Effective methods for bulk RNA-Seq deconvolution using scnRNA-Seq transcriptomes

Francisco Avila Cobos[1,*], Mohammad Javad Najaf Panah[2,*], Jessica Epps[2], Xiaochen Long[2,3], Tsz-Kwong Man[2], Hua-Sheng Chiu[2], Elad Chomsky[4], Evgeny Kiner[4], Michael J Krueger[2], Diego di Bernardo[5], Luis Voloch[4], Jan Molenaar[6], Sander R. van Hooff[6], Frank Westermann[7], Selina Jansky[7], Michele L. Redell[2], Pieter Mestdagh[1,#], Pavel Sumazin[2,#]

[1]Department of Biomolecular Medicine, Ghent University, Ghent, Belgium; Cancer Research Institute Ghent, Ghent, Belgium.
[2]Department of Pediatrics, Baylor College of Medicine; Texas Children's Hospital and Cancer Center, Houston, Texas, USA.
[3]Department of Statistics, Rice University, Houston, Texas, 77251, USA.
[4]ImmunAi, New York, New York, USA.
[5]Telethon Institute of Genetics and Medicine, Via Campi Flegrei 34, 80078 Pozzuoli; Dept. Chemical, Materials and Industrial Engineering, University of Naples "Federico II", Naples, Italy.
[6]Princess Maxima Center for Pediatric Oncology, Utrecht, The Netherlands.
[7]German Cancer Research Center, DKFZ, Heidelberg, Germany.
[*]Equal contribution. [#]Corresponding authors.

ABSTRACT

RNA profiling technologies at single-cell resolutions, including single-cell and single-nuclei RNA sequencing (scRNA-Seq and snRNA-Seq, scnRNA-Seq for short), can help characterize the composition of tissues and reveal cells that influence key functions in both healthy and disease tissues. However, the use of these technologies is operationally challenging because of high costs and stringent sample-collection requirements. Computational deconvolution methods that infer the composition of bulk-profiled samples using scnRNA-Seq-characterized cell types can broaden scnRNA-Seq applications, but their effectiveness remains controversial. We produced the first systematic evaluation of deconvolution methods on datasets with either known or scnRNA-Seq-estimated compositions. Our analyses revealed biases that are common to scnRNA-Seq 10X Genomics assays and illustrated the importance of accurate and properly controlled data preprocessing and method selection and optimization. Moreover, our results suggested that concurrent RNA-Seq and scnRNA-Seq profiles can help improve the accuracy of both scnRNA-Seq preprocessing and the deconvolution methods that employ them. Indeed, our proposed method, Single-cell RNA Quantity Informed Deconvolution (SQUID), combined RNA-Seq transformation and dampened weighted least-squares deconvolution approaches to consistently outperform other methods in predicting the composition of cell mixtures and tissue samples. Furthermore, our analysis suggested that only SQUID could identify outcomes-predictive cancer cell subclones in pediatric acute myeloid leukemia and neuroblastoma datasets, suggesting that deconvolution accuracy improvements are vital to enabling its applications in the life sciences.

INTRODUCTION

Single-cell and single-nuclei RNA-sequencing (scnRNA-Seq) technologies have revolutionized our ability to quantify cell types and cell states in healthy and disease tissues. scnRNA-Seq technologies generate cell-type specific transcriptomes, with individual cells labeled, enumerated, and molecularly characterized. This, in turn, allows for comparing the cell composition of tissues

and for associating changes in tissue cell-type abundances and both their molecular and clinical parameters. Examples include scnRNA-Seq assays that helped identify programs for tissue development[1] and regeneration[2] and associated patient outcomes with tumor subclones[3]. scnRNA-Seq assays helped reveal immune-cell-type composition differences that may dictate responses to immune checkpoint inhibition therapies[4], identified tumor subclones that acquired drug resistance during treatments[5], and identified cancer cells that adapted to evade targeted therapies[6]. scnRNA-Seq assays are increasingly enabling research to identify therapeutic targets and diagnostic biomarkers in efforts to improve therapies for cancer and other diseases.

