DeepDeconUQ Estimates Malignant Cell Fraction Prediction Intervals in Bulk RNA-seq Tissue

Jiawei Huang[1], Yuxuan Du[1,4], Kevin R. Kelly[3], Jinchi Lv[5], Yingying Fan[5], Jiang F. Zhong[2*], and Fengzhu Sun [†1*]

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, 90089, USA
[2]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA, 92350, USA
[3]Division of Hematology, University of Southern California, Los Angeles, CA, 90089, USA
[4]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX, 78249, USA
[5]Data Sciences and Operations Department, University of Southern California, Los Angeles, CA, 90089, USA
[*]Correspondence: jzhong@llu.edu, fsun@usc.edu

Abstract

Accurate estimation of malignant cell fractions in tissues plays a critical role in cancer diagnosis, prognosis, and subsequent treatment decisions. However, most currently available methods provide only point estimates, neglecting the quantification of uncertainties, which is essential for both clinical and research applications. This study introduces DeepDeconUQ, a deep neural network model developed to estimate prediction intervals for malignant cell fractions based on bulk RNA-seq data. This approach addresses limitations in current malignant cell fraction estimation methods by integrating uncertainty quantification into predictions of cancer cell fractions. DeepDeconUQ leverages single-cell RNA sequencing (scRNA-seq) data in conjunction with conformalized quantile regression to produce reliable prediction intervals. The model trains a quantile regression neural network to establish upper and lower bounds for cancer cell proportions, followed by a calibration step that refines these intervals to ensure both statistical validity (coverage probability) and discrimination (narrow intervals). Benchmark analyses indicate that DeepDeconUQ consistently surpasses existing methods, achieving high coverage accuracy with tight prediction intervals across simulated and real cancer datasets. The robustness of DeepDeconUQ is further demonstrated by its resilience to various gene expression perturbations. The DeepDeconUQ method is publicly accessible at `https://github.com/jiaweih14/DeepDeconUQ`.

Keywords

Malignant Cell Fraction Estimation, Uncertainty Quantification, Prediction Interval, Deep Neural Network

Author Summary

Accurately determining the proportion of malignant cells in tumor tissues is crucial for cancer diagnosis and treatment planning. Current methods often provide single estimates without indicating the uncertainty, which can lead to overconfidence in clinical decisions. Here, we present

[†]Lead Contact

DeepDeconUQ, a deep learning tool that not only predicts the fraction of malignant cells in bulk RNA sequencing data but also quantifies the uncertainty around these estimates. By leveraging single-cell RNA sequencing data to simulate realistic tumor samples, DeepDeconUQ trains a neural network to generate prediction intervals—ranges within which the true malignant cell fraction is likely to lie with high probability. This approach combines quantile regression and statistical calibration to ensure reliability without restrictive assumptions about data distribution. When tested on both simulated and real-world datasets, DeepDeconUQ consistently outperformed existing methods, delivering precise intervals that reliably capture true values while remaining robust against technical noise in gene expression measurements. Our tool addresses a critical gap in cancer genomics by providing clinicians and researchers with confidence intervals that enhance the interpretability of bulk tissue analyses. This advancement could improve personalized treatment strategies and reduce errors in downstream research applications.

Introduction

Recent advancements in next-generation sequencing methodologies, particularly bulk RNA sequencing (RNA-seq) and single-cell RNA sequencing (scRNA-seq), have substantially driven progress across biological and medical research domains[1–4]. One prominent application is to estimate malignant cell fraction from bulk RNA-seq samples[5–9]. This process typically involves using regression-based methods that leverage malignant and normal expression data (e.g., scRNA-seq) as a reference profile[10]. Most available estimation methods merely provide point estimates of cell-type proportions from bulk RNA-seq data[5,6]. The accuracy of these methods often depends on the choice and quality of the reference profile[8]. Furthermore, limited efforts have been made to investigate and quantify the impacts of uncertainties in estimated cell-type proportions, which can critically impact downstream analyses in malignant-cell-associated disease research, leading to potential errors in findings[11]. Uncertainty quantification of the estimated malignant cell fraction is thus essential, as is the quantification of prediction accuracy.

Uncertainty in malignant cell fraction estimation can be quantified through prediction intervals, which provide a range within which the true cell-type composition is likely to fall with a high probability[12,13]. An ideal procedure for generating prediction intervals should satisfy two properties. The first property is validity[14]. It should provide valid coverage in finite samples without making strong distributional assumptions, such as normality. The second property is discrimination[12]. The predicted intervals should be as narrow as possible at each point in the input space so that the predictions will be informative. When the data is heteroscedastic, getting valid but narrow prediction intervals requires adjusting the lengths of the intervals according to the local variability at each query point in the predictor space.

RNA-Sieve[9] and MEAD[7] are two statistical methods that have been proposed recently that can be used to estimate cell-type proportions and, in the meantime, quantify the uncertainties of the estimated cell proportions. RNA-Sieve[9] is a likelihood-based deconvolution method. It assumes that the estimates of cell-type fractions are normally distributed around the true fractions. Meanwhile, the errors arising from the gene expression profile and observed bulk gene expressions are independent. Therefore, the confidence intervals of the cell proportions can be calculated through likelihood estimation. However, these assumptions may not hold consistently in practice, as gene expression levels within samples (either bulk or single-cell) often exhibit inter-gene dependencies due to coregulation mechanisms[15]. MEAD[7], another statistical inference approach, incorporates a gene-gene dependency structure to improve the accuracy of cell proportion estimates. MEAD asserts that the estimated proportions follow asymptotic normal distributions, with solutions constrained to non-negative values. While MEAD considers the correlation across different genes, the assumption that individuals in the bulk and reference

data are from the same population may not hold universally, especially in contexts like cancer research, where gene expression levels vary greatly in different populations. Moreover, the dependence matrix used in MEAD is highly dependent on the choice of bulk samples and cannot be generated when there is only one single bulk sample to decompose.

In this study, we introduce DeepDeconUQ, a deep learning model that is distribution-agnostic and designed to estimate prediction intervals for malignant cell compositions in bulk RNA-seq data. DeepDeconUQ trains a neural network on simulated bulk RNA-seq data, avoiding parametric assumptions about bulk gene expression distributions. Through conformalized quantile regression[14], it provides both valid and precise prediction intervals for malignant cell fractions. Specifically, DeepDeconUQ employs scRNA-seq data to simulate artificial bulk RNA-seq datasets with predefined malignant cell proportions. These simulated datasets are then used to train a quantile regression neural network, which predicts the lower and upper bounds of malignant cell proportions in new cancer tissue samples. Following this, a conformal prediction process is applied to a separate calibration dataset of artificial bulk RNA-seq to adjust the intervals generated by the neural network. This conformalization step ensures that the estimated malignant cell proportions achieve stronger coverage guarantees. Benchmarking with both simulated and real datasets demonstrates that DeepDeconUQ surpasses existing methods in performance and remains robust against perturbations in gene expression levels. By leveraging scRNA-seq data, employing deep neural networks, and utilizing conformalized quantile regression, DeepDeconUQ achieves superior performance in cancer cell deconvolution analysis with uncertainty quantification.

Results

Methods overview

Fig 1 provides a schematic representation of DeepDeconUQ. The framework begins with single-cell RNA sequencing (scRNA-seq) datasets, where the cells from each subject are assumed to have labeled cell types (malignant or normal) and known gene expression profiles. The scRNA-seq data is a gene expression matrix where each row is a single cell sample, and each column is a gene. To simulate bulk RNA-seq data, first, we randomly select certain numbers of malignant and normal cells with replacement. Second, the bulk gene expression profile can be generated by summing up the gene expression values of the selected cells (Fig 1A). These processes are repeated many times to generate a large number of simulated bulk sequencing data. These simulated bulk RNA-seq datasets are then divided into two disjoint groups: a training set and a calibration set. Specifically, 70% of the data is randomly selected for training a highly accurate quantile function, while the remaining 30% is reserved for conformal calibration. After the TF-IDF transformation and MinMax normalization, the trained model uses bulk RNA-seq data x and a predefined significance level α as input and outputs predictions of the lower and upper bounds for malignant cell fractions, $\{\hat{\mathrm{q}}_{\alpha_{lo}}(x), \hat{\mathrm{q}}_{\alpha_{hi}}(x)\}$ (Fig 1B). Following model training, the calibration set is employed to compute conformity scores using equation 10. The adjustment minimizes both the risk of overly conservative predictions (over-coverage) and the potential for overly narrow intervals that miss true values (under-coverage) (Fig 1C). Finally, for a real bulk sample, DeepDeconUQ firstly uses the neural network to get an estimate of the prediction interval and then makes use of the conformity score to adjust the prediction interval $\hat{C}(X_{n+1})$ (Fig 1D). This prediction interval provides a measure of uncertainty, offering a more reliable estimate of the malignant cell fractions within a bulk RNA-seq sample.

Our model was constructed using artificial bulk RNA-seq samples and evaluated through the leave-one-out cross-validation. The evaluation is based on validity and discrimination. For

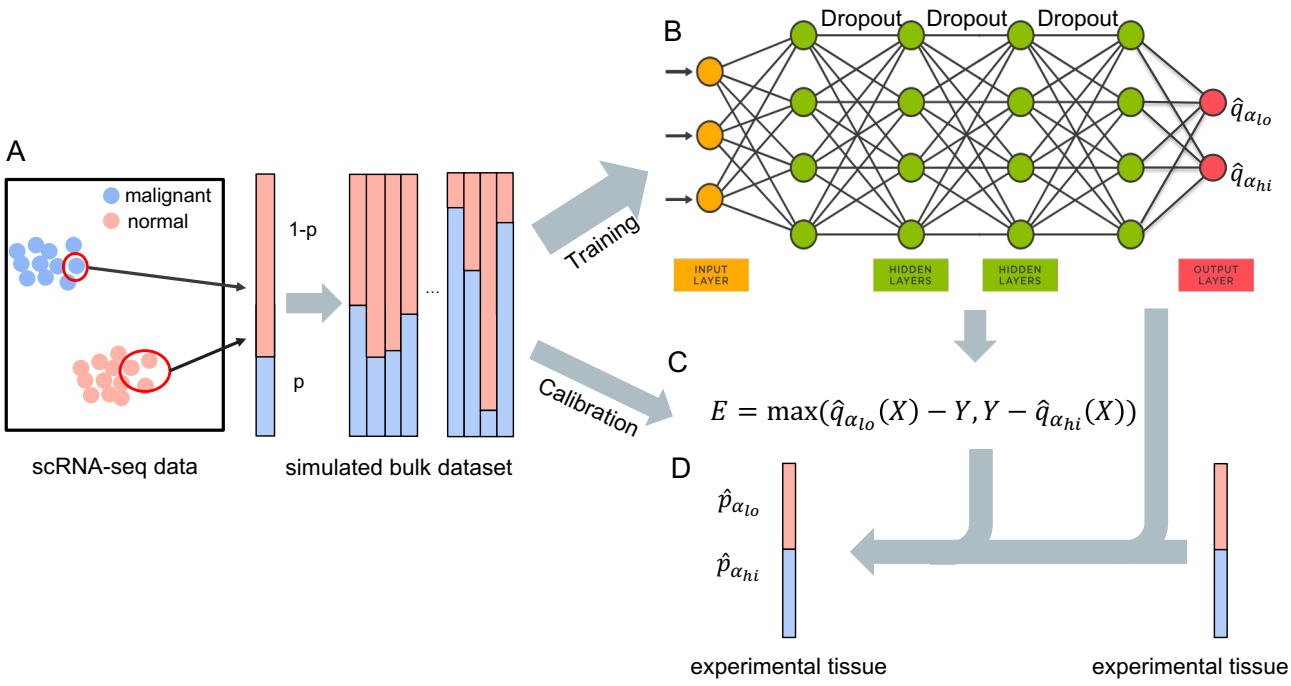

Fig 1: **Overview of DeepDeconUQ. A**: Constructing simulated bulk RNA-seq samples with different fractions of malignant cells. p is the fraction of malignant cells in a simulated bulk sample. **B**: Model structure used to train DeepDeconUQ. It consists of four fully connected layers with dropout layers. Seventy percent of the simulated data are used for training. The output is two quantile functions at a given significance level α. **C**: Conformity scores are calculated on the remaining 30% of the simulated dataset. **D**: Estimating the prediction interval of malignant cells from a real bulk sample. The trained model is used to calculate the lower and upper bounds, and the conformity scores are used to adjust the quantiles, which finally outputs the prediction interval $\{\hat{\mathrm{p}}_{\alpha_{lo}}, \hat{\mathrm{p}}_{\alpha_{hi}}\}$.

validity, we check the coverage rate, defined as the frequency of true malignant cell fraction within the prediction interval of the testing dataset (see Formula 6). For discrimination, we use the average length of prediction interval of the testing datasets as an evaluation metric (see Formula 2).

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{\mathrm{p}}_{i,\alpha_{lo}} \leq y_i \leq \hat{\mathrm{p}}_{i,\alpha_{hi}}), \tag{1}$$

$$\mathrm{L}_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{\mathrm{p}}_{i,\alpha_{hi}} - \hat{\mathrm{p}}_{i,\alpha_{lo}}|, \tag{2}$$

where y_i is the true malignant cell fraction of the ith sample in the testing dataset. $\hat{\mathrm{p}}_{i,\alpha_{lo}}$ and $\hat{\mathrm{p}}_{i,\alpha_{hi}}$ are the corresponding lower and upper bounds of the ith sample's prediction interval. n is the total number of samples in the testing dataset, and $\mathbb{1}(x)$ is an indicator function of 1 when x is true and 0 otherwise.

For each subject, we generated the simulated bulk datasets as described in the Dataset simulation subsection separately. Leave-one-out cross-validation was used to evaluate model performance across subjects during simulation. Specifically, we selected one of the k artificial bulk RNA-seq datasets as the testing dataset, while the remaining $k-1$ datasets served as the training set. This process was repeated k times to fully evaluate the performance of our model.

For real-world dataset applications, we aggregated all k artificial bulk RNA-seq datasets to train a unified model, which was subsequently validated using real data.

