Hu and Chikina *Genome Biology* (2024) 25:169
https://doi.org/10.1186/s13059-024-03292-w

Genome Biology

RESEARCH

Open Access

Heterogeneous pseudobulk simulation enables realistic benchmarking of cell-type deconvolution methods

Mengying Hu[1,2] and Maria Chikina[1,2*]

*Correspondence: mchikina@pitt.edu

[1] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA
[2] Joint Carnegie Mellon - University of Pittsburgh Computational Biology PhD Program, University of Pittsburgh, Pittsburgh, USA

Abstract

Background: Computational cell type deconvolution enables the estimation of cell type abundance from bulk tissues and is important for understanding tissue microenviroment, especially in tumor tissues. With rapid development of deconvolution methods, many benchmarking studies have been published aiming for a comprehensive evaluation for these methods. Benchmarking studies rely on cell-type resolved single-cell RNA-seq data to create simulated pseudobulk datasets by adding individual cells-types in controlled proportions.

Results: In our work, we show that the standard application of this approach, which uses randomly selected single cells, regardless of the intrinsic difference between them, generates synthetic bulk expression values that lack appropriate biological variance. We demonstrate why and how the current bulk simulation pipeline with random cells is unrealistic and propose a heterogeneous simulation strategy as a solution. The heterogeneously simulated bulk samples match up with the variance observed in real bulk datasets and therefore provide concrete benefits for benchmarking in several ways. We demonstrate that conceptual classes of deconvolution methods differ dramatically in their robustness to heterogeneity with reference-free methods performing particularly poorly. For regression-based methods, the heterogeneous simulation provides an explicit framework to disentangle the contributions of reference construction and regression methods to performance. Finally, we perform an extensive benchmark of diverse methods across eight different datasets and find BayesPrism and a hybrid MuSiC/CIBERSORTx approach to be the top performers.

Conclusions: Our heterogeneous bulk simulation method and the entire benchmarking framework is implemented in a user friendly package https://github.com/humengying0907/deconvBenchmarking and https://doi.org/10.5281/zenodo.8206516, enabling further developments in deconvolution methods.

Keywords: Deconvolution, Benchmark, Bulk simulation

Background

Bulk RNA-sequencing experiments reveal average gene expression values for all cells present in a sample mixture. Computational deconvolution methods separate the mixed signals from the aggregated expression and provide estimation of cellular components without physical isolations. The inferred cellular proportions are important to understand the ecosystem of the tissue and can be used as covariates in differential expression, reducing false positives and false negatives [1, 2]. Moreover, for heterogeneous bulk samples like tumor [3], deconvolution enables identification and quantification of the infiltrating immune populations, which provides rich prognostic values and can guide targeted therapy (e.g., in immunotherapy) [4–8].

Numerous deconvolution methodologies have been developed (see [1], for review), aiming at estimation of cell-type abundance from bulk transcriptomic data. Depending on if and how priori knowledge used, these methods can be broadly classified into four categories: regression-based, marker-based, and reference-free methods. Regression-based methods require an expression matrix as input, which consists of a cell type-specific expression profile for selected genes. These methods then solve the deconvolution as a regression problem. A comprehensive evaluation of factors involved in regression-based methods, like data transformation, normalization, and regression algorithms, can be found elsewhere [9]. Marker-based methods require a set of genes that characterize the expression patterns in different cell types and return either an enrichment score [10] that is unitless or abundance estimates [11, 12].

Reference-free methods [12–14] are completely unsupervised and do not require any prior knowledge as input. Such methods are based on finding a simplex, which is a geometric data structure expected under ideal mixture proportion scenarios. Finally, we note that a recent advance in deconvolution uses a Bayesian framework that relies on a reference matrix but uses it in a way that is distinct from reference based approaches [15].

Rapid development of deconvolution methodologies now raises another challenge of evaluating their performance across diverse realistic settings. Many benchmarking studies have been undertaken to meet this demand [2, 9, 16, 17]. Regardless of the focus of their evaluation, all benchmarking efforts rely on datasets with known ground truth. To acquire such data, one traditional approach involves using real bulk data with paired cell type fraction information, which can be derived from fluorescence-activated cell sorting (FACS) or immunohistochemistry (IHC) staining [17, 18]. However, this approach is restricted by the extensive experimental labor and limited sample availability, making it less practical for large-scale benchmarking studies. An alternative approach is computational mixing where purified expressions of different cell-types are mixed in controlled proportions [10, 19, 20]. While the purely computational strategy can generate large datasets, this approach has the clear limitation that it makes the strong assumption that proportion variation and random noise are the only source of variance in the data.

Increasing availability of single-cell data [21] offers the opportunity to create more realistic simulations. Instead of computational mixing of pure expression states, individual single cell profiles are added together in controlled proportions [9, 17, 18, 22]. This has the explicit advantage over pure computational mixing as it introduces more variations in the simulated samples. However, as we will show in this work, while this approach has rapidly become the standard method for bulk simulation, the problem with unrealistic biological variance is only partially resolved. To simulate data compatible with bulk measurements, a large number of cells (typically hundreds) are added for each simulated sample. As such the pure cell type-specific expression in each sample, while not exactly identical, tends towards the global mean of that cell type in the source scRNA data, enforcing the unrealistic assumption that there is no systematic variation beyond cell-type proportions. One possible solution is to take into account intra-sample heterogeneity in the simulated bulk mixtures. In Chu et al.'s [15] study, they created such simulated bulk mixtures by restricting that the malignant cells aggregated to form a simulated bulk sample originating from the same biological sample. Dong et al. [23] and Menden et al. [24] implemented a simulation strategy that involved repetitively sub-sampling cells from the same patient, ensuring proper inter-sample heterogeneity. However, there is currently no general evaluation of how heterogeneity affects the deconvolution results, as compared with bulk simulation using random cells.

In this study, we introduce a novel heterogeneous simulation approach that aims at capturing accurate biological variance. Through systematic comparison of these simulation methods, we demonstrated that bulk simulation methods using random cells do not reflect realistic biological heterogeneity while our newly proposed approach does. Leveraging the varying heterogeneity levels in the simulated bulk samples, we provided an in-depth comparison of different categories of deconvolution methods using our systematic benchmarking frameworks (Additional file 1: Fig. S1), aiming to elucidate the impact of heterogeneity on the results. By summarizing deconvolution performance across experimental repeats, we found that introducing biological heterogeneity has a notable effect on the deconvolution results, with reference-free methods being most affected. Our study can guide researchers in choosing the most appropriate deconvolution methods, and the highly realistic simulation framework we proposed can facilitate further methodological development.

Results

Exploring biological variance in simulated bulk data: influence of different simulation strategies

In previous benchmarking studies [2, 9, 17], the evaluation of deconvolution performance relied on simulated bulk expression using predefined cell-type fractions as ground truth. These studies employed a "homogeneous" simulation approach, where single-cell profiles from single-cell data were combined randomly within each cell type and aggregated in proportions. However, this approach only accounts for cell-type proportion level variance and overlooks other sources of biological variance.

To address this limitation, we explored alternative simulation methods aiming at introducing more biological variance within the simulated samples. We proposed a "heterogeneous" simulation setting where cells used to compose the cell type components of a simulated bulk sample are constrained to come from the same biological samples (Additional file 1: Fig. S2), thus capturing the sample-level heterogeneity [25, 26]. We also considered a less "heterogeneous" setting where only malignant cells are originated from the same sample in the simulation, which we referred to as "semi-heterogeneous" simulation, inspired from Chu's benchmarking work [15].

We postulated that these three simulation methods—homogeneous, semi-heterogeneous, and heterogeneous—will inherently produce samples with distinct levels of variance, reflecting varied capacity to mimic real biological complexity. To test this, we applied these simulation strategies on four distinct single-cell datasets (Table 1, Additional file 2: Table S1), resulting in a total of 12 simulated expression profiles for evaluation. For cell type fraction simulation, we adopted a beta distribution-based strategy, allowing the mean and variances of each cell-type fractions to be approximately matched to those of real data (see the "Methods" section; Additional file 1: Fig. S3). To create baseline bulk expression for variance comparison, we aggregated single cells from the same patients and used them as approximations of real bulk samples (see the "Methods" section). We also collected expression profiles from the TCGA datasets [27] when the relevant tumor type is available. A detailed description of bulk datasets used in variance comparison can be found in Additional file 2: Table S2.

We illustrate our framework with bulk data simulated from Jerby_Arnon2018_SKCM. Utilizing the coefficient of variation (CV) of gene expressions as a measure of intra-sample variance, we first compared gene-level CV between simulated bulk samples and baseline bulk expression. Our findings revealed that the heter-simulated bulk samples exhibited variance closely aligned with that of actual bulk samples, while the homo-simulated samples displayed generally lower variability and semi-heter simulated samples failing between (Fig. 1a). Summarized gene-CV at pathway levels [43] further confirmed this finding, and by extending the CV analysis to include real TCGA bulk samples from the same tumor type, we showed that heter-simulated samples retained proper biological variance compared with real bulk samples (Fig. 1b).