While scnRNA-Seq assays can provide cell-type-specific information at unprecedented resolutions, their implementation is associated with challenges that prevent their widespread adoption in clinical settings. These challenges include the high cost of library preparation and sequencing, and the stringent requirements for sample collection, processing, and storage. Namely, the current cost of scnRNA-Seq assays is 10-30-fold greater than the cost of bulk RNA sequencing (RNA-Seq), which effectively prevents their adoption at scales previously seen for RNA-Seq. Importantly, specialized facilities for sample collection and tissue processing are required for accurate profiling and these are not readily available at most hospitals or academic institutions. For example, accurate scRNA-Seq profiles require fresh tissue dissociation and cell suspension generation at carefully controlled temperatures. Moreover, tissue preservation and cell sorting are known to alter scnRNA-Seq estimates, with some commonly used methods shown to introduce bias by selectively depleting genes and cell types[7-9].

RNA-Seq is less challenging to implement in clinical settings, but it only provides mean gene expression abundance estimates across cell types. Recently, computational deconvolution methods were proposed to infer cell-type abundances from RNA-Seq profiles using either reference matrices composed of cell-type-specific gene expression signatures[10-12] or scnRNA-Seq data from the same tissue type[13-15]. In various benchmarking efforts, we and others have shown that multiple factors, including data transformation, data normalization, and the composition of the reference matrix can impact the performance of deconvolution methods[10]. However, given the potential impact of scnRNA-Seq-based deconvolution on advances in the life sciences, there remains a need to systematically compare and quantify the absolute accuracies of deconvolution methods.

Here, we evaluated deconvolution methods in 8 datasets of concurrent bulk RNA-Seq and scnRNA-Seq profiles (see Table S4). These datasets included cell mixtures, where cell type abundances and expression profiles are known with high accuracy and that could be used to quantify both deconvolution and scRNA-Seq expression estimates, as well as tissues that allow comparing the effects of common preservation protocols. When evaluating deconvolutions of bulk RNA-Seq profiles, accuracy was determined by comparing deconvolution-predicted cell abundances to gold-standard estimates, where gold-standard estimates were derived from either validated counts of the composing cells or the analyses of scnRNA-Seq profiles. Surprisingly, our results suggested that some methods consistently produced the most accurate cell-abundance estimates, irrespective of datasets or data processing.

We hypothesized that concurrent RNA-Seq and scnRNA-Seq profiling could be used to not only evaluate deconvolution methods but also improve deconvolution accuracy. To test this, we

developed the R package Single-cell RNA Quantity Informed Deconvolution (SQUID), which combines bulk RNA-Seq transformation and dampened weighted least squares deconvolution approaches. Analyses of SQUID accuracy suggested that methods that harness the power of concurrent RNA-Seq and scnRNA-Seq profiling can consistently outperform other methods in predicting the composition of cell mixtures and tissue samples. Finally, to evaluate the benefit of improved deconvolution accuracy for applications in cancer research, we concurrently profiled pediatric acute myeloid leukemia (AML) and neuroblastoma samples by RNA-Seq and scRNA-Seq and tested whether deconvolution methods can predict risk, based on the abundance of potential high-risk cancer subclones in diagnostic samples. Our results indicated that only SQUID subclone-abundance estimates were predictive of outcomes in RNA-Seq-profiled AML and neuroblastoma diagnostic samples. Thus, we concluded that SQUID's deconvolution-accuracy improvement is key to enabling its potential applications in diagnostic protocols for these cancers.

METHODS

Deconvolution benchmarking framework

For those scnRNA-Seq datasets for which no metadata or cell label information was available, cells were clustered together in an unsupervised fashion using Monocle3. Specifically, we sequentially applied the "preprocess_cds" (num_dim = 100, norm_method = "log", method = "PCA", scaling = TRUE), "reduce_dimension" (max_components = 2, umap.metric = "cosine", umap.fast_sgd = FALSE, preprocess_method = 'PCA') and "cluster_cells" (k = 20, resolution = NULL, partition_qval = 0.05, num_iter = 1) functions; see our GitHub repository for detailed code, functions, and parameters. During quality control and preprocessing, we removed cells with extreme mitochondrial or ribosomal content (top 0.5% and bottom 0.5%) and we kept detectable genes that were expressed in at least 10 cells or 1% of the cells in any of the clusters. Next, cluster-specific gene expression profiles were obtained by averaging raw gene expression values across all cells from a given cluster, and cluster-specific markers were obtained using the FindAllMarkers function from Seurat v4.0.4 with the threshold of log2(1.5) and using a Wilcoxon test on TMM normalized scnRNA-Seq data. Gold standard abundance estimates were obtained either as the sum of individual cells or nuclei present in each cluster or by immunohistochemistry/ Fluorescence-activated Cell Sorting (FACSymphony) cell counts; see Figure 1 for a schematic representation of the benchmarking framework.