DeepDeconUQ outperforms other methods for estimating the prediction interval of malignant cell fraction

To assess the performance of DeepDeconUQ, we conducted a comparative analysis against two alternative methods, RNA-Sieve (v. 0.1.4)[9] and MEAD (v. 1.0.1)[7], both of which have been proposed in the literature to quantify uncertainties in estimated cell-type proportions. This evaluation was performed on both simulated and real bulk RNA-seq datasets. Since RNA-Sieve and MEAD are statistical inference methods and do not include a step for simulating artificial bulk RNA-seq datasets for model training, we utilized the scRNA-seq data directly as the reference for these methods. The same scRNA-seq data were also employed to generate the synthetic bulk RNA-seq datasets for DeepDeconUQ. All benchmarking methods were executed using their default configurations, ensuring a consistent basis for comparison. Additionally, the methods were evaluated on identical test datasets, which were kept separate from the training datasets used to develop the models. Details of implementations of these compared methods are explained in Supplementary Information, Section 6.

Fig 2 presents boxplots illustrating coverage and average prediction interval lengths for 15 simulated bulk RNA-seq datasets at three significance levels (15%, 10%, and 5%). Although RNA-Sieve maintains relatively narrow prediction intervals, it often fails to meet the coverage criterion across the datasets, indicating a tendency toward marked undercoverage. This suggests that RNA-Sieve's intervals may be too narrow to reliably contain the true malignant fraction. In contrast, MEAD achieves the coverage criterion for some datasets but exhibits considerable variability in prediction interval lengths, with some interval lengths extending beyond 0.6. Such substantial intervals lead to overcoverage, reducing interpretability by producing intervals that are too broad to offer precise estimates. DeepDeconUQ demonstrates superior performance across all three methods on the simulation datasets, consistently satisfying the coverage requirement while maintaining tight prediction intervals. This performance advantage is attributed to two primary factors: first, the neural network's effective quantile learning enables it to meet the coverage criterion; second, the well-trained model generates low conformity scores on the calibration set, ensuring that the quantile of these scores remains sufficiently small to yield narrow prediction intervals.

We further evaluated the performance of these three methods on real AML datasets, including 'primary,' 'recurrent,' and 'BeatAML' samples, one real Neuroblastoma dataset, and one real HNSCC dataset. As illustrated in Table 1, RNA-Sieve consistently has the worst performance, with its average prediction interval length fixed at 1.0, indicating it predicts 0.0 as the lower bound and 1.0 as the upper bound for every real sample. This likely stems from RNA-Sieve's limitations in handling gene expression data sourced from diverse sequencing protocols. Consequently, while RNA-Sieve can provide an estimate of malignant cell fraction, the results lack reliability. MEAD, conversely, accounts for variations in sequencing depth and tissue sample size, thus yielding relatively robust performance on real datasets. DeepDeconUQ demonstrates an even higher capability by addressing batch effects and sequencing biases via TF-IDF transformation and Min-Max normalization, achieving superior performance relative to MEAD, with more consistent coverage and narrower prediction intervals across the real datasets. Fig 3 depicts the prediction intervals generated by DeepDeconUQ and MEAD on the real primary dataset at α = 0.05 (95% confidence level). Given that RNA-Sieve consistently generated maximum-width prediction intervals (0.0-1.0) on real AML datasets, we restricted our visualization analysis to DeepDeconUQ and MEAD. The visualization clearly demonstrates that MEAD failed to encompass

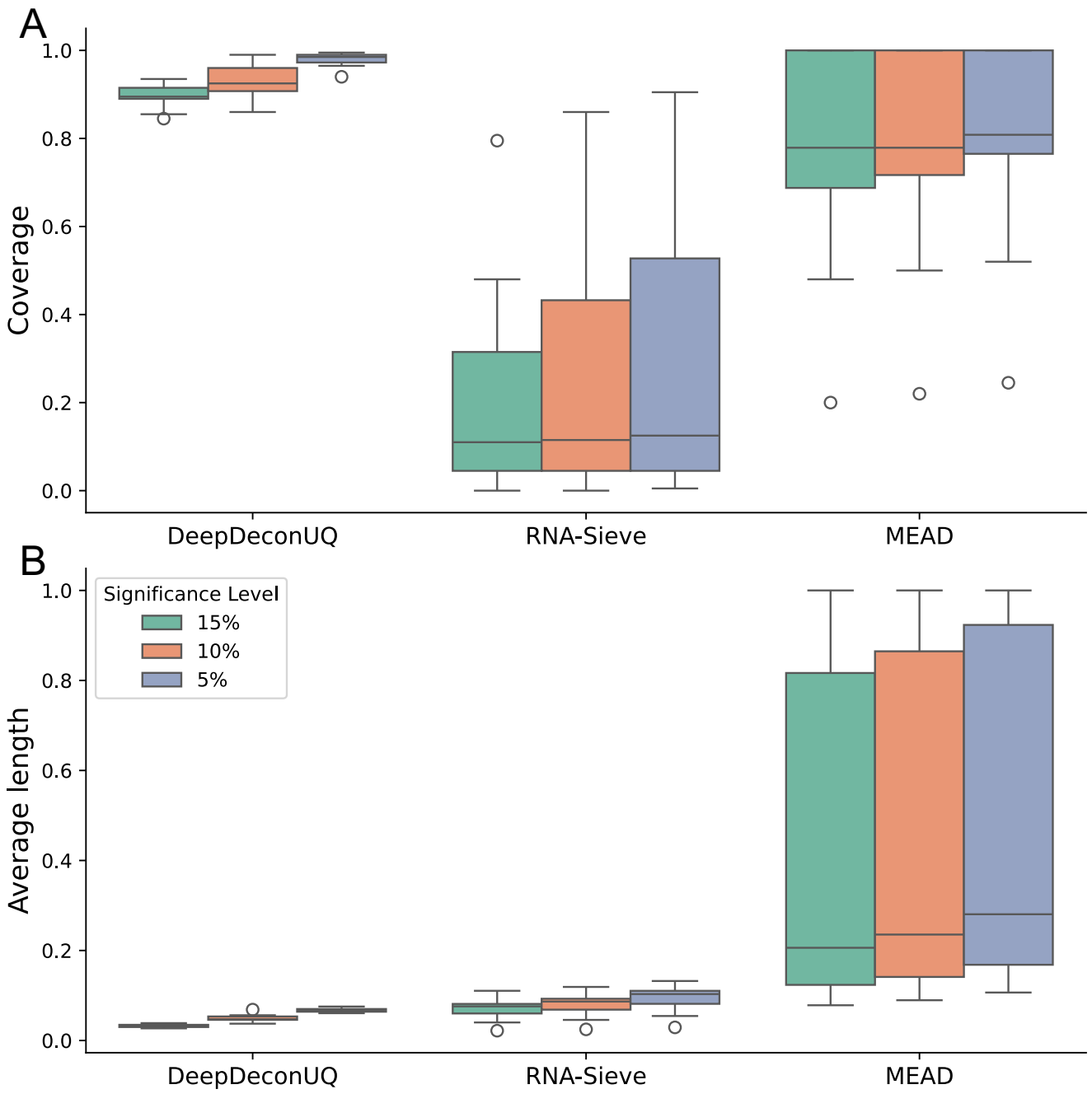

Fig 2: **DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on AML simulated bulk RNA-seq datasets.** Boxplots of coverage (**A**) and average prediction interval length (**B**) on 15 AML simulated bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Each bar in the boxplot comprises 15 data points, each corresponding to one of 15 simulated AML datasets. Significance levels are indicated with different colors.

several real samples with true malignant cell fractions in the range of 0.5-0.8, whereas DeepDeconUQ successfully captured all samples within this range. Although both DeepDeconUQ and MEAD exhibited coverage failures for samples with true malignant cell fractions below 0.4,

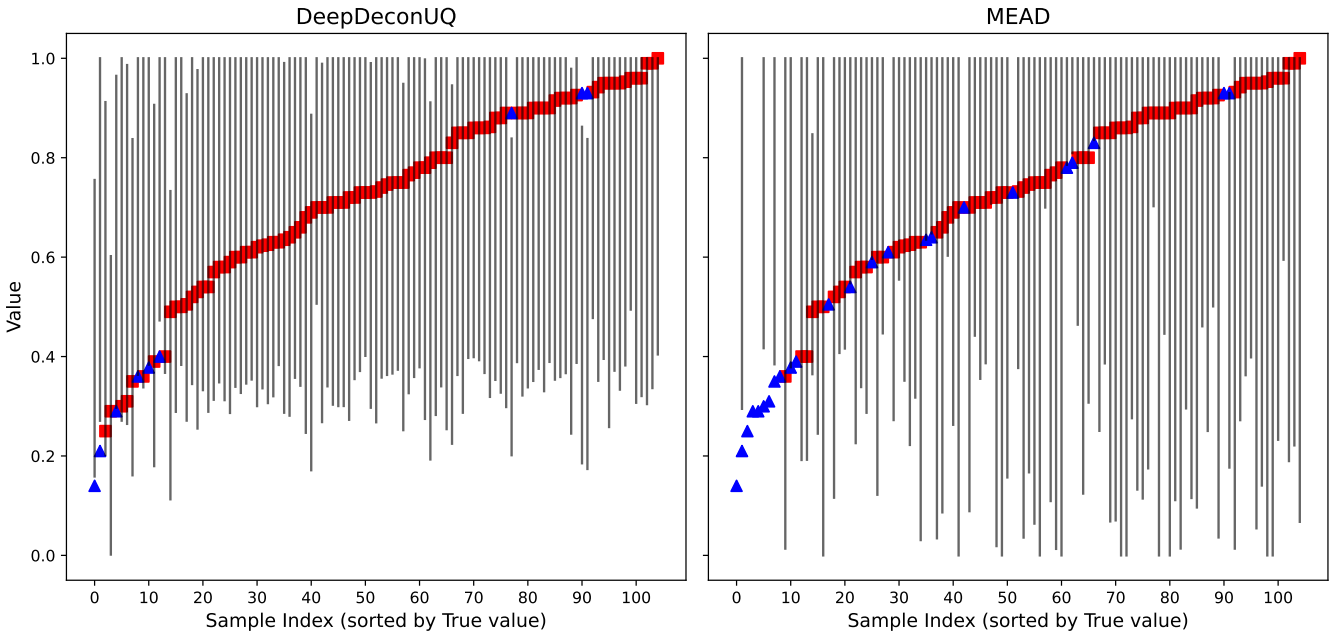

Fig 3: **Visualization of prediction intervals on the real primary dataset of DeepDeconUQ and MEAD at α = 0.05 (95% confidence level).** Comparison of uncertainty intervals generated by DeepDeconUQ (left) and MEAD (right) methods. Each vertical line represents the prediction interval (lower to upper bound) for an individual sample, with samples sorted by their true malignant fraction values in ascending order along the x-axis. The true values are marked with either red squares (when contained within the prediction interval) or blue triangles (when falling outside the prediction interval).

DeepDeconUQ demonstrated superior performance with significantly fewer coverage failures in this lower range. Results for other significance levels can be accessed in Fig S8, and Fig S9. It should be noted that the malignant cell fractions given by flow cytometry most likely deviate from a true fraction of malignant cells, resulting in under coverage compared to the prespecified coverage levels, which is expected. Despite these caveats, the results show that the coverages of the prediction intervals from DeepDeconUQ are generally higher than those from MEAD, while the lengths of the prediction intervals from DeepDeconUQ are shorter than those based on MEAD.

Additionally, We further compared coverages of the prediction intervals based on DeepDeconUQ and MEAD using McNemar's statistical test[16]. We also compared the lengths of the prediction intervals based on DeepDeconUQ and MEAD using the Wilcoxon signed-rank test. For the coverage analysis, each sample in the dataset was assigned a label of 1 if its true malignant cell fraction fell within the predicted interval; otherwise, it was labeled as 0. This approach enabled the generation of binary outcome pairs for each sample between DeepDeconUQ and MEAD, thereby providing paired nominal data suitable for McNemar's statistical test. Furthermore, we aggregated all samples across the three AML datasets into a consolidated dataset to perform a statistical assessment of this unified sample set. The resulting p-values from McNemar's test are 1.4035×10^{-6}, 1.2438×10^{-13}, and 1.3977×10^{-8} at significance levels 15%, 10%, 5%, respectively. Moreover, the p-value of the Wilcoxon signed-rank test on the prediction lengths are 3.33×10^{-6}, 9.75×10^{-9}, and 0.0013 at the same significance levels. These findings underscore a statistically significant performance distinction between DeepDeconUQ and MEAD.

We also tested DeepDeconUQ's performance on two other cancer types, neuroblastoma and head and neck squamous cell carcinoma (HNSCC), as evaluated in DeepDecon. DeepDeconUQ consistently achieved the highest coverage and the narrowest prediction intervals across all three

Table 1: **DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on real cancer bulk RNA-seq datasets.** Coverage and average prediction interval length ($\mathrm{L_{avg}}$) are shown under different significance levels on three real AML bulk RNA-seq datasets ('primary,' 'recurrent,' and 'BeatAML'), one real Neuroblastoma dataset and one real HNSCC dataset.

Methods	**Dataset**	**15%**		**10%**		**5%**	
		Coverage	$\mathrm{L_{avg}}$	**Coverage**	$\mathrm{L_{avg}}$	**Coverage**	$\mathrm{L_{avg}}$
RNA-Sieve	primary	1.0	1.0	1.0	1.0	1.0	1.0
	recurrent	1.0	1.0	1.0	1.0	1.0	1.0
	beat	1.0	1.0	1.0	1.0	1.0	1.0
	Neuroblastoma	1.0	1.0	1.0	1.0	1.0	1.0
	HNSCC	1.0	1.0	1.0	1.0	1.0	1.0
MEAD	primary	0.667	0.553	0.705	0.630	0.771	0.738
	recurrent	0.676	0.520	0.706	0.591	0.735	0.694
	beat	0.496	0.386	0.544	0.433	0.663	0.515
	Neuroblastoma	0.333	0.211	0.397	0.224	0.428	0.231
	HNSCC	0.139	0.035	0.270	0.059	0.283	0.062
DeepDeconUQ	primary	0.800	0.434	0.876	0.572	0.912	0.662
	recurrent	0.824	0.606	0.853	0.604	0.882	0.685
	beat	0.592	0.409	0.730	0.554	0.781	0.611
	Neuroblastoma	0.349	0.280	0.460	0.293	0.556	0.309
	HNSCC	0.361	0.081	0.433	0.103	0.635	0.151

datasets at different significance levels. The results are presented in Table 1. Moreover, DeepDeconUQ is also robust in complex tumor microenvironments (TME) when tested with epithelial datasets. Details can be accessed in Supplementary Information, Section 4.