Moving forward, we calculated pairwise correlations between the simulated samples to gain further insight into biological variance (Fig. 1c). The distribution of pairwise correlations across different simulation settings revealed distinct variance levels, with homo-simulated samples showing the highest pairwise correlations, and the heter-simulated samples exhibiting the least pairwise similarities, aligning more closely with real-world settings. Moreover, as we visualized pairwise correlation between genes under different simulation settings, we found that the heter-simulated samples effectively captured proper gene correlations (Fig. 1d). They maintained appropriate gene clusters, similar to those seen in baseline bulk expression, and demonstrated reasonable

Table 1 Single-cell datasets used in benchmarking

Dataset	Tumor type	# of cells	Publication	Data
Puram2017_HNSCC[a]	HNSCC	5901	Puram et al. 2017 [25]	[28]
Tirosh2016_SKCM[a]	SKCM	4645	Tirosh et al. 2016 [29]	[30]
Riemondy2022_MB[a]	MB	39,946	Riemondy et al. 2022 [31]	[32]
Jerby_Arnon2018_SKCM[a]	SKCM	7186	Jerby-Arnon et al. 2018 [33]	[34]
Lee2020_CRC	CRC	21,657	Lee et al. 2020 [35]	[36]
Qian2020_BRCA	BRCA	16, 537	Qian et al. 2020 [37]	[38]
Kim2020_LUAD	LUAD	32,493	Kim et al. 2020 [39]	[40]
Izar2020_OV	OV	10,788	Izar et al. 2020 [41]	[42]

HNSCC head and neck squamous cell carcinomas, *SKCM* skin cutaneous melanoma, *MB* medulloblastoma, *CRC* colorectal cancer, *BRCA* breast cancer, *LUAD* lung adenocarcinoma, *OV* ovarian cancer

[a] scRNA datasets marked with a superscript (a) are also utilized for simulation strategy comparisons

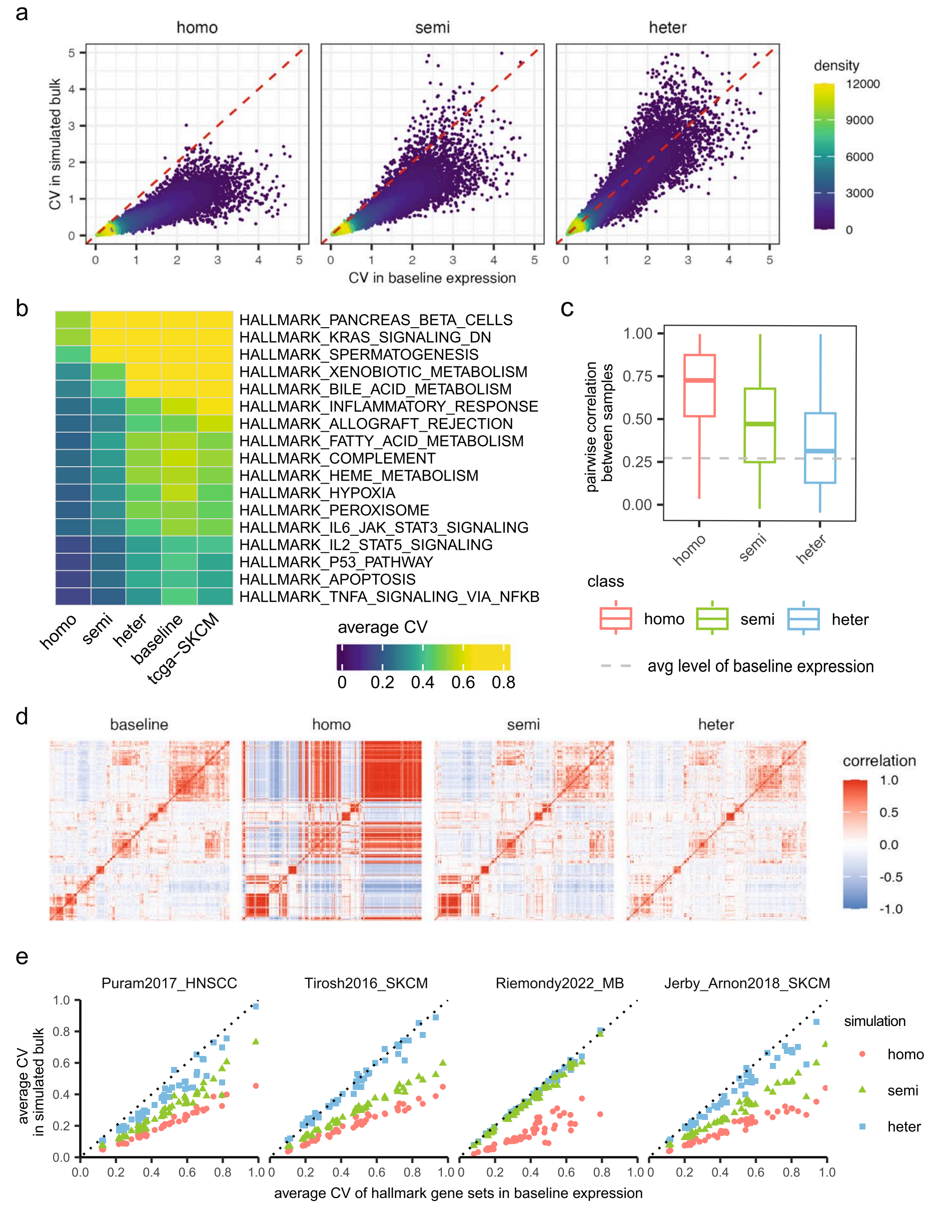

Fig. 1 Comparison of different bulk simulation strategies. **a** Scatter plots comparing coefficient of variation (CV) for all genes between the simulated samples and baseline expression. **b** Heatmap comparing average CV of genes from different biological pathways. **c** Boxplot comparing pairwise correlations between simulated bulk samples, with the dashed line indicating the average pairwise correlation in baseline expression. **d** Heatmaps showing gene correlations in bulk samples, using the top 300 most variable genes from the baseline expression. **e** Scatter plots comparing variance of 50 hallmark gene sets between simulated and baseline bulk expression, with each dot representing the averaged CV of genes from a hallmark gene set. In **a**–**d**, all simulated datasets and baseline expression are derived from the Jerby_Arnon2018_SKCM dataset. Throughout the comparison, each simulated dataset comprises 100 simulated samples, and baseline expression is generated by aggregating single-cell expression from the same biological sample, serving as an approximation to real bulk samples

coefficient correlations. In contrast, homo-simulated samples resulted in false-positive gene clustering structures and spuriously high gene correlations. We note that in this comparison we only focused on the top 300 most variant genes due to the immense scale of the gene-gene correlation matrix; the same patterns hold when specifically considering cell-type marker genes as well (Additional file 1: Fig. S4, Additional file 2: Table S3).

In Fig. 1e, we extended the variance comparison to all datasets by comparing pathway-level variance in simulated samples against real bulk expressions. Our results showed that across each dataset, the heter-simulated samples closely mirrored the actual variance observed in real data, exhibiting pathway-level CVs that align closely with those of real bulk samples, whereas the homo-simulated samples are in general less variable. We note that for bulk data simulated from Riemondy2022_MB dataset, the semi-simulated and heter-simulated samples show little difference with respect to pathway variances; this is because the simulated MB samples are mainly dominated by malignant cells, making it less distinguishable between two strategies.

In addition to the three simulation methods mentioned above, we also investigated four additional bulk simulation approaches: the "favilaco" method [9], the "immunedeconv" method, [17], the "SCDC" method [23], and one "heterogeneous" simulation method that does not rely on single-cell sampleID information, which we referred to as the "heterogeneous-sampleIDfree" method (see the "Methods" section). These expanded methodologies, along with our original simulation strategies, have been integrated into our published package, offering a comprehensive toolkit for bulk simulations.

An extensive comparison of all seven simulation methods is detailed in Additional file 1: Fig. S4-S7. Our analysis revealed that the "heterogeneous-sampleIDfree" method successfully maintained comparable variance without the constraints of sample ID dependency, offering a significant advantage for single-cell datasets where sample IDs may be limited or absent. Conversely, the other two methods, "immunedeconv" and "favilaco," which were originally implemented for deconvolution benchmarking studies [9, 17], exhibited a notable lack of variance in simulated samples. Finally, the "SCDC" method [23], although it achieved variance comparable to that of real bulk samples, carries the risk of generating empty gene expression values. This is due to its reliance on repeated sub-sampling from the same patient, which can easily suffer from the sparsity issue in single cells [44] when aggregating over only a limited number of cells.

Together, our results suggested that different simulation strategies result in simulated bulk samples with diverse levels of biological variance. Specially, stepping through homogeneous, semi-heterogeneous, and heterogeneous simulation, the heterogeneity level inside samples is increasing with the final heterogeneous simulation closely retaining the characteristics observed in real bulk samples.

Bulk simulation using random cells ignores heterogeneity within constituent cell types

Heterogeneity of tumors between different patients with the same tumor type has long been recognized [45]. Despite similar histological appearance, different patients can have intrinsically different genomic landscapes. In clinical practice, this heterogeneity motivates molecular subtyping and enables personalized treatment protocols [46, 47]. Retaining biological heterogeneity within simulated bulk samples is essential for realistic bulk simulation.

To illustrate the limitation of bulk simulation using randomly selected cells (namely homogeneous simulation), we considered a simulation setup using single cell Medulloblastoma (MB) dataset from Riemondy et al. [31]. Medulloblastoma is a well-recognized heterogeneous brain cancer with four distinct subtypes based on genetic characteristics: WNT, SHH, Group 3, and Group 4 [48, 49]. Analyzing the tSNE clustering of malignant populations (Fig. 2a), we found that cells from the same subtype predominantly clustered together, revealing marked differences between subtypes. Additionally, within each subgroup, further patient-specific heterogeneity was observed. We note that such intra-heterogeneity of malignant cells extends beyond this scRNA dataset and is also found in other tumor types (Additional file 1: Fig. S8). Additionally, beyond just malignant populations, non-malignant cells may also exhibit diverse patterns across different samples (Additional file 1: Fig. S9). Together, these findings suggested even among cells identified as the same cell type, intra-tumor heterogeneity is not uncommon.

Despite this prevalent intra-heterogeneity, random selection of cells results in an evenly distributed selection of cells (Fig. 2b), and such selection, if performed repeatedly, will create a homogeneous expression profile with low variance. In Fig. 2c, we simulated a bulk expression dataset in this manner and compared it to actual MB patient profiles (Additional file 2: Table S2). By analyzing the expression levels of 22 MB-subtype specific genes (Additional file 2: Table S4) [46], we observed distinct expression patterns across different MB subtypes in real patient profiles, whereas the simulated samples exhibited minimal heterogeneity in expression values of these genes, with the 10% and 90% expression quantiles fluctuate around the average level.

