


REVIEW

Open Access



Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets

Sean K. Maden¹, Sang Ho Kwon^{2,3}, Louise A. Huuki-Myers², Leonardo Collado-Torres^{1,2}, Stephanie C. Hicks^{1,4,5,6*} and Kristen R. Maynard^{2,3,7*} 

[†]Stephanie C. Hicks and Kristen R. Maynard share equal contributions and are co-corresponding senior authors.

*Correspondence: shicks19@jhu.edu; kristen.maynard@libd.org

¹ Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

² Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA

³ The Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, USA

⁴ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

⁵ Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

⁶ Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA

⁷ Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA

Abstract

Deconvolution of cell mixtures in “bulk” transcriptomic samples from homogenate human tissue is important for understanding disease pathologies. However, several experimental and computational challenges impede transcriptomics-based deconvolution approaches using single-cell/nucleus RNA-seq reference atlases. Cells from the brain and blood have substantially different sizes, total mRNA, and transcriptional activities, and existing approaches may quantify total mRNA instead of cell type proportions. Further, standards are lacking for the use of cell reference atlases and integrative analyses of single-cell and spatial transcriptomics data. We discuss how to approach these key challenges with orthogonal “gold standard” datasets for evaluating deconvolution methods.

Keywords: Deconvolution, Single-cell RNA-sequencing, Single-nucleus RNA-sequencing, Cell sizes

Introduction

An important challenge in the analysis of gene expression data from complex tissue homogenates measured with RNA-sequencing (bulk RNA-seq) is to reconcile cellular heterogeneity or unique gene expression profiles of distinct cell types in the sample. A prime example is bulk RNA-seq data from human brain tissue, which consists of two major categories of cell types, neurons and glia, both of which have distinct morphologies, cell sizes, and functions across brain regions and sub-regions [1–3]. Failing to account for biases driven by molecular and biological characteristics of distinct cell types can lead to inaccurate cell type proportion estimates from deconvolution of complex tissue such as the brain [3].



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Broadly, methods that computationally estimate cell proportions from bulk tissue “-omics” data, such as gene expression or DNA methylation (DNAm) data, are referred to as “deconvolution algorithms” [4, 5]. Deconvolution commonly uses three terms: (1) a cell type signatures reference matrix, called Z ; (2) a convoluted signals matrix, Y ; and (3) a vector of the proportions of cell types in Y , called P . Here, we focus on gene expression reference-based algorithms that predict P given Z and Y (Fig. 1). With these standard terms, deconvolution is often mathematically described using the equation $Y = Z * P$, where the goal is to estimate the set of proportions P (i.e., where each $p \in P$ satisfies $0 \leq p \leq 1$ and P sums to 1). Approaches for estimating P have been widely reviewed in the literature [6, 7] and are outside the scope of this review. Nonetheless, recent work has described important challenges (Fig. 2) for deconvolution with various tissues including blood, kidney, and pancreas [8, 9]. However, tissues with notably different cell sizes, total mRNA expression, and transcriptional activity levels, such as brain or immune cell populations, present additional challenges for deconvolution that have not been previously described. It is important to be able to accurately estimate the cell type proportions of these complex tissues, as cell composition has been shown to change with disease [10–15].

In computational methods development, gold standard datasets are used to set baseline performance expectations and provide a well-characterized reference against which

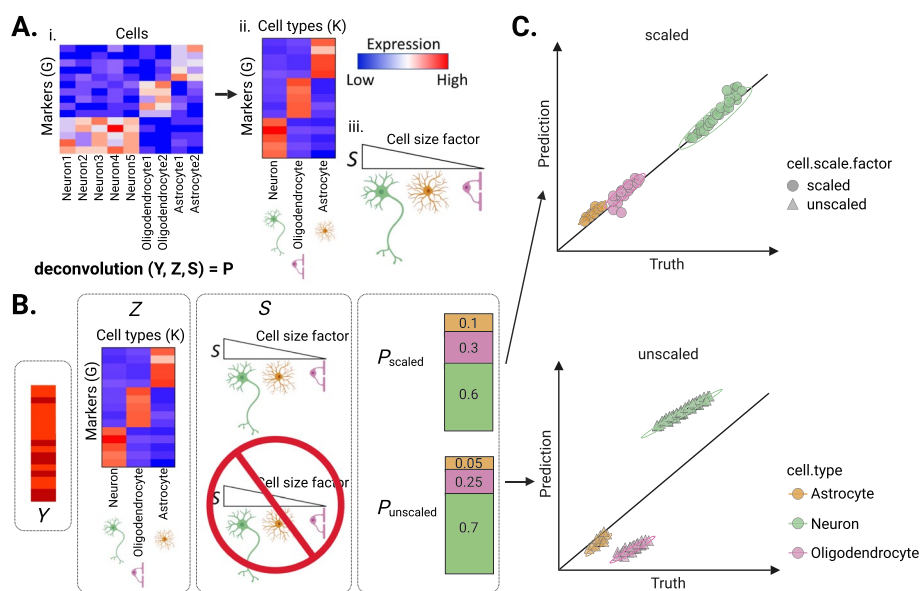


Fig. 1 Diagram of example deconvolution experiment using cell scale factors. **A** Heatmaps of gene expression: (i) for the (y-axis) marker genes G by cell labels for each of (x-axis) neurons, oligodendrocytes, or astrocytes, (ii) the (y-axis) G marker genes by (x-axis) cell types (K). Expression value colors: blue = low, white = intermediate, red = high. (iii) Wedge diagram of (S) cell scale factors, where wedge size is the value and cartoons indicate each cell type. **B** (left-to-right) Heatmaps of bulk expression Y , and marker expression Z , cell scale factors S , and cell type proportions P for either (top) scaled or (bottom) unscaled expression, where bar plot values show cell type proportions with colors as in panel C. **C** Scatterplot of example experiment results for multiple bulk samples Y , showing the (x-axis) true cell proportions and (y-axis) predicted cell proportions, where points are outcomes for a sample and cell type, and shapes show whether the cell scale factor transformation was applied. Plots were created using the ggplot2 v3.4.1 [16] and ComplexHeatmap v2.12.1 [17] software; data used to reproduce these plots are available from GitHub (Data Availability)

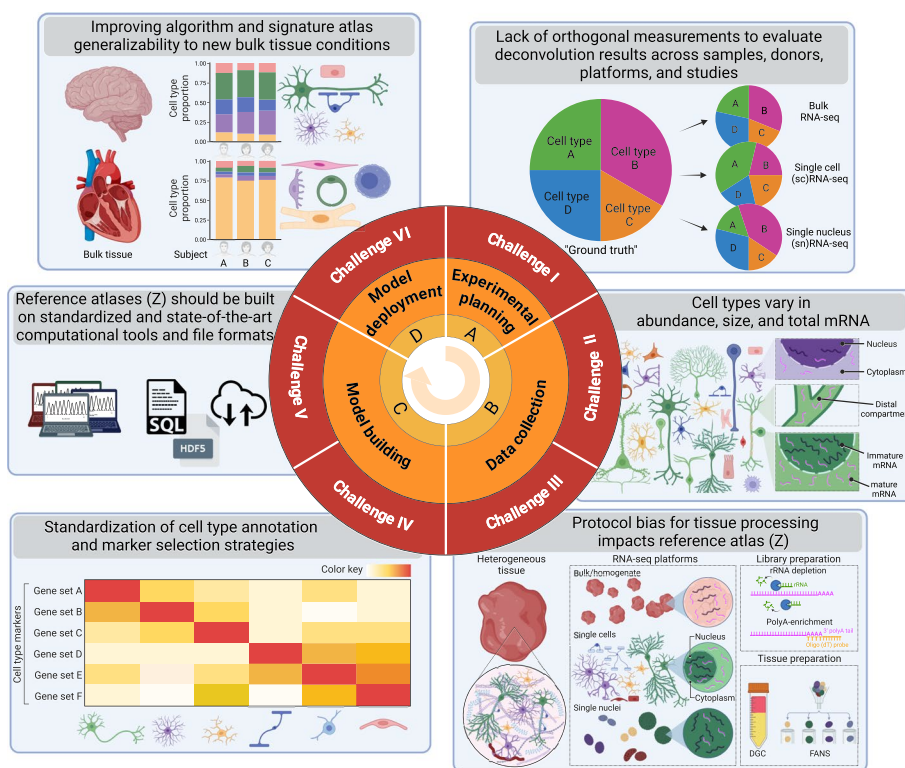


Fig. 2 Six challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets. Direction of experimental process (middle arrow), experiment phases (orange labels), challenge number (red labels), challenge titles (gray panel titles), and depictions of key challenge concepts (box graphics)

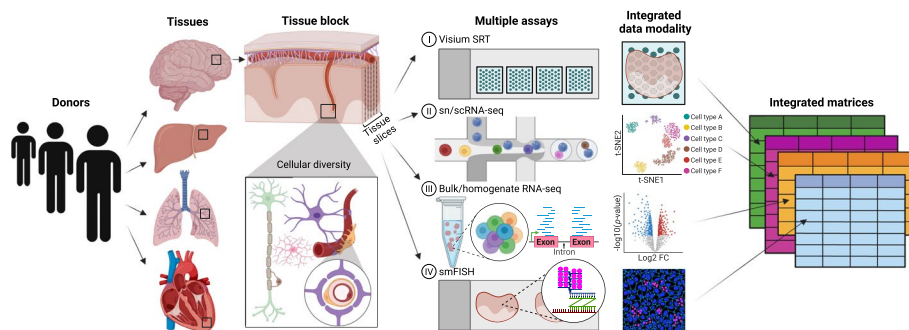


Fig. 3 Collecting an integrated dataset of orthogonal assays from the same tissue block across donors and tissues. The development and benchmarking of deconvolution algorithms can be improved with gold standard reference datasets. Gold standards are developed across donors and tissues on which multiple assays are performed on the same tissue block. For example, adjacent sections of a tissue block could be used for spatial transcriptomics, sc/snRNA-seq, bulk/homogenate RNA-seq, and single-molecule FISH (smFISH) to generate orthogonal cell type proportion and transcriptomic profile measurements. These assays generate data with distinct features (i.e., gene expression, cell size/shape, isoform diversity, etc.) that can also be incorporated into deconvolution models to improve accuracy

new outputs can be evaluated. For example, Sanger sequencing is used as a gold standard platform for validation of genetic sequencing data [18, 19]. Similarly, in deconvolution, independent or orthogonal measurements (Fig. 3) from different platforms of

cell composition can be used to validate algorithm-based estimates from bulk tissue expression.