We refer to cell clusters as cell types throughout the manuscript, even when no annotations are available. Cell-type specific gene signatures were used to establish reference matrices for the deconvolution of their matching bulk RNA-Seq data using CIBERSORT[11,16], FARDEEP[17], RLR[18], and NNLS[19]. Alternatively, deconvolution of bulk RNA-Seq data was performed with Ordinary Least Squares regression (OLS), dampened weighted least squares (DWLS)[14], and MuSiC[15], which directly use the scnRNA-Seq data as the reference. Of note, MuSiC was tested in two different ways: (1) using the markers found by FindAllMarkers described above and (2) without including any prior marker information (markers = NULL). We used OLS, and OLS with a non-negativity constraint (NNLS) as naïve deconvolution tools to benchmark all other methods. Performance was quantified by calculating the Pearson correlation coefficient and root mean squared error (RMSE) between the cell-type proportions observed by deconvolution and the expected cell-type proportions that were either known or derived from scnRNA-Seq.

Deconvolution with ordinary least squares regression (OLS)

OLS was used to solve a simple set of linear equations that seeks to find the optimal composition P of a set of mixtures with bulk RNA profiles Z to minimize the difference between the observed bulk RNA-Seq profiles and the abundance-weighted sums of the expression profiles of composing cell types X. Namely, given bulk expression profiles $\{z_i \in Z\}$ of each mixture, and expression estimates for each cell type j $\{x_j \in X\}$, we sought to identify $p_{ij} \in P$ across all mixtures i and cell types j to minimize the difference between the bulk RNA-Seq profile and the mean expression of profiles of the composing cells (Equation 1).

$$\underset{P}{\operatorname{argmin}} \sum_i \left(z_i - \sum_j p_{ij} x_j\right)^2 \text{ for mixture } i \text{ and cell type } j \qquad \textbf{Equation 1}$$

Transforming and deconvolving bulk RNA-Seq profiles with SQUID

We present SQUID, a conversion-dampened weighted least squares strategy to transform and deconvolve bulk RNA-Seq data into scnRNA-Seq vector spaces. SQUID was intended to test the potential of using concurrent RNA-Seq and scnRNA-Seq profiling to improve deconvolution accuracy. Similar to Bisque[20], SQUID learns a transformation function of bulk RNA-Seq profiles Z to the concurrent pseudobulk profiles $\hat{Z}$, where pseudobulk scnRNA-Seq profiles are estimated as mean abundance across cells and samples. Then, the bulk RNA-Seq expression profile of each gene g with non-zero expression in both the bulk and scnRNA-Seq profiles is mapped to its pseudobulk profile according to Equation 2, where $\hat{z}_{g,i}$ and $z_{g,i}$ are the pseudobulk and bulk profiles of gene g in sample i, respectively. The coefficient a_g and constant b_g form the linear transformation for each gene g.

$$\underset{a,b}{\operatorname{argmin}} \sum_i \left(\hat{z}_{g,i} - \left(a_g z_{g,i} + b_g\right)\right)^2 \qquad \textbf{Equation 2}$$