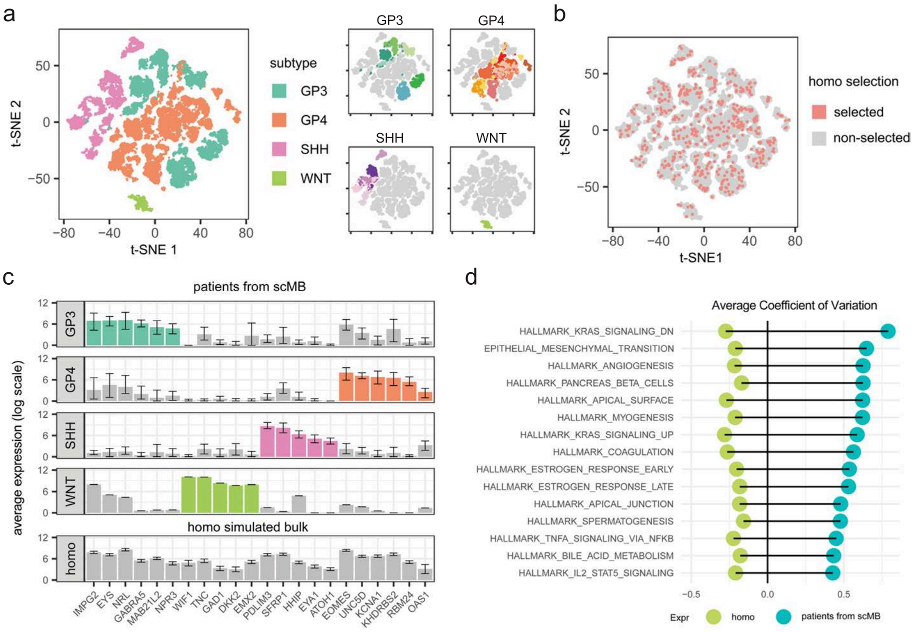

Fig. 2 Bulk simulation using random cells failed to retain intra-tumor variations. **a** tSNE plot of $n =$ 31,823 malignant cells from 28 medulloblastoma patients in Riemondy2022_MB dataset, colored by MB subtypes (left) and patient ID (right). **b** tSNE plot showing that 500 randomly selected malignant cells are evenly dispersed among 31,823 malignant cells. **c** Barplot showing the average expression levels of 22 MB subtype-specific genes in different groups of samples, colored by different MB subtypes. Upper panel: average expression levels of these genes across different MB subtype patients from Riemondy2022_MB; bottom panel: average expression values in 100 homogeneous simulated bulk samples, with error bars indicating the 10th and 90th quantiles. **d** Diverging bar plot comparing variance of hallmark gene-sets between homo-simulated and patient-specific bulk expression from Riemondy2022_MB, with patient-specific expression defined as the aggregated expression of single cells from the same patient

Moreover, analyzing the variation at the pathway level using the hallmark genes [43], we found that systematic pathway-level variance is pervasive in real data but is not recapitulated in the random cell simulation (Fig. 2d). Together, these results suggested that bulk simulation methods employing random single-cell selection overlook meaningful biological variability, resulting in a "homogeneous" profile characterized by low variance.

Benchmarking framework

To systematically evaluate the performance of different categories of deconvolution methods and examine how different bulk simulation strategies will impact the performance, we designed a benchmarking framework as depicted in Additional file 1: Fig. S1. The deconvolution methods we included span four categories of deconvolution methodologies as mentioned in previous sections: reference-free, regression-based, marker-based, and Bayesian method.

For reference-free methods, we selected debCAM (referred to as CAMfree) [12] and linseed [14]. For regression-based methods, we included five regression algorithms that have been previously developed or implemented for deconvolution: MuSiC [50], Robust Partial Correlations (RPC) [51], weighted robust linear regression (wRLM) [52], CIBERSORT [53], and non-negative least squares (nnls) [54]. For marker-based method, we considered debCAM-marker [12], TOAST-marker [55], and gsva [56]. For Bayesian method, we explored the recently published BayesPrism [15].

Procedures to evaluate deconvolution results vary in terms of whether the agreement between ground truth and inferred proportions is assessed by correlation or squared error and whether performance is evaluated per-cell type or globally. We focused our evaluation on per cell-type Pearson correlation, which reflects the accuracy of downstream inference such as the difference in proportions between two groups. We also calculated root mean square error (RMSE) values, which evaluate if the inferred proportions are correct on the absolute scale across different cell types, with smaller RMSE indicating better performance. The deconvolution pipelines including simulation, deconvolution, and evaluation are then applied to eight published single-cell cohorts (Table 1) and repeated 10 times for each cohort.

Regression-based methods differ in their robustness to heterogeneity

Regression based approaches dominate the deconvolution field with many available methods and extensive independent benchmarking [9, 50, 57]. All regression methods fit a model that assumes that bulk expression matrix $Y_{g\times s}$ with g genes and s samples, can be expressed as a $Y_{g\times s} = X_{g\times k}P_{k\times s} + E_{g\times s}$, where P is the proportion matrix and k is the number of cell-types. X is the given reference matrix and the task is to fit P, which is a general regression problem.

A first decision in regression-based approaches involves the construction of the reference matrix X, a process termed as hyper-parameter reference construction. Selecting an optimal subset of genes is crucial for effective performance [58], while utilizing the entire gene set often results in poor outcomes (data not shown). The second methodological choice is the form of the regression problem itself, whether to use constraints, feature weights, and how to formulate the loss. For example: squared loss with constraints gives the basic non-negative least squares formulation (nnls). Alternative approaches involve weighted regression, robust regression, epsilon-insensitive loss (CIBERSORT), etc. Importantly, since the feature selection and regression are decoupled, we can combine different methods arbitrarily.

In our benchmarking work, we investigated four different reference matrix construction methods: CIBERSORTx [59], autogeneS [60], and two marker selection methods derived from differential expression (DE) analysis: limma [61] and scran (see the "Methods" section for details). Additionally, we considered an "all genes" setting, which is applicable exclusively to MuSiC, where all genes are utilized by default and no feature selection is performed.

Implementing these reference construction approaches to different regression methods, we systematically examined all possible combinations of these methodologies across eight distinct cohorts (Table 1), with 10 simulations for each cohorts (excluding "all genes" from methods other than MuSiC). A representative result using bulk data simulated from the Puram2017_HNSCC dataset is presented in Fig. 3a. We found that

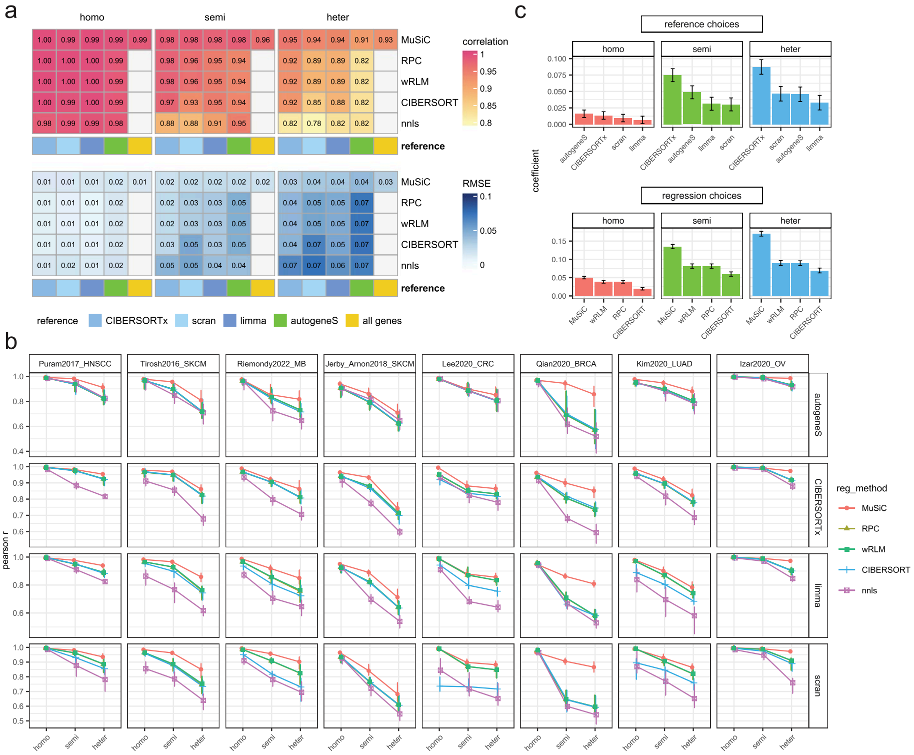

Fig. 3 Impact of bulk simulation strategies on regression based deconvolution. **a** Heatmap comparing the deconvolution performance of regression models (rows) under different bulk simulation strategies, as evaluated by average Pearson correlation (upper panel) and average RMSE values (lower panel) over 10 experimental repeats for the Puram2017_HNSCC dataset. The columns of the heatmaps correspond to different reference construction methods, with the "all genes" column being specifically applicable to MuSiC, suggesting that all genes will be used as the input, which is the default setting for MuSiC. **b** Line plot comparing Pearson *r* of regression-based methods under various simulation strategies across eight different datasets, with each row representing a reference construction choice and the error bars indicating the min and max level of Pearson *r* over 10 experimental repeats. **c** Box plot showing the effect size (estimated by coefficients from the multivariate regression model) of different methodological choices in deconvolution performance across different simulated bulk conditions. The error bars indicate the 95% confidence intervals of the coefficients

in the homogeneous setting, the results from different methodological combinations are remarkably consistent, with correlations spanning from 0.98 to 1. However, the performance of methods diverges with increasing heterogeneity. The results on all datasets is presented in Fig. 3b, where we showed that the performance of all methods decays as heterogeneity is added, while some methods decay less rapidly. We further summarized the influence of various methodological decisions by fitting a multi-linear regression to the Pearson correlation results with dataset and cell-type as a covariate (Fig. 3c; see the "Methods" section).

Altogether these results revealed that effect of methodological choices to be more pronounced with increased heterogeneity levels. Considering the summary in Fig. 3c, we find that in the homogeneous setting the reference choice method contributes relatively little (0.0067 to 0.016) and the coefficients are not significantly different. However, as heterogeneity levels in the bulk samples increase, the choice of reference becomes increasingly important and the CIBERSORTx reference dominates compared to other methods. We also note that the recently proposed optimization-based method autogeneS is indeed best in the homogeneous setting but is no better than differential expression-based method under heterogeneous settings.