DeepDeconUQ is robust to gene expression perturbations

In the Methods section, we discussed how perturbations in bulk RNA-seq gene expression data can affect the accuracy of the estimation algorithms. Fig 4, Fig S4, and Fig S5 illustrate the impact of various perturbation levels on the performance of these methods under different significance levels. For RNA-Sieve, the performance remains comparable to prior results without noise interference, with the prediction interval coverage consistently low. For MEAD, increasing noise levels results in decreased coverage and increased variability in the intervals. In the case of DeepDeconUQ, while coverage decreases as noise levels rise, the majority of coverage values still meet the required threshold. Notably, the average length of DeepDeconUQ's prediction intervals remains stable across different noise levels. DeepDeconUQ achieves the highest coverage and smallest average interval length across all methods under various noise conditions, demonstrating its robustness to expression perturbations.

Ablation study

To understand the contribution of key architectural components to model performance, we conducted a systematic ablation study. We focused on two critical elements: conformal calibration

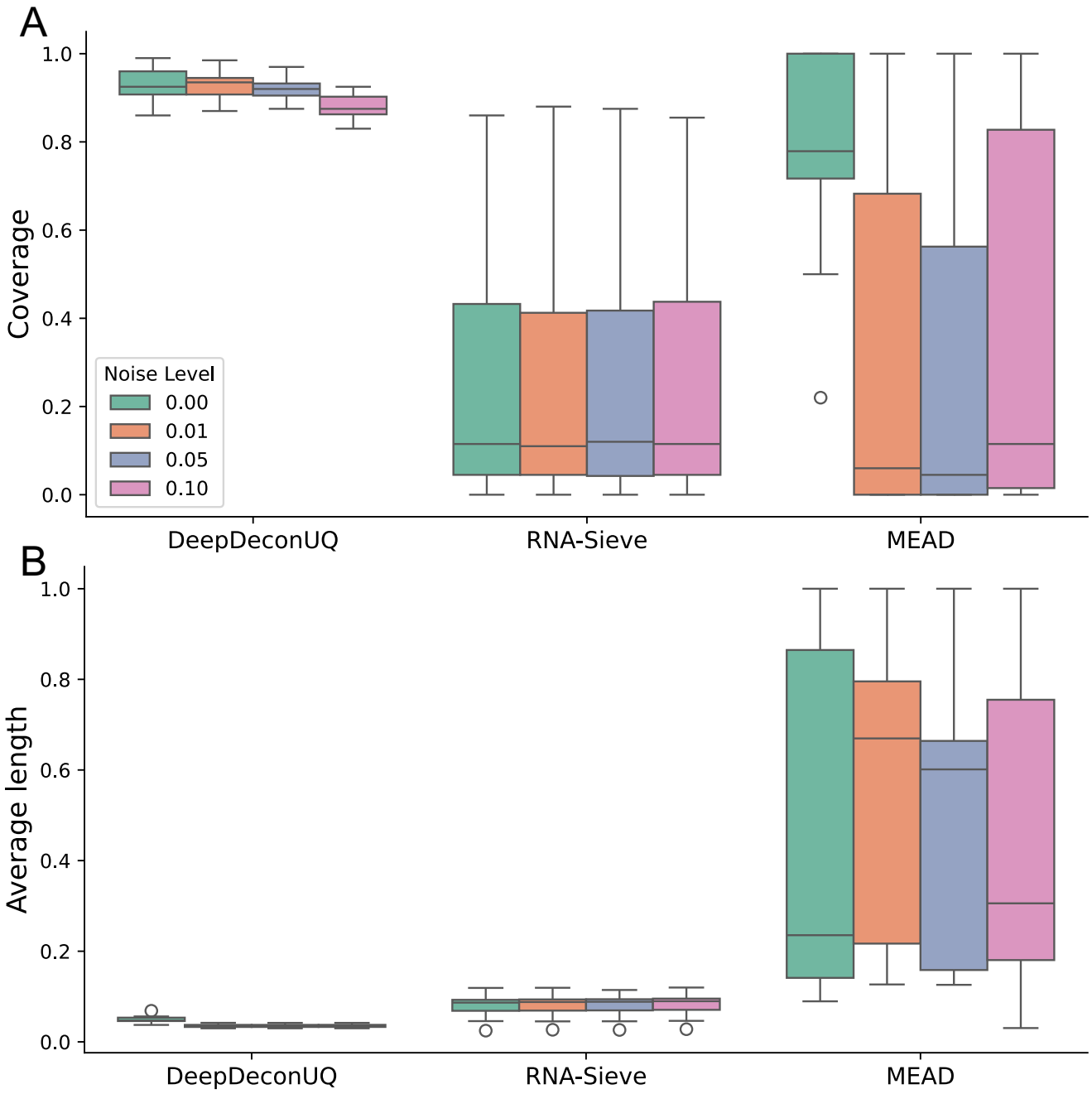

Fig 4: **DeepDeconUQ is robust to gene expression perturbations.** Boxplots of coverage and average prediction interval length on 15 AML simulated bulk RNA-seq datasets under different noise levels. We added random noise generated from a Gaussian distribution with zero mean and variance that equals $\lambda(\lambda = 0.01, 0.05, 0.1)$ times the gene expression level for each gene in each sample. Each bar contains a total of 15 points, representing 15 separate AML datasets. The color represents different levels of noise level λ. Significance level $\alpha = 0.1$.

and TF-IDF transformation. Quantile regression was preserved throughout this analysis as it provides the fundamental mechanism for generating lower and upper prediction interval bounds.

In the conformal calibration ablation experiment, we eliminated the calibration phase and allocated the entire training dataset to neural network training. For the TF-IDF transformation ablation, we removed this feature engineering step while retaining MinMax normalization, which

is essential for stabilizing gradient-based optimization in deep learning frameworks.

The result is shown in Fig 5. When conformal calibration was removed, DeepDeconUQ demonstrated systematic over-coverage with expanded interval widths compared to the original implementation. This finding confirms that conformal calibration plays a crucial role in optimizing prediction intervals by balancing coverage precision and interval width. The elimination of TF-IDF transformation had more pronounced consequences, resulting in a markedly degraded performance characterized by insufficient coverage (substantially below prescribed confidence levels) and wider prediction intervals. The severity of this performance deterioration highlights the fundamental importance of TF-IDF transformation in enabling effective neural network learning.

Collectively, these ablation experiments validate the necessity of both components in the DeepDeconUQ architecture, with each contributing significantly to the model's overall predictive capabilities and uncertainty quantification accuracy.

Time and memory usage

DeepDeconUQ was trained and tested on a High-Performance-Cluster (HPC) with a xeon-2640 6-core CPU node. It is the only algorithm that requires the generation of *in silico* training data, which takes 20 min for 3000 samples with a peak memory usage of 10 GB. Additionally, it took $\sim$20 minutes to train a model and took $\sim$3s to predict on one bulk tissue.

Discussion

DeepDeconUQ is an advanced deep neural network-based algorithm designed to leverage single-cell RNA sequencing (scRNA-seq) data to generate prediction intervals for malignant cancer cell fractions. Building on our earlier method, DeepDecon, DeepDeconUQ retains all its foundational advantages, such as the ability to automatically extract complex nonlinear features within its hidden layers and to accurately estimate the quantile function by integrating a comprehensive input of genes ($\sim 10^4$). To address intrinsic variability in RNA-seq data, DeepDeconUQ employs TF-IDF transformation and Min-Max normalization, which enables it to yield prediction intervals that account for both biological and technical sources of noise. Additionally, it utilizes a calibration dataset to fine-tune the prediction interval, effectively mitigating risks of overcoverage and undercoverage. Integrating training and calibration datasets in DeepDeconUQ represents a significant advancement in malignant cancer cell fraction estimation, allowing for more accurate and interpretable predictions. By leveraging quantile regression and conformal inference, DeepDeconUQ not only enhances confidence in the malignant cell prediction interval results but also facilitates the application of the method to real-world datasets with minimal adjustments. The framework's ability to generate reliable uncertainty estimates positions DeepDeconUQ as a valuable tool for the analysis of bulk RNA-seq data, particularly in contexts where precise quantification of cell type proportions is critical for downstream analyses and clinical decision-making.

While DeepDeconUQ can achieve good performance on AML cancer tissues, we note that this method still has limitations. First of all, the quality of training data is very important. DeepDeconUQ is a neural network-based method, which means it needs a large amount of data to train. Currently, we use single-cell data from 15 AML subjects to construct simulation bulk RNA-seq datasets. If the number of subjects is small or the single-cell data is dominated by one specific cell type, DeepDeconUQ can learn less information from the data and cannot generalize and represent the latent features well. In theory, the UQ approach may also work for previous decomposition methods with or without single-cell data, provided we have sufficient bulk RNA-seq

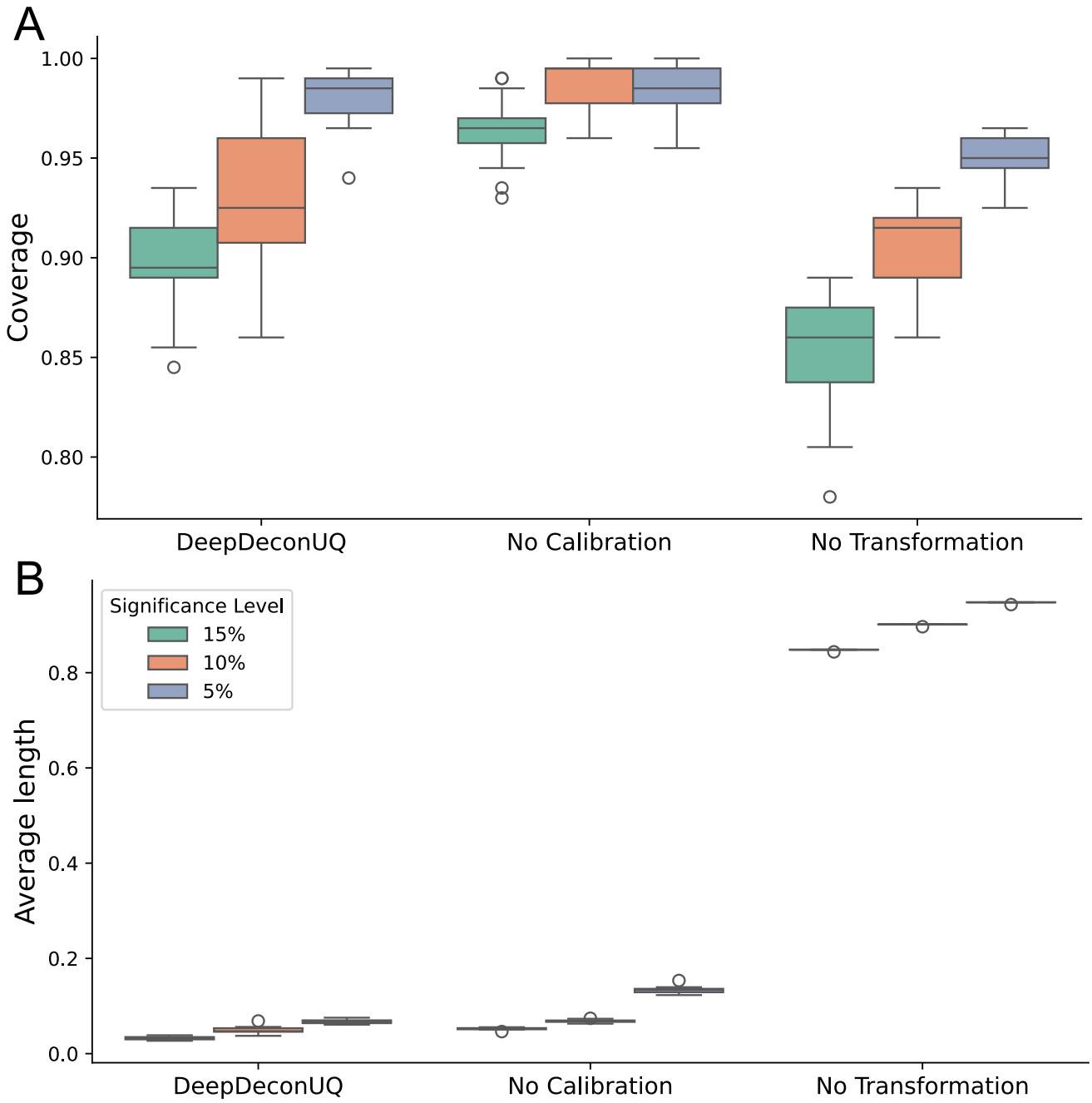

Fig 5: **Ablation study of DeepDeconUQ.** DeepDeconUQ is the original model. No Calibration removes the calibration part of the DeepDeconUQ model and uses all the training data to train the neural network. No Transformation removes the TF-IDF transformation and uses MinMax normalization for data preprocessing. Each point in the boxplot is an artificial bulk RNA-seq dataset.

data with corresponding malignant cell fractions. A critical prerequisite is that these annotated fractions should span the complete range from 0.0 to 1.0. However, from a practical perspective, such comprehensively annotated bulk RNA-seq datasets remain scarce. Secondly, experimental bias and noise can greatly affect the estimate performance, even though we take different ways such as TF-IDF transformation and Min-Max normalization to mitigate batch effects and bias. The complexity and difficulties of real RNA-seq can still affect DeepDeconUQ's performance.

Thirdly, DeepDeconUQ can only estimate the prediction interval of malignant cell fraction. In practice, tissues usually consist of multiple cell types, and some tissues even contain unknown sub-cell types.

We plan to further improve the performance and applicability of DeepDeconUQ by implementing several key modifications to the existing methodology. Firstly, we want to extend DeepDeconUQ's capacity to include multiple cell types or subtypes. The current method avoids the statistical complexity of handling multivariate prediction regions, which are required when deconvolving bulk RNA-seq data into more than two cell types. Prediction regions, unlike univariate intervals, must account for dependencies between cell type proportions (e.g., sum-to-one constraints and correlations), necessitating advanced methods like multivariate conformal prediction. Secondly, DeepDeconUQ's capability to detect technical bias and diverse sequencing protocols should be improved. In addition to current normalization processing, methods like autoencoder[17,18], transfer learning[19] and transformers[20] can be used to generate latent embeddings to reduce these biases (see Section 3 in the supplementary information for preliminary results). Thirdly, the current DeepDeconUQ model takes all genes into account. Whether selective incorporation of cell type-specific genes could enhance prediction accuracy is an interesting topic. To investigate this issue, we selected differentially expressed genes between normal and malignant cells using MAST[21], a widely used method for single cell differential gene analysis. DeepDeconUQ was trained and validated based on the selected genes and the detailed results are given in Section 5 in the supplementary information. The preliminary study shows that gene selection does not markedly impact the performance of DeeDeconUQ. More complete and extensive studies on the impacts of gene selection using other software packages on the performance of DeepDeconUQ will be studied in the future.

