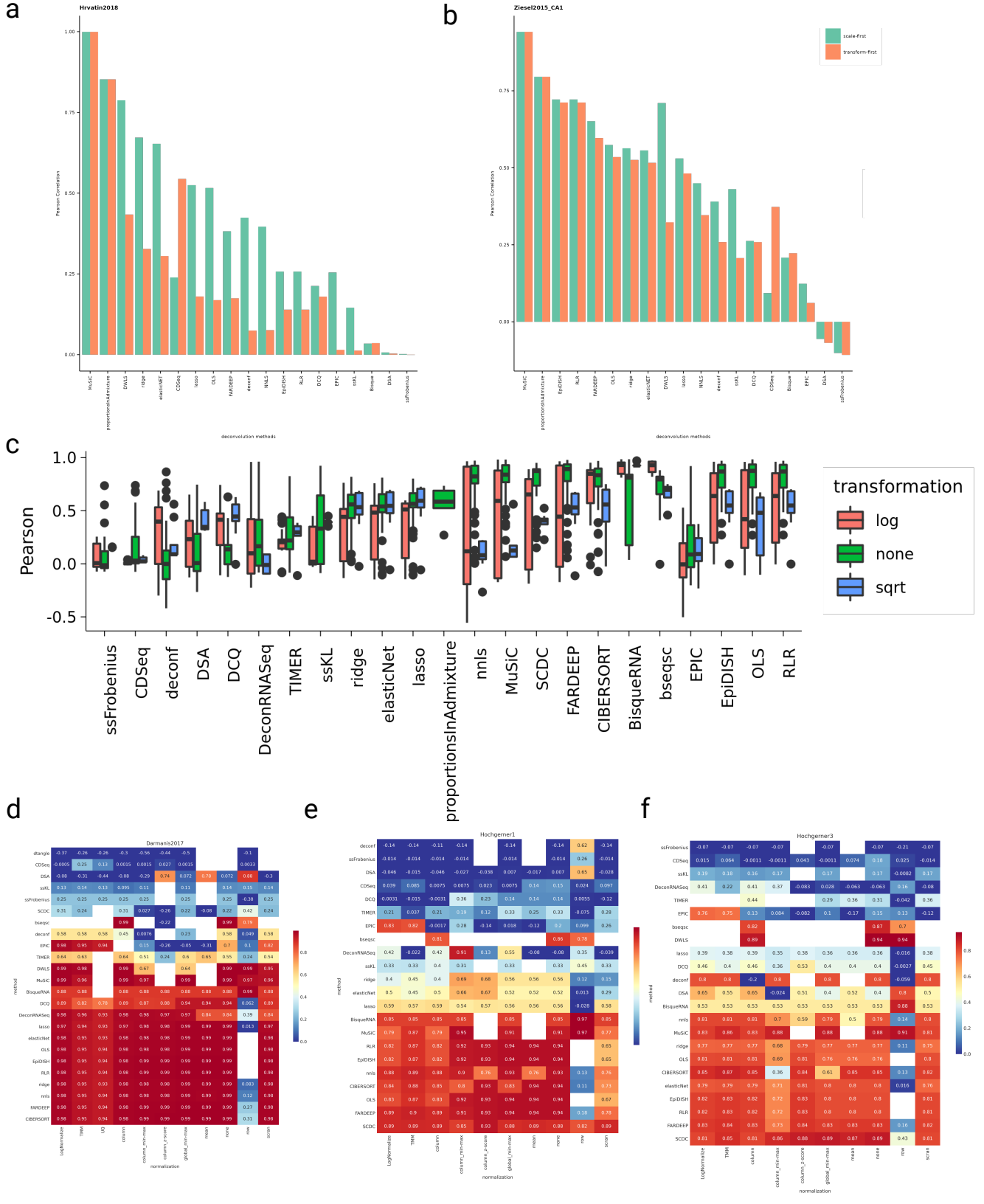

a
Hrvatin2018
Pearson Correlation
deconvolution methods
b
Ziesel2015_CA1
scale-first
transform-first
Pearson Correlation
deconvolution methods
c
Pearson
transformation
log
none
sqrt
ssFrobenius
CDSeq
deconf
DSA
DCQ
DeconRNASeq
TIMER
ssKL
ridge
elasticNet
lasso
proportionsInAdmixture
nnls
MuSiC
SCDC
FARDEEP
CIBERSORT
BisqueRNA
bseqsc
EPIC
EpiDISH
OLS
RLR
d
Darmanis2017
method
normalization
e
Hochgerner1
method
normalization
f
Hochgerner3
method
normalization