Regarding regression choices, MuSiC consistently outperforms the others, irrespective of the chosen reference, under heterogeneous settings (Fig. 3b, c). On the other hand, nnls demonstrates the highest sensitivity to heterogeneity, exhibiting a sharp decline in performance as heterogeneity escalates and consistently ranking as the least effective among the regression models. The method of intermediate performance wRLM, RPC, and CIBERSORT are conceptually similar in that they make the regression problem robust in the technical statistical sense of robustness to outliers. This achieved by altering the loss function from squared loss to a function that grows less rapidly: epsilon insensitive loss for CIBERSORT and Huber loss for both RPC and wRLM (both use R based "rlm" function with default parameters). Overall these three conceptually similar approaches are indistinguishable when using sophisticated reference algorithms (CIBERSORTx or autogeneS, Fig. 3b top two rows), except that CIBERSORT regression lags behind the other two on DE-based references in four out of eight datasets. In Additional file 1: Fig. S11, we expanded our analysis to include another robust regression-based method, FARDEEP [62], which utilizes adaptive least trimmed squares in its optimization. Similar to the results discussed above, all robust regression-based methods demonstrated comparable sensitivity to changes in heterogeneity levels, each exhibiting decreased deconvolution capability in heterogeneous settings.

Notably, the MuSiC weighted regression which performs best differs conceptually from other regression approaches. Rather than altering the loss MuSiC weights the features based on the variance/covariance proprieties in the reference data. Our analysis showed that while the improvement afforded by this more complex approach is negligible in the homogeneous setting, its advantages are clearly evident in the heterogeneous one. Results were consistent across correlation and RMSE (Additional file 1: Fig. S10),

Finally, we note that the effects of the two regression methodological choices are additive. While MuSiC is originally designed to work with all genes—the only method capable of yielding satisfactory outcomes in an unfiltered context, adding an additional feature selection step further increases its performance. Specifically, the MuSiC (regression choice) and CIBERSORTx (reference construction choice) combination stood out as the best overall. We will refer to this combination as “MuSiC_CIBERSORTx” to highlight the difference from “MuSiC_default.”

Marker-based methods are robust to heterogeneity

Marker-based methods represent a conceptually different class. Instead of solving a regression problem, these methods infer the cell type proportions based on the aggregate behavior of cell-type specific genes. The approaches can be broken down into two steps: selection of cell-type specific gene sets and the summarization of these gene sets. Depending on the summarization method the output may be either unitless scores (e.g., gsva [56]) or adhere to a sum-to-one constraint, providing a direct estimate of proportions that can be assessed using RMSE.

Since lack of proportion estimates is a major criticism of marker-based methods, we focused our analysis on two methods that report proportions: debCAM-marker (referred to as debCAM) [12] and TOAST-marker (referred to as TOAST) [55]. Additionally, we considered gsva [56], which is a widely used score-based method that does not provide proportions. For gene set selection process, we employed the same approaches that were applied in the reference construction for reference-based methods. Instead of taking quantitative expression values, in this step we only considered list of genes associated with each cell type: given a reference, we assigned each gene from the reference to the cell-type with the highest expression (see the “Methods” section).

Overall, our findings revealed that performance varies significantly across different combinations of marker selection and summarization methods, with debCAM emerging as the best summarization technique (Fig. 4a, Additional file 1: Fig. S12). Specifically, the combination of debCAM and scran-based gene set selection consistently delivered the best overall results. Similar to regression-based approaches, we observed a decline in performance with the introduction of heterogeneity in simulations. However, unlike regression-based methods where performance disparities are only apparent under heterogeneous conditions, the relative performance of marker-based methods remains stable across both homogeneous and heterogeneous settings, with debCAM consistently excelling.

Finally, we compared best-in-class marker-based method (debCAM_scran) with two regression-based methods, MuSiC_CIBERSORTx, the best composite method we proposed and CIBERSORT_CIBERSORTx, the default CIBERSOTx method that has been widely used (Fig. 4b). Using Pearson correlation as the evaluation metric, our results indicated that while debCAM does not outperform regression-based methods in homogeneous simulation settings, it is highly competitive in heterogeneous settings, matching or even exceeding the performance of CIBERSORTx. This suggested that the gene set scoring technique employed by debCAM shows robustness against sample heterogeneity, even if it uses less information compared with regression-based methods. However, we also observed that debCAM demonstrates more variation in the RMSE values across experimental repeats, and generally exhibits higher RMSE compared to regression-based methods. This suggested that while debCAM accurately captures correlations, it is less effective at capturing the correct magnitude of

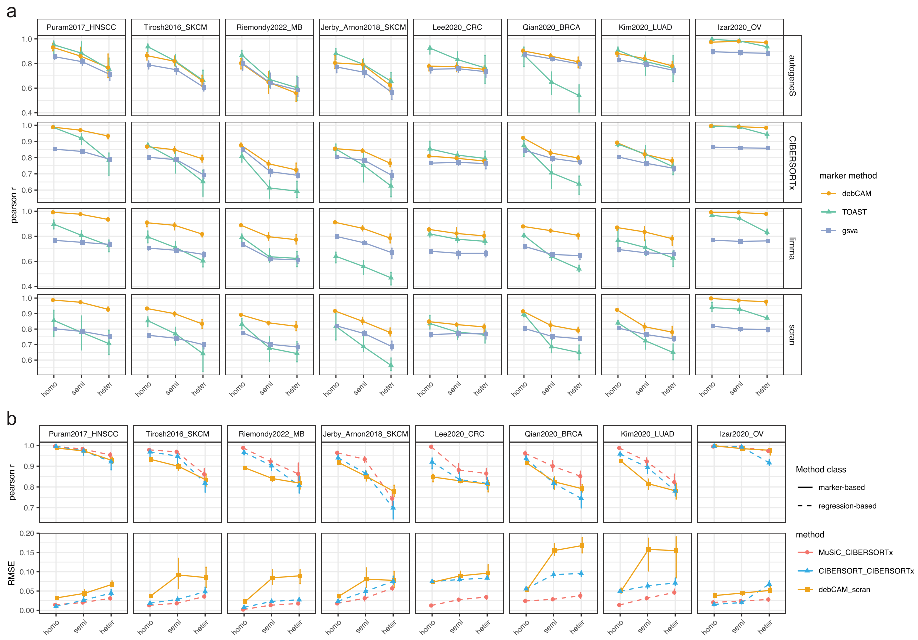

Fig. 4 Impact of bulk simulation strategies on marker based deconvolution. **a** Line plot comparing Pearson *r* of marker-based methods under different simulation strategies across eight different datasets, with each row representing a marker construction method and the error bars indicating the min and max level of Pearson *r* over 10 experimental repeats. **b** Line plot comparing the best-in-class marker based method (debCAM_scran) with two regression based methods (MuSiC_CIBERSORTx and CIBERSORT_CIBERSORTx), with the performance being evaluated by Pearson *r* and the error bars indicating the min and max level of Pearson *r* over 10 experimental repeats

the cell type fractions. Indeed, for debCAM, the absolute fraction inference is performed as a post-hoc adjustment to the scores; therefore, it does not consistently yield accurate estimations.

We note that the unitless marker-based method “gsva,” which typically exhibits the lowest performance when assessed through Pearson correlation, actually correlates non-linearly with the actual cell fractions. When evaluated using the Spearman correlation metric, the performance of “gsva” becomes considerably comparable to other marker-based methods, underscoring its potential for preliminary assessments of relative cell type proportions (Additional file 1: Fig. S13).

Additionally, we investigated the marker-based method “xCell” [10], which employs a built-in reference for estimating cell-type abundance. We compared the ground truth fractions with the matched cell-type signatures and observed that xCell signatures effectively predicts the true abundance within each immune cell type (Additional file 1: Fig. S14), Furthermore, most abundance estimates are highly coordinated between homogeneous and heterogeneous simulated samples, indicating its applicability as a preliminary investigation of bulk samples with unknown composition or unavailable single cell data. However, it is worth noting that xCell does not provide malignant fraction estimation and multiple xCell signatures can be mapped to the same cell type, making it difficult to distinguish and interpret the relevant signatures. For example, multiple B-cell related signatures are found to be closely correlated with B cell fractions (Additional file 1: Fig. S14).

Overall, we found that some marker-based methods are competitive with regression-based approaches and in some cases can offer advantages such as not requiring precise knowledge of reference values.

Comprehensive assessment of deconvolution performance across conceptual classes

So far, we have performed a detailed evaluation of two major categories of deconvolution methods: the regression-based and marker-based approaches. In our final evaluation, we also included two additional classes: reference-free and BayesPrism, which uses a quantitative reference in a unique way that sets it apart from conventional regression-based approaches and thus constitutes its own class. While regression approaches fit an equation of the form $Y = XP + E$, where E represents error, BayesPrism solves a Latent Dirichlet Allocation (LDA) problem allocating all of the observed gene expression to a cell type so there is no residual [15]. Moreover, BayesPrism differs from conventional regression-based methods in that it uses all genes by default and its performance is not improved by subsetting (data not shown).

Since regression-based and marker-based methods are dependent on the feature selection procedure, for this evaluation we used the best performing choices. We selected reference matrices constructed with CIBERSORTx and markers generated from scran-derived DE analysis. For MuSiC, we included both the default implementation which uses all genes and the composite approach MuSiC_CIBERSORTx.

We summarized the final results both in terms of average Pearson correlations and relative rankings. The comparison between homogeneous and heterogeneous conditions is insightful for assessing shifts in performance under varying simulation scenarios and determining whether performance is influenced by changes in heterogeneity levels. Focusing on the rank-based comparison (Fig. 5a), methods that demonstrate consistent performance in both settings are located diagonally, while methods with significant performance disparities are positioned off-diagonal, for example, methods on the top left excel in homogeneous conditions but underperform in heterogeneous scenarios. Across dataset being tested, the relative rankings of different methods can be indeed drastically different under different simulation settings.

Specifically, the reference-free methods (purple) appear to perform well on some homogeneous datasets but they are ranked towards the bottom in the heterogeneous settings in *all cases*. This pattern suggested that their performance is highly context dependent and can be easily impacted by the increased heterogeneity level. Reference-free methods are conceptually attractive and continue to be developed and our simulation approach thus provides a valuable evaluation platform.

One the other hand, marker-based methods (green) are clearly overrepresented among those methods whose ranking improved in the heterogeneous setting. Maker-based methods are widely used [63] despite the advance of more sophisticated and more accurate approaches. Our analysis showed that they are particularly resilient to heterogeneity, which is likely an important property contributing to their popularity.

Out of all the methods being evaluated, we found that the rankings of BayesPrism and MuSiC (with CIBERSORTx reference) are consistently either unaffected or improved under heterogeneous settings.