This linear transformation was applied to all bulk RNA-Seq profiles to transform them to scnRNA-Seq space. This transformation minimizes the deviation between a sample's pseudobulk and bulk RNA-Seq profiles by mapping the bulk RNA-Seq expression profile of each gene to the magnitude and deviation of pseudobulk scnRNA-Seq values. Equation 2 also applies when converting bulk RNA-Seq profiles with no concurrent scnRNA-Seq profiles. However, when testing deconvolution by SQUID on our datasets, which included concurrent bulk RNA-Seq and scnRNA-Seq profiles for each sample, we used a left-one-out strategy. Namely, the linear transformation was optimized using all but one sample and was then used to transform the bulk RNA-Seq profile of the remaining sample. This transformed profile was then used to predict the composition of the sample with a dampened weighted least squares strategy like DWLS[14]. Deconvolution performance was determined using cell counts for our cell mixtures and estimates from single-cell profiles for patient samples with concurrent bulk and scnRNA-Seq profiles (gold standard). We note that cell counts are the most accurate and unbiased estimates for our cell mixtures, and single-cell estimates are our only estimates for the true composition of patient samples. Comparisons of SQUID and other deconvolution method accuracy without cross validation are given in Supplementary Figures S2-6.

We note that the proposed linear transformation in Equation 2 is not unique. Indeed, Bisque proposed an alternative transformation that could be used more generally (Equation 3). We tested

this formulation and found that it performed equivalently to the formulation given in Equation 2. Namely, let $\overline{\hat{z}_g}$ denote the average expression estimate of gene g in pseudobulk profiles $\hat{Z}$ and $\overline{z_g}$ the average expression of this gene in all bulk RNA-Seq profiles—including the matching and other bulk RNA-Seq and profiles Z, and let $\widehat{\sigma_g}$ and σ_g denote their respective standard deviations. Then the transformed profile for gene g in sample i ($\vec{z}_{g,i}$) is given in Equation 3. Note that this formulation does not require RNA-Seq and scnRNA-Seq profiling to be concurrent.

$$\vec{z}_{g,i} = \overline{\hat{z}_g} + \frac{\widehat{\sigma_g}}{\sigma_g}(z_{g,i} - \overline{z_g}) \qquad \textbf{Equation 3}$$

Following transformation using Equations 2 or 3, SQUID adopts a simplification of DWLS's strategy to deconvolve transformed bulk profiles. SQUID doesn't require signature gene selections, and instead uses all genes with nonzero expression in both the transformed bulk and scnRNA-Seq profiles. The objective function is identical to the one employed by OLS (Equation 1), however, here, the SQUID process seeks to identify $\tilde{p}_{ij} \in \tilde{P}$ that minimizes the discrepancy between transformed bulk RNA-Seq profiles $\vec{Z}$ and the abundance-weighted sums of the expression profiles of composing cell types X. Consequently, following the iterative process proposed by Tsoucas et al., SQUID minimizes this dampened weighted discrepancy until convergence is reached at iteration l, so that $\left\|\tilde{P}^{(l)} - \tilde{P}^{(l-1)}\right\| \leq 0.01$.[14]

A five-step approach to determine the number of clusters in scRNA-Seq data

For those datasets for which no metadata was available, we performed the following five-step iterative process.

(1) Use Monocle3 clustering (see Figure S1A), which does an internal log transformation and library size normalization, to assign initial labels to all cells in each scRNA-Seq dataset.
(2) Compute a mean expression profile per cluster using log-transformed and library-size normalized data from Monocle3.
(3) Compute all pairwise Pearson correlations across the mean expression profiles.
(4) Combine all non-overlapping cluster pairs with the highest correlation where Pearson correlation r>= 0.95 (see Figure S1B).
(5) Modify the clustering information inside the metadata file (that we labeled as "phenoDataC").

Cell mixture construction

Tissue culture. MCF7 cells were purchased from the Tissue Culture Core at Baylor College of Medicine. BT474, T47D, and THP1 cells were purchased from ATCC; Jurkat (J32) cells were a gift from Dr. Andras Heczey; hMSC cells were purchased from Lonza (PT-2501). MCF7 cells were cultured in DMEM with 10% FBS; T47D cells were cultured in RPMI with 10% FBS; BT474 cells were cultured in DMEM with 10% FBS and 15µg/ml insulin; Thp1 and Jurkat cells were cultured in RPMI with 10% FBS and 1% L-glutamine; hMSCs were cultured using the MSCGM BulletKit from Lonza. All cell lines were cultured with 1% penicillin-streptomycin (ThermoFisher Scientific) and maintained at 37ºC in a humidified incubator with 5% CO_2. All cell lines were confirmed to be free of mycoplasma contamination by DNA staining with Hoechst (ThermoFisher Scientific) or Syto 82 (ThermoFisher Scientific). DMEM, RPMI, and L-glutamine were purchased from ThermoFisher Scientific, and FBS and bovine insulin were purchased from Sigma.