In this paper, we summarize a set of challenges for performing deconvolution in highly heterogeneous tissues, using human brain tissue as a motivating example. We also present a set of recommendations and future opportunities for how to address these challenges to more accurately estimate tissue cell composition and better understand human disease. This poses an opportunity to set a higher bar for biological discovery and publication practices including increased computational reproducibility [9]. The ability to iteratively implement and optimize new methods and benchmark workflows in heterogeneous tissues will enable deconvolution tools to further our understanding of the role of changes in cell type composition with disease risk and progression.

Challenge 1: lack of orthogonal measurements to evaluate deconvolution results across samples, donors, platforms, and studies

Need for orthogonal measurements from matched tissue samples for bulk and single-cell data

When developing a deconvolution method, using matched bulk and single-cell/nucleus RNA-seq (sc/snRNA-seq) datasets from the same tissue samples (Fig. 3) enables controlling for potential variation beyond cell type variation, observed from unwanted factors [18–21], such as donor-to-donor variation [22, 23]. For example, confounding variation may come from factors relating to donor demographics (i.e., sex [20], genetics [21], diagnosis [22]), tissue dissection (i.e., tissue microenvironment representation [23, 24]), and/or sample quality (i.e., tissue pH, post-mortem interval, RNA quality [25–27]), where certain sources need to be evaluated in specific tissues, such as the expected proportions of white and gray matter in brain specimens [13]. Excess variation from such sources can cause challenges downstream to accurately estimate the cell type reference matrix, Z [24, 25], leading to inaccurate estimates of cell composition, P . This concept is further supported by Wang et al. [28], who studied errors from using a sc/snRNA-seq reference dataset from source A to deconvolve a RNA-seq sample from source B and showed these errors can lead to inaccurate estimates of cell composition P for source B, where sources could be distinct donors or studies. Considering the potential limitations of using existing reference atlas datasets, specific sample sourcing schemes, such as generation of multiple orthogonal assays matched to the same tissue block, could alleviate some of these issues. As orthogonal datasets generated from specimens gathered from the same tissue block (a.k.a. “source-matched” samples) are replicates for important clinical and demographic factors, their greater utilization will limit the influence of excess variation of unwanted factors beyond what is possible with modeling strategies alone. We would also advocate the use of these orthogonal matched assays in the development and benchmarking of new algorithms to better evaluate algorithm performance while controlling unwanted confounds such as technical and biological variation.

Need for orthogonal measurements from health and disease samples

Deconvolution algorithms are commonly used to investigate whether changes in cell composition of tissue samples are associated with a phenotype or outcome, such as in case–control study designs. This poses a potential generalizability challenge when

Table 1 Deconvolution algorithms developed for bulk transcriptomics with sc/snRNA-seq reference datasets. The table includes the name and reference (column 1) along with the year published (column 2) and a description (column 3) of the algorithm. The primary tissues used in the publication associated with the algorithm are also provided (column 4)

Algorithm	Citation	Year	Description	Primary publication tissues
BayesPrism	[29]	2022	Bayesian approach, joint posterior inference and posterior summing over cell states, explicit cell type expression modeling	Blood, multiple cancer types
Coex	[30]	2022	Marker co-expression networks and network module attribution	Brain
MuSiC2	[22]	2021	Differential marker weighting and filtering on condition-specific differential expression	Pancreas and retina
SCDC	[31]	2021	Ensemble framework to integrate references across sources	Pancreas and mammary gland
Bisque	[32]	2020	Gene-specific transformations to address assay-specific biases	Adipose and brain
DWLS	[33]	2019	Dampened weighted least squares, rare cell type detection	Blood, tumor/melanoma (human); kidney, lung, liver, small intestine (mouse)
MuSiC	[28]	2019	Differential marker weighting to address marker expression confounding	Pancreas and kidney
dtangle	[34]	2019	Marker selection with linear mixed modeling	Blood, breast, brain, liver, lung, muscle, cancer
ABIS	[35]	2019	Absolute deconvolution with cell scale factors on TPM-normalized marker expression	Blood and immune cells
quanTIseq	[36]	2019	Non-negative regression with cell factor scaling and unknown cell type estimation	Blood and tumor
Fardeep	[37]	2019	Machine learning with adaptive trimmed least squares	Immune cells [38], tumor cells (GSM269529)
BrainInAblender	[20]	2018	Prediction with mean marker expression across references	Brain, pyramidal neurons, stem cells, immune cells, blood cells
xCell	[39]	2017	Linear scaling of marker enrichment scores	Immune, stem, epithelial, and tumor cells
EPIC	[40]	2017	Renormalization of reference markers by cell scale factors, quantification of unknown types	Cancer and blood
MCP-counter	[41]	2016	Cell type amount scoring for heterogeneous tissues, numerous cell types, and multiple clinical conditions	Immune, stromal, and tumor cells and cell lines
TIMER	[42]	2016	Batch effects removal from tumor purity markers; constrained least squares with orthogonal validation	Multiple tumor types
CIBERSORT	[43]	2015	Machine learning-based dimension reduction and permutation optimization	Blood
DCQ	[44]	2014	Whole transcriptome regularized regression followed by ensemble selection, with focus on cell surface marker genes	Lung and immune cells
DeconRNASeq	[45]	2013	Linear modeling, non-negative least squares, and quadratic programming	Brain, heart, skeletal muscle, lung and liver

algorithms (Table 1) are only trained on one type of tissue sample (e.g., healthy/control samples) and not on tissues with the observed phenotype or outcome (e.g., disease samples). It was previously shown [22] that differential expression (DE) between group conditions can limit the utility of a normal tissue reference to accurately deconvolve cell type abundances in a disease condition. Including multiple phenotypes can also avoid algorithm overfitting, encourage the selection of better cell type markers, and boost the overall generalizability of findings. Ideally, cases should be matched to the reference samples on potentially confounding factors like subject demographics, tissue collection procedures, and specimen handling strategies.

Need for orthogonal measurements to form a reference atlas (Z) across multiple donors

A key experimental design consideration is to select the sc/snRNA-seq samples used to build a reference atlas (Z). For example, a reference atlas (Z) could contain data from multiple donors or from only tissue samples that have matched bulk and sc/snRNA-seq samples. This decision depends on the specific research question, the statistical power to detect cell types [46], the availability of previously published data [5], and the cost of generating new data [47]. Multi-group references can mitigate the low reliability of cell type proportion estimates from a single sc/snRNA-seq sample [22]. As sc/snRNA-seq data is characteristically sparse, pooling cells across groups can further boost power to characterize rare, small, or less active cell types [46, 48].

Need for measurements of cell type composition from orthogonal platforms

The primary gold standard measurement to evaluate the accuracy of estimated cell compositions from a deconvolution algorithm is an orthogonal cell type fraction measurement (Table 2) in the tissue sample, and this should ideally be known with high accuracy and reliability. In multiple tissues including blood and brain, fluorescence-activated cell sorted (FACS) RNA-seq [6, 49] and DNAm microarray data [3, 50] have been used as orthogonal measurements of “true” cell composition. Cell type proportion estimates based on relative yields from sc/snRNA-seq data are not likely to be reliable [6] because of dissociation bias [26] and incomplete representation of sequenced cells (i.e., only a subset of the sample is sequenced). This bias impacts the “true” cell composition yield in a cell type-specific manner [51], is not present in bulk RNA-seq data, and can explain systematic expression differences between bulk RNA-seq data [32]. However, orthogonal cell type measures could be extracted from many different data types (Table 2), including microscopy images from molecular marker-based protocols such as single-molecule fluorescent in situ hybridization (smFISH) [3]. These image-based technologies could offer an opportunity to characterize cell type proportions, as well as other size/shape measurements directly from the tissue. Furthermore, these images can then be integrated with transcriptome-wide gene expression measurements based on emerging spatial transcriptomics technologies [52–55].

Challenge 2: cell types vary in abundance, size, and total mRNA

Cell types exhibit a wide range in size and function within and across human tissues

Most eukaryotic cells are between 10–100 μm in diameter, for example ranging from red blood cells (8 μm), skin cells (30 μm), and neurons (up to 1 m long) [64]. In particular,

Table 2 Orthogonal cell type amount measurements used for bulk transcriptomics deconvolution. Table describes the name (column 1) and a description (column 2) of the type of measurement, the type of assay used to capture the measurement (column 3), and example citations for these measurements (column 4)

Name	Description	Assays	Citations
Fluorescent in situ hybridization (FISH)	Labeling and imaging of DNA-based cell type markers	In situ labeling, imaging	[3, 56]
Immunohistochemistry (IHC)	Antibody-based cell marker labeling and imaging	In situ labeling, imaging	[40, 57]
Immunofluorescence (IF)	Antibody-based fluorescent labeling of cell markers	In situ labeling, imaging	[3, 58]
In vitro cell mixtures	Sequencing of manually mixed cells from dissociated bulk tissues or cell lines	Bulk RNA-seq	[30, 31, 38, 44, 45, 59]
Fluorescence-activated cell sorting (FACS)	Sequencing of cells isolated by cytometric sorting	Flow cytometry; bulk RNA-seq	[6, 35, 40]
Genetic panel	DNA marker-based differentiation of tissues, esp. tumor from non-tumor	Genetic marker assay; microarray	[39, 60]
DNA methylation	Deconvolution using DNA methylation cell type markers	Microarray; bisulfite sequencing	[3, 42, 50, 61–63]
Hematoxylin and eosin staining	Clinical tissue slide staining procedure	In situ staining; imaging	[42, 56]

the brain is an excellent example of a tissue exhibiting a wide range of cell types with different sizes and morphologies [9, 65]. Within the brain, there are a diversity of cell types that fall into several broad categories, including neurons, glia, and vasculature-related cells. These cell types have distinct functions reflected by differences in morphology, physiology, cell body size, and molecular identity. For example, neurons are larger and more transcriptionally active than glial cells [2]. Vasculature-related cells, including endothelial cells, smooth muscle cells, and pericytes that comprise the building blocks of blood vessels and are also smaller in size than neurons [66]. These cell types have specific genetic programs that facilitate distinct functions [66]. For example, neurons (larger excitatory glutamatergic neurons and smaller inhibitory GABAergic neurons [67]) are bigger in size and less numerous than glial cells, a heterogeneous group of cells comprised of oligodendrocytes (Oligo) (20–200 μm) [68], oligodendrocyte precursor cells (OPC) (50 μm) [69], microglia (15–30 μm) [70], and astrocytes (Astro) (40–60 μm) [71], which serve many roles, such as myelination, immune signaling, and physical and metabolic support. This extensive cell type diversity found in the brain, and other tissues, underscores the motivation for adjusting for differences in cell sizes prior to performing deconvolution (see data sources in Table 2).