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-f. Results from deconvolution of 3 different simulated data from brain tissues (Darmanis2017,Hochgerner1,Hochgerner3) across 11 normalisation methods and 24 tested deconvolution methods.

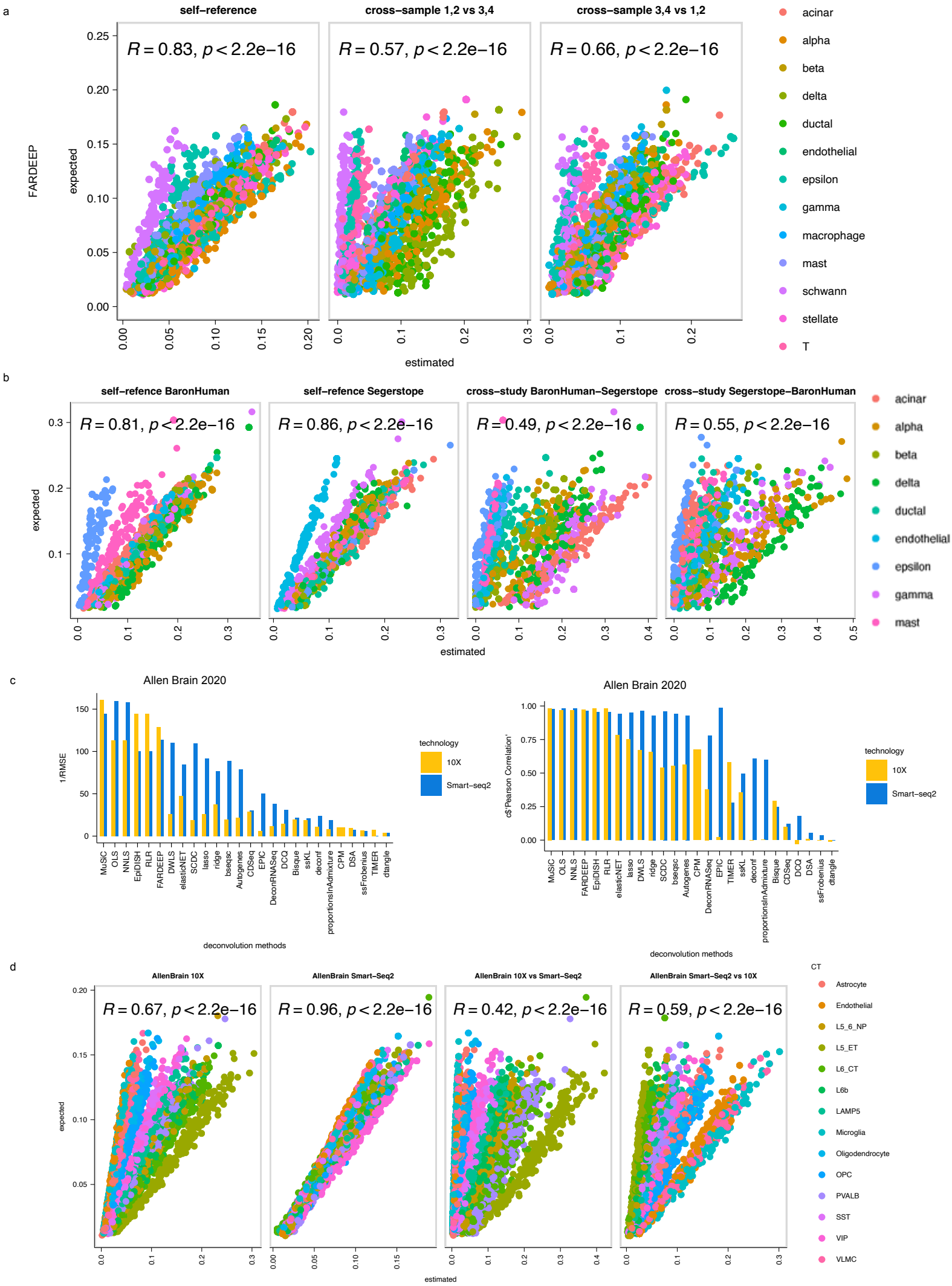
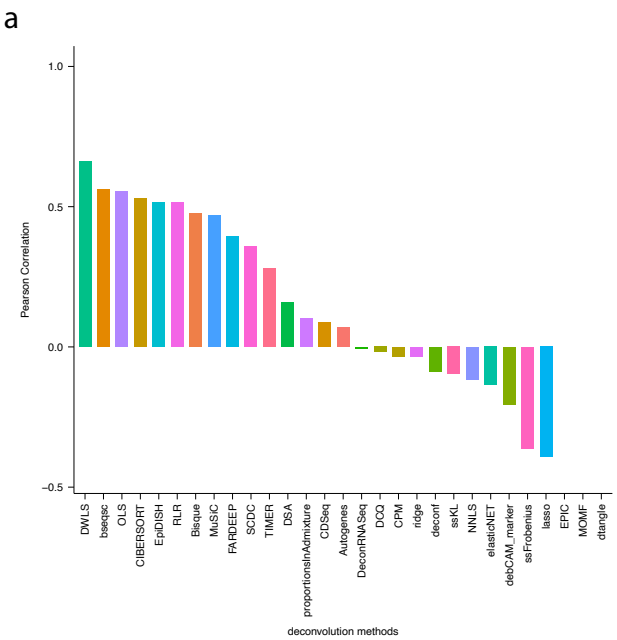
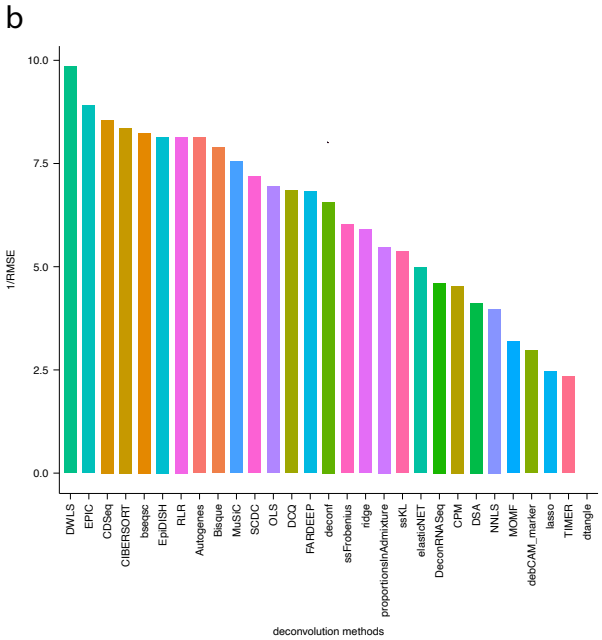
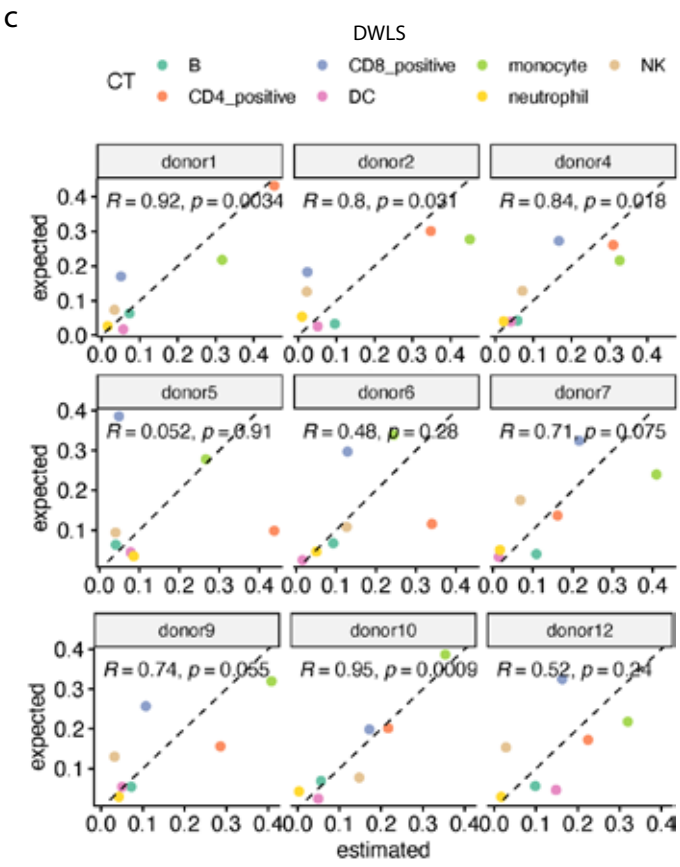