Mixture assay. Adherent cells were harvested in a proliferative state. Cells were washed in PBS, trypsinized, collected, and resuspended in HBSS (ThermoFisher Scientific) with 10% FBS. Suspension cells were collected during the log growth phase and resuspended in HBSS with 10% FBS. All cells were maintained on ice after harvest and counted on a Countess II FL (Life Technologies). Viability was high for all cell lines, and the average of three counts was used to calculate cell concentrations. Per mixture, 16K cells were submitted for scRNA-Seq, 500K cells were prepared for bulk RNA sequencing, and 50K cells were prepared for flow cytometry.

Bulk RNA isolation and sequencing. Cell pellets of approximately 500K cells were prepared in triplicates and flash frozen at the time of the experiment. RNA was extracted using the Qiagen RNeasy Plus mini kit with a genomic DNA elimination column (74104). RNA quality was confirmed based on RIN, and 150bp paired-end mRNA libraries were prepared by Novogene (Sacramento, California, USA), who also sequenced libraries at a depth of 20M reads per sample on the NovaSeq 6000 platform (Illumina).

Single-cell RNA library preparation and sequencing. Single-cell samples were submitted to the Baylor College of Medicine Single Cell Genomics Core immediately after preparation. Per sample, 16K cells were loaded, with an expected return of 10K cells. Single-cell gene expression libraries were prepared according to the Chromium NextGEM Single Cell Gene Expression 3v3.1 kit (10x Genomics). Briefly, cells, reverse transcription reagents, gel beads containing barcoded oligonucleotides, and oil were loaded on a Chromium controller (10x Genomics) to generate single-cell GEMs (Gel Bead-In-Emulsion). Full-length cDNA was synthesized and barcoded within each GEM. Subsequently, GEMs were broken, and cDNA was pooled. Following cleanup using Dynabeads MyOne Silane Beads, cDNA was amplified by PCR. The amplified product was fragmented prior to end-repair, A-tailing, and adaptor ligation. Final libraries were generated by amplification. Sequencing of single-cell libraries was performed by the Genomics and RNA Profiling Core at Baylor College of Medicine. To reach an estimated 20K reads per cell, samples were sequenced at a depth of 200M reads on the NovaSeq 6000 platform (Illumina).

Flow cytometry. Immediately after cell collection, a portion of each cell suspension was stained with Hoechst (10µM in HBSS) or Syto 82 (5µM in HBSS) for 10 minutes at 37ºC. After staining, cells were washed and resuspended in HBSS containing 10% FBS. Stained cells were counted twice, and staining efficiency was assessed using Countess II FL. Staining efficiency was nearly 100% in all cell lines, and viability remained high. Count averages were used to calculate the number of cells added to each mixture, and 50K cells were targeted for each flow sample. All flow samples were prepared and analyzed in triplicate. For each of the six mixtures (1–6), three flow samples (1A–1C, 2A–2C, etc.) containing one Hoechst-stained cell line (either T47D, BT474, or MCF7), one Syto 82-stained cell line (either Jurkat, THP1, or hMSC), and four unstained cell lines at identical proportions were generated. Single-stained cells from these samples represented the proportion of that cell line in the corresponding mixture. This strategy (Figure S7) was developed to avoid spectral overlap and to increase our ability to accurately quantify positive cells. For each cell line, unstained and single-stained samples were used as controls to set voltages and define positive and negative gates. Flow cytometry was performed on a FACSymphony (BD Biosciences). Forward and side scatter areas were compared to select cells and exclude debris. Then, forward scatter height and area were compared to select single cells and exclude doublets. Single cells were sub-gated using positive and negative cut-offs for Hoechst (405nm laser, BV421 channel) and Syto 82 (561nm laser, PE channel). Gates were set independently for each cell line due to large differences in cell sizes and to maximize the number of single-stained cells. Once

set, these gates were applied universally to all mixtures. Comparison of BV421 and PE areas demonstrated few double-positive cells and three distinct populations: unstained cells, BV421-positive cells, and PE-positive cells. Data were exported and analyzed using FlowJo v10.8.0 (BD Biosciences). Average flow proportions were compared to expected cell counts and showed a high correlation.