Cell-type scale factor transformations can improve the performance of deconvolution algorithms

While bulk transcriptomics deconvolution commonly predicts cell type proportions from expression data, it was noted that this approach may instead quantify total mRNA content in the absence of an adjustment for systematic differences in size and expression activity at the cell type level [3]. This adjustment, which we will call a “cell type scale

factor transformation” (or cell scale factors for short), is used to transform the cell type reference matrix (Z) data prior to deconvolution [3, 72]. Consider the following standard mathematical formula $Y_{G \times J} = Z_{G \times K} * P_{K \times J}$ with dimensions for G marker genes, J bulk sample(s), and K cell types, which we drop the dimensions after this point for brevity. Assume we have $S_{K \times K} = I_{K \times K} * s_K$, where S is a matrix, $I_{K \times K}$ is an identity matrix, and s_K is a vector of scalars $s_{1...K}$ that refer to the size, such as the average mRNA molecules in a cell, for each k^{th} cell type, which are often experimentally derived (Supplemental Fig. 1). Then, the formulation to deconvolve Y with cell type scale factors S is described as $Y = Z * S * P$, where we can define a new $Z'_{G \times K} = Z_{G \times K} * S_{K \times K}$, and then we see a formulation similar to the standard formula as before: $Y = Z' * P$. It is worth noting that without this transformation, the assumption made by existing deconvolution methods is that cell types are equal sizes, but incorporating this transformation enables models to assume cells have different sizes. Deconvolution accuracies improved when S was calculated from a tissue-matched independent reference [35] and even if cell size estimates were from distinct organisms [3].

Cell size scaling was initially introduced for microarray-based expression data [72, 73] and later used for scRNA-seq data in multiple tissues [3, 35, 40]. Cell scale factors are frequently used to generate sc/snRNA-seq-based data that resemble real bulk RNA-seq data based on “pseudobulking” or aggregating molecular profiles across sc/snRNA-seq data [74]. Reference atlas transformation using orthogonal and non-orthogonal cell scale factors reduced errors from deconvolution-based cell proportion predictions. This may be because estimates without this transformation quantify total RNA rather than cell proportion [3]. Cell scale factors may be estimated from either expression or expression-orthogonal data, such as sorted or purified populations of immune cells, which are used in existing deconvolution algorithms such as *EPIC* and *ABIS* [35, 40]. The algorithms *MuSiC* and *MuSiC2* [22, 28] can use either expression-based or user-defined scale factors. Similar algorithms using variance-based marker weighting, such as *SCDC* [31], do not incorporate cell scale factors, but have been made compatible with outputs from algorithms that do (Table 1). Importantly, there are currently no standards for applying cell scale factors prior to deconvolution, and users may need to transform the reference atlas (Z) prior to calling certain algorithms. Further, many algorithms have not been extensively tested in complex tissues, such as brain, that show large differences in size and transcriptomic activity across cell types. Ultimately, more reliable cell scale factor estimation and standardized transformation procedures can facilitate future deconvolution research [3, 72].

Different approaches to obtain cell scale factors can influence cell composition estimates

There are several approaches to estimate and scale cell types in application of deconvolution. Expression-orthogonal cell size estimation methods can come from, for example, fluorescent in situ hybridization (FISH) or immunohistochemistry (IHC) [3, 32, 67] (Table 3). Image processing softwares such as ImageJ/FIJI [75] and HALO (Indica Labs) can provide cell body or nucleus measurements, including diameter, area, perimeter, among other size features (Table 4). However, cell segmentation presents a key obstacle limiting the accuracy of imaging-based approaches, especially for cells with complex morphologies [76]. Expression-based cell size estimates are commonly calculated from

Table 3 Experimental data platforms to estimate cell sizes and calculate cell size scaling factors to adjust for systematic differences in size and transcriptomic activity between cell types. The table contains the type of experimental data (column 1), the metric used for cell size (column 2), a set of standards (gold, silver, and bronze) introduced by Dietrich et al. [20] (column 3), the format for how the data are captured (column 4), example data analysis challenges when using these data (column 5), and if the experimental data are orthogonal to using sc/snRNA-seq (column 6)

Experimental data	Cell size metric	Standard [74]	Data format	Data analysis challenges	Orthogonal to sc/snRNA-seq
FISH [4, 78–80]	Label intensity	Gold	Image	Label performance; cell segmentation; image artifact removal [22, 28, 35, 40, 74]	Yes
IHQ/IHC [36]	Label intensity	Gold			Yes
Labeled expression marker [79, 80]	Expression/label intensity	Silver			Yes
sc/snRNA-seq	mRNA spike-in expression	Silver	Gene expression counts	Embedding alignment, batch effects, dissociation biases, platform biases [26, 48, 81]	Yes
sc/snRNA-seq	Housekeeping gene expression	Silver			No
sc/snRNA-seq	Library size [36, 78, 82]	Bronze			No
sc/snRNA-seq	Expressed genes [36, 78, 82]	Bronze			No

total mRNA counts, often referred to as “library size factor” [77], which are typically unique to each cell, but could also be considered distinct for each cell type (Table 1). However, these estimates may be confounded by either the total sequenced RNA or genes with outlying high expression [35]. For this reason, total expressed genes may be a good alternative robust to this type of confound. Cell scale factors from sc/snRNA-seq data are further subject to bias from tissue dissociation, cell compartment isolation, and other factors that have cell type-specific impacts [22, 28, 31]. Another consideration is the application of cell scale factor transformations, as published deconvolution algorithms apply scale factors before [28] or after [72] prediction of cell type proportions. Application of cell scale factor transformation to the reference atlas (Z) may prevent quantification of total RNA rather than cell proportions [3]. In summary, cell scale factor transformations can improve bulk transcriptomics deconvolution across multiple species, tissues, and sequencing platforms.

Challenge 3: protocol bias for tissue processing impacts reference atlas (Z)

Acquisition of data with single-nucleus (sn) versus single-cell (sc) RNA-seq protocols

Similar to donor-to-donor variation leading to unwanted confounds in estimating the cell type reference matrix Z , sampling RNA from different cellular compartments can also introduce unwanted variation. For example, experimenters can perform “single cell” sequencing by isolating either whole cells (containing both nuclear and cytoplasmic RNA, often performed from fresh tissue) or just the nuclear compartment (containing only nuclear RNA, performed from frozen tissue). While it has been demonstrated that nuclear RNA is representative of RNA from the whole cell [83, 84], there can be substantial differences for certain transcripts thereby introducing variability into the data contained in Z . In the human brain, the majority of studies are conducted on fresh frozen post-mortem tissue rather than fresh tissue. When post-mortem brain tissues are flash

Table 4 Cell scale factor estimates from the literature, with focus on deconvolution studies that use sequencing references. Values for blood cell types are from the SimBu R package (v1.2.0), and values for brain cell types are from Table 1 in (3). The Scale factor value (column 3) can be used in existing deconvolution algorithms leading to less biased results for estimating cell composition

Cell type	Tissue	Scale factor value	Scale factor type	Scale factor data source	Citation(s)
Glial	Brain	91	Cell area	osmFISH	[3, 80]
Neuron	Brain	123	Cell area	osmFISH	[3, 80]
Glial	Brain	180	Nuclear mRNA	osmFISH	[3, 80]
Neuron	Brain	198	Nuclear mRNA	osmFISH	[3, 80]
Glial	Brain	12,879	Library size	expression	[1, 3]
Neuron	Brain	18,924	Library size	expression	[1, 3]
B cells	Multiple	65.66	Median expression	Housekeeping gene expression	[36, 74]
Macrophages	Multiple	138.12	Median expression	Housekeeping gene expression	[36, 74]
Macrophages (M2)	Multiple	119.35	Median expression	Housekeeping gene expression	[36, 74]
Monocytes	Multiple	130.65	Median expression	Housekeeping gene expression	[36, 74]
Neutrophils	Multiple	27.74	Median expression	Housekeeping gene expression	[36, 74]
NK cells	Multiple	117.72	Median expression	Housekeeping gene expression	[36, 74]
T cells CD4	Multiple	63.87	Median expression	Housekeeping gene expression	[36, 74]
T cells CD8	Multiple	70.26	Median expression	Housekeeping gene expression	[36, 74]
T regulatory cells	Multiple	72.55	Median expression	Housekeeping gene expression	[36, 74]
Dendritic cells	Multiple	140.76	Median expression	Housekeeping gene expression	[36, 74]
T cells	Multiple	68.89	Median expression	Housekeeping gene expression	[36, 74]
B cells	Multiple	0.40	Intensity	FACS	[40, 74]
Macrophages	Multiple	1.42	Intensity	FACS	[40, 74]
Monocytes	Multiple	1.42	Intensity	FACS	[40, 74]
Neutrophils	Multiple	0.13	Intensity	FACS	[40, 74]
NK cells	Multiple	0.44	Intensity	FACS	[40, 74]
T cells	Multiple	0.40	Intensity	FACS	[40, 74]
T cells CD4	Multiple	0.40	Intensity	FACS	[40, 74]
T cells CD8	Multiple	0.40	Intensity	FACS	[40, 74]
T helper cells	Multiple	0.40	Intensity	FACS	[40, 74]
T regulatory cells	Multiple	0.40	Intensity	FACS	[40, 74]
B cells	Multiple	20837.57	Intensity	FACS	[35, 74]
Monocytes	Multiple	22824.32	Intensity	FACS	[35, 74]
Neutrophils	Multiple	9546.74	Intensity	FACS	[35, 74]
NK cells	Multiple	21456.91	Intensity	FACS	[35, 74]
T cells CD4	Multiple	14262.07	Intensity	FACS	[35, 74]
T cells CD8	Multiple	10660.95	Intensity	FACS	[35, 74]
Plasma cells	Multiple	325800.99	Intensity	FACS	[35, 74]
Dendritic cells	Multiple	57322.18	Intensity	FACS	[35, 74]

frozen during the preservation process, cells are lysed prohibiting the molecular profiling of whole single cells using scRNA-seq approaches. Instead, only nuclei are accessible for profiling using snRNA-seq approaches. While the nuclear transcriptome can serve as a proxy for the whole cell transcriptome [85–87] nuclear transcripts include more intron-containing pre-mature mRNA and may not include transcripts locally expressed in cytoplasmic compartments, such as neuronal axons and dendrites, or transcripts rapidly exported out of the nucleus [2]. On the other hand, compared to whole cells, nuclei are less sensitive to mechanical/enzymatic tissue dissociation procedures, which may artificially impact gene expression [26], and are suitable for multi-omic profiling such as combined RNA-seq and ATAC-seq from the same nucleus [88]. In fact, dissociation protocol differences can help explain variation in average nuclei per donor observed across brain snRNA-seq reference datasets [12]. While prior work showed only a small impact from cell compartment DE between bulk and snRNA-seq data, accounting for this slightly improves deconvolution accuracy [30]. However, new computational methods are being developed to remove these protocol-specific biases [28].