Methods

Datasets

To initially train and test DeepDeconUQ, we utilized simulated datasets derived from Acute Myeloid Leukemia (AML) single-cell data previously used in DeepDecon[22]. The single-cell AML datasets were downloaded from Gene Expression Omnibus (GEO) with accession number GSE116256[23]. We selected 15 subjects, totaling 38,410 cells, to simulate artificial bulk RNA-seq datasets, employing the same preprocessing and simulation procedures established in DeepDecon. Preprocessing of scRNA-seq data followed the workflow of Scanpy (v.1.7.2), a widely-adopted Python package for single-cell gene expression analysis[24]. Initially, cells with fewer than 500 detected genes and genes expressed in fewer than five cells were filtered out (Fig S1). Further, gene expression count matrices were processed to remove extreme outliers (Table S1). Gene expression values were normalized using Scanpy's 'normalize_total' function to ensure uniform total counts across cells. This will mitigate discrepancies arising from varying library sizes. This produced a normalized matrix of all filtered cells and genes, ready for the generation of simulated bulk data. Ultimately, 30,000 simulated bulk samples (2,000 per subject) were generated for training and testing DeepDeconUQ.

We further assessed DeepDeconUQ using real AML bulk RNA-seq datasets. Real AML data were collected from the GDC Data Portal(`https://portal.gdc.cancer.gov/`) with the project name "TARGET-AML". The AML samples were further divided into primary and recurrent AML categories according to different cancer stages. As a result, there were a total of 117 primary AML samples and 38 recurrent AML samples. For these bulk RNA-seq datasets, ground-truth cancer cell fractions via flow cytometry are available. Additionally, an independent real AML dataset, "BeatAML"[25], was collected from cBioportal[26]. "BeatAML" contains a total of 451 bulk

RNA-seq samples and 300 of them have corresponding ground-truth cancer cell fractions. This dataset used the "SureSelect" sequencing platform, which is different from the sequencing platform for the single-cell data in "TARGET-AML" dataset (Table S2). The inclusion of these diverse datasets allowed us to evaluate DeepDeconUQ's performance across different sequencing platforms and data sources.

To test DeepDeconUQ's performance on other cancer tissues, we also collected 19,173 single cells from 9 neuroblastoma cancer patients[27] and 184,868 single cells from 27 Head and neck squamous cell carcinoma (HNSCC) cancer patients[28]. They were used to simulate artificial RNA-seq bulk samples to build and evaluate DeepDeconUQ. Additionally, a real neuroblastoma bulk RNA-seq dataset consisting of 99 bulk RNA-seq samples with known cancer cell fractions was collected from cBioportal[26] and another real HNSCC bulk RNA-seq dataset, 'TCGA-HNSC', consisting of 518 bulk RNA-seq samples with known cancer cell fractions was collected from LinkedOmics[29]. These two real datasets were used for testing. Moreover, the above datasets have the knowledge of malignant and normal cells. However, in practice, cancer tissues usually exhibit a complex tumor microenvironment (TME). A total of 18,062 single cells derived from four individuals were collected[30], It contains epithelial cells (tumor), T-cells, B-cells, plasma cells, macrophage, fibroblast cells, and so on. Experiments were conducted to test the capacity of DeepDeconUQ to estimate epithelial cell proportion regarding heterogeneity.

Generating artificial bulk RNA-seq datasets

To generate artificial bulk RNA-seq samples, we used the previously described scRNA-seq datasets, simulating each sample with predetermined malignant cell fractions for training the DeepDeconUQ model. Specifically, for each artificial bulk sample, we set a fixed total cell count, N, and a malignant cell number n_m was randomly sampled from a uniform distribution between 0 and N. Subsequently, n_m malignant cells and $N - n_m$ normal cells were randomly drawn from the same scRNA-seq dataset. If the available malignant or normal cells were fewer than n_m or $N - n_m$, respectively, cells were sampled with replacement, meaning that each cell was uniformly drawn from all single cells in the dataset; otherwise, cells were sampled without replacement to ensure no duplicates. Importantly, cells from different subjects (i.e., individuals) were not combined within a single artificial sample to maintain individual-specific gene expression profiles. This principle was motivated by two reasons. Firstly, the aim was to safeguard within-subject relationships among genes by preserving the unique gene expression patterns inherent to each subject. Secondly, the intention was to capture the variability between subjects, commonly referred to as cross-subject heterogeneity[8]. After generating an artificial bulk sample by summing the expression values of all selected cells, it was labeled according to the malignant cell fraction, n_m/N. This process was repeated for each scRNA-seq dataset, resulting in a corresponding artificial bulk RNA-seq dataset with T samples, each tagged with a known malignant cell proportion. Here, we set $N = 3,000$ and $T = 200$, consistent with the configuration in DeepDecon[22]. This sampling strategy serves as a substantial data generation resource for training and evaluating DeepDeconUQ.

Data Processing

Before training, the artificial bulk RNA-seq samples were preprocessed to ensure alignment between training and prediction data. Only genes present in both the training and testing datasets were retained, and genes with low expression variance (below 0.1) were excluded. To further standardize the data, a TF-IDF transformation was applied to the raw RNA-seq count matrix. This transformation, commonly used in information retrieval and text mining[31,32], starts by calcu-

lating the 'term frequency (TF)' for each gene in each sample by normalizing the gene expression profile (see Formula 3). The 'inverse document frequency (IDF)' was then calculated by dividing the total number of bulk samples by the total gene expression values of the gene across all samples (see Formula 4), followed by log-transformation and multiplication by the TF value. The TF-IDF transformation weights genes with lower expression levels more heavily, which helps to adjust for the imbalanced expression levels across genes[33].

$$\mathrm{TF}(X_{i,j}) = \frac{X_{i,j}}{\sum_j X_{i,j}}, \tag{3}$$

$$\mathrm{IDF}(G_j) = \log\left(\frac{T}{\sum_i X_{i,j}} + 1\right), \tag{4}$$

where $X_{i,j}$ is the expression level of the jth gene in the ith sample, G_j indicates the jth gene, and T is the number of bulk samples.

Let X' denote the gene expression matrix after TF-IDF transformation. A MinMax normalization was applied to the resulting expression matrix X' to scale the expression values to the [0, 1] range (see Formula 5). This is a common practice in deep learning models that use gradient-based optimization algorithms[8,17].

$$X_i^{\mathrm{norm}} = \frac{X_i' - \min(X_i')}{\max(X_i') - \min(X_i')}, \tag{5}$$

where X_i' is the ith row of X' and X_i^{norm} is the ith row of the resulting expression matrix after the MinMax transformation.

TF-IDF transformation and MinMax normalization are important steps in ensuring the quality and consistency of the data used to train deep learning models. Although the input datasets varied between platforms and protocols, we utilized the same processing workflow to make it easy to apply DeepDeconUQ to other datasets.

DeepDeconUQ

Problem formulation

Suppose we are given n bulk RNA-seq gene expression samples $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^p$ represents the ith bulk RNA-seq gene expression vector with $p > 0$ features (genes) and $Y_i = (y_i, 1 - y_i)$ is the corresponding ith cell fraction vector of malignant and normal cells. Our aim is to construct a distribution-agnostic prediction interval $\hat{C}(X_{n+1})$ that contains the malignant cell fraction y_{n+1} for a new bulk RNA-seq sample X_{n+1}. Specifically, given a desired significance level α, the prediction interval $\hat{C}(X_{n+1})$ is likely to contain the true malignant cell fraction vector y_{n+1} with a user-specified coverage probability $1 - \alpha$:

$$\mathbb{P}\{y_{n+1} \in \hat{C}(X_{n+1}))\} \geq 1 - \alpha, \tag{6}$$

for any joint distribution P_{XY} and any sample size n. Meanwhile, the estimated prediction interval $\hat{C}(X_{n+1})$ should be as narrow as possible while achieving the desired coverage level.

Quantile regression

Methods like DeepDecon[22] formulate the problem as a regression task, typically addressed using variations of non-negative least squares or more advanced machine learning methodologies.

The estimation of cell type proportions is often solved by minimizing squared residuals over the n training points $\{(X_i, Y_i)\}_{i=1}^{n}$ (see Formula 7):

$$\hat{\mu}(x) = \mu(x; \hat{\theta}), \qquad \hat{\theta} = \text{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(X_i; \theta))^2 + R(\theta), \tag{7}$$

where θ are the parameters of the regression model, $\mu(x; \theta)$ is the learned regression model, and $R(\theta)$ is a regularization module.

Similarly, quantile regression estimates the conditional quantiles of cell type proportions, assuming that the τth conditional quantile is associated with gene expression profiles. A conditional quantile function q_α is learned from n training samples $\{(X_i, Y_i)\}_{i=1}^{n}$ at a specified quantile (or significance) level α (see Formula 8).

$$\hat{\text{q}}_\alpha(x) = f(x; \hat{\theta}), \qquad \hat{\theta} = \text{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i, f(X_i; \theta)) + R(\theta), \tag{8}$$

where $f(x; \theta)$ is the quantile regression function and can be learned through neural networks. ρ_α is the quantile (pinball) loss[34], defined as,

$$\rho_\alpha(y, \hat{y}) = \begin{cases} \alpha(y - \hat{y}) & \text{if} \quad y - \hat{y} > 0, \\ (1 - \alpha)(\hat{y} - y) & \text{otherwise}, \end{cases} \tag{9}$$

where y and $\hat{y}$ are the observed and predicted cell type fraction, and $\alpha \in (0, 1)$ is the corresponding quantile (significance) level. Pinball loss is a skewed transformation of the absolute value function and is commonly used in quantile regression[14].

Given a significance level α, we can get the lower bound and upper bound prediction $\hat{\text{q}}_{\alpha_{lo}}, \hat{\text{q}}_{\alpha_{hi}}$ through quantile regression. Here, $\alpha_{lo} = \frac{\alpha}{2}, \alpha_{hi} = 1 - \frac{\alpha}{2}$. Then, $\hat{C}(X_{n+1}) = [\hat{\text{q}}_{\alpha_{lo}}, \hat{\text{q}}_{\alpha_{hi}}]$ can be used as the estimate of the true prediction interval $C(X_{n+1})$. The simplicity and generality of this approach make quantile regression highly versatile, allowing for the integration of various machine learning techniques to model and learn q_α[14,35,36].

Conformal prediction

The quantile regression method is widely applicable and often works well in practice, yielding intervals that are adaptive to heteroscedasticity. However, it is not guaranteed to satisfy the validity property when the true prediction interval $C(X_{n+1})$ is estimated by the prediction interval $\hat{C}(X_{n+1})$. Fortunately, conformal prediction[37] was then brought out to solve this problem. Specifically, split (inductive) conformal prediction[38,39], which is general and whose computational cost is a small fraction of the full conformal prediction, helps construct prediction intervals that are valid and discriminative. We borrowed the idea from Romano et al.[14] and combined DeepDecon with conformal quantile regression (CQR) to obtain valid and discriminative cell fraction prediction intervals on bulk RNA-seq samples. We refer the resulting algorithm as DeepDeconUQ.

The split conformal method begins by splitting the training data into two disjoint subsets: a proper training set $\{(X_i, Y_i) : i \in I_1\}$ and a calibration set $\{(X_i, Y_i) : i \in I_2\}$. We then apply a neural network to estimate the lower and upper quantile functions, $\hat{\text{q}}_{\alpha_{lo}}$ and $\hat{\text{q}}_{\alpha_{hi}}$, as described in Equation 8. This model's architecture is similar to our previously developed cell fraction estimation framework, DeepDecon[22], and will be further explained in the model structure subsection.

Next, we compute conformity scores that quantify the error made by the prediction interval. The scores are evaluated on the calibration set as follows:

$$E_i := \max(\hat{\text{q}}_{\alpha_{lo}}(X_i) - Y_i, Y_i - \hat{\text{q}}_{\alpha_{hi}}(X_i)) \quad i \in I_2, \tag{10}$$

Finally, given new input data X_{n+1}, we construct the prediction interval of Y_{n+1} as:

$$\hat{C}(X_{n+1}) = [\hat{\mathrm{q}}_{\alpha_{lo}}(X_{n+1}) - Q_{1-\alpha}(E, I_2), \hat{\mathrm{q}}_{\alpha_{hi}}(X_{n+1}) + Q_{1-\alpha}(E, I_2)], \tag{11}$$

where $Q_{1-\alpha}(E, I_2)$ is the $(1-\alpha/2)(1+\frac{1}{|I_2|})$th quantile of $\{E_i : i \in I_2\}$. In this context, we select $\alpha/2$ due to the presence of two distinct cell types within the dataset—malignant and normal—as suggested in multivariate quantile regression[40]. Moreover, Romano et al. demonstrated that when conformity scores E_i are almost surely unique, the prediction interval achieves an approximate state of perfect calibration[14].

The specific steps of DeepDeconUQ are given in Algorithm 1.

Algorithm 1 DeepDeconUQ

Require: Bulk RNA-seq samples with labels $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}^2, 1 \leq i \leq n$
Significance level α
Testing bulk sample X_{n+1}

Ensure: Cell fraction prediction interval $C(X_{n+1})$ for X_{n+1}.

1: Randomly split n bulk RNA-seq samples into two disjoint sets, I_1 and I_2.
2: Fit two conditional quantile functions $\{\hat{\mathrm{q}}_{\alpha_{lo}}, \hat{\mathrm{q}}_{\alpha_{hi}}\}$ according to Equation 8 on training set I_1
3: Compute conformity scores E_i according to formula 10 on calibration set I_2
4: Compute $Q_{1-\alpha}(E, I_2)$, the $(1-\alpha/2)(1+\frac{1}{|I_2|})$th quantile of $\{E_i : i \in I_2\}$.
5: Compute prediction interval $\hat{C}(X_{n+1})$ according to formula 11 for X_{n+1}.

Lei et al. advocated for selecting a larger I_1 compared to I_2 to improve the accuracy of estimated quantile functions[41]. Given the size of our training dataset (30,000 simulated samples), we opted for a 7:3 split ratio between the training and calibration sets to optimize the model performance.