Pediatric AML and neuroblastoma profiling

Paired diagnosis-relapse samples from 6 Pediatric AML patients that were enrolled in AAML1031 were profiled by CITE-Seq, including scRNA-Seq (Immunai), labeling RNAs with a 10x Geonomics Chromium controller and sequencing with Illumina Novaseq 600. In total, we profiled a total of 15,857 genes in 27,687 cells, with an average of 4,644 UMIs and 1,432 gene features per cell (RNA only). Cells with mitochondrial gene content above 10% and fewer than 500 UMIs were excluded. AML samples were treated with RNAlater and profiled using Illumina Novaseq 600 with 25M reads per sample. Similarly, patients in the NB1 dataset were profiled by bulk RNA-Seq with Illumina Novaseq 600 with 25M reads per sample.

RESULTS

To quantify absolute deconvolution performance, we established a framework based on concurrent bulk RNA-Seq and scRNA-Seq or snRNA-Seq data across different human and murine tissues. In parallel, we evaluated the impact of RNA-Seq and scnRNA-Seq data normalization strategies on deconvolution performance (Figure 1). While concurrent RNA-Seq and scnRNA-Seq assays can be used to evaluate deconvolution accuracy, they lack controls for both true composition and cell-type expression estimates. Namely, divergent estimates from the two assays cannot be resolved, and technical analysis errors may not be identified due to missing information. Consequently, accurate and fully resolved deconvolution-strategy evaluations require fully characterized datasets, where the expression profiles and composition of each cell type are known with high degrees of accuracy. To accomplish this, we developed a solid tumor model that includes multiple solid-tumor cell types, immune cells, and lower-abundance stem cells. We then generated and concurrently profiled cell mixtures that conform to this model by flow cytometry, RNA-Seq, and scRNA-Seq. Here, we present the results of our efforts to evaluate deconvolution methods on cell mixtures and tissue samples and evaluate whether improved deconvolution accuracy can benefit its potential applications in diagnosing cancer patients.

Cell mixtures characterization

We established six *in vitro* cell mixtures that are composed of varying proportions of cells from 3 breast cancer lines (T47D, BT474, MCF7), monocytes (Thp1), lymphocytes (Jurkat), and stem cells (hMSC). Mixture composition was recorded based on input cell counts. Cells from each cell line, and cells from each mixture were profiled by bulk RNA-Seq in triplicates. Cell mixtures were profiled by flow cytometry in triplicates to independently evaluate their composition (Figure 2A, Supplemental Table 1). The proportions of breast cancer cell lines varied across mixtures, with some mixtures composed of predominantly one cell type (e.g., 66% of Mixture 1 were T47D cells) and others having a balanced composition (e.g., Mixture 4). Monocytes and lymphocytes accounted for 15% of the mixtures, and hMSC abundances varied from 0.5% to 2% (Figure 2B, Supplemental Table S1). See Methods for detailed experimental descriptions.

UMAP analysis of mixture scRNA-Seq profiles verified the existence of 6 clusters with biomarkers that correspond to their 6 composing cell types (Figure 2C, Supplemental Table S2). We confirmed breast-cancer cell type identities by integrating 7 scRNA-Seq profiles of breast-cancer cell samples[21] including T47D, BT474, MCF7, and 4 cell lines that were not used in our mixtures (BT483, AU565, HCC70, DU4475); see Figure 2D. Cellular composition estimates based on absolute cell counts that were determined during mixture assembly showed high correlations with composition estimates by flow cytometry and scRNA-Seq: r=0.97 and r=0.96 respectively, Figures 2E and 2F. However, the correlation between estimates by flow cytometry and scRNA-Seq clusters was significantly lower (r=0.92, $p<0.05$ by Fisher's transformation). This suggested that composition estimates by cell counts are the most accurate, and flow cytometry and scRNA-Seq introduce independent errors to composition estimates. Overall, however, these results confirmed the mixture composition as estimated by cell counting and demonstrated that it is reflected in scRNA-Seq data with good accuracy. Note that the accuracy of scRNA-Seq-derived expression estimates of individual cell types was not as good as the corresponding mixture composition estimates. Specifically, Pearson correlation of the profiles of the predicted T47D, BT474, and MCF7 cells and their respective bulk RNA-Seq profiles were r=0.53, r=0.53, and r=0.55, respectively; Jurkat and Thp1 had Pearson correlations of r=0.66 and r=0.63, respectively; hMSCs, which were the least abundant cells in each mixture, were correlated at r=0.16 with their bulk profiles. Moreover, restricting comparisons to the top expressed genes did not improve these correlations (Table S3).