Tissue preparation protocols can impact the diversity and quality of cells profiled during sc/snRNA-seq

Cell type-specific associations between dissociation treatment and gene expression were observed from sc/snRNA-seq data across multiple tissues and species [26]. Expression patterns may further be influenced by the specific cell/nucleus isolation protocol utilized [26, 89]. There are several approaches for isolating nuclei from frozen tissues and removing debris from homogenization steps. While some studies employ a centrifugation-based approach with gradients of sucrose or iodixanol to purify nuclei from debris [90, 91], others use fluorescence-activated nuclear sorting (FANS) to label and mechanically isolate single nuclei [92, 93]. FANS also allows for enrichment of distinct cell types by implementing an immunolabeling procedure for populations of interest prior to sorting. There are pros and cons to each of these nuclei preparation approaches. FANS gating strategies may bias towards certain cell sizes and influence the final population of profiled cells. In the brain, recent work highlighted advantages for sorting approaches that remove non-nuclear ambient RNA contaminating glial cell populations [94]. Ultimately, tissue dissociation protocols can drive variation among and between sc/snRNA-seq populations.

Choice of sc/snRNA-seq platforms can impact reference gene expression profiles

There are several sequencing platform technologies to generate sc/snRNA-seq reference profiles. While these have been previously reviewed [47, 81], it is important to note that the different sample preparations and chemistries required for each of these platforms impact the downstream gene expression data. For example, the widely used single-cell gene expression platform from 10 × Genomics is a droplet-based approach offering a 3′ or 5′ assay for up to 10,000 nuclei/cells in a single pooled reaction [95]. While the 10 × Genomics platform allows profiling a large number of cells in a single experiment, a major limitation is the sparsity of data and restriction of coverage to one end of the transcript. This is in contrast to approaches such as SMART-seq [96] from Takara, which offers full-length transcriptome analysis, but requires isolation of nuclei into individual

tubes for separate reactions, thereby often resulting in fewer total cells profiled. Other technologies are rapidly becoming available for *sc/snRNA-seq* approaches, and each of these can introduce different biases into reference data. Importantly, recently published deconvolution algorithms use data transformation strategies to adjust for these biases [28, 32].

Potential differences in library preparation strategies for bulk RNA-seq and *sc/snRNA-seq* data

Library preparation is a crucial protocol step impacting RNA profiles in RNA-seq data. Factors of library bias in RNAseq expression have been well documented and include library prep base composition [97], fragmentation bias [98], 3' direction bias [99], and lacking template DNA [100]. The two most popular library preparation strategies are ribosomal RNA (rRNA) depletion [101, 102], where rRNA is removed and remaining RNA sequenced, and polyA-enrichment [103], where polyA mRNA is isolated and sequenced. The former strategy can isolate a more diverse RNA population, including pre-mature and alternatively spliced mRNAs lacking polyA tails, and non-protein encoding RNAs [104, 105]. This difference may drive protocol bias that needs to be accounted for [106]. Library preparation strategies may differ between bulk and *sc/snRNA-seq* data used for deconvolution. While polyA-enrichment was initially common for bulk RNA-seq, many newly available datasets now use rRNA depletion. By contrast, with the accessibility and popularity of the *sc/sn* droplet-based technologies [95], many reference atlases (Z) are based on polyA-enrichment. Further, marker genes may not be consistently expressed across different library preparation conditions, which can reduce deconvolution accuracy. A recent benchmark showed library size normalization was crucial for RNAseq deconvolution [107]. Computational tools [108–111] for correcting library preparation biases include specialized tools for particular bias types, such as DNA sequence-specific bias correction [112, 113]. As newer deconvolution algorithms accept large marker gene sets, systematic RNA population differences between library preparation strategies likely need to be accounted for, warranting further investigation.

Assay-specific biases between bulk and *sc/snRNA-seq* data

Systematic differences between bulk RNA-seq and *sc/snRNA-seq* assays can increase errors and reduce the utility of estimated cell type abundances from deconvolution algorithms. Assay-specific bias is more generally defined than library preparation bias and may arise from differences in sample processing protocol (e.g., cDNA synthesis, PCR amplification, UMI versus full-length transcript), sequencing platform (e.g., short- versus long-read, droplet- or microfluidics-based), and cell compartment isolation (e.g., whole cell, only cytoplasm, or only nucleus) [25, 114]. Different sequencing technologies also show varying transcript length bias, which increases the power to detect highly expressed long transcripts over low expressed short transcripts [115, 116]. This bias can impact the genes and pathways identified from DE analyses [117, 118]. While the use of unique molecular identifiers (UMIs) protocols [116, 119] may reduce the extent of transcript length bias in *sc/snRNA-seq* data relative to bulk, it may persist from internal priming, a type of off-target polyA primer binding [120]. Furthermore, unlike bulk RNA-seq datasets, *sc/snRNA-seq* data are highly sensitive to both cDNA synthesis and PCR

protocols [25]. Great improvements to both protocols have been made in recent years [121, 122]. Finally, bulk and sc/snRNA-seq data show distinct distributional properties that may impact downstream analyses and the utility of simulation approaches [74, 123]. Dispersion, or the extent of inequality between expression variances and means, is among the most important of these [124]. Bulk RNA-seq expression may show less dispersion, and thus may be modeled either using a Poisson or negative binomial [125] distribution, while expression sparsity and heterogeneity in sc/snRNA-seq data increases dispersion and often motivates the use of the negative binomial distribution [126, 127]. Orthogonal sample collection protocol can limit bias between bulk and sc/snRNA-seq data (Challenge 1). Computational methods are available that have been specifically tested across assays [113, 118, 128], data modalities [129], and preparation protocols [130]. Finally, new analyses of existing datasets can reveal new assay bias sources like cell stress from hypoxia [131].

Differences in detectability of rare cell types across batches and assays

Because cell type detection from sc/snRNA-seq data is confounded by low expression levels, downsampling sc/snRNA-seq profiles on library size is often performed prior to downstream analyses [132]. Recently introduced normalization strategies can further increase the reliability of rare cell type quantification [18], and similar approaches are already being applied to newer spatial sc/snRNA-seq datasets [133]. This may be especially useful for complex heterogeneous tissues like brain, where previously noted protocol biases limit the amount of available reference data for rare cell types [9]. In general, uncommon or rare cell types do not have a large impact on abundant cell type predictions unless there is high expression collinearity between gene markers of rare and abundant cell types [8]. In the human brain, deconvolution accuracy decreased substantially with the exclusion of neurons, but not less common glial cell types [30]. Importantly, the low-end limit for reliable cell type proportion predictions was found to vary across deconvolution algorithms [7]. Computational tools for rare cell type identification have been developed for rare immune cell populations [134], myeloid progenitor cells [135], and rare brain cells [136]. These used a variety of statistical modeling techniques, including latent variable models, such as scLVM [134], and dampened weighted least squares (DWLS) [33], which outperformed several other methods in a recent benchmark [8].

Challenge 4: standardization of cell type annotation and marker selection strategies

Standard brain cell type definitions and nomenclature are complex and emerging

As new cell type-specific molecular and functional datasets rapidly come online, our understanding and definition of cell type diversity is evolving. In the context of the brain, key factors impacting our understanding of distinct cell populations [137] include (1) discovery and improved molecular characterization of functionally distinct cell types in brain regions and subregions, (2) new insights into how physiology and connectivity impact neuronal identity, and (3) an improved understanding of how cells change during development and aging. Anatomical and spatial position also influences cell type gene expression. For example, while virtually all excitatory populations in the cortex are glutamatergic pyramidal neurons, they show strong molecular and morphological differences

across cortical layers [27] and still further differences with glutamatergic populations in other brain regions such as the hippocampus and amygdala [92]. This underscores the necessity for a common cell type nomenclature to organize cell type labels and pair these with key contextual features like tissue microenvironment [137, 138]. Reviews of cell type label management strategies highlight challenges with reconciling types and subtypes [139] as well as with tracing cell identities to anatomic and spatial regions (Fig. 1 in [129]). Specialized computational tools automate [140, 141] cell type label assignments across data sources, including scType [142], scAnno [143], scReClassify [144], and neural network-based tool NeuCA [145]. Further, as new data emerge and cell type nomenclature evolves, reference datasets will likely need to be revisited and modified accordingly to ensure their utility.

Cell-type resolution should be experimentally driven

Given that cell type definitions can be complex and defined at multiple resolutions (i.e., as either broad cell classes or as fine subpopulations), the resolution for a given deconvolution analysis needs to be experimentally motivated. That is, the ideal cell type resolution may differ depending on the biological question under investigation. For certain applications, such as distinguishing the contribution of two adjacent brain regions to a given bulk RNA-seq sample, relatively coarse definitions of neurons and glial cells may be adequate. For other applications, such as understanding the contribution of different neuronal cell types to differential gene expression between healthy and disease samples, fine-resolution cell types may be required. An important first step for deconvolution is deciding the appropriate cell type resolution to address the underlying biological question. Prior work in human blood utilized an optimization procedure to identify the 17 most optimal blood and immune cell types for deconvolution from 29 total candidate cell types [35]. In the human brain, it was found that the definition of the reference atlas (Z) is more important than the choice of deconvolution algorithm, and accordingly, the target cell types should have expression data of sufficient quality to select the most optimal marker genes possible [30]. Cell-type resolutions are systematically set and benchmarked to understand deconvolution algorithm performances and generalizability [6]. Higher resolutions typically show worse performance [8], while the exclusion of abundant cell types often has a more detrimental impact than the exclusion of rare cell types [30, 35]. Algorithms such as BayesPrism [29] were designed to handle multiple cell subtypes implicitly and automatically without requiring the user to specify K dimensions for each resolution, and these could be preferred when robust cell type markers are lacking or cell type identity is particularly uncertain or heterogeneous.

Cell type definitions should be based on robust and identifiable expression data

One of the key conditions of a successful deconvolution experiment is that the cell types of interest are identifiable in the sample type(s) of interest. For a cell type to be identifiable, it should be sufficiently abundant and have clear gene markers. Gene markers should have a sufficient expression to be distinguishable from the background (i.e., relative high expression and sufficient read depth), as well as from other cell types of interest (i.e., sufficient DE from other cell types, with other cell types ideally having none or very low expression) [7]. While reference-free deconvolution algorithms [23, 24, 146] do

not rely on specific reference marker genes to the same degree as reference-based algorithms, the suitability of available expression data to perform deconvolution with high accuracy is a key issue across algorithm types and needs to be carefully considered.