Model structure

The main neural network architecture of DeepDeconUQ is similar to DeepDecon, which consists of two main components. The first component consists of four fully connected layers with a dropout regularization between each layer, and the rectified linear unit (ReLU) is used as the activation function in every internal layer. The second component differs from DeepDecon, which uses a softmax function to predict the malignant and normal cell fractions. To reduce the computational cost, instead of fitting two separate neural networks to estimate the lower and upper quantile functions, we replaced the original one-dimensional estimate of the malignant cell fraction with a two-dimensional estimate of the lower and upper quantiles. In this way, most of the network parameters are shared between the two quantile estimators. All model parameters were optimized using the Adam optimization algorithm[42] with a learning rate of 0.0001 and a batch size of 128. The model was trained as a regression task, with the pinball loss (see Formula 9) as the loss function. Hyperparameters that are tested and tuned in DeepDecon were also used in DeepDeconUQ.

The impact of gene expression perturbations on DeepDeconUQ

To test the model's robustness to gene expression perturbations, we introduced varying levels of Gaussian noise to the expression levels within the simulated datasets. Specifically, for each gene in each sample, random noise was added, drawn from a Gaussian distribution with a mean of zero. The variance of this noise was proportional to the expression level of each gene,

set at λ times the gene expression level, where λ was assigned values of 0.01, 0.05, and 0.1 (see Formula 12). This approach allowed us to systematically examine the model stability and predictive accuracy under controlled levels of expression variability.

$$X_{ij}^{\text{noise}} = \max(0, X_{ij} + \mathrm{N}(0, \lambda X_{ij})), \tag{12}$$

where X_{ij} is the gene expression value of gene j in simulated bulk sample i and λ is the noise level.

Following this processing, we applied the previously trained DeepDeconUQ models to each simulated bulk RNA-seq dataset to estimate the prediction intervals. This enabled us to systematically evaluate the model's robustness under various gene expression perturbations, providing insights into its stability and reliability in producing accurate intervals when gene expression data is subject to different levels of noise.

Acknowledgments

Author contributions

J.H, J.F.Z, and F.S conceived the study. J.H. designed the DeepDeconUQ method conceptually, designed the neural network architecture and developed code for implementing, training and evaluating models. Y.D., Y.F, and J.L helped with the finalization of the manuscript and applications of the model. J.F.Z and K.R.K. helped with the applications to real data sets, explanations of the implications of the computational results, and the finalization of the manuscript. F.S. supervised the project and helped design the DeepDeconUQ method, data analysis, and the finalization of the manuscript. All authors contributed to the writing of the manuscript.

Declaration of interests

The authors declare no competing interests. The source code and data used to produce the results and analyses presented in this manuscript are available from a GitHub repository at: `https://github.com/jiaweih14/DeepDeconUQ`

References

1. Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using rna-seq. Nature Methods *8*, 469–477. doi:10.1038/nmeth.1613.

2. Finotello, F., and Di Camillo, B. (2015). Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. Briefings in Functional Genomics *14*, 130–142. doi:10.1093/bfgp/elu035.

3. Qin, Y., Zhang, W., Sun, X., Nan, S., Wei, N., Wu, H.-J., and Zheng, X. (2020). Deconvolution of heterogeneous tumor samples using partial reference signals. PLOS Computational Biology *16(11)*, e1008452. doi:10.1371/journal.pcbi.1008452.

4. Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P. S., Povinelli, B. J., Booth, C. A. G., Sopp, P., Norfo, R., Rodriguez-Meira, A., Ashley, N., Jamieson, L., Vyas, P., Anderson, K., Segerstolpe, Å., Qian, H., Olsson-Strömberg, U., Mustjoki, S., Sandberg, R., Jacobsen,

S. E. W., and Mead, A. J. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nature Medicine *23*, 692–702. doi:10.1038/nm.4336.

5. Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., and Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology *37*, 773–782. doi:10.1038/s41587-019-0114-2.

6. Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications *10*, 380. doi:10.1038/s41467-018-08023-x.

7. Xie, D., and Wang, J. (2022). Robust statistical inference for cell type deconvolution. arXiv preprint arXiv:2202.06420. doi:10.48550/arXiv.2202.06420.

8. Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D. S., Kloiber, K., Heutink, P., and Bonn, S. (2020). Deep learning–based cell composition analysis from tissue expression profiles. Science Advances *6*, eaba2619. doi:10.1126/sciadv.aba2619.

9. Erdmann-Pham, D. D., Fischer, J., Hong, J., and Song, Y. S. (2021). A likelihood-based deconvolution of bulk gene expression data using single-cell references. Genome Research (gr.272344.120). doi:10.1101/gr.272344.120.

10. Mohammadi, S., Zuckerman, N., Goldsmith, A., and Grama, A. (2017). A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proceedings of the IEEE *105*, 340–366. doi:10.1109/JPROC.2016.2607121.

11. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun *11*, 5650. doi:10.1038/s41467-020-19015-1.

12. Lin, Z., Trivedi, S., and Sun, J. (2021). Locally valid and discriminative prediction intervals for deep learning models. Advances in Neural Information Processing Systems *34*, 8378–8391. doi:10.48550/arXiv.2106.00225.

13. Cai, B., Zhang, J., Li, H., Su, C., and Zhao, H. (2022). Statistical inference of cell-type proportions estimated from bulk expression data. arXiv preprint arXiv:2209.04038. doi:10.48550/arXiv.2209.04038.

14. Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. Advances in neural information processing systems *32*. doi:10.48550/arXiv.1905.03222.

15. Su, C., Xu, Z., Shan, X., Cai, B., Zhao, H., and Zhang, J. (2023). Cell-type-specific co-expression inference from single cell rna-sequencing data. Nature Communications *14*, 4846. doi:10.1038/s41467-023-40503-7.

16. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika *12*, 153–157. doi:10.1007/BF02295996.

17. Chen, Y., Wang, Y., Chen, Y., Cheng, Y., Wei, Y., Li, Y., Wang, J., Wei, Y., Chan, T.-F., and Li, Y. (2022). Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nature Communications *13*, 6735. doi:10.1038/s41467-022-34550-9.

18. Sagendorf, J. M., Mitra, R., Huang, J., Chen, X. S., and Rohs, R. (2024). Structure-based prediction of protein-nucleic acid binding using graph neural networks. Biophysical Reviews *16*, 297–314. doi:10.1007/s12551-024-01201-w.

19. Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In: International conference on machine learning. PMLR (2208–2217). doi:10.48550/arXiv.1605.06636.

20. Castro, E., Godavarthi, A., Rubinfien, J., Givechian, K., Bhaskar, D., and Krishnaswamy, S. (2022). Transformer-based protein generation with regularized latent space optimization. Nature Machine Intelligence *4*, 840–851. doi:10.1038/s42256-022-00532-1.

21. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M. et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology *16*, 1–13. doi:10.1186/s13059-015-0844-5.

22. Jiawei Huang, A. S. K. R. K. J. F. Z., Yuxuan Du, and Sun, F. (2020). Deepdecon accurately estimates cancer cell fractions in bulk rna-seq data. Patterns *38*, 716–733. doi:10.1016/j.ccell.2020.08.014.

23. van Galen, P., Hovestadt, V., Wadsworth, M., Hughes, T., Griffin, G. K., Battaglia, S., Verga, J. A., Stephansky, J., Pastika, T. J., Lombardi Story, J., Pinkus, G. S., Pozdnyakova, O., Galinsky, I., Stone, R. M., Graubert, T. A., Shalek, A. K., Aster, J. C., Lane, A. A., and Bernstein, B. E. (2019). Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. Cell *176*, 1265–1281.e24. doi:10.1016/j.cell.2019.01.031.

24. Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. Genome Biology *19*, 15. doi:10.1186/s13059-017-1382-0.

25. Tyner, J. W., Tognon, C. E., Bottomly, D., Wilmot, B., Kurtz, S. E., Savage, S. L., Long, N., Schultz, A. R., Traer, E., Abel, M. et al. (2018). Functional genomic landscape of acute myeloid leukaemia. Nature *562*, 526–531. doi:10.1038/s41586-018-0623-z.

26. Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E. et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery *2*, 401–404. doi:10.1158/2159-8290.CD-12-0095.

27. Dong, R., Yang, R., Zhan, Y., Lai, H.-D., Ye, C.-J., Yao, X.-Y., Luo, W.-Q., Cheng, X.-M., Miao, J.-J., Wang, J.-F. et al. (2020). Single-cell characterization of malignant phenotypes and developmental trajectories of adrenal neuroblastoma. Cancer Cell *38*, 716–733. doi:10.1016/j.ccell.2020.08.014.

28. Sun, D., Wang, J., Han, Y., Dong, X., Ge, J., Zheng, R., Shi, X., Wang, B., Li, Z., Ren, P. et al. (2021). Tisch: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Research *49*, D1420–D1430. doi:10.1093/nar/gkaa1020.

29. Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). Linkedomics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Research *46*, D956–D963. doi:10.1093/nar/gkx1090.

30. Sathe, A., Mason, K., Grimes, S. M., Zhou, Z., Lau, B. T., Bai, X., Su, A., Tan, X., Lee, H., Suarez, C. J. et al. (2023). Colorectal cancer metastases in the liver establish immunosuppressive spatial networking between tumor-associated spp1+ macrophages and fibroblasts. Clinical Cancer Research *29*, 244–260.

31. Teller, V. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Computational Linguistics *26*, 638–641. doi:10.1162/089120100750105975.

32. Chowdhury, G. G. Introduction to modern information retrieval. Facet publishing (2010). ISBN 185604694X.

33. Moussa, M., and Măndoiu, I. I. (2018). Single cell RNA-seq data clustering using TF-IDF based methods. BMC Genomics *19*, 569. doi:10.1186/s12864-018-4922-4.

34. Steinwart, I., and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. Bernoulli *17*, 211 – 225. doi:10.3150/10-BEJ267.

35. Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. Journal of forecasting *19*, 299–311. doi:10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V.

36. Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. Journal of Machine Learning Research *7*, 1231–1264. URL: http://jmlr.org/papers/v7/takeuchi06a.html.

37. Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In: Proceedings of the Sixteenth International Conference on Machine Learning. ICML '99 San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558606122 (444–453).

38. Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In: Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13. Springer (345–356). doi:10.1007/3-540-36755-1_29.

39. Vovk, V., Gammerman, A., and Shafer, G. Algorithmic learning in a random world vol. 29. Springer (2005). doi:10.1007/b106715.

40. Feldman, S., Bates, S., and Romano, Y. (2023). Calibrated multiple-output quantile regression with representation learning. Journal of Machine Learning Research *24*, 1–48. doi:10.48550/arXiv.2110.00816.

41. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association *113*, 1094–1111. doi:10.48550/arXiv.1604.04173.

42. Kingma, D. P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv. doi:10.48550/arXiv.1412.6980. arXiv:1412.6980.

Supporting information

S1 Text. Preprocessing of single-cell gene expression data
(PDF)
S2 Text. Artificial bulk dataset simulation
(PDF)
S3 Text. The influence of feature embedding to DeepDeconUQ
(PDF)
S4 Text. Extention to complex tumor microenvironment (TME)
(PDF)
S5 Text. Influence of gene selection on DeepDeconUQ
(PDF)
S6 Text. Software comparison and settings
(PDF)
S1 Fig. Barplots of the numbers of malignant and normal cells in each scRNA-seq subject. Subjects with at least 100 malignant and 100 normal cells were selected for this study.
(TIF)
S2 Fig. DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on simulated HNSCC bulk RNA-seq datasets. Boxplots of coverage (A) and average prediction interval length (B) on simulated HNSCC bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Significance levels are indicated with different colors.
(TIF)
S3 Fig. DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on simulated Neuroblastoma bulk RNA-seq datasets. Boxplots of coverage (A) and average prediction interval length (B) on simulated Neuroblastoma bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Significance levels are indicated with different colors.
(TIF)
S4 Fig. DeepDeconUQ is robust to gene expression perturbations. Boxplots of coverage and average prediction interval length on 15 AML simulated bulk RNA-seq datasets under different noise levels. We added random noise generated from a Gaussian distribution with zero mean and variance that equals $\lambda(\lambda = 0.01, 0.05, 0.1)$ times the gene expression level for each gene in each sample. Each bar contains a total of 15 points, representing 15 separate AML datasets. The color represents different levels of noise level λ. Significance level $\alpha = 0.15$.
(TIF)
S5 Fig. DeepDeconUQ is robust to gene expression perturbations. Boxplots of coverage and average prediction interval length on 15 AML simulated bulk RNA-seq datasets under different noise levels. We added random noise generated from a Gaussian distribution with zero mean and variance that equals $\lambda(\lambda = 0.01, 0.05, 0.1)$ times the gene expression level for each gene in each sample. Each bar contains a total of 15 points, representing 15 separate AML datasets. The color represents different levels of noise level λ. Significance level $\alpha = 0.05$.
(TIF)
S6 Fig. Comparison of DeepDeconUQ with different embedding methods. Two embedding methods are used to compare with DeepDeconUQ. Principal Component Analysis (PCA) selects

the top 100 principal components as neural network inputs. DAN makes use of transfer learning and generates a latent embedding layer for both training and testing datasets. Each point in the boxplot is an artificial bulk RNA-seq dataset.

(TIF)

S7 Fig. DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on simulated epithelial bulk RNA-seq datasets. Boxplots of coverage (A) and average prediction interval length (B) on four simulated epithelial bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Each bar in the boxplot comprises 4 data points, each corresponding to one of 4 simulated epithelial datasets (except one dataset in MEAD that gives NA values and, therefore, doesn't show). Significance levels are indicated with different colors.

(TIF)

S8 Fig. Visualization of prediction intervals on the real primary dataset of DeepDeconUQ and MEAD at α = 0.10 (90% confidence level). Comparison of uncertainty intervals generated by DeepDeconUQ (left) and MEAD (right) methods. Each vertical line represents the prediction interval (lower to upper bound) for an individual sample, with samples sorted by their true malignant fraction values in ascending order along the x-axis. The true values are marked with either red squares (when contained within the prediction interval) or blue triangles (when falling outside the prediction interval).

(TIF)

S9 Fig. Visualization of prediction intervals on the real primary dataset of DeepDeconUQ and MEAD at α = 0.15 (85% confidence level). Comparison of uncertainty intervals generated by DeepDeconUQ (left) and MEAD (right) methods. Each vertical line represents the prediction interval (lower to upper bound) for an individual sample, with samples sorted by their true malignant fraction values in ascending order along the x-axis. The true values are marked with either red squares (when contained within the prediction interval) or blue triangles (when falling outside the prediction interval).

(TIF)

S10 Fig. Comparison of DeepDeconUQ with/without gene selection on simulated AML datasets. Each bar in the boxplot comprises 15 data points, each corresponding to one of 15 simulated epithelial datasets. Significance levels are indicated with different colors.