Cell mixtures reveal differences in deconvolution accuracy

To evaluate the effects of expression-estimate inaccuracies on the quality of deconvolution, we tested the accuracy of OLS in predicting mixture composition from its bulk profiles and using either scRNA-Seq or bulk-derived expression profile estimates for each cell type. Our results suggested that OLS can estimate mixture composition with high accuracy when input expression profile estimates are accurate. Namely, using bulk RNA-Seq profiles of each cell type, OLS composition predictions had Pearson correlations of r=0.95 with mixture composition estimates by cell counts (Figure 2G). However, when using scRNA-Seq-based expression estimates for each cell type, this correlation declined to r=0.78 (Figure 2H). Note that the correlation r=0.78 is significant at $p<1E-5$, suggesting that, overall, OLS can predict composition in our mixtures using scRNA-Seq-based expression estimates. However, its deconvolution accuracy using scRNA-Seq-based expression estimates was significantly lower than when using bulk RNA-Seq-based expression estimates. As expected, hMSC composition estimates were the least accurate (Figure 2H).

Having confirmed the quality and validity of the *in vitro* cell mixtures and associated data, we used our benchmarking framework to evaluate deconvolution methods on these data (Figure 3A). We observed substantial differences in performance (i.e., predicted abundances versus gold standard) between methods, with DWLS outperforming the other 5 methods, irrespective of the bulk RNA-Seq and scRNA-Seq normalization strategy. Overall, normalization of the bulk RNA-Seq data with TPM resulted in better performance compared to TMM, LogNormalize, or when no normalization was applied. Normalization of the scRNA-Seq-derived reference matrix had a lower impact on deconvolution accuracy. All methods performed poorly in predicting the abundance of hMSC cells (Figure 3B). All methods also underestimated the fraction of Jurkat cells in several mixtures, but this was most pronounced for CIBERSORT and NNLS. In addition, MuSiC

underestimated the fraction of THP1 cells. Together, these observations demonstrated that, in an ideal setting, with concordant scRNA-Seq and bulk RNA-Seq, deconvolution with DWLS leads to the most accurate cell-type abundance estimates.

Variable deconvolution accuracy across human tissues

We studied 7 human and murine tissue datasets with concurrent RNA-Seq and scnRNA-Seq profiles. DWLS outperformed the other methods in 6/7 datasets with higher Pearson correlation coefficients and lower RMSE (Figures 4A and 4B). The absolute performance of all methods was very high in the remaining dataset (Brain, see Figures 4A and 4B). Despite DWLS outperforming the other methods, its absolute performance differed substantially across datasets. DWLS performance was high in the fresh kidney, AML, NB1, and brain datasets, but lower in the NB2, breast cancer, and synapse datasets, with average Pearson correlation coefficients above 0.67 and below 0.4, respectively. The choice of data normalization method impacted deconvolution performance in a subset of datasets, but the impact on performance was typically modest, and none of the normalization methods consistently performed better or worse across datasets.

Single-cell storage procedures impact deconvolution accuracy

Procedures to store single-cell suspensions are known to alter cell type abundance estimates by scRNA-Seq[7]. To evaluate the impact of cell storage procedures on deconvolution accuracy we compared deconvolution performance on two datasets—mouse kidney and human breast cancer—with concurrent bulk profiles and technical replicate scRNA-Seq profiles of single-cell suspensions derived from alternative tissue preservation methods. The kidney dataset included scRNA-Seq profiles of methanol fixed, cryopreserved, and fresh tissues[22], and the breast cancer dataset included profiles of fresh and cryopreserved tissues[23,24]. We applied deconvolution on the