a
self−reference
cross−sample 1,2 vs 3,4
cross−sample 3,4 vs 1,2
R = 0.83, p < 2.2e−16
R = 0.57, p < 2.2e−16
R = 0.66, p < 2.2e−16
FARDEEP
expected
estimated
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
schwann
stellate
T
b
self−refence BaronHuman
self−refence Segerstope
cross−study 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
MuSiC
OLS
NNLS
EpiDISH
RLR
FARDEEP
DWLS
elasticNET
SCDC
lasso
ridge
bseqsc
Autogenes
CDSeq
EPIC
DeconRNASeq
DCQ
Bisque
ssKL
deconf
proportionsInAdmixture
CPM
DSA
ssFrobenius
TIMER
dtangle
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.


a
Pearson Correlation
1.0
0.5
0.0
−0.5
DWLS
bseqsc
OLS
CIBERSORT
EpiDISH
RLR
Bisque
MuSiC
FARDEEP
SCDC
TIMER
DSA
proportionsInAdmixture
CDSeq
Autogenes
DeconRNASeq
DCQ
CPM
ridge
deconf
ssKL
NNLS
elasticNET
debCAM_marker
ssFrobenius
lasso
EPIC
MOMF
dtangle
deconvolution methods


b
1/RMSE
10.0
7.5
5.0
2.5
0.0
DWLS
EPIC
CDSeq
CIBERSORT
bseqsc
EpiDISH
RLR
Autogenes
Bisque
MuSiC
SCDC
OLS
DCQ
FARDEEP
deconf
ssFrobenius
ridge
proportionsInAdmixture
ssKL
elasticNET
DeconRNASeq
CPM
DSA
NNLS
MOMF
debCAM_marker
lasso
TIMER
dtangle
deconvolution methods


c
DWLS
CT
B
CD4_positive
CD8_positive
DC
monocyte
neutrophil
NK
donor1
R = 0.92, p = 0.0034
donor2
R = 0.8, p = 0.031
donor4
R = 0.84, p = 0.018
donor5
R = 0.052, p = 0.91
donor6
R = 0.48, p = 0.28
donor7
R = 0.71, p = 0.075
donor9
R = 0.74, p = 0.055
donor10
R = 0.95, p = 0.0009
donor12
R = 0.52, p = 0.24
expected
estimated