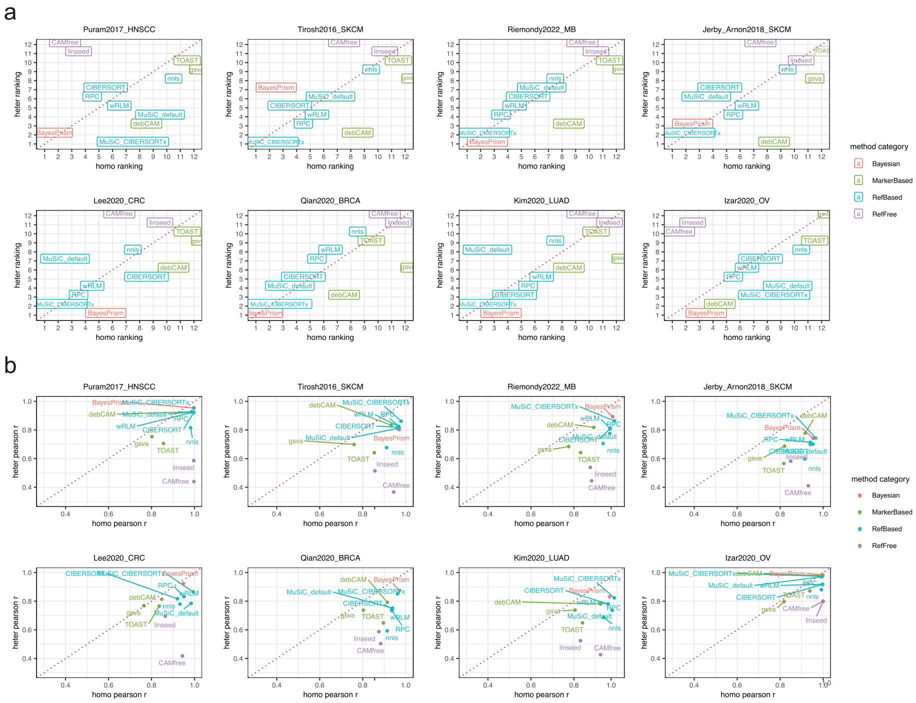

Fig. 5 Deconvolution performance comparison under homogeneous and heterogeneous simulation. Scatter plot comparing **a** the ranking and **b** the average Pearson correlations of different deconvolution methods under homo and heter simulations, where the correlations are averaged over multiple experimental repeats. Different colors indicate different categories of deconvolution methods. All the regression-based methods are using CIBERSORTx-derived reference and all the marker-based methods are using scran-derived markers, while "MuSiC_default" means the default MuSiC setting where all the genes are being used as input

We also provided the quantitative correlation plot as ranking can artificially inflate small differences. In the correlation based visualization (Fig. 5b), as expected we observed a systematic decrease in performance in the heterogeneous setting across all methods. It is not unusual to get correlations near 1 in the homogeneous setting, reflecting performance levels that can be unrealistically high, whereas in heterogeneous environments, correlations noticeably decline. This observation is further supported by a quantitative comparison of RMSE values under both homogeneous and heterogeneous settings (Additional file 1: Fig. S15). We expect that this indeed reflects real world performance. When deconvolution methods are tested against independent measurements of proportion as can be done with the malignant cell fractions in TCGA samples, even the best-performing methods BayesPrism does not achieve a correlation of > 0.8 [64]. Thus, beyond ranking methods the heterogeneous simulation provides value in producing realistic performance estimates.

Overall, we found that BaysePrism and MuSiC_CIBERSORTx stand out as best overall. BayesPrism ranks first in the heterogeneous setting in 5 out of 8 datasets and is in the top 3 in another 2. MuSiC_CIBERSORTx is another top performer ranking in the top 2 for 5 datasets and always in the top 3. We note that in the ranking performance of BayesPrism on Tirosh2016_SKCM is ranked 7 and appears inconsistent with this assessment. However, from quantitative correlation plot the low performance of BayesPrism on Tirosh2016_SKCM is less remarkable as it is quantitatively similar to other top performing methods. A detailed performance summary for all methods under heterogeneous setting can be found in Additional file 1: Fig. S16.

Taken together, our findings emphasized the impact of heterogeneity on deconvolution results and indicated that benchmarking using homogeneous samples may not accurately represent real-world performance. In realistic bulk simulation scenarios, the Bayesian method BayesPrism and the regression methods MuSiC stand out as the top-performing approaches. These results shed light on the importance of considering heterogeneity when evaluating and selecting deconvolution methods for practical applications.

Discussion

In this study, we introduced the importance of heterogeneity in bulk sample simulation and examined how heterogeneity could influence the deconvolution results. We investigated four major categories of deconvolution methods by applying them to simulated samples with different heterogeneity levels and identified the top performing ones. Our results showed that introducing biological heterogeneity has a notable effect on deconvolution performance, leading to a global performance drop as heterogeneity level increases, while some methods being more robust to this change.

Specifically, BayesPrism is one of the top-performed method across all the datasets we tested. Our results align with a recent study [65] that benchmarked deconvolution methods on real bulk and single cell data finding that BayesPrism strongly outperforms all tested reference based methods when evaluated for consistency across different biochemical and bioinformatic processing pipelines for the same biological sample. We reasoned that this can be attributed to its explicit consideration of heterogeneity within the malignant cells. The novelty of incorporating "cell-state" information within the same cell-type and reporting the posterior sum over various cell states makes BayesPrism particularly effective in handling heterogeneous settings.

Another method that exhibits high robustness in our evaluation is MuSiC, which also accounts for heterogeneity within a cell type. Specifically, MuSiC employs a weighting scheme that prioritizes genes with low cross-subject variance. It then solves the deconvolution task as a weighted non-negative least squares regression problem. Although it has been proposed that pre-selecting marker genes for the weighted regression is not necessary, our findings indicated that incorporating marker genes pre-filtered from CIBERSORTx can significantly enhance the deconvolution performance of MuSiC.

In addition to the aforementioned methods, it is worth noting the outstanding performance of the marker-based method debCAM. It ranked in the top 3 in 6 out 8 datasets surpassing many reference-based methods. It may appear counter-intuitive that maker-based methods can outperform regression-based ones as marker-based methods seemingly use less of the available prior information. However, in the heterogeneous setting, this may be an advantage. In the heterogeneous setting, the regression objective becomes only an approximation as the true cell-type means are altered and genes have considerable residual covariance. The various feature selection and weighing schemes using by top performing methods serve to account for this but may do so imperfectly.

On the other hand, marker-based methods are highly robust to these effects as the residual covariance of marker genes is low by construction and the exact mean values are not relevant.

In line with the view, BayesPrism presents an interesting case of a method that is fully quantitative but has some conceptual similarity with marker-based approach. BayesPrism uses the full reference matrix but does so in a way that does not rely on the exact reference values. One of the sampling steps of BayesPrism involves distributing the counts in the observed bulk expression for a gene over the current estimate of cell-type specific contributions with a multinomial distribution [66]. As such the absolute scale of gene expression in the reference matrix is not relevant, as the values are interpreted as probabilities and normalized to sum to 1. Moreover, in this setting, the contribution of a single gene to the final proportion estimate is directly proportional to its relative cell-type specificity times its expression value in the bulk sample. Consequently, the BayesPrism approach to a large degree negates the model misspecification sensitivity of regression-based methods.

Finally, our finding suggested that reference-free methods perform poorly under the heterogeneous scenario. Reference-free methods are attractive as they require no prior knowledge and it has been repeatedly suggested that these methods produce reliable proportion estimates [13, 14]. However, we showed that the more realistic the simulation strategy the accuracy is much lower than would be expected from the previously reported results. Reference-free methods rely on fitting a simplex structure which is the expected data geometry if the only source of variation is cell-type proportions. However, adding variation beyond cell-type proportions introduces additional lower dimensional structure making the proportion associated simplex difficult to isolate.

Our analysis provides valuable insights into the performance and tradeoffs of different conceptual approaches in a highly realistic simulation scenario thus establishing a framework for future methodological development. Beyond the specific deconvolution problem addressed in this work, the heterogeneous simulation strategy can be employed in other simulation pipelines to produce more realistic performance benchmarks for additional tasks such as cell-type specific differential expression [67].

We also acknowledge some limitations of our approach. While we demonstrated that our heterogeneous simulation strategy matches the variance observed in real bulk samples, not all aspects of real data will be preserved. For example, the dependencies between cell types can be violated as we randomly combine cell types from different patients. Chu et al. [15] found that certain biological pathway activation in malignant cells could be negatively correlated with cell type fractions of other non-malignant cells and the heterogeneous simulation we propose does not take into account such correlations. Methods that overcome this limitation would need to take the ground truth cell-type covariance into account necessitating development of new proportion sampling strategies.

Additionally, our study did not directly account for potential batch effects between the single-cell data used to generate references and the bulk data undergoing deconvolution. The benchmarking framework we introduced was based on a simplified scenario where matched single-cell data is available for reference construction. In real practice, there could be technical variation between signature matrices and bulk mixtures due to differences in assay platform [59, 68]. A complete evaluation for complex prior mis-specification will be the subject of future work.

Conclusions

Overall, our work suggests specific recommendations for creating realistic bulk simulations and highlights counterintuitive findings regarding the performance of deconvolution approaches from different conceptual classes. Together, we expect that these contributions will provide the groundwork for future methodological improvements.

Methods

Single-cell RNA seq datasets and quality control

A total of eight single-cell RNA sequencing datasets from seven tumor types are used in this paper (Table 1): (i) head and neck squamous cell carcinomas (HNSCC) from Puram et al. [25], (ii) melanoma (SKCM) from Tirosh et al. [29] and Jerby-Arnon et al. [33], (iii) medulloblastoma (MB) from Riemondy et al. [31], (iv) colorectal cancer from Lee et al. [35], (v) breast cancer from Qian et al. [37], (vi) lung adenocarcinoma from Kim et al. [39], and (vii) ovarian cancer from Izar et al. [41]. A detailed description of single-cell datasets used in this paper can be found in Additional file 2: Table S1.

UMI counts were converted to counts per million (CPM) prior to downstream simulation tasks. We removed genes that are expressed in less than 5 cells and discarded genes from mitochondrial or ribosomal content. All expression matrices are in linear (non-log) scale.