(TIF)

S1 Table. Preprocessing criteria for each subject in AML and neuroblastoma datasets. The gene expression threshold means the maximum gene expression value of a cell. The gene number threshold means the maximum number of expressed genes. This is to avoid gene expressions that do not represent a single cell. The criteria are based on Scanpy (v. 1.7.2) functions 'filter_cells' and 'filter_genes'.

S2 Table. Real bulk AML RNA-seq datasets used in DeepDecon.

S3 Table. Performance of DeepDeconUQ with and without PCA embedding and with gene selection on real AML datasets.

Supplementary Information
DeepDeconUQ Estimates Malignant Cell Fraction Prediction Intervals in Bulk RNA-seq Tissue

Jiawei Huang[1], Yuxuan Du[1,4], Kevin R. Kelly[3], Jinchi Lv[5], Yingying Fan[5], Jiang F. Zhong [*2] and Fengzhu Sun [†1]

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, United States of America
[2]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, California, United States of America
[3]Division of Hematology, University of Southern California, Los Angeles, California, United States of America
[4]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, Texas, United States of America
[5]Data Sciences and Operations Department, University of Southern California, Los Angeles, California, United States of America

1 Preprocessing of single-cell gene expression data

The AML single cell RNA-seq data used in DeepDeconUQ is the same as the data used in DeepDecon [1]. AML data was obtained from the Gene Expression Omnibus (GEO) under accession number GSE116256 [2]. To ensure data quality, we utilized single-cell RNA sequencing (scRNA-seq) data from subjects with at least 100 normal and 100 malignant cells, respectively (Fig J). This criterion was employed to avoid extreme scenarios in which very few normal or malignant cells were selected. A total of 15 AML subjects were selected. For each subject, we filtered out cells with less than 500 detected genes and genes expressed in less than five cells. The resulting gene expression profile for each subject was further filtered for extreme outliers in gene expression values. The filtering criteria for each subject were given in Table B. Finally, gene expression was normalized to library size by total counts across all genes. This will counteract the effect of different library sizes. Finally, the resulting normalized matrix of all filtered cells and genes was saved for subsequent pseudo bulk data generation. We used the tag 'PredictionRefined' [2] given in the dataset as the final label (malignant/normal) to annotate the cell types. This tag was a manual reclassification of cells by close inspection of mutations/expression profiles.

2 Artificial bulk dataset simulation

The simulated artificial bulk datasets were generated by subsampling within each scRNA-seq subject. Cells from different subjects were not merged into one bulk sample to preserve potential

[*]Corresponding author: jzhong@llu.edu
[†]Corresponding author: fsun@usc.edu

correlations between the expression levels of different genes within subjects. The true cell-type proportion for each bulk sample was calculated by dividing the number of single cells with a specific cell type by the total number of cells in the bulk sample.

First, in each bulk scRNA-seq data, N cells from different cell types (malignant/normal) were generated where "1" and "2" correspond to malignant and normal cell types, respectively. Let

$$N = n_1 + n_2, \tag{1}$$

$$f_i = \frac{n_i}{N}, i = 1, 2 \tag{2}$$

where n_i and f_i are the number and fraction of cells of type i, respectively, and N is the total number of cells in one simulated bulk sample. Here n_i was generated uniformly from 0 to N through the python random module [3]. When n_i was determined, n_i cells were sampled from the scRNA-seq gene expression matrix for each cell type i (if n_i is bigger than the total number of cells of type i in one particular subject, the cells were chosen with replacement. Otherwise, the cells were chosen without replacement). Next, the selected single-cell expression profiles for every cell type were aggregated by summing their expression values,

$$G = \sum_i \sum_j X_{ij}, \tag{3}$$

where X_{ij} is the jth gene expression vector of cell type i and G is the final bulk RNA-seq expression profile. Repeating the above steps T times to construct a simulated bulk dataset with T samples. In our simulations, T was chosen as 200 for each subject. Finally, we had 15 simulated bulk RNA-seq datasets, each with 200 bulk samples with known cell type proportions.

3 The influence of feature embedding to DeepDeconUQ

Although batch effects can be mitigated through TF-IDF transformation and Min-Max normalization in DeepDeconUQ, some denoising methods can still be tried to test DeepDeconUQ's performance. One of such is feature embedding.

To test more advanced denoising methods, we employed both classical dimensionality reduction via Principal Component Analysis (PCA) and an advanced transfer learning method, Deep Adaptation Network (DAN) [4]. DAN represents a neural architecture that aligns feature distributions between training and testing domains. DAN uses both the labels and samples of training data and only uses the samples of testing data to generate domain-invariant latent embeddings with theoretical guarantees. It will generate a latent embedding for both the training and testing datasets.

For PCA-based dimensionality reduction, data underwent standard preprocessing (TF-IDF transformation followed by MinMax normalization) before extracting the top 100 principal components as neural network inputs. The DAN implementation faithfully reproduced the methodology described by Long et al. [4], generating latent embeddings for both training and testing datasets that subsequently served as inputs for the neural network training pipeline. All downstream computational processes remained consistent with the original DeepDeconUQ framework.

Performance comparisons revealed distinct trade-offs between methods (Fig H). PCA embeddings demonstrated systematic over-coverage in simulation datasets, with coverage rates exceeding corresponding significance thresholds. Even though it has narrower prediction intervals than baseline DeepDeconUQ. However, when evaluated on real-world datasets, PCA embeddings exhibited reduced coverage relative to the original model, suggesting potential overfitting (Table A). While

achieving coverage rates appropriately aligned with significance levels, DAN embeddings produced substantially wider prediction intervals than the baseline model. We hypothesize that inherent biological heterogeneity across cancer patients presents fundamental challenges to transfer learning in this context, limiting the effectiveness of domain adaptation techniques.

4 Extention to complex tumor microenvironment (TME)

DeepDeconUQ requires the knowledge of malignant and normal cells. However, in practice, cancer tissues usually exhibit a complex tumor microenvironment (TME). There are usually multiple subtypes for either malignant or normal cells. In this case, we can merge the malignant subtypes into one malignant type and the normal subtypes into one normal type. DeepDeconUQ can then be used to estimate the prediction interval of the malignant cell fraction. Specifically, we have collected single-cell data from [5], which comprises scRNA-seq profiles from seven individuals. Following data filtering procedures identical to those applied to the AML dataset (detailed in Section 1), we retained 18,062 single cells derived from four individuals. Cell classifications were established using the 'condition' parameter from Sathe, et al. 2023 [5]. These individuals contain tumor epithelial cells as malignant labels and NK, T-cell, B-cell as normal labels. We subsequently conducted analogous experiments to those performed with the AML dataset. This involved initially constructing synthetic bulk RNA-seq datasets with variable malignant cell proportions based on the available scRNA-seq data. We then trained a DeepDeconUQ model to estimate prediction intervals for malignant cell fractions.

Fig E illustrates the performance of DeepDeconUQ on the epithelial datasets compared to existing methodologies. RNA-Sieve demonstrated a low performance, with coverage probabilities substantially below threshold values across different significance levels, indicating under-coverage. MEAD exhibited complete performance failure on the epithelial dataset, generating prediction intervals approximating 0-1 for all samples and, in some instances, failing to produce valid prediction intervals entirely (returning NA values). In contrast, DeepDeconUQ exhibited superior performance, yielding statistically valid coverage with comparatively narrow prediction intervals. The modest over-coverage observed is likely attributed to the limited training cohort (n=4) and could potentially be mitigated by incorporating additional high-quality datasets.

5 Influence of gene selection on DeepDeconUQ

In the current implementation, DeepDeconUQ utilizes the complete gene set for prediction. We investigated whether selective incorporation of cell type-specific genes could enhance prediction accuracy. To address this question, we employed MAST (Model-based Analysis of Single-cell Transcriptomics) [6] as a differential gene expression analysis tool to evaluate the impact of feature selection on DeepDeconUQ performance.

We applied MAST to identify differentially expressed genes between malignant and normal cell populations within the AML scRNA-seq datasets. MAST takes the AML scRNA-seq datasets as input and runs a likelihood ratio test to get differential genes between two conditions (malignant/normal). This approach yielded $1,414$ genes with adjusted p-values below the significance threshold of 0.05. Subsequently, we executed DeepDeconUQ using only this subset of differentially expressed genes and compared the results against the original implementation utilizing the complete gene set.

Fig I illustrates DeepDeconUQ performance with and without gene selection on simulated AML datasets, while Table A presents corresponding results on real AML datasets. In simulated datasets,

DeepDeconUQ demonstrated comparable performance regardless of gene selection strategy, highlighting the model's inherent robustness. However, in real AML datasets, the considerable heterogeneity of cancer tissue potentially resulted in differential gene signatures that did not effectively generalize across real samples. Consequently, DeepDeconUQ with gene selection exhibited reduced coverage rates at all significance levels when applied to recurrent and beat AML datasets compared to the original implementation.

6 Software comparison and settings

We compared DeepDeconUQ with other deconvolution methods, including MEAD (v. 1.0.1) [7], RNA-Sieve (v. 0.1.4) [8].

For MEAD [7], we installed the R package given in the manuscript and ran it with default settings. In the leave-one-out cross-validation, the single-cell profile was constructed using the single cells of all subjects, excluding the subject itself, while in the real bulk testing data, the single-cell profile was constructed by combining all available single-cell data. Subject information was included in the single-cell reference. Then, we ran MEAD with default settings by following the example provided by the authors.

For RNA-Sieve [8], we executed it by following the example code provided. In the leave-one-out cross-validation, the single-cell profile was constructed by combining the single cells of all subjects, excluding the data itself, while in the real bulk testing data, the single-cell profile was constructed by combining all available single-cell data.

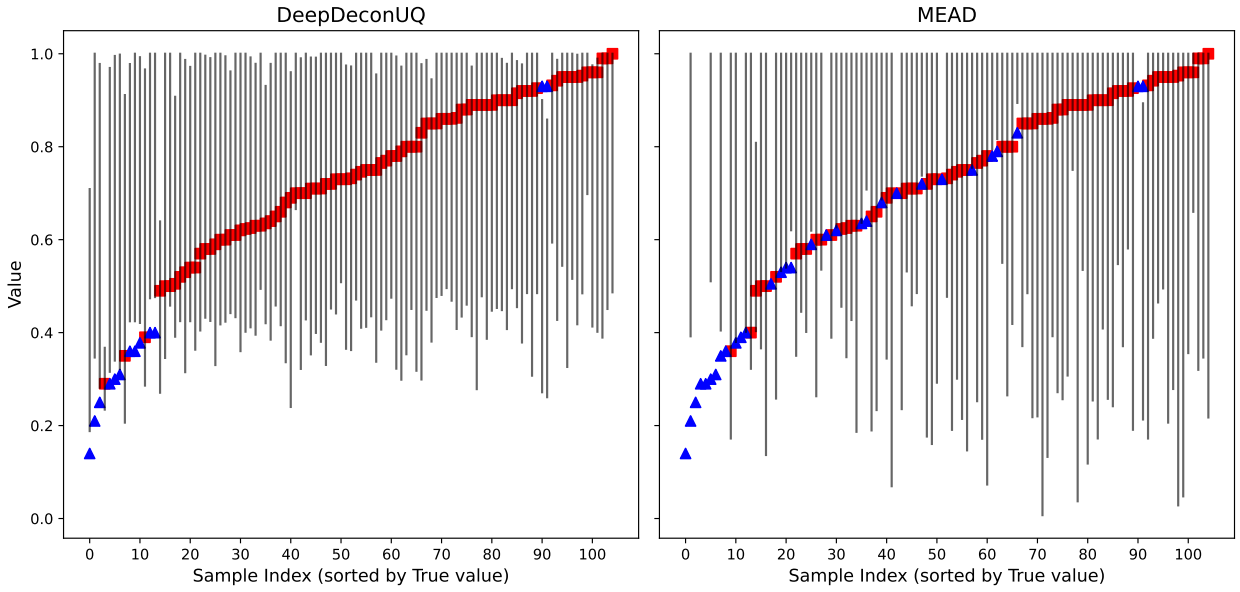

Fig A: **Visualization of prediction intervals on the real primary dataset of DeepDeconUQ and MEAD at $\alpha = 0.10$ (90% confidence level).** Comparison of uncertainty intervals generated by DeepDeconUQ (left) and MEAD (right) methods. Each vertical line represents the prediction interval (lower to upper bound) for an individual sample, with samples sorted by their true malignant fraction values in ascending order along the x-axis. The true values are marked with either red squares (when contained within the prediction interval) or blue triangles (when falling outside the prediction interval).

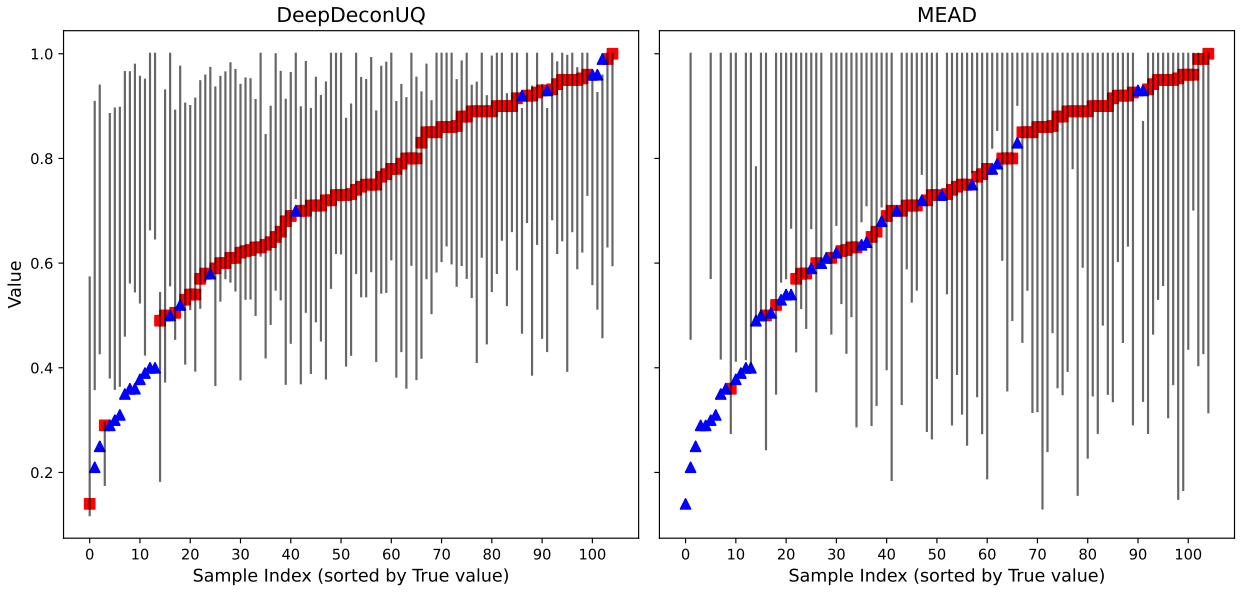

Fig B: **Visualization of prediction intervals on the real primary dataset of DeepDeconUQ and MEAD at $\alpha = 0.15$ (85% confidence level).** Comparison of uncertainty intervals generated by DeepDeconUQ (left) and MEAD (right) methods. Each vertical line represents the prediction interval (lower to upper bound) for an individual sample, with samples sorted by their true malignant fraction values in ascending order along the x-axis. The true values are marked with either red squares (when contained within the prediction interval) or blue triangles (when falling outside the prediction interval).