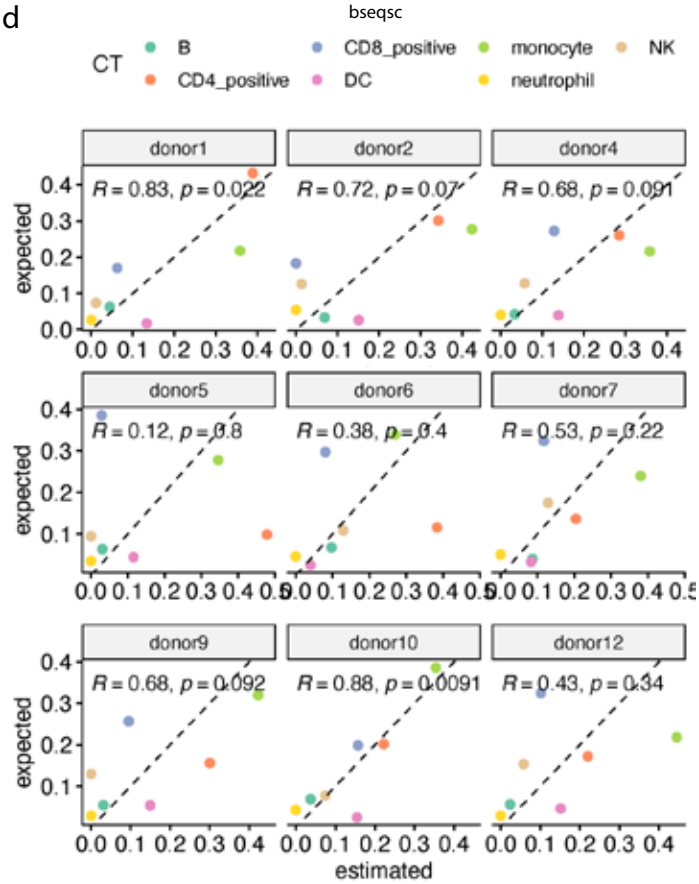

d
bseqsc
CT
B
CD4_positive
CD8_positive
DC
monocyte
neutrophil
NK
donor1
R = 0.83, p = 0.022
donor2
R = 0.72, p = 0.07
donor4
R = 0.68, p = 0.091
donor5
R = 0.12, p = 0.8
donor6
R = 0.38, p = 0.4
donor7
R = 0.53, p = 0.22
donor9
R = 0.68, p = 0.092
donor10
R = 0.88, p = 0.0091
donor12
R = 0.43, p = 0.34
expected
estimated

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.