For melanoma dataset Tirosh2016_SKCM, we used cell-type labels re-classified in Schelker et al.'s [69, 70] study. For MB dataset Riemondy2022_MB, we re-annotated the immune population based on immune cell subtyping information from the interactive website of the original paper [71]. We included major immune cell types from their annotations for further study: DC, Neutrophil, NK cell, T cells. For all the macrophage subpopulations: chemokine myeloid, complement myeloid, M2-activated myeloid, and non-activated microglia, we relabeled them into macrophages to ensure a reasonable resolution of cell types. Immune cells that are classified as "Proliferate" or do not have any subtyping label are excluded from further study. For ovarian cancer dataset Izar2020_OV, we excluded cell type "Erythrocyte" due to limited single-cell size. For the remaining scRNA datasets, we used their original cell-type labels.

Bulk data used for variance comparison

When comparing the variance of the simulated bulk expression, we considered two types of "real" bulk data for comparison. First, baseline expression: we aggregated single cells from the same biological samples within the scRNA-seq data, which serves as an approximation of actual bulk samples [72, 73]. Specifically, this was achieved by calculating the mean expression from the CPM-normalized expression of cells from the same biological samples. The resulting profiles maintain consistent library sizes and are ready for variance analysis. Second, real bulk expression data from The Cancer Genome Atlas (TCGA): we accessed the TCGA cohorts from https://xenabrowser.net/. The gene expression data was acquired in HTSeq-FPKM format and subsequently transformed to

Transcript Per Million (TPM) for variance comparison. Only samples from primary and metastatic sites were selected for the analysis. A detailed description of the datasets used for variance comparison can be found in Additional file 2: Table S2.

Simulation of cell-type frequencies

Cell-type fractions simulation from beta distribution

To introduce variances into the cellular compositions of simulated bulk samples, we simulated cell-type frequencies that are close to that in real bulk samples. The cell-type proportions of each patient from the scRNA dataset were used as an approximation to the cell-type frequencies of real bulk samples. We fitted a beta distribution for each cell type and drew random values from the fitted distribution as the simulated frequencies. Randomly selected frequencies for different cell types are then scaled and summed to one for each simulated bulk sample. This fraction simulation method, refereed to as "beta" method in Additional file 1: Fig. S3, is implemented throughout the simulated bulk expression in this study.

Cell-type fractions simulation from external resources

In Additional file 1: Fig. S3, we evaluated the performance of our fraction simulation approach by comparing it with an external method employed by Avila Cobos et al. [9] (referred to as "favilaco"), and a basic fraction simulation utilizing the Dirichlet distribution. The "favilaco" method, which is part of the bulk simulation pipeline proposed by Avila et al., is now available as the *bulkSimulator_favilaco*() function in the deconvBenchmarking package. This function simultaneously simulates bulk expression data and fractions, which we used to extract and compare the simulated fraction distributions with other methods.

For the Dirichlet-based simulation, we set the shape parameter α to reflect the relative abundance of each cell type. Specifically, α is set to the cell type frequencies from the single-cell dataset used for bulk simulation and adjusted by a dispersion parameter to modulate the spread of the distribution. We considered a set of varying levels of dispersion parameters 0.01, 0.05, 0.001 in the simulation. The Dirichlet-based simulation was conducted using the *fracSimulator_Dirichlet*() function from the deconvBenchmarking package.

By comparing the distribution of the simulated fractions under different settings, we showed that "favilaco"-based method significantly deviates from the baseline's mean and variance, and while Dirichlet distribution exhibits comparable distributions, it requires additional tuning of the dispersion parameter, making it less efficient than the beta distribution based method we proposed.

Bulk simulation strategies

Using the same source scRNA-seq dataset and simulated cell-type fractions, each simulated bulk expression in this study was comprised of 100 simulated samples, created using the following strategies:

Homogeneous simulation

We standardized the widely used bulk simulation method that aggregates over random cells as homogeneous simulation. Specifically, in each simulated sample, n single cells are aggregated linearly, with their proportions aligned to the simulated cell-type frequencies. The value of n is set to approximate the typical number of single cells of a biological sample from the scRNA data (Additional file 2: Table S1).

Semi-heterogeneous simulation

We restricted that the malignant parts of each simulated bulk sample come from the same patient, while the non-malignant parts are randomly selected regardless of where they are from. Specifically, for each simulated bulk sample i, the malignant expression signal come exclusively from a randomly selected patient's malignant profile $C_{malignant}$ and is weighted according to the simulated malignant fraction, and the non-malignant single cells are randomly selected and weighted according to the corresponding simulated frequencies.

Heterogeneous simulation

We restricted that both malignant and non-malignant parts of each simulated bulk sample come from the same biological sample. Specifically, for each simulated bulk sample i, given a cell-type k, the expression signal of cell-type k comes exclusively from a randomly selected patient's k profile C_k and is weighted according to the simulated fraction (Additional file 1: Fig. S2).

Note that in both semi-heterogeneous and heterogeneous simulation settings, we employed additional randomization to prevent using the exact same cells across different simulated samples. Specifically, when aggregating patient-specific single cells to create a unique patient profile, we randomly select between 50 and 100% of the single cells. Moreover, we introduced a threshold parameter to specify the limited number of cells used for aggregation. If the number of patient-specific single cells falls below this threshold, we aggregate across multiple samples to prevent sparsity issues [44] within the cell-type specific profile. These two methods are now implemented as the *bulkSimulator_semi*() and *bulkSimulator_heter*() function in the deconvBenchmarking package.

sampleID independent heterogeneous simulation

We restricted that each cell-type component in the simulated bulk sample is constrained to originating from the same sub-cluster. Specifically, sub-cluster information for each cell type is obtained using the *quickCluster*() function from scran R package, with the *min.size* parameter set to 10 [74].

Bulk simulation from external resources

We incorporated three additional published approaches for bulk simulation comparison: the "favilaco" method [9], the "immunedeconv" method [17], and the "SCDC" method [23]. Note that only the "immunedeconv" method supports user-provided fractions;

therefore, we passed the simulated fractions to this method; the “favilaco” and “SCDC” methods do not support user-provided fractions, so we retained their default settings in the bulk simulation. These methods are implemented as *bulkSimulator_favilaco*(), *bulkSimulator_SCDC*(), and *bulkSimulator_immunedeconv*() function in the deconvBenchmarking package, which we used for bulk simulation.

Calculation of biological variance in bulk samples

The following statistics are calculated as indicators of biological variance. Note that the baseline expression referred to below represents pseudobulk samples from single-cell expression, which is obtained by aggregating single cells from the same patients, as an approximation of real bulk sample.

Coefficient of variation (CV)

For each gene i in the simulated and baseline bulk samples, we calculated CV values on the log transformed expression using the following formula:

$$CV_i = \frac{\sigma_i}{\mu_i}$$

Average coefficient of variation (CV) for biological pathways

We downloaded the hallmark gene list from https://www.gsea-msigdb.org/gsea/msigdb/ and calculated the average CV values for genes included in each genelist, which is used as indicators for pathway-level variance.

Pairwise correlations between genes

We considered two sets of genes for calculating the gene-gene correlation matrix: (1) the top 300 most variable genes from the baseline bulk expression and (2) the cell type marker genes derived from limma-based differential expression analysis. The detailed listing of the genes used can be found in Additional file 2: Table S3. Within each simulated bulk expression and the baseline expression, we calculated the Pearson correlations between these genes and visualized them in a heatmap [75, 76].

Pairwise similarities between samples

Using the top 300 most variable genes from the baseline bulk expression, we calculated the pairwise Pearson correlation between samples for each simulated bulk expression. We then visualized the distribution of these statistics using a boxplot (Additional file 1: Fig. S7).

Reference construction

To generate necessary input for reference-based methods (regression-based, marker-based, and BayesPrism), we applied the following reference construction methods using the training cells.

In particular, for interchangeability between the signature matrices and cell type markers, we applied the *refMarkers_sigMatrixList*() and *refMatrix_markerList*()

functions we developed in the deconvBenchmarking R package. Specifically, to create a signature matrix from a set of gene lists, we first averaged the expression values of single cells from the same cell type, resulting in a raw gene-by-cell-type matrix; we then refined this matrix by subsetting it with the marker genes. To identify cell-type markers from a given reference matrix, we assigned each gene to the cell type with the highest expression.

We note that for MuSiC, the regression-based method in our benchmarking, the values generated in the reference matrix are not directly used by MuSiC itself. Instead, they serve as a feature selection step, where the genes identified in the reference matrix are used as the *marker* input for MuSiC.

CIBERSORTx signature matrix

We used the "Create Signature Matrix" module from the online CIBERSORTx portal (https://CIBERSORTx.stanford.edu) to generate the signature matrix, with the training expression as input and all the parameters set to default values. We note that CIBERSORTx has a size limit for the input, so we downsampled the training cells and shrunk the input size when necessary. The resulting signature matrices are used directly as input for regression-based methods. Note that the CIBERSORTx-based reference typically contains thousands of genes in its signature. When converting this reference matrix to cell-type marker genes using the $refMatrix\_markerList()$ function, we employed the $maximum\_n = 100$ parameter to ensure that each cell type can have at most 100 marker genes, prioritized by the fold change in the reference matrix.

autogeneS signature matrix

The python package autogeneS [60] is utilized for signature matrix construction. We used the following parameters in the *optimize* function: $ngen = 3000, seed = 0, mode = fixed, nfeatures = 400$. The resulting optimized reference matrix with pareto index 0 is selected as autogeneS signature matrix.

limma derived cell-type specific markers

R package limma [61] is used to identify cell-type-specific markers that are differentially expressed within each cell-type. Specifically, the one-against-rest comparison is performed for the statistical test, comparing each cell-type against all other cell-types combined. Genes with a log fold change greater than 2 are considered as cell-type specific markers, and a parameter $minimum\_n = 5$ is introduced to ensure a sufficient number of markers for each cell type. Cell types with fewer than $minimum\_n$ genes passing the log2 fold change threshold will be excluded from the marker list. Note that for the Riemondy2022_MB and Jerby_Arnon2018_SKCM dataset, the log fold change threshold is relaxed to 1 to ensure proper number of cell-type specific markers.

scran-derived cell-type-specific markers

The $BayesPrism :: get.exp.stat()$ function, which incorporates the $pairwiseTTests()$ function from the scran package, is used to identify cell-type specific markers. Specifically, pairwise comparison between cell types is performed for the statistical test [74].