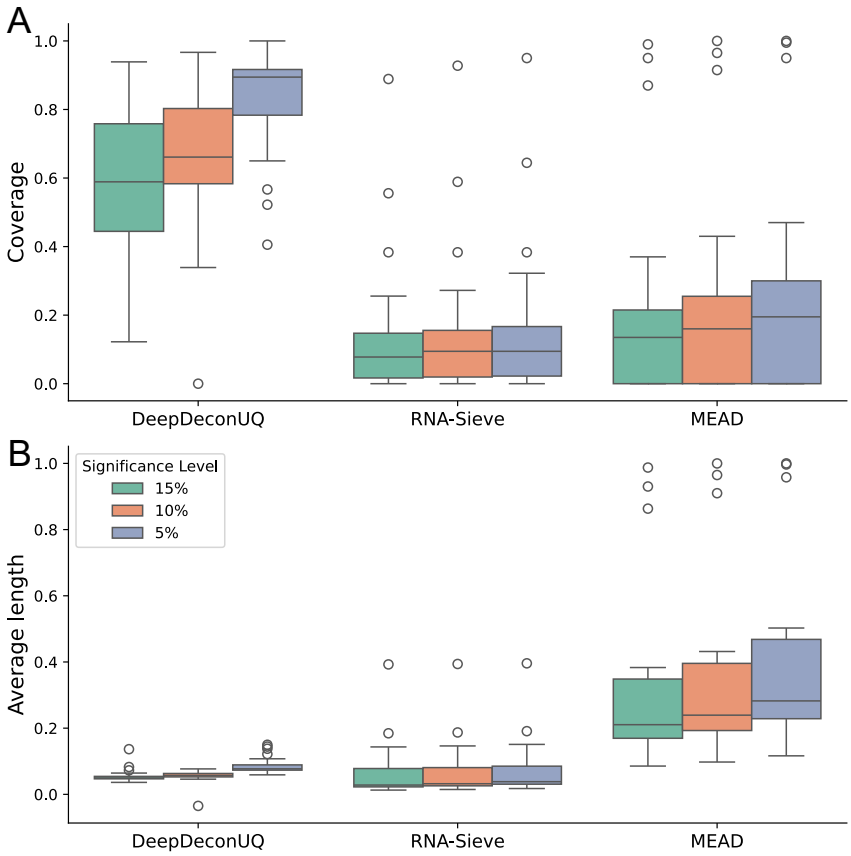

Fig C: **DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on simulated HNSCC bulk RNA-seq datasets.** Boxplots of coverage (A) and average prediction interval length (B) on simulated HNSCC bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Significance levels are indicated with different colors.

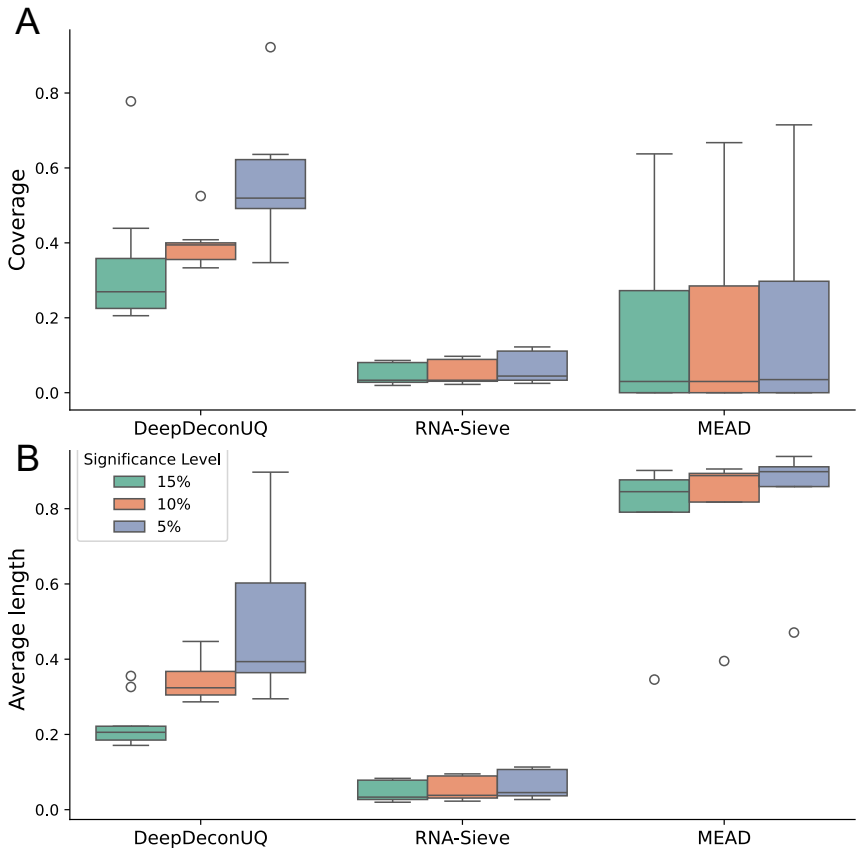

Fig D: **DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on simulated Neuroblastoma bulk RNA-seq datasets.** Boxplots of coverage (A) and average prediction interval length (B) on simulated Neuroblastoma bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Significance levels are indicated with different colors.

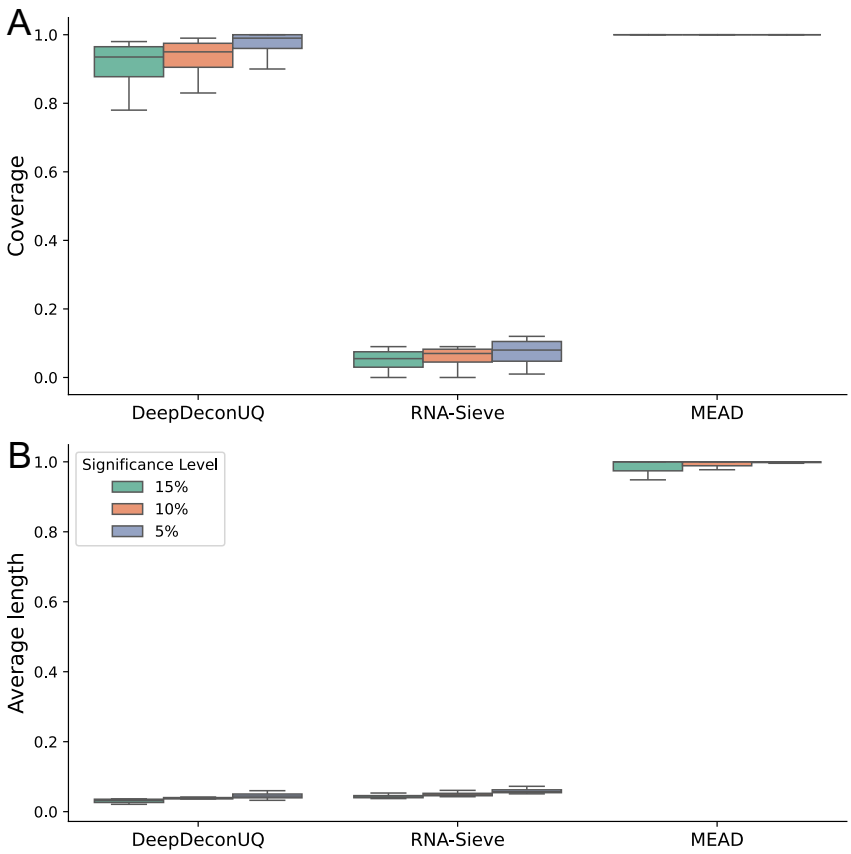

Fig E: **DeepDeconUQ outperforms other methods in predicting malignant cell type prediction interval on simulated epithelial bulk RNA-seq datasets.** Boxplots of coverage (A) and average prediction interval length (B) on four simulated epithelial bulk RNA-seq datasets. Coverage is defined as the proportion of instances in which the true fraction of malignant cells falls within the prediction interval for the testing dataset. The average length represents the mean length of the prediction intervals across the testing datasets. Each bar in the boxplot comprises 4 data points, each corresponding to one of 4 simulated epithelial datasets (except one dataset in MEAD that gives NA values and, therefore, doesn't show). Significance levels are indicated with different colors.

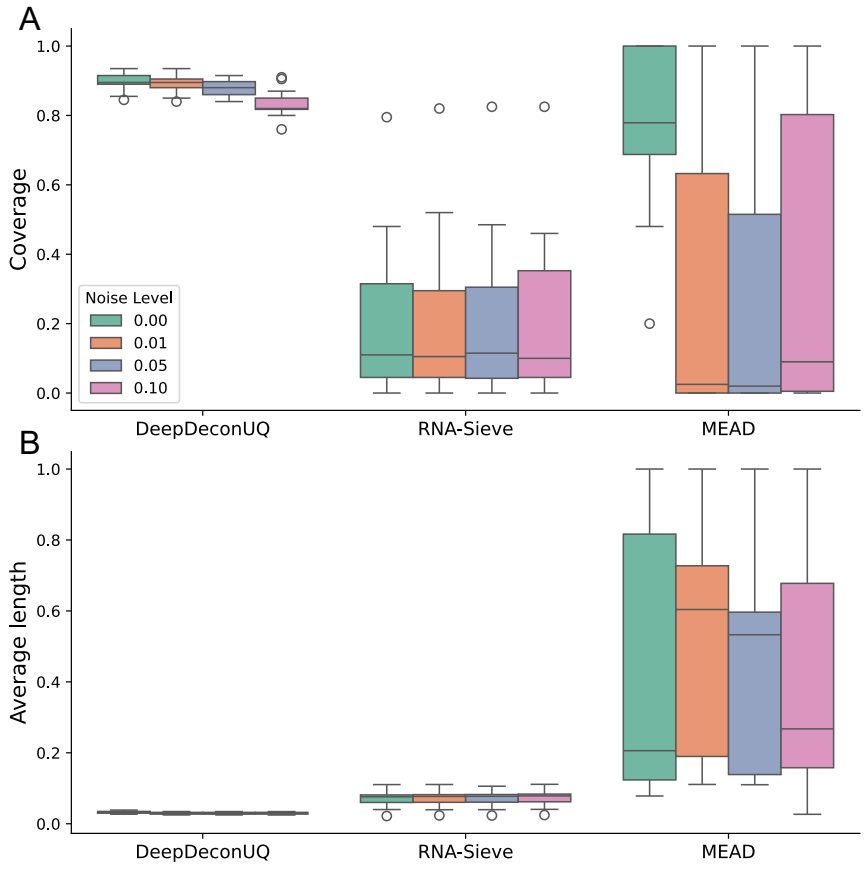

Fig F: **DeepDeconUQ is robust to gene expression perturbations.** Boxplots of coverage and average prediction interval length on 15 AML simulated bulk RNA-seq datasets under different noise levels. We added random noise generated from a Gaussian distribution with zero mean and variance that equals $\lambda(\lambda = 0.01, 0.05, 0.1)$ times the gene expression level for each gene in each sample. Each bar contains a total of 15 points, representing 15 separate AML datasets. The color represents different levels of noise level λ. Significance level $\alpha = 0.15$.

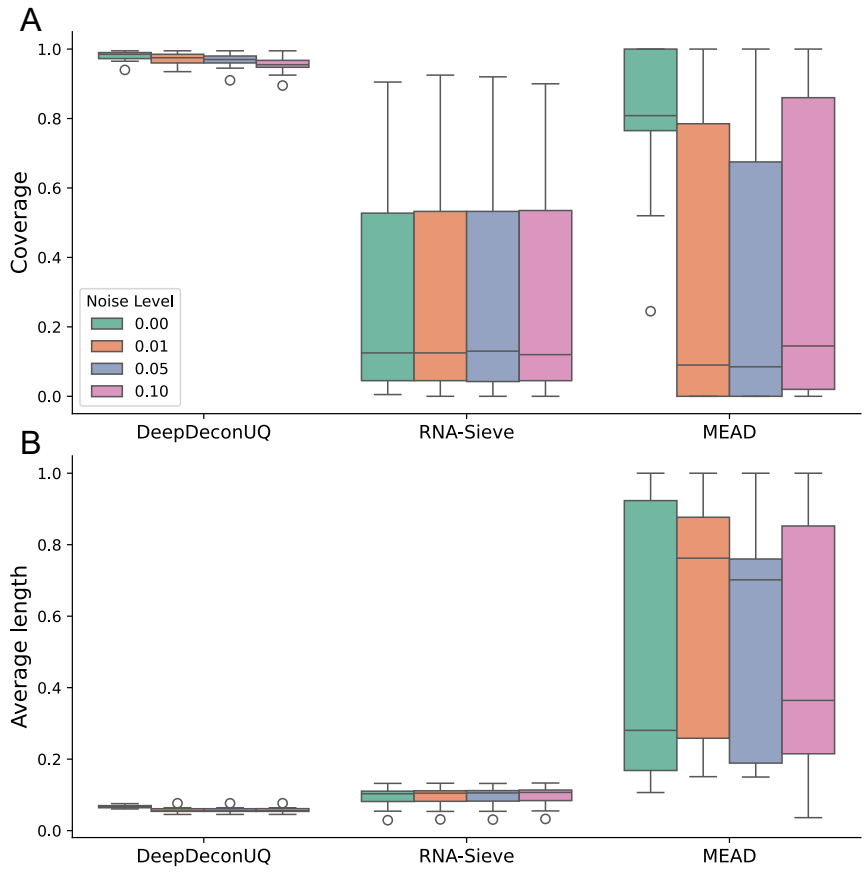

Fig G: **DeepDeconUQ is robust to gene expression perturbations.** Boxplots of coverage and average prediction interval length on 15 AML simulated bulk RNA-seq datasets under different noise levels. We added random noise generated from a Gaussian distribution with zero mean and variance that equals $\lambda(\lambda = 0.01, 0.05, 0.1)$ times the gene expression level for each gene in each sample. Each bar contains a total of 15 points, representing 15 separate AML datasets. The color represents different levels of noise level λ. Significance level $\alpha = 0.05$.