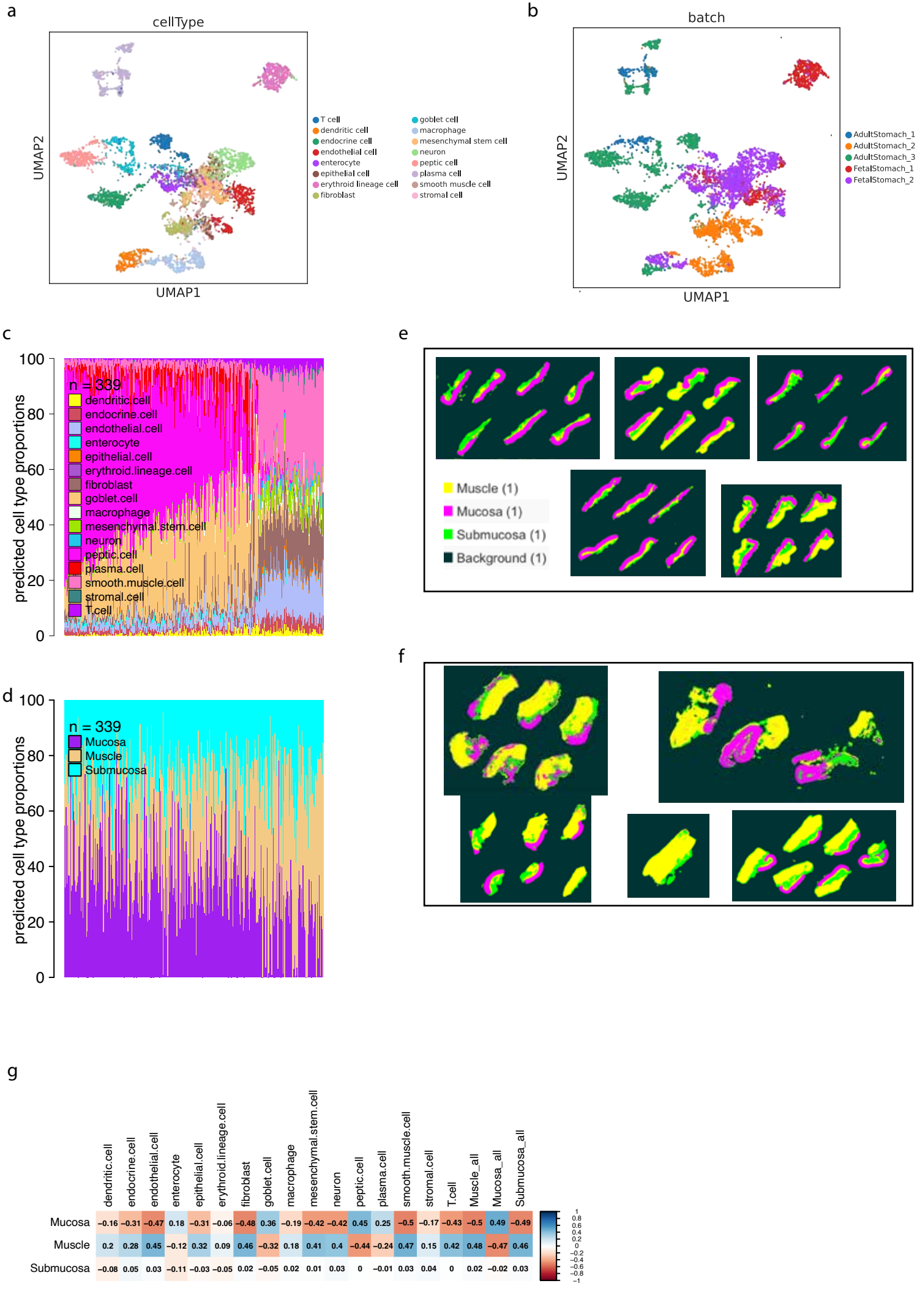

a
cellType
UMAP2
UMAP1
T cell
dendritic cell
endocrine cell
endothelial cell
enterocyte
epithelial cell
erythroid lineage cell
fibroblast
goblet cell
macrophage
mesenchymal stem cell
neuron
peptic cell
plasma cell
smooth muscle cell
stromal cell
b
batch
UMAP2
UMAP1
AdultStomach_1
AdultStomach_2
AdultStomach_3
FetalStomach_1
FetalStomach_2
c
predicted cell type proportions
n = 339
dendritic.cell
endocrine.cell
endothelial.cell
enterocyte
epithelial.cell
erythroid.lineage.cell
fibroblast
goblet.cell
macrophage
mesenchymal.stem.cell
neuron
peptic.cell
plasma.cell
smooth.muscle.cell
stromal.cell
T.cell
d
predicted cell type proportions
n = 339
Mucosa
Muscle
Submucosa
e
Muscle (1)
Mucosa (1)
Submucosa (1)
Background (1)
f
g
dendritic.cell
endocrine.cell
endothelial.cell
enterocyte
epithelial.cell
erythroid.lineage.cell
fibroblast
goblet.cell
macrophage
mesenchymal.stem.cell
neuron
peptic.cell
plasma.cell
smooth.muscle.cell
stromal.cell
T.cell
Muscle_all
Mucosa_all
Submucosa_all
Mucosa −0.16 −0.31 −0.47 0.18 −0.31 −0.06 −0.48 0.36 −0.19 −0.42 −0.42 0.45 0.25 −0.5 −0.17 −0.43 −0.5 0.49 −0.49
Muscle 0.2 0.28 0.45 −0.12 0.32 0.09 0.46 −0.32 0.18 0.41 0.4 −0.44 −0.24 0.47 0.15 0.42 0.48 −0.47 0.46
Submucosa −0.08 0.05 0.03 −0.11 −0.03 −0.05 0.02 −0.05 0.02 0.01 0.03 0 −0.01 0.03 0.04 0 0.02 −0.02 0.03