Even with appropriate cell type definitions and evidence from expression data, the issue of defining the total cell types (K) to predict in a sample presents its own challenge. If the cell types in the reference do not reflect the cell types in the bulk or pseudobulk sample, deconvolution accuracy can suffer [8]. Given a set of more than two well-defined cell type labels, it is also reasonable to ask whether we should deconvolve all cell types together, or whether similar cell types should be binned prior to attempting deconvolution. For example, suppose an expression dataset contains cells with the Excit, Inhib, Oligo, and Astro cell type labels. From these, we could define the following $K=4$ types, each with its own reference atlas: (1) neuronal (i.e., excitatory and inhibitory) and non-neuronal (i.e., Oligo and Astro); (2) Excit, Inhib, and non-neuronal; or (3) Excit, Inhib, Oligo, and Astro. Recent deconvolution studies have advanced our understanding of how cell type label definitions impact deconvolution outcomes. In both blood [35] and brain [30], iterative assessments may lead to the effective quantification of relatively specific cell types and the exclusion of others. Efforts to bin and evaluate cell type definitions should be considered alongside strategies to identify the cell type-specific gene markers for the reference. Marker identification methods may be based on differences in differentially expressed genes, such as Wilcoxon rank-sum statistics, and clustering, to name a few [147].

Expression markers of disease may confound signature atlas reliability

A further consideration for bulk deconvolution methods is heterogeneity introduced by disease state that may influence marker gene expression. As many algorithms are intended for use in bulk tissue samples from disease states, it is important to understand how illness may uniquely impact cell types and their expression of marker genes. Gene expression atlases [148, 149] can be used to identify cell-type marker genes. Development of these atlases parallels technological advances and led to the generation of the first whole-organism atlases in an attempt to collect more assays matched to an individual organism than ever before [148, 150]. However, many new tissue- [148, 149, 151] and condition-specific [11, 152, 153] atlases have been generated from sc/snRNA-seq data that may be limited by differential expression across conditions [13, 14, 57]. For example, in samples from individuals with Alzheimer's disease (AD) relative to neurotypical control subjects, neurons show marker gene repression, while glial cells generally show up-regulation of marker genes [11]. Changes in gene expression have also been reported for psychiatric disorders such as major depression, where prior work showed 16 cell types with altered expression including excitatory and OPC cell types [10]. Computational interfaces [154] and tools, including scGen [155], scVI [156], and scANVI [157], enable single-cell RNA-seq reference use by managing unwanted variation and between-source marker performance and reliability. Further, new tools [129] and standards [137] facilitate the management of reference atlases from newer technologies measuring combined modalities, such as transcriptomes and proteomes from the same cell. Given that disease-specific differential expression [22, 153] can interfere with the effectiveness of cell-type signature matrices, cell-type marker genes selected for deconvolution should

show equivalent expression between healthy and disease conditions. If expression is not equivalent between conditions, further adjustments to either the reference marker or bulk expression data may be necessary and collection of orthogonal matched assays facilitates these efforts (Challenge 1, Fig. 3).

Challenge 5: reference atlases (Z) should be built on standardized and state-of-the-art computational tools and file formats

Standardized data-driven cell type labels can facilitate deconvolution advances

As discussed above, effective cell type definitions are crucial for deconvolution success. As more data comes online (Fig. 4), there is an increasing need for uniform labeling of cell types [9] and careful documentation of study metadata, including cell type enrichment methods [158, 159]. For example, in the brain, anti-NeuN antibodies are commonly used to enrich neuronal cell populations during FANS [160]. Cataloging cell markers and the reagents used to select specific cell types will be important for standardizing data collection practices. On the data analysis side, sc/snRNA-seq cell type labels

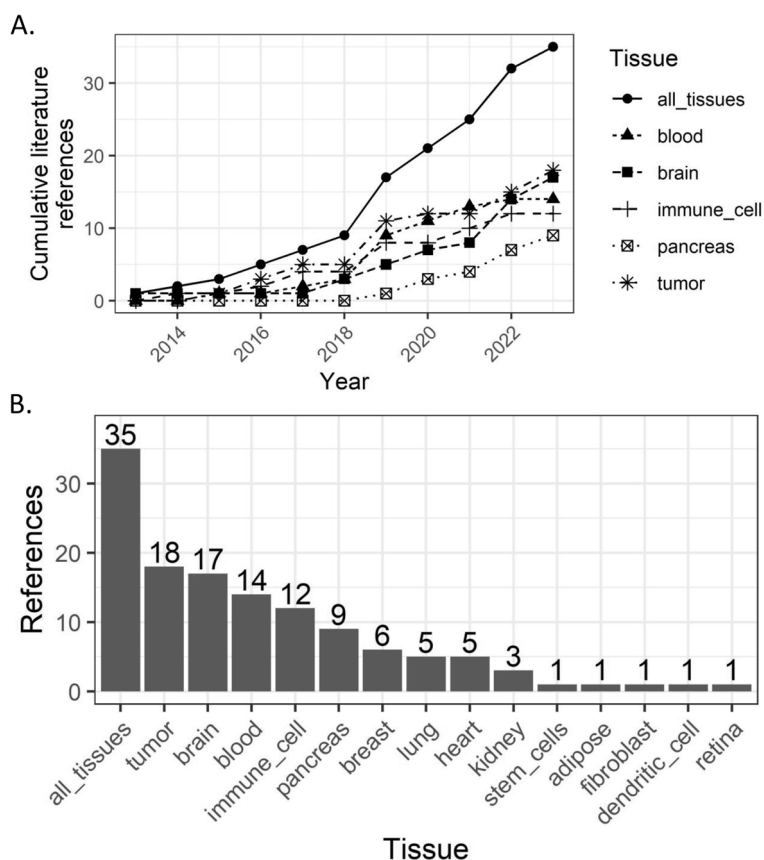


Fig. 4 Summary of tissues by literature reference from bulk transcriptomics deconvolution literature. **A** Dot and line plot of (x-axis) yearly (y-axis) cumulative references by (shape, line type, label) tissue, including (circle, solid line, "all_tissues" label) the combined set of all tissues, (triangle, short dashed line, "blood") blood tissue, (square, middle dashed line, "brain") brain tissue, (plus, long dashed line, "immune_cell") immune cell tissue, (box, dotted line, "pancreas") pancreas tissue, (asterisk, dotted line, "tumor") tumor tissue. **B** Barplot showing (y-axis) the number of literature references (x-axis) per tissue, including ("all_tissues" label) the combined set of all tissues. Plots were created using the ggplot2 (v3.4.1; [16] software; data used to reproduce these plots are available from GitHub (Data Availability))

may be derived from clustering [35, 92, 154], reference-based tools [20, 161], or other analytical approaches [7, 78, 162]. In these cases, cell type labels could be indexed with a link to their originating annotation method. Further, hierarchical organization of cell type descriptors can facilitate insights into their molecular and physiological properties. Examples of this practice include term ontologies from the ENCODE project (<https://www.encodeproject.org>), common cell type nomenclature [137], and the Human Cell Atlas (HCA) [163], and it can be leveraged for cell type marker selection [162]. Combining key analysis and definitional metadata with standardized cell type labels can encourage reproducibility and new analyses. An individual sc/snRNA-seq experiment can use either manual [42, 56], computational [155–157, 164], or combined [165] approaches to assign cell type labels or ascertain their abundances (Table 2). For certain tissues [166] and conditions [152], consulting external references can narrow the cell type or subtype definition according to its expected properties [151]. Despite the availability of existing protocols, recommendations may disagree, protocols may perform suboptimally in a particular experiment, and it may be difficult to reconcile conflicting recommendations. The choice of marker genes impacts downstream analyses [143, 145, 155, 156]. For these reasons, more standard approaches [138] and flexible analysis strategies are needed.

Expression data needs to be published using state-of-the-art data science formats

Publishing key datasets and results with essential documentation using standard data formats is an important part of reproducible computational research [167–170]. Generating, hosting, and distributing large volumes of transcriptomics data and metadata at scale and in a reproducible manner demands substantial time and resources [150, 171], and specialized technologies facilitate this effort [148, 172, 173]. While flat table files (e.g., files with.csv or.tsv extension) are most common, many other data formats allow rapid and memory-efficient access [174, 175] to reduce computing time and resources for access and analysis. Some important examples include relational database formats (e.g., structured query language [SQL], hierarchical data format 5 [HDF5]). For example, the active memory required to load expression counts from 77,604 cells and 36,601 genes is 9.32 MB as an HDF5 file using the DelayedArray framework in Bioconductor table versus 22.72 GB as a standard matrix (Data Availability). Specialized data formats are compatible with increasingly used cloud servers and remote computing environments [176], and large-scale data efforts [5, 177] to comprehensively compile and analyze microarray and sequencing data from the Sequence Read Archive [172], Gene Expression Omnibus [173], Array Express [178], and other public repositories has led to publication of HDF5-based data compilations and Bioconductor packages. Further, many of the 246 [179] actively maintained Bioconductor software packages with the SingleCell descriptor use specialized data formats [77, 180]. These include the *SummarizedExperiment* format for most omics data types [180], and the *SingleCellExperiment* format for sc/snRNA-seq expression data [77, 181], which is being extended for use with image coordinate information from spatial transcriptomics experiments [27, 65, 182, 183]. Further, the Azimuth project [149] provides cell reference datasets for multiple tissues as SeuratObjects [48], another data class specialized for sc/snRNA-seq data. Newer data formats may be subject to updates that introduce errors or conflicts with other data classes, and

resolving data class conflicts frequently demands a high degree of technical knowledge. This is one reason it is important to publish versions along with packages and object classes, in case an older version needs to be used while a newer version is updated. In summary, sequencing data may be published in a variety of formats to facilitate access, methods should include details like versions for computational tools that were used, and researchers should be aware of the many available technologies for reproducible transcriptomics analyses and data sharing.

Challenge 6: improving algorithm and signature atlas generalizability to new bulk tissue conditions

Cross-validation can limit algorithm overfitting and improve algorithm generalizability

Developers of new deconvolution algorithms and studies seeking to benchmark existing approaches must consider statistical power [184] and generalizability [185]. Here, power refers to the ability to detect cell type markers from DE analysis and differentiate between significantly different cell type proportions [46] and generalizability refers to the replicability of the experiment [167, 186]. For example, an experiment showing good algorithm performance in terms of accurate cell composition estimates and reliable cross-group comparisons could also perform well when analyzing additional data from an independent data source or new participant population. To encourage generalizability and reduce chances of algorithm overfitting to training data, cross-validation should be performed whenever possible, even if sc/snRNA-seq reference data is only available from relatively few sources [186, 187]. As mentioned previously, subjects and sample characteristics should further be balanced across experimental groups, as imbalances could bias the results or undercut their generalizability [13].