The marker gene filtering process here utilizes the same log fold change threshold and *minimum_n* parameter as applied in limma derived marker identification.

Reference for BayesPrism

We designated cell type labels from the single-cell dataset as *cell.type.label* and subclassified malignant populations based on their biological origins as *cell.state.label* in the *BayesPrism* :: *new.prism*() function, the same BayesPrism reference construction strategy used in Hippen et al.'s study [65]. We adhered to the reference derived directly from the scRNA data, bypassing the reference update step in BayesPrism, which entails generating a new reference from initial deconvolution results. This decision was based on our observation that utilizing the updated reference contributes minimally to performance enhancement (Additional file 1: Fig. S17).

Highly variable (hv) genes

For computational efficiency, we selected highly variable genes as candidates to run autogeneS and DE analysis. We used the *plot.scRNA.outlier*() function from R package BayesPrism to calculate the maximum cell-type specificity score for each gene. Genes with *max.spec* greater than a threshold value (0.5 for autogeneS and 0.3 for DE analysis) are selected for downstream analysis. This filtering narrows down the gene candidates from more than 10 thousands to thousands.

Deconvolution methods

Deconvolution methods are applied to the simulated bulk samples in their linear scale (non-log transformed) following recommendations from previous benchmarking studies [9, 16]. All methods evaluated in the study are wrapped in the deconvBenchmarking R package we developed: https://github.com/humengying0907/deconvBenchmarking.

Evaluation of deconvolution performance

Pearson correlation and root mean square error (RMES) values are used to evaluate the accuracy of different deconvolution methods. Specifically, for Pearson correlation, we calculated per cell-type correlations by comparing estimated fractions to known fractions within each cell type, with higher Pearson *r* corresponding to better performance. This process results in a set of per cell-type correlations for each simulated bulk dataset. These correlation values are then averaged to yield the overall correlation performance score for a given deconvolution method. For RMSE values, we focus on a global comparison between the estimated and known fractions for all cell types altogether, with smaller RMSE values indicating a lower absolute difference and thus better performance.

The averaged Pearson *r* across experimental repeats is used to represent the overall performance of a method. Variation in Pearson *r* and RMSE values across experimental repeats is visually inspected by comparing the minimum and maximum level across 10 experimental repeats, which help evaluate the reproducibility and stability of each method.

For reference-free deconvolution methods, where cell-type labels are not explicitly provided, we calculated pairwise Pearson correlations between the estimated and the known fractions and assign the unnamed cell type to the cell type with which it showed the highest correlation (Additional file 1: Fig. S18).

Multi-linear regression model of deconvolution performance

In order to evaluate the impact of methodological choices on deconvolution performance for regression-based methods, we utilized a linear modeling approach. Specifically, we fitted a linear model using the *lm*() function in R to predict Pearson correlation coefficients based on various predictors, incorporating dataset and cell types as covariates. The model formula employed was:

$$\text{model} = lm(\text{correlation} \sim 1 + \text{regression\_choice} + \text{reference\_choice} + \text{dataset} \times \text{cell\_type})$$

We designated nnls as the baseline regression choice, and no selection with MuSiC regression as the baseline reference choice for comparison. The coefficients obtained from the model indicate, conditional on all other effects, the relative impact of a specific methodological choice over the baseline. Higher coefficients indicate a more significant impact of this methodological choice on overall performance. The model was fitted separately for bulk data simulated from different approaches, in order to explore the effect shift of methodological choices under different conditions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03292-w.

Additional file 1. Supplementary figures S1-S18 referenced in the main text.

Additional file 2. Supplementary tables S1-S4 referenced in the main text.

Additional file 3. Review history.

Acknowledgements
We acknowledge the TCGA Research Network for data generated in https://www.cancer.gov/tcga. We acknowledge Casey Greene and Ariel Hippen for helpful discussion.

Review history
The review history is available as Additional file 3.

Peer review information
Andrew Cosgrove 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
All authors contributed equally to this work.

Funding
This study was supported by NSF 2238125, NIH R01 HG 009299-6A1, NIH R01 EY 030546-01A1, and DARPA W911NF-23-2-0023.

Availability of data and materials
All source code, simulated datasets, and detailed benchmarking results are available in Zenodo with DOI: 10.5281/zenodo.11238300 [77], under the open-source GPL-3.0 license.

The deconvBenchmarking R package developed in this study is deposited at the GitHub repository: https://github.com/humengying0907/deconvBenchmarking [78], under the open-source GPL-3.0 license.

All data used in this study were obtained from publicly available repositories:

1 Single cell head and neck cancer dataset Puram2017_HNSCC is downloaded from the GEO database under the accession code GSE103322 [28].

2 Single cell melanoma dataset Tirosh2016_SKCM is downloaded from the GEO database under the accession code GSE72056 [30]; the single cell annotation data is downloaded from https://github.com/icbi-lab/immune_deconvolution_benchmark [69, 70].

3 Single cell medulloblastoma dataset Riemondy2022_MB is downloaded from GEO database under the accession code GSE155446 [32]; the immune cells meta data is downloaded from the online interactive browser [71].

4 Single cell melanoma dataset Jerby_Arnon2018_SKCM is downloaded from GEO database under the accession code GSE115978 [34].

5 Single cell colorectal cancer dataset Lee2020_CRC dataset is downloaded from the 3CA database (https://www.weizmann.ac.il/sites/3CA/colorectal) under the Title identifier "Lee et al. 2020" [36].
6 Single cell breast cancer dataset Qian2020_BRCA is downloaded from the 3CA database (https://www.weizmann.ac.il/sites/3CA/breast) under the Title identifier "Qian et al. 2020" [38].
7 Single cell lung cancer dataset is downloaded from the 3CA database (https://www.weizmann.ac.il/sites/3CA/lung) under the Title identifier "Kim et al. 2020" [40].
8 Single cell ovarian cancer dataset is downloaded from the 3CA database (https://www.weizmann.ac.il/sites/3CA/ovarian) under the Title identifier "Izar et al. 2020" [42].
9 TCGA Melanoma cohort is downloaded from UCSC Xena Browser under the cohort identifier "GDC TCGA Melanoma (SKCM)" [79].
10 TCGA Head and Neck Cancer cohort is downloaded from UCSC Xena Browser under the cohort identifier "GDC TCGA Head and Neck Cancer (HNSC)" [80].
11 Hallmark gene sets used in this paper are downloaded from The Molecular Signatures Database (MSigDB): https://www.gsea-msigdb.org/gsea/msigdb/ [81].

Declarations

Ethics approval and consent to participate

Not applicable for this study.

Competing interests

The authors declare that they have no competing interests.

Received: 20 January 2023 Accepted: 29 May 2024
Published online: 01 July 2024

References

1. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics (Oxford, England). 2018;34(11):1969–79. https://doi.org/10.1093/bioinformatics/bty019.
2. 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. https://doi.org/10.1038/s41467-022-28655-4.
3. Zhang H, Lu X, Lu B, Gullo G, Chen L. Measuring the composition of the tumor microenvironment with transcriptome analysis: past, present and future. Futur Oncol. 2024. https://doi.org/10.2217/fon-2023-0658.
4. Bruno TC. B cells to the forefront of immunotherapy. Nature. 2020;577(7791):474–6. https://doi.org/10.1038/d41586-019-03943-0.
5. Barrett RL, Puré E. Cancer-associated fibroblasts and their influence on tumor immunity and immunotherapy. eLife. 2020;9:e57243. https://doi.org/10.7554/eLife.57243.
6. Raskov H, Orhan A, Christensen JP, Gögenur I. Cytotoxic CD8+ T cells in cancer and cancer immunotherapy. Br J Cancer. 2021;124(2):359–67. https://doi.org/10.1038/s41416-020-01048-4.
7. González-Navajas JM, Fan DD, Yang S, Yang FM, Lozano-Ruiz B, Shen L, et al. The Impact of Tregs on the Anticancer Immunity and the Efficacy of Immune Checkpoint Inhibitor Therapies. Front Immunol. 2021;12:625783. https://doi.org/10.3389/fimmu.2021.625783.
8. Darvin P, Toor SM, Sasidharan Nair V, Elkord E. Immune checkpoint inhibitors: recent progress and potential biomarkers. Exp Mol Med. 2018;50(12):1–11. https://doi.org/10.1038/s12276-018-0191-1.
9. 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. https://doi.org/10.1038/s41467-020-19015-1.
10. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220. https://doi.org/10.1186/s13059-017-1349-1.
11. Zhong Y, Wan YW, Pang K, Chow LML, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. 2013;14:89. https://doi.org/10.1186/1471-2105-14-89.
12. Chen L, Wu CT, Wang N, Herrington DM, Clarke R, Wang Y. debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. Bioinformatics. 2020;36(12):3927–9. https://doi.org/10.1093/bioinformatics/btaa205.
13. Wu CT, Chen L, Herrington DM, Shen M, Yu G, Clarke R, et al. Determining molecular archetype composition and expression from bulk tissues with unsupervised deconvolution. bioRxiv. 2021. https://doi.org/10.1101/2021.07.12.452047.
14. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat Commun. 2019;10(1):2209. https://doi.org/10.1038/s41467-019-09990-5.
15. 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. https://doi.org/10.1038/s43018-022-00356-3.
16. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. Genome Biol. 2021;22(1):102. https://doi.org/10.1186/s13059-021-02290-6.
17. 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 (Oxford, England). 2019;35(14):i436–45. https://doi.org/10.1093/bioinformatics/btz363.