Figure 6 : Application of consensus approach on stomach GTEx bulk expression data
a,b UMAP plots of the stomach subset of Human Cell Landscapes single-cell reference used for the stomach GTEx deconvolution showing the different annotated cell-types and the distribution of donors in the datasets. **c**. Bar plot showing the prediction of cell-type proportions from the consensus approach across 339 selected stomach samples that have matched tissue slides. **d.** Quantification of the different stomach layers (mucosa, submucosa, muscle) from the matched tissue slides using QuPath. Area µm2 is measured for each layer in each tissue slide. **e.** 5 randomly selected tissue slides from the samples on the left side of the barplot c. Submucosa is shown in green, mucosa in magenta and muscle compartment in green **f.** 5 randomly selected tissue slides from samples originating from the right side of barplot c **g.** correlation plot showing the Pearson Correlation values between the proportions measured from tissue slides and the predicted proportions from deconvolution of the bulk RNA-seq samples

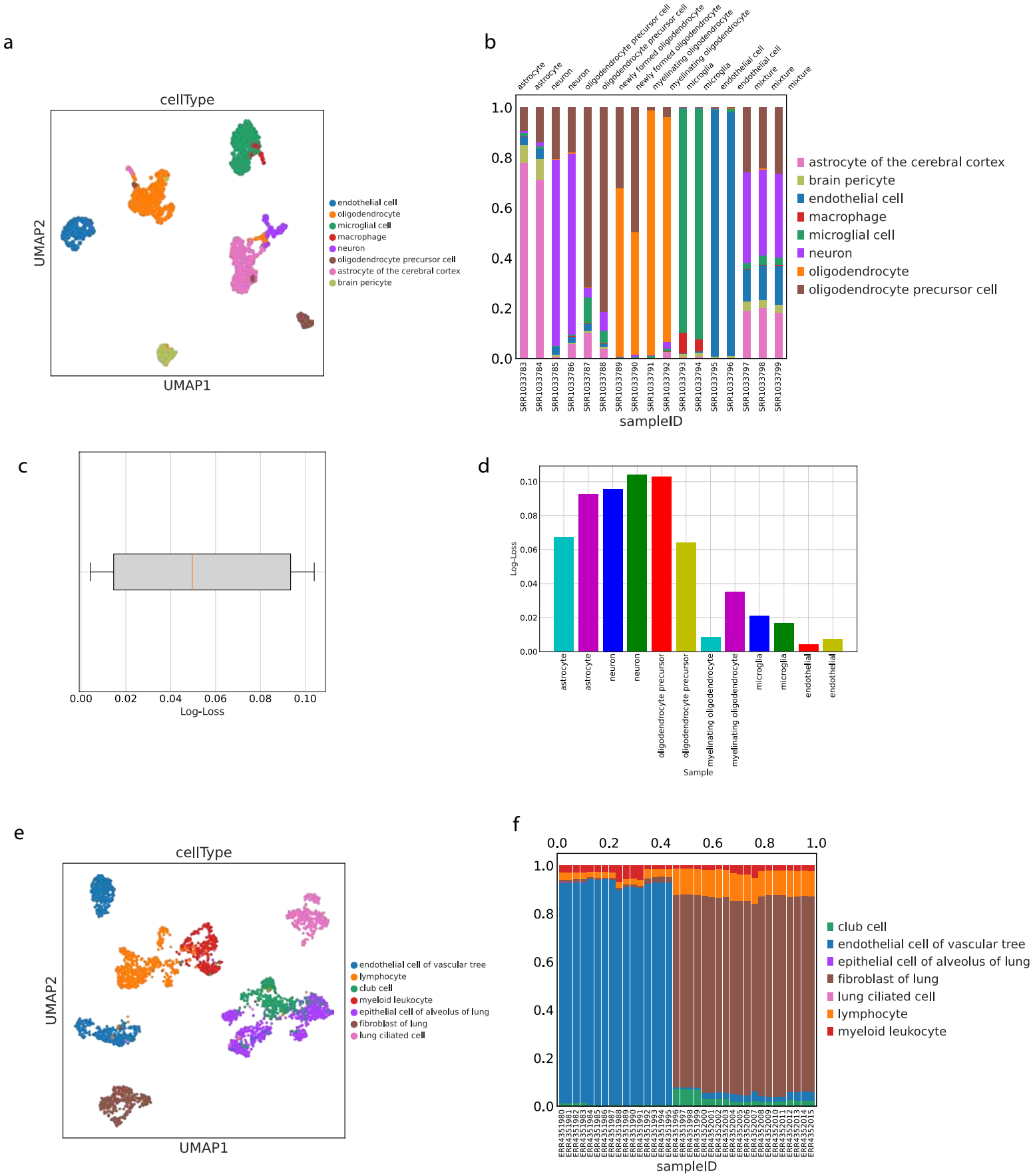

a
cellType
UMAP2
UMAP1
endothelial cell
oligodendrocyte
microglial cell
macrophage
neuron
oligodendrocyte precursor cell
astrocyte of the cerebral cortex
brain pericyte
b
astrocyte
astrocyte
neuron
neuron
oligodendrocyte precursor cell
oligodendrocyte precursor cell
newly formed oligodendrocyte
newly formed oligodendrocyte
myelinating oligodendrocyte
myelinating oligodendrocyte
microglia
microglia
endothelial cell
endothelial cell
mixture
mixture
mixture
1.0
0.8
0.6
0.4
0.2
0.0
SRR1033783
SRR1033784
SRR1033785
SRR1033786
SRR1033787
SRR1033788
SRR1033789
SRR1033790
SRR1033791
SRR1033792
SRR1033793
SRR1033794
SRR1033795
SRR1033796
SRR1033797
SRR1033798
SRR1033799
sampleID
astrocyte of the cerebral cortex
brain pericyte
endothelial cell
macrophage
microglial cell
neuron
oligodendrocyte