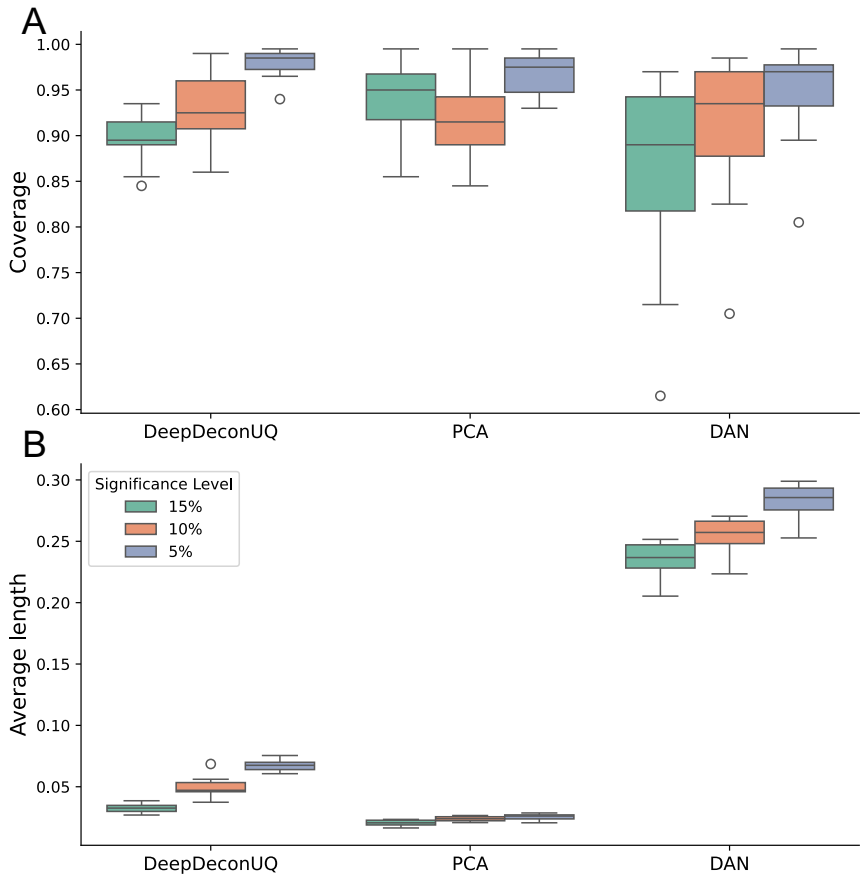

Fig H: **Comparison of DeepDeconUQ with different embedding methods.** Two embedding methods are used to compare with DeepDeconUQ. Principal Component Analysis (PCA) selects the top 100 principal components as neural network inputs. DAN makes use of transfer learning and generates a latent embedding layer for both training and testing datasets. Each point in the boxplot is an artificial bulk RNA-seq dataset.

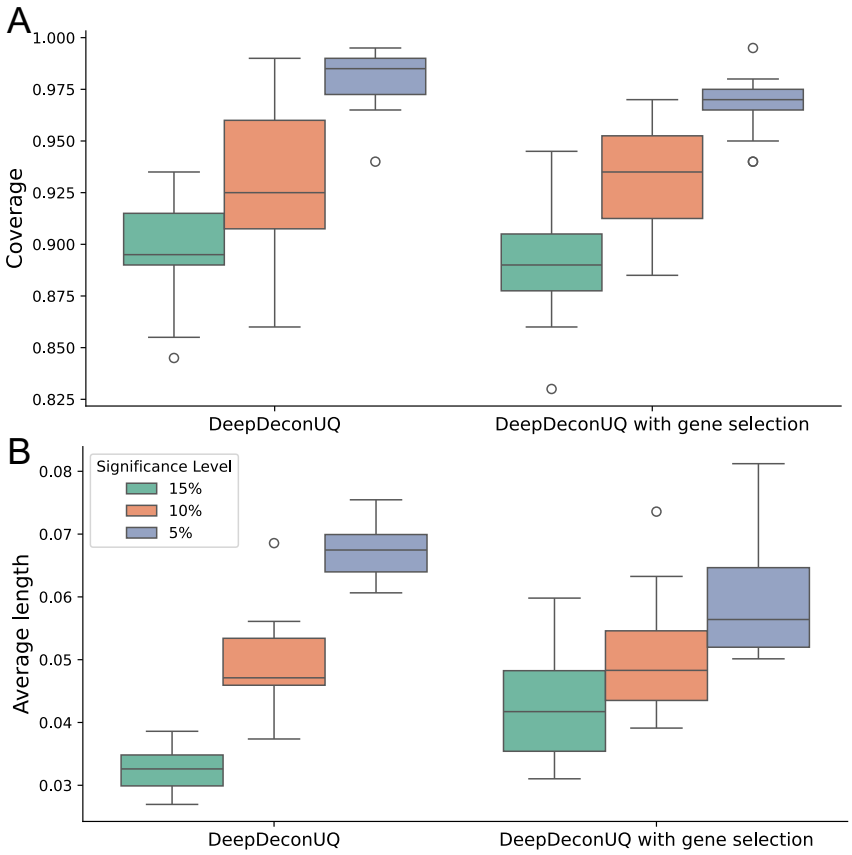

Fig I: **Comparison of DeepDeconUQ with/without gene selection on simulated AML datasets.** Each bar in the boxplot comprises 15 data points, each corresponding to one of 15 simulated epithelial datasets. Significance levels are indicated with different colors.

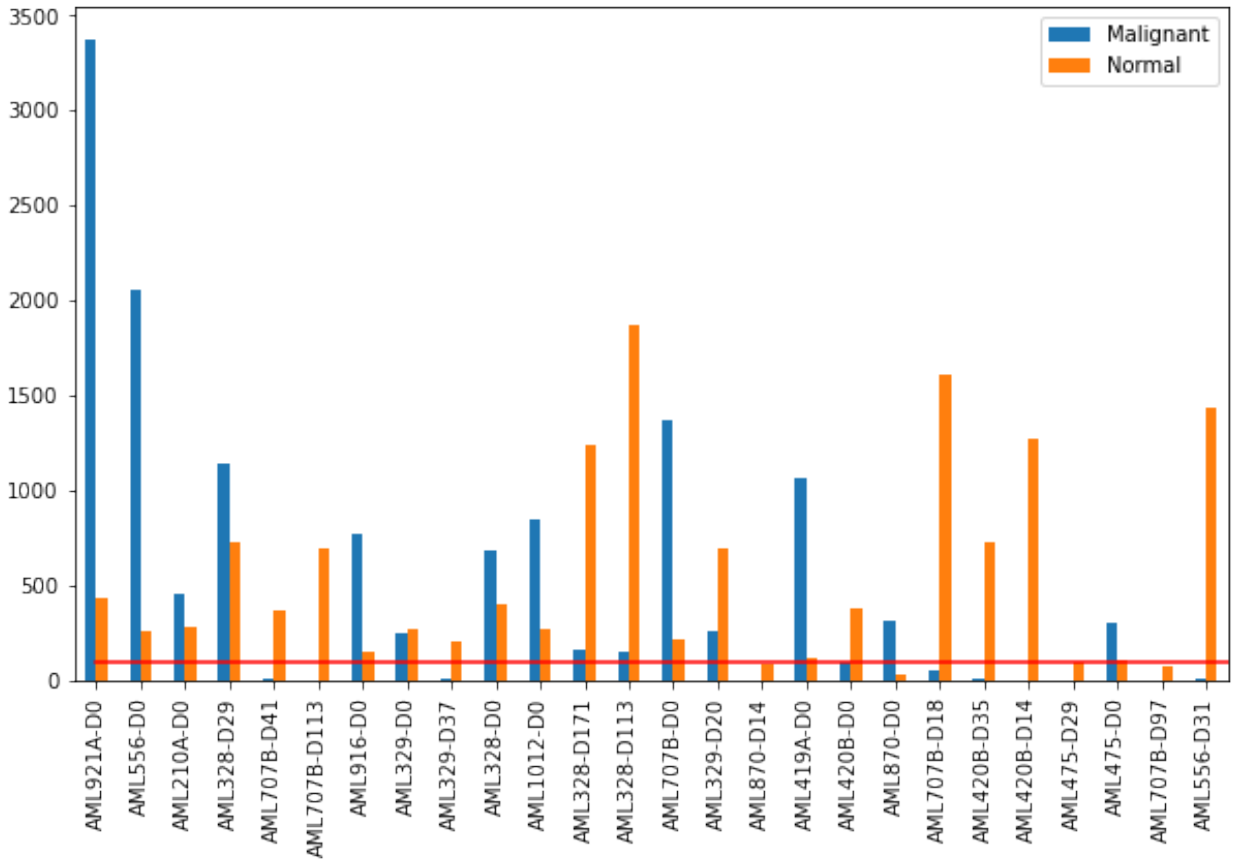

Fig J: **Barplots of the numbers of malignant and normal cells in each scRNA-seq subject.** Subjects with at least 100 malignant and 100 normal cells were selected for this study.

Table A: Performance of DeepDeconUQ with and without PCA embedding and with gene selection on real AML datasets.

Methods	**Dataset**	**15%**		**10%**		**5%**	
		Coverage	L_{avg}	**Coverage**	L_{avg}	**Coverage**	L_{avg}
DeepDeconUQ	primary	0.800	0.434	0.876	0.572	0.912	0.662
	recurrent	0.824	0.606	0.853	0.604	0.882	0.685
	beat	0.592	0.409	0.730	0.554	0.781	0.611
PCA embedding	primary	0.714	0.202	0.800	0.356	0.867	0.310
	recurrent	0.706	0.236	0.735	0.324	0.853	0.433
	beat	0.233	0.149	0.237	0.151	0.344	0.215
Gene Selection	primary	0.829	0.511	0.905	0.608	0.952	0.680
	recurrent	0.676	0.490	0.794	0.641	0.853	0.678
	beat	0.537	0.460	0.619	0.465	0.648	0.482

Table B: Preprocessing criteria for each subject in AML and neuroblastoma datasets. The gene expression threshold means the maximum gene expression value of a cell. The gene number threshold means the maximum number of expressed genes. This is to avoid gene expressions that do not represent a single cell. The criteria are based on Scanpy (v. 1.7.2) functions 'filter_cells' and 'filter_genes'.

Subject	Gene expression threshold	Gene number threshold
AML328-D29	7000	2500
AML1012-D0	5000	1600
AML556-D0	10000	3000
AML328-D171	5000	2000
AML210A-D0	6000	2000
AML419A-D0	7000	2500
AML328-D0	5000	2000
AML707B-D0	6000	2000
AML916-D0	5000	2000
AML328-D113	6000	2000
AML329-D0	8000	2000
AML420B-D0	7000	2000
AML329-D20	7000	2200
AML921A-D0	8000	2500
AML475-D0	4800	1500

Table C: Real bulk AML RNA-seq datasets used in DeepDecon

Name	Number of samples	Sequencing platform	Normalization method	Source
primary	117	Affymetrix Gene ST Array [9]	FPKM	GDC Data Portal [10]
recurrent	38	Affymetrix Gene ST Array	FPKM	GDC Data Portal
BeatAML	300	SureSelect [11]	CPM	Tyner, et al. 2018 [12]
Pediatric Neuroblastoma	99	Illumina Hi-Seq 2000 [13]	RPKM	cBioPortal [14]
TCGA-HNSC	518	Illumina Hi-Seq 2000	RPKM	LinkedOmics [15]

References

[1] Huang J, Du Y, Stucky A, Kelly KR, Zhong JF, Sun F. DeepDecon accurately estimates cancer cell fractions in bulk RNA-seq data. Patterns. 2024;5(5):100969. doi:10.1016/j.patter.2024.100969.

[2] van Galen P, Hovestadt V, Wadsworth M, Hughes T, Griffin GK, Battaglia S, et al. Single-Cell RNA-seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019;176(6):1265–1281.e24. doi:10.1016/j.cell.2019.01.031.

[3] Matsumoto M, Nishimura T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. ACM Trans Model Comput Simul. 1998;8(1):3–30. doi:10.1145/272991.272995.

[4] Long M, Zhu H, Wang J, Jordan MI. Deep transfer learning with joint adaptation networks. In: International conference on machine learning. PMLR; 2017. p. 2208–2217.

[5] Sathe A, Mason K, Grimes SM, Zhou Z, Lau BT, Bai X, et al. Colorectal cancer metastases in the liver establish immunosuppressive spatial networking between tumor-associated SPP1+ macrophages and fibroblasts. Clinical Cancer Research. 2023;29(1):244–260. doi:10.1158/1078-0432.CCR-22-2041.

[6] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology. 2015;16:1–13. doi:10.1186/s13059-015-0844-5.

[7] Xie D, Wang J. Robust Statistical Inference for Cell Type Deconvolution. arXiv preprint arXiv:220206420. 2022;doi:10.48550/arXiv.2202.06420.

[8] Erdmann-Pham DD, Fischer J, Hong J, Song YS. Likelihood-based deconvolution of bulk gene expression data using single-cell references. Genome Research. 2021;31(10):1794–1806. doi:10.1101/gr.272344.120.

[9] Array AGS. TARGET's Study of Acute Myeloid Leukemia.;. `https://www.cancer.gov/ccg/research/genome-sequencing/target/using-target-data/technology#aml`.

[10] Portal GD. TARGET's Study of Acute Myeloid Leukemia.;. `https://gdc.cancer.gov/content/target-aml-publication-summary`.

[11] 38Mb S. Functional Genomic Landscape of Acute Myeloid Leukemia.;. `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001657.v1.p1`.

[12] Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018;562(7728):526–531. doi:10.1038/s41586-018-0623-z.

[13] 2000 IHS. Protocols used for TARGET's study of Neuroblastoma (NBL);. `https://www.cancer.gov/ccg/research/genome-sequencing/target/using-target-data/technology`.

[14] cBioPortal Pediatric Neuroblastoma. Neuroblastoma dataset from cBioPortal;. `https://www.cbioportal.org/study/summary?id=nbl_target_2018_pub`.

[15] Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Research. 2018;46(D1):D956–D963. doi:10.1093/nar/gkx1090.