Developers should account for the tissues and conditions in which new algorithms will be applied

Deconvolution algorithms have varying performance across tissues and conditions, which we will call “domains,” and algorithms may be considered either general (e.g., good performance across domains) or domain-specific (e.g., good performance in a specific domain). Further, algorithm assumptions may vary depending on their intended domains of use. For example, algorithms often assume good markers are known for each type when developed with normal tissues [7] but algorithms for bulk tumor deconvolution may assume no tumor cell type markers are available [36, 40, 72]. As algorithms are often developed in a single or constrained domain set and then benchmarked in new domains, certain programming practices can facilitate algorithm testing across domains. For example, functions for algorithms like EPIC [40] and MuSiC [28, 72] flexibly support either default or user-specified cell scale factors, which may encourage more standard application of these adjustments in deconvolution experiments. Deconvolution algorithms (Table 1) such as dtangle [34], SCDC [31], Bisque [32], and DWLS [33] that do not support user-specified cell scale factors could instead use a transformed or rescaled reference (Challenge 2). Ultimately, developers should carefully consider the scope and nature of the domain(s) in which an algorithm will be applied.

Deconvolution algorithms should be optimized for prediction across conditions of interest

Beyond understanding normal tissue expression dynamics, effective deconvolution can allow new hypothesis-testing to elucidate relationships between cell types and disease mechanisms. Of particular interest in brain research is the prospect of studying significant changes in the abundances of neurons and/or glial cells between neurotypical samples and neurodevelopmental, neuropsychiatric, and neurodegenerative disorders, including autism spectrum disorder (ASD), Parkinson's disease (PD), and AD. Glia-specific inflammation in AD is detectable from snRNA-seq data, and further studies could reveal biomarker candidates and risk factors with utility for patient prognosis or diagnosis [14]. Microglial activation has been correlated with AD severity, illuminating mechanisms related to disease progression [32]. Total neuron proportion may decline in AD brains and reflect neuronal death as a hallmark symptom of AD; this trend was detectable in bulk tissue using multiple deconvolution methods [32]. Finally, accurate cell type quantification in case/control studies of bulk tissues revealed 29 novel differentially expressed genes in ASD that were independent of cell composition differences [30]. As new data and algorithms are published, more practical guidelines [6, 7] will be needed to match the most appropriate strategies to their specific biological questions. Specific protocols may be consulted to effectively deconvolve specific cell types across conditions, such as for deconvolution of reactive microglia across brain tissue from donors having neuropsychiatric conditions including Alzheimer's or epilepsy [187]. Marker gene reference atlases across more conditions [11, 87, 152] could also be consulted and utilized, though new standards for systematic cross-condition atlas utilizations may be needed to reconcile expression differences across conditions [13, 14, 57] (Challenge 4). Finally, a likelihood-based approach that utilizes confidence intervals for cell proportion predictions [34] could be extended to quantify prediction uncertainty across tissues and/or tissue donor conditions.

Future opportunities and recommendations

We wish to highlight several opportunities for bulk transcriptomics deconvolution in heterogeneous tissues, including the human brain. First, new reference datasets featuring multiple orthogonal assays from matched samples have huge potential to shape and inform new studies. Second, aggregation of published data into centralized repositories using standard data formats paired with structured and comprehensive metadata will increase the impact of new reference datasets and the reproducibility of analyses based on these reference data. Finally, mitigating biases and improving statistical rigor in sample collection, experimental design, and training new deconvolution methods should greatly improve the efficacy of new deconvolution algorithms and benchmarking of existing and emerging algorithms. Applying a transformation reference atlas (Z) matrix using cell scale factors, such as in Table 4, may reduce errors in deconvolution predictions due to improved quantification of cell proportions rather than RNA amounts [3].

Researchers can take several steps to act on these opportunities. First, even studies with a small number of donors can improve their rigor by running technical replicates (i.e., multiple runs of the same assay) and biological replicates (i.e., multiple distinct samples or tissue blocks from the same donor). Further, deconvolution algorithms

can be deployed as high-quality open-access software packages and made available in centralized curated repositories such as CRAN or Bioconductor [180]. Finally, new research efforts can utilize existing references to perform validation and inform the collection of new samples.

Conclusions

While the rapidly evolving future of transcriptomics is promising, it will be important to not only address existing experimental and computational challenges in this field, but also anticipate future challenges. Orthogonal assays are important for deconvolution experiments (Table 2, Supplemental Fig. 1), allow for biological variation to be systematically studied and modeled (Challenge 1), and are more readily managed and analyzed thanks to specialized open-access technologies (Challenge 5). We have drawn on our collective research experience to detail the key challenges of designing experiments with technical and biological replicates, effective use and integration of orthogonal assays, performance of data analyses to improve statistical rigor and generalizability of findings, and publication of datasets with comprehensive and structured metadata and methods with runnable and versioned code. Taking proactive steps to address these challenges will facilitate studies of increasing scale and complexity while encouraging greater reproducibility.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-03123-4>.

Additional file 1: Supplemental Figure 1. Schematic of collecting orthogonal assays from the same tissue block across donors and tissues.

Additional file 2. Review history.

Acknowledgements

We would like to thank Kelsey Montgomery, Sophia Cinquemani, and Keri Martinowich for the discussions and feedback of this manuscript. While an Investigator at LIBD, Andrew E. Jaffe helped secure funding for this work. Schematic illustrations were generated using Biorender.

Review history

The review history is available as Additional file 2.

Peer review information

Veronique van den Berghe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

SKM, KRM, and SCH wrote the initial draft and edited the manuscript. SHK and SKM prepared the figures. SKM prepared the tables. LCT and LAHM contributed to the conceptualization of the manuscript and provided comments on the draft. All authors approved the final manuscript.

Funding

This project was supported by the Lieber Institute for Brain Development, and National Institutes of Health grant R01 MH123183. All funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Code and data tables to reproduce panels in Figs. 1 and 4 and the memory usage example from Challenge 5 are available on GitHub (https://github.com/LieberInstitute/deconvo_review_paper, [188]) and Zenodo (<https://zenodo.org/records/10179283>, [189]). Cell size scale factors were compiled and provided as an R/Bioconductor annotations package on GitHub (<https://github.com/metamaden/cellScaleFactors>, [190]) and Zenodo (<https://zenodo.org/records/10149934>, [191]) to facilitate their reuse.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 May 2023 Accepted: 24 November 2023

Published online: 14 December 2023

References

- Darmanis S, Sloan SA, Zhang Y, Engle M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci USA*. 2015;112(23):7285–90.
- Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods*. 2017;14(10):955–8.
- Sosina OA, Tran MN, Maynard KR, Tao R, Taub MA, Martinowich K, et al. Strategies for cellular deconvolution in human brain RNA sequencing data. *F1000Res*. 2021;10:750.
- PsychENCODE Consortium, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, et al. The PsychENCODE project. *Nat Neurosci*. 2015;18(12):1707–12.
- Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol*. 2021;22(1):323.
- Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*. 2019;35(14):i436–45.
- Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*. 2018;34(11):1969–79.
- Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*. 2020;11(1):5650.
- Huang Q, Li Y, Xu C, Teichmann S, Kaminski N, Pellegrini M, et al. Challenges and perspectives in computational deconvolution in genomics data. *arXiv*. 2022. <https://doi.org/10.48550/arXiv.2211.11808>.
- Nagy C, Maitra M, Tanti A, Suderman M, Th eroux J-F, Davoli MA, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat Neurosci*. 2020;23(6):771–81.
- Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*. 2019;570(7761):332–7.
- Van den Oord EJCG, Aberg KA. Fine-grained deconvolution of cell-type effects from human bulk brain data using a large single-nucleus RNA sequencing based reference panel. *bioRxiv*. 2022. <https://doi.org/10.1101/2022.06.23.497397>.
- Lipska BK, Deep-Soboslay A, Weickert CS, Hyde TM, Martin CE, Herman MM, et al. Critical factors in gene expression in postmortem human brain: focus on studies in schizophrenia. *Biol Psychiatry*. 2006;60(6):650–8.
- Zhu Y, Webster MJ, Murphy CE, Middleton FA, Massa PT, Liu C, et al. Distinct phenotypes of inflammation associated macrophages and microglia in the prefrontal cortex schizophrenia compared to controls. *Front Neurosci*. 2022;30(16):858989.
- Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 2014;15(2):R31.
- Wickham H. *ggplot2: elegant graphics for data analysis (Use R!)*. 2nd ed. Cham: Springer; 2016.
- Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18):2847–9.
- Beck TF, Mullikin JC, NISC Comparative Sequencing Program, Biesecker LG. Systematic evaluation of sanger validation of next-generation sequencing variants. *Clin Chem*. 2016;62(4):647–54.
- Arteche-L pez A,  vila-Fern ndez A, Romero R, Riveiro- lvarez R, L pez-Mart nez MA, Gim nez-Pardo A, et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. *Sci Rep*. 2021;11(1):5697.
- Hagenauer MH, Schulmann A, Li JZ, Vawter MP, Walsh DM, Thompson RC, et al. Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *PLoS One*. 2018;13(7):e0200003.
- Doostparast Torshizi A, Duan J, Wang K. A computational method for direct imputation of cell type-specific expression profiles and cellular compositions from bulk-tissue RNA-Seq in brain disorders. *NAR Genom Bioinform*. 2021;3(2):lqab056.
- Fan J, Lyu Y, Zhang Q, Wang X, Li M, Xiao R. MuSIC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief Bioinformatics*. 2022;23(6):bbac430.
- Wang L, Sebra RP, Sfakianos JP, Allette K, Wang W, Yoo S, et al. A reference profile-free deconvolution method to infer cancer cell-intrinsic subtypes and tumor-type-specific stromal profiles. *Genome Med*. 2020;12(1):24.
- Charytonowicz D, Brody R, Sebra R. Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with UniCell deconvolve. *Nat Commun*. 2023;14(1):1350.

25. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet.* 2015;16(3):133–45.
26. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* 2020;21(1):130.
27. Maynard KR, Collado-Torres L, Weber LM, Uyttingco C, Barry BK, Williams SR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci.* 2021;24(3):425–36.
28. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun.* 2019;10(1):380.
29. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer.* 2022;3(4):505–17.
30. Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R, et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat Commun.* 2022;13(1):1358.
31. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform.* 2020;22:416–27.
32. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun.* 2020;11(1):1971.
33. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. Accurate estimation of cell-type composition from gene expression data. *Nat Commun.* 2019;10(1):2975.
34. Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. dtangle: accurate and robust cell type deconvolution. *Bioinformatics.* 2019;35(12):2093–9.
35. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.* 2019;26(6):1627–1640.e7.
36. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* 2019;11(1):34.
37. Hao Y, Yan M, Heath BR, Lei YL, Xie Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS Comput Biol.* 2019;15(5):e1006976.
38. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One.* 2009;4(7):e6098.
39. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 2017;18(1):220.
40. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife.* 2017;6:e26476.
41. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* 2016;17(1):218.
42. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 2016;17(1):174.
43. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015;12(5):453–7.
44. Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol Syst Biol.* 2014;10(2):720.
45. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics.* 2013;29(8):1083–5.
46. Schmid KT, Höllbacher B, Cruceanu C, Böttcher A, Lickert H, Binder EB, et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat Commun.* 2021;12(1):6625.
47. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-Cell RNA sequencing methods. *Mol Cell.* 2017;65(4):631–643.e4.
48. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018;36(5):411–20.
49. Hardy LW, Peet NP. The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. *Drug Discov Today.* 2004;9(3):117–26.
50. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13:86.
51. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med.* 2018;24(8):1277–89.
52. Huuki-Myers L, Spangler A, Eagles N, Montgomery KD, Kwon SH, Guo B, et al. Integrated single cell and unsupervised spatial transcriptomic analysis defines molecular anatomy of the human dorsolateral prefrontal cortex. *BioRxiv.* 2023. <https://doi.org/10.1101/2023.02.15.528722>.
53. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science.* 2015;348(6233):aaa6090.
54. Janesick A, Shelansky R, Gottscho A, Wagner F, Rouault M, Beliakoff G, et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. *BioRxiv.* 2022. <https://doi.org/10.1101/2022.10.06.510405>.
55. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* 2022;14(1):68.
56. Miller BF, Huang F, Atta L, Sahoo A, Fan J. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat Commun.* 2022;13(1):2339.
57. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science.* 2019;364(6441):685–9.
58. Tu J-J, Li H-S, Yan H, Zhang X-F. EnDecon: cell type deconvolution of spatially resolved transcriptomics data via ensemble learning. *Bioinformatics.* 2023;39(1):btac825.

59. Kang K, Meng Q, Shats I, Umbach DM, Li M, Li Y, et al. CDSseq: a novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLoS Comput Biol*. 2019;15(12):e1007510.
60. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res*. 2020;48(W1):W509–14.
61. Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat Commun*. 2018;9(1):3220.
62. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, et al. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun*. 2022;13(1):761.
63. Teschendorff AE, Zhu T, Breeze CE, Beck S. EPISCOPE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol*. 2020;21(1):221.
64. Goodsell DS. *The machinery of life*. New York: Springer, New York; 1993.
65. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Grayback LT, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*. 2019;573(7772):61–8.
66. Garcia FJ, Sun N, Lee H, Godlewski B, Mathys H, Galani K, et al. Single-cell dissection of the human brain vasculature. *Nature*. 2022;603(7903):893–9.
67. Huuki-Myers LA, Montgomery KD, Kwon SH, Page SC, Hicks SC, Maynard KR, et al. Data-driven identification of total RNA expression genes (TREGs) for estimation of RNA abundance in heterogeneous cell types. *bioRxiv*. 2022. <https://doi.org/10.1101/2022.04.28.489923>.
68. Simons M, Nave K-A. Oligodendrocytes: myelination and axonal support. *Cold Spring Harb Perspect Biol*. 2015;8(1):a020479.
69. Hughes EG, Kang SH, Fukaya M, Bergles DE. Oligodendrocyte progenitors balance growth with self-repulsion to achieve homeostasis in the adult brain. *Nat Neurosci*. 2013;16(6):668–76.
70. Leyh J, Paeschke S, Mages B, Michalski D, Nowicki M, Bechmann I, et al. Classification of microglial morphological phenotypes using machine learning. *Front Cell Neurosci*. 2021;29(15):701673.
71. Khakh BS, Sofroniew MV. Diversity of astrocyte functions and phenotypes in neural circuits. *Nat Neurosci*. 2015;18(7):942–52.
72. Racle J, Gfeller D. EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data. *Methods Mol Biol*. 2020;2120:233–48.
73. Liebner DA, Huang K, Parvin JD. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*. 2014;30(5):682–9.
74. Dietrich A, Sturm G, Merotto L, Marini F, Finotello F, List M. SimBu: bias-aware simulation of bulk RNA-seq data with variable cell-type composition. *Bioinformatics*. 2022;38(Suppl_2):ii141–7.
75. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods*. 2012;9(7):676–82.
76. Hu J, Schroeder A, Coleman K, Chen C, Auerbach BJ, Li M. Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Comput Struct Biotechnol J*. 2021;19:3829–41.
77. Amezcua RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with bioconductor. *Nat Methods*. 2020;17(2):137–45.
78. Kumar V, Krolewski DM, Hebda-Bauer EK, Parsegian A, Martin B, Foltz M, et al. Optimization and evaluation of fluorescence in situ hybridization chain reaction in cleared fresh-frozen brain tissues. *Brain Struct Funct*. 2021;226(2):481–99.
79. Kernohan KD, Bérubé NG. Three dimensional dual labelled DNA fluorescent in situ hybridization analysis in fixed tissue sections. *MethodsX*. 2014;1:30–5.
80. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods*. 2018;15(11):932–5.
81. Natarajan KN, Miao Z, Jiang M, Huang X, Zhou H, Xie J, et al. Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol*. 2019;20(1):70.
82. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct comparative analyses of 10X genomics chromium and smart-seq2. *Genomics Proteomics Bioinformatics*. 2021;19(2):253–66.
83. Chamberlin J, Lee Y, Marth G, Quinlan A. Variable RNA sampling biases mediate concordance of single-cell and nucleus sequencing across cell types. *BioRxiv*. 2022. <https://doi.org/10.1101/2022.08.01.502392>.
84. Gupta A, Shamsi F, Altemose N, Dorlhiac GF, Cypess AM, White AP, et al. Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. *Genome Res*. 2022;32(2):242–57.
85. Lake BB, Codeluppi S, Yung YC, Gao D, Chun J, Kharchenko PV, et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci Rep*. 2017;7(1):6031.
86. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One*. 2018;13(12):e0209648.
87. Price AJ, Hwang T, Tao R, Burke EE, Rajpurohit A, Shin JH, et al. Characterizing the nuclear and cytoplasmic transcriptomes in developing and mature human cortex uncovers new insight into psychiatric disease gene regulation. *Genome Res*. 2020;30(1):1–11.
88. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol*. 2019;37(12):1452–7.
89. Hippen AA, Omran DK, Weber LM, Jung E, Drapkin R, Doherty JA, et al. Performance of computational algorithms to deconvolve heterogeneous bulk tumor tissue depends on experimental factors. *BioRxiv*. 2022. <https://doi.org/10.1101/2022.12.04.519045>.
90. Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*. 2018;362(6420):eaat7615.

91. Mitroi DN, Tian M, Kawaguchi R, Lowry WE, Carmichael ST. Single-nucleus transcriptome analysis reveals disease- and regeneration-associated endothelial cells in white matter vascular dementia. *J Cell Mol Med*. 2022;26(11):3183–95.
92. Tran MN, Maynard KR, Spangler A, Huuki LA, Montgomery KD, Sadashivaiah V, et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. *Neuron*. 2021;109(19):3088–3103.e5.
93. Zhu K, Bendl J, Rahman S, Vicari JM, Coleman C, Clarence T, et al. Multi-omic profiling of the developing human cerebral cortex at the single cell level. *BioRxiv*. 2022. <https://doi.org/10.1126/sciadv.adg3754>.
94. Caglayan E, Liu Y, Konopka G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*. 2022;110(24):4043–4056.e5.
95. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
96. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*. 2013;10(11):1096–8.
97. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 2011;12(2):R18.
98. Wery M, Descrimes M, Thermes C, Gautheret D, Morillon A. Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. *Methods*. 2013;63(1):25–31.
99. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008;320(5881):1344–9.
100. Akbari M, Hansen MD, Halgunset J, Skorpen F, Krokan HE. Low copy number DNA template can render polymerase chain reaction error prone in a sequence-dependent manner. *J Mol Diagn*. 2005;7(1):36–9.
101. Herbert ZT, Kershner JP, Butty VL, Thimmapuram J, Choudhari S, Alekseyev YO, et al. Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. *BMC Genomics*. 2018;19(1):199.
102. Haile S, Corbett RD, Bilobram S, Mungall K, Grande BM, Kirk H, et al. Evaluation of protocols for rRNA depletion-based RNA sequencing of nanogram inputs of mammalian total RNA. *PLoS One*. 2019;14(10):e0224578.
103. Viscardi MJ, Arribere JA. Poly(a) selection introduces bias and undue noise in direct RNA-sequencing. *BMC Genomics*. 2022;23(1):530.
104. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*. 2014;15(1):419.
105. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep*. 2018;8(1):4781.
106. Bush SJ, McCulloch MEB, Summers KM, Hume DA, Clark EL. Integration of quantitated expression estimates from polyA-selected and rRNA-depleted RNA-seq libraries. *BMC Bioinformatics*. 2017;18(1):301.
107. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol*. 2021;22(1):102.
108. Katayama S, Skoog T, Söderhäll C, Einarsdóttir E, Krjutškov K, Kere J. Guide for library design and bias correction for large-scale transcriptome studies using highly multiplexed RNAseq methods. *BMC Bioinformatics*. 2019;20(1):418.
109. Tuerk A, Wiktorin G, Güler S. Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. *PLoS Comput Biol*. 2017;13(5):e1005515.
110. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*. 2014;32(5):462–4.
111. Chen L, Zheng S. BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res*. 2018;46(14):e82.
112. Ni Z, Chen S, Brown J, Kendziorski C. CB2 improves power of cell detection in droplet-based single-cell RNA sequencing data. *Genome Biol*. 2020;21(1):137.
113. Zhang Y-Z, Yamaguchi R, Imoto S, Miyano S. Sequence-specific bias correction for RNA-seq data using recurrent neural networks. *BMC Genomics*. 2017;18(Suppl 1):1044.
114. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods*. 2017;14(6):565–71.
115. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*. 2009;4:14.
116. Phipson B, Zappia L, Oshlack A. Gene length and detection bias in single cell RNA sequencing protocols. [version 1; peer review: 4 approved]. *F1000Res*. 2017;6:595.
117. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14.
118. Gao L, Fang Z, Zhang K, Zhi D, Cui X. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*. 2011;27(5):662–9.
119. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*. 2014;11(2):163–6.
120. 10X Genomics. Interpreting intronic and antisense reads in 10x genomics single cell gene expression data. 10. Available from: <https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/interpreting-intronic-and-antisense-reads-in-10-x-genomics-single-cell-gene-expression-data>. [cited 2023 Feb 24].
121. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*. 2011;21(7):1160–7.
122. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009;6(5):377–82.
123. You Y, Dong X, Wee YK, Maxwell MJ, Alhamdoosh M, Smyth G, et al. Modelling group heteroscedasticity in single-cell RNA-seq pseudo-bulk data. *BioRxiv*. 2022. <https://doi.org/10.1186/s13059-023-02949-2>.
124. Yu X, Abbas-Aghababazadeh F, Chen YA, Fridley BL. Statistical and bioinformatics analysis of data from bulk and single-cell RNA sequencing experiments. *Methods Mol Biol*. 2021;2194:143–75.

125. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
126. Svensson V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol*. 2020;38(2):147–50.
127. Kuo A, Hansen KD, Hicks SC. Quantification and statistical modeling of Chromium-based single-nucleus RNA-sequencing data. *BioRxiv*. 2022. <https://doi.org/10.1101/2022.05.20.492835>.
128. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics*. 2011;19(12):290.
129. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet*. 2023;24(8):550–72.
130. Jones DC, Ruzzo WL, Peng X, Katze MG. A new approach to bias correction in RNA-Seq. *Bioinformatics*. 2012;28(7):921–8.
131. Ascensión AM, Araúz-Bravo MJ, Izeta A. The need to reassess single-cell RNA sequencing datasets: the importance of biological sample processing. *F1000Res*. 2021;10:767.
132. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun*. 2019;10(1):4667.
133. Liu W, Liao X, Luo Z, Yang Y, Lau MC, Jiao Y, et al. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST. *Nat Commun*. 2023;14(1):296.
134. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol*. 2015;33(2):155–60.
135. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*. 2015;163(7):1663–77.
136. Llorens-Bobadilla E, Zhao S, Baser A, Saiz-Castro G, Zwadlo K, Martin-Villalba A. Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell*. 2015;17(3):329–40.
137. Miller JA, Gouwens NW, Tasic B, Collman F, van Velthoven CT, Bakken TE, et al. Common cell type nomenclature for the mammalian brain. *eLife*. 2020;9:e59928.
138. Wang Y, Sarfraz I, Teh WK, Sokolov A, Herb BR, Creasy HH, et al. Matrix and analysis metadata standards (MAMS) to facilitate harmonization and reproducibility of single-cell data. *BioRxiv*. 2023. <https://doi.org/10.1101/2023.03.06.531314>.
139. Diaz-Mejia JJ, Meng EC, Pico AR, MacParland SA, Ketela T, Pugh TJ, et al. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. [version 3; peer review: 2 approved, 1 approved with reservations]. *F1000Res*. 2019;8:562082.
140. Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J*. 2021;19:961–9.
141. Huang Q, Liu Y, Du Y, Garmire LX. Evaluation of cell type annotation R packages on single-cell RNA-seq data. *Genomics Proteomics Bioinformatics*. 2021;19(2):267–81.
142. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun*. 2022;13(1):1246.
143. Liu H, Li H, Sharma A, Huang W, Pan D, Gu Y, et al. scAnno: a deconvolution strategy-based automatic cell type annotation tool for single-cell RNA-sequencing data sets. *Brief Bioinform*. 2023;24(3):bbad179.
144. Kim T, Lo K, Geddes TA, Kim HJ, Yang JYH, Yang P. scReClassify: post hoc cell type classification of single-cell rRNA-seq data. *BMC Genomics*. 2019;20(Suppl 9):913.
145. Li Z, Feng H. A neural network-based method for exhaustive cell label assignment using single cell RNA-seq data. *Sci Rep*. 2022;12(1):910.
146. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, et al. Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv*. 2020;6(30):eaba2619.
147. Pullin JM, McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. *BioRxiv*. 2022. <https://doi.org/10.1101/2022.05.09.490241>.
148. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. *eLife*. 2017;6:e27041.
149. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573–87.
150. Quake SR. A decade of molecular cell atlases. *Trends Genet*. 2022;38(8):805–10.
151. Wang Q, Ding S-L, Li Y, Royall J, Feng D, Lesnar P, et al. The allen mouse brain common coordinate framework: a 3D reference atlas. *Cell*. 2020;181(4):936–953.e20.
152. Gabitto MI, Travaglini KJ, Rachleff VM, Kaplan ES, Long B, Ariza J, et al. Integrated multimodal cell atlas of Alzheimer's disease. *BioRxiv*. 2023. <https://doi.org/10.21203/rs.3.rs-2921860/v1>.
153. Rood JE, Maartens A, Hupalowska A, Teichmann SA, Regev A. Impact of the human cell Atlas on medicine. *Nat Med*. 2022;28(12):2486–96.
154. Delaney C, Schnell A, Cammarata LV, Yao-Smith A, Regev A, Kuchroo VK, et al. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Mol Syst Biol*. 2019;15(10):e9005.
155. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods*. 2019;16(8):715–21.
156. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat Methods*. 2018;15(12):1053–8.
157. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol*. 2021;17(1):e9620.
158. Hu P, Zhang W, Xin H, Deng G. Single cell isolation and analysis. *Front Cell Dev Biol*. 2016;4:116.
159. Baron CS, Barve A, Muraro MJ, van der Linden R, Dharmadhikari G, Lyubimova A, et al. Cell type purification by single-cell transcriptome-trained sorting. *Cell*. 2019;179(2):527–542.e19.
160. Nott A, Schlachetzki JCM, Fixsen BR, Glass CK. Nuclei isolation of multiple brain cell types for omics interrogation. *Nat Protoc*. 2021;16(3):1629–46.

161. Li R, Banjanin B, Schneider RK, Costa IG. Detection of cell markers from single cell RNA-seq with sc2marker. *BMC Bioinformatics*. 2022;23(1):276.
162. Dumitrascu B, Villar S, Mixon DG, Engelhardt BE. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nat Commun*. 2021;12(1):1186.
163. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, et al. Cell type ontologies of the human cell Atlas. *Nat Cell Biol*. 2021;23(11):1129–35.
164. Christensen E, Luo P, Turinsky A, Husić M, Mahalanabis A, Naidas A, et al. Evaluation of single-cell RNAseq labelling algorithms using cancer datasets. *Brief Bioinform*. 2023;24(1):bbac561.
165. Clarke ZA, Andrews TS, Atif J, Pouyababar D, Innes BT, MacParland SA, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat Protoc*. 2021;16(6):2749–64.
166. Winnubst J, Bas E, Ferreira TA, Wu Z, Economo MN, Edson P, et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell*. 2019;179(1):268–281. e13.
167. Peng RD. Reproducible research in computational science. *Science*. 2011;334(6060):1226–7.
168. Bjornson E. Reproducible research: best practices and potential misuse [perspectives]. *IEEE Signal Process Mag*. 2019;36(3):106–23.
169. Seirup M, Chu L-F, Sengupta S, Leng N, Browder H, Kapadia K, et al. Reproducibility across single-cell RNA-seq protocols for spatial ordering analysis. *PLoS One*. 2020;15(9):e0239711.
170. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. *Gigascience*. 2016;5(1):30.
171. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence read archive: a decade more of explosive growth. *Nucleic Acids Res*. 2022;50(D1):D387–90.
172. Leinonen R, Sugawara H, Shumway M. International nucleotide sequence database collaboration. The sequence read archive. *Nucleic Acids Res*. 2011;39(Database issue):D19–21.
173. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207–10.
174. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. *Nat Methods*. 2022;19(2):171–8.
175. Hu J, Chen M, Zhou X. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic Acids Res*. 2022;50(4):e21.
176. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet*. 2018;19(4):208–19.
177. Maden SK, Thompson RF, Hansen KD, Nellore A. Human methylome variation across Infinium 450K data on the Gene Expression Omnibus. *NAR Genom Bioinform*. 2021;3(2):lqab025.
178. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res*. 2019;47(D1):D711–5.
179. Bioconductor - BioViews. Available from: https://www.bioconductor.org/packages/devel/BiocViews.html#___Software. [cited 2023 Aug 21].
180. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12(2):115–21.
181. Davide Risso [Aut C Cph], Michael Cole. scRNAseq. Bioconductor. 2017.
182. Pardo B, Spangler A, Weber LM, Page SC, Hicks SC, Jaffe AE, et al. spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics*. 2022;23(1):434.
183. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, et al. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using bioconductor. *Bioinformatics*. 2022;38(11):3128–31.
184. Jeon H, Xie J, Jeon Y, Jung KJ, Gupta A, Chang W, et al. Statistical power analysis for designing bulk, single-cell, and spatial transcriptomics experiments: review, tutorial, and perspectives. *Biomolecules*. 2023;13(2):221.
185. Cui W, Xue H, Wei L, Jin J, Tian X, Wang Q. High heterogeneity undermines generalization of differential expression results in RNA-Seq analysis. *Hum Genomics*. 2021;15(1):7.
186. Gibson G. Perspectives on rigor and reproducibility in single cell genomics. *PLoS Genet*. 2022;18(5):e1010210.
187. Wang M, Song W-M, Ming C, Wang Q, Zhou X, Xu P, et al. Guidelines for bioinformatics of single-cell sequencing data analysis in Alzheimer's disease: review, recommendation, implementation and application. *Mol Neurodegener*. 2022;17(1):17.
188. Maden S, Kwan SH, Huuki-Meyers LA, Collado-Torres L, Hicks SC, Maynard KR. deconvo_review-paper. GitHub; 2023. Available from: https://github.com/LieberInstitute/deconvo_review-paper/tree/master. [cited 2023 Nov 9].
189. Maden S, Kwan SH, Huuki-Meyers LA, Collado-Torres L, Hicks SC, Maynard KR. deconvo_review-paper. Zenodo; 2023. Available from: <https://zenodo.org/records/10085497>. [cited 2023 Nov 9].
190. Maden S, Kwan SH, Huuki-Meyers LA, Collado-Torres L, Hicks SC, Maynard KR. cellScaleFactors. GitHub; 2023; <https://github.com/metamaden/cellScaleFactors>.
191. Maden S, Hicks S. cellScaleFactors v0.0.1. Zenodo. 2023. <https://doi.org/10.5281/zenodo.10149934>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.