oligodendrocyte precursor cell
c
0.00
0.02
0.04
0.06
0.08
0.10
Log-Loss
d
0.10
0.08
0.06
0.04
0.02
0.00
Log-Loss
astrocyte
astrocyte
neuron
neuron
oligodendrocyte precursor
oligodendrocyte precursor
myelinating oligodendrocyte
myelinating oligodendrocyte
microglia
microglia
endothelial
endothelial
Sample
e
cellType
UMAP2
UMAP1
endothelial cell of vascular tree
lymphocyte
club cell
myeloid leukocyte
epithelial cell of alveolus of lung
fibroblast of lung
lung ciliated cell
f
0.0
0.2
0.4
0.6
0.8
1.0
sampleID
club cell
endothelial cell of vascular tree
epithelial cell of alveolus of lung
fibroblast of lung
lung ciliated cell
lymphocyte
myeloid leukocyte

Figure 7: Application of consensus deconvolution approach in samples of purified cell types

a. UMAP plot of downsampled single-cell RNA-seq dataset from TabulaMuris(cerebral cortex subset) composed of 8 cell-types. Original authors' labels have been reduced(Methods). b. Barplot showing the proportions of cell-types for each sample in the bulk mouse cerebral cortex dataset as predicted by the consensus deconvolution approach. c. boxplot illustrating the distribution of log-loss values across samples when comparing the predicted cell types from deconvolution with the known presence of a cell type in a purified sample. d. The barplot shows the log-loss values for 12 samples in the dataset(SRR1033789,SRR1033789-newly formed oligodendrocytes and the mixtures- SRR1033797,SRR1033798,SRR1033799 were excluded from this analysis). Lower the log-loss indicates more accurate predictions of the method. e. UMAP plot of the downsampled single-cell RNA-seq dataset from the lung subset of GTEx composed of 7 cell-types. Original authors' labels have been reduced(Methods). f. Barplot showing the proportions of cell-types for each sample in the bulk human fetal lung dataset[known purified endothelial(left) and non endothelial (right)samples] as predicted by the consensus deconvolution approach.