18. Chen Y, Wang Y, Chen Y, Cheng Y, Wei Y, Li Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nat Commun. 2022;13(1):6735. https://doi.org/10.1038/s41467-022-34550-9.
19. 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. https://doi.org/10.1186/s13073-019-0638-6.
20. Andrade Barbosa B, van Asten SD, Oh JW, Farina-Sarasqueta A, Verheij J, Dijk F, et al. Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. Nat Commun. 2021;12(1):6106. https://doi.org/10.1038/s41467-021-26328-2.
21. Gavish A, Tyler M, Greenwald AC, Hoefflin R, Simkin D, Tschernichovsky R, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. Nature. 2023;618(7965):598–606. https://doi.org/10.1038/s41586-023-06130-4.
22. Torroja C, Sanchez-Cabo F. Digitaldlsorter: Deep-Learning on scRNA-Seq to Deconvolute Gene Expression Data. Front Genet. 2019;10:978. https://doi.org/10.3389/fgene.2019.00978.
23. 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 Bioinforma. 2021;22(1):416–27. https://doi.org/10.1093/bib/bbz166.
24. 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. https://doi.org/10.1126/sciadv.aba2619.
25. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171(7):1611-1624.e24. https://doi.org/10.1016/j.cell.2017.10.044.
26. Boyeau P, Hong J, Gayoso A, Jordan MI, Azizi E, Yosef N. Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics. bioRxiv. 2022. https://doi.org/10.1101/2022.10.04.510898.
27. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8. https://doi.org/10.1038/s41587-020-0546-8.
28. Tirosh I, Puram S, Parikh A. Single cell RNA-seq analysis of head and neck cancer. Gene Expr Omnibus. 2017. https://identifiers.org/geo:GSE103322. Accessed Apr 2022.
29. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189–96. https://doi.org/10.1126/science.aad0501.
30. Tirosh I, Izar B. Single cell RNA-seq analysis of melanoma. Gene Expr Omnibus. 2016. https://identifiers.org/geo:GSE72056. Accessed Apr 2022.
31. Riemondy KA, Venkataraman S, Willard N, Nellan A, Sanford B, Griesinger AM, et al. Neoplastic and immune single-cell transcriptomics define subgroup-specific intra-tumoral heterogeneity of childhood medulloblastoma. Neuro-Oncol. 2022;24(2):273–86. https://doi.org/10.1093/neuonc/noab135.
32. Riemondy K, Donson A. Single-cell RNAseq of childhood medulloblastoma. Gene Expr Omnibus. 2021. https://identifiers.org/geo:GSE155446. Accessed May 2022.
33. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175(4):984-997.e24. https://doi.org/10.1016/j.cell.2018.09.006.
34. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al.. Single-cell RNA-seq of melanoma ecosystems reveals sources of T cells exclusion linked to immunotherapy clinical outcomes. Gene Expr Omnibus. 2018. https://identifiers.org/geo:GSE115978. Accessed May 2022.
35. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52(6):594–603. https://doi.org/10.1038/s41588-020-0636-z.
36. Curated Cancer Cell Atlas Research Group. The Curated Cancer Cell Atlas: Colorectal Cancer Data from Lee et al., 2020. Curated Cancer Cell Atlas. 2023. https://www.weizmann.ac.il/sites/3CA/colorectal. Accessed Sept 2023.
37. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res. 2020;30(9):745–62. https://doi.org/10.1038/s41422-020-0355-0.
38. Curated Cancer Cell Atlas Research Group. The Curated Cancer Cell Atlas: Breaset Cancer Data from Qian et al., 2020. Curated Cancer Cell Atlas. 2023. https://www.weizmann.ac.il/sites/3CA/breast. Accessed Sept 2023.
39. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11(1):2285. https://doi.org/10.1038/s41467-020-16164-1.
40. Curated Cancer Cell Atlas Research Group. The Curated Cancer Cell Atlas: Lung Cancer Data from Kim et al., 2020. Curated Cancer Cell Atlas. 2023. https://www.weizmann.ac.il/sites/3CA/lung. Accessed Sep 2023.
41. Izar B, Tirosh I, Stover EH, Wakiro I, Cuoco MS, Alter I, et al. A single-cell landscape of high-grade serous ovarian cancer. Nat Med. 2020;26(8):1271–9. https://doi.org/10.1038/s41591-020-0926-0.
42. Curated Cancer Cell Atlas Research Group. The Curated Cancer Cell Atlas: Ovarian Cancer Data from Izar et al., 2020. Curated Cancer Cell Atlas. 2023. https://www.weizmann.ac.il/sites/3CA/ovarian. Accessed Sep 2023.
43. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004.
44. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. 2022;23(1):31. https://doi.org/10.1186/s13059-022-02601-5.
45. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018;15(2):81–94. https://doi.org/10.1038/nrclinonc.2017.166.
46. Northcott PA, Robinson GW, Kratz CP, Mabbott DJ, Pomeroy SL, Clifford SC, et al. Medulloblastoma. Nat Rev Dis Prim. 2019;5(1):1–20. https://doi.org/10.1038/s41572-019-0063-6.

47. Turashvili G, Brogi E. Tumor Heterogeneity in Breast Cancer. Front Med. 2017;4. https://www.frontiersin.org/articles/10.3389/fmed.2017.00227.
48. Gershanov S, Madiwale S, Feinberg-Gorenshtein G, Vainer I, Nehushtan T, Michowiz S, et al. Classifying Medulloblastoma Subgroups Based on Small, Clinically Achievable Gene Sets. Front Oncol. 2021;11:637482. https://doi.org/10.3389/fonc.2021.637482.
49. Northcott PA, Shih DJH, Remke M, Cho YJ, Kool M, Hawkins C, et al. Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. Acta Neuropathol. 2012;123(4):615–26. https://doi.org/10.1007/s00401-011-0899-7.
50. 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. https://doi.org/10.1038/s41467-018-08023-x.
51. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics. 2017;18(1):105. https://doi.org/10.1186/s12859-017-1511-5.
52. Li H, Sharma A, Ming W, Sun X, Liu H. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. BMC Genomics. 2020;21(1):652. https://doi.org/10.1186/s12864-020-06888-1.
53. 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. https://doi.org/10.1038/nmeth.3337.
54. Mullen KM, van Stokkum IHM. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). 2012. R package version 1.4. https://CRAN.R-project.org/package=nnls. Accessed Apr 2022.
55. Li Z, Wu H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome Biol. 2019;20(1):190. https://doi.org/10.1186/s13059-019-1778-0.
56. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14(1):7. https://doi.org/10.1186/1471-2105-14-7.
57. Huang P, Cai M, Lu X, McKennan C, Wang J. Accurate estimation of rare cell type fractions from tissue omics data via hierarchical deconvolution. Ann Appl Stat. 2024;18(2):1178-94. https://doi.org/10.1214/23-AOAS1829.
58. Garmire LX, Li Y, Huang Q, Xu C, Teichmann SA, Kaminski N, et al. Challenges and perspectives in computational deconvolution of genomics data. Nat Methods. 2024;21(3):391–400. https://doi.org/10.1038/s41592-023-02166-6.
59. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82. https://doi.org/10.1038/s41587-019-0114-2.
60. Aliee H, Theis FJ. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. Cell Syst. 2021;12(7):706-715.e4. https://doi.org/10.1016/j.cels.2021.05.006.
61. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. https://doi.org/10.1093/nar/gkv007.
62. 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. https://doi.org/10.1371/journal.pcbi.1006976.
63. Aran D. Cell-Type Enrichment Analysis of Bulk Transcriptomes Using xCell. Methods Mol Biol. 2020;2120:263-276. https://doi.org/10.1007/978-1-0716-0327-7_19.
64. Chu T, Wang Z, Pe'er D, Danko CG. BayesPrism Extended Data Figure 9: Comparison between tumor purity inferred by BayesPrism, CIBERSORTx, ABSOLUTE, ESTIMATE and IHC. 2022. (Extended Data Figure 9). https://doi.org/10.1038/s43018-022-00356-3. Accessed Sept 2023.
65. Hippen AA, Omran DK, Weber LM, Jung E, Drapkin R, Doherty JA, et al. Performance of computational algorithms to deconvolve heterogeneous bulk ovarian tumor tissue depends on experimental factors. Genome Biol. 2023;24(1):239. https://doi.org/10.1186/s13059-023-03077-7.
66. Chu T, Wang Z, Pe'er D, Danko CG. BayesPrism Source code. 2022. https://github.com/Danko-Lab/BayesPrism/blob/36e6d26f50d402876ee4e64c35e181465b58dda7/BayesPrism/R/run_gibbs.R#L57-L59. Accessed Mar 2023.
67. Meng G, Tang W, Huang E, Li Z, Feng H. A comprehensive assessment of cell type-specific differential expression methods in bulk data. Brief Bioinforma. 2023;24(1):bbac516. https://doi.org/10.1093/bib/bbac516.
68. 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. https://doi.org/10.1038/s41467-020-15816-6.
69. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8(1):2032. https://doi.org/10.1038/s41467-017-02289-3.
70. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Single cell dataset curated by Schelker et al. and Sturm et al.; 2018. Github Repository. https://github.com/icbi-lab/immune_deconvolution_benchmark. Accessed Apr 2022.
71. Riemondy KA, Venkataraman S, Willard N, Nellan A, Sanford B, Griesinger AM, et al. Interactive Visualization of singlg-cell medulloblastoma data using UCSC CellBrowser. 2021. https://d33sxa6bpqwi51.cloudfront.net. Accessed Aug 2022.
72. Murphy AE, Skene NG. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun. 2022;13(1):7851. https://doi.org/10.1038/s41467-022-35519-4.
73. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12(1):5692. https://doi.org/10.1038/s41467-021-25960-2.
74. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. F1000Res. 2016;5:2122. https://doi.org/10.12688/f1000research.9501.2.
75. Gu Z. Complex heatmap visualization. iMeta. 2022;1(3):e43. https://doi.org/10.1002/imt2.43.
76. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. Bioinformatics (Oxford, England). 2014;30(19):2811–2. https://doi.org/10.1093/bioinformatics/btu393.
77. Hu M, Chikina M. Source code and materials in deconvolution benchmarking study. 2024. Zenodo. https://doi.org/10.5281/zenodo.8206516.

78. Hu M, Chikina M. deconvBenchmarking R package. 2024. Github. https://github.com/humengying0907/deconvBenchmarking. Accessed Aug 2023.
79. The Cancer Genome Atlas Research Network. GDC TCGA Melanoma (SKCM) cohort. Xena browswer. https://xenabrowser.net/datapages/. Accessed May 2023.
80. The Cancer Genome Atlas Research Network. GDC TCGA Head and Neck Cancer (HNSC) cohort. Xena browswer. https://xenabrowser.net/datapages/. Accessed May 2023.
81. Broad Institute of MIT and Harvard. Molecular Signatures Database (MSigDB). https://www.gsea-msigdb.org/gsea/msigdb/. Accessed Aug 2022.

Publisher's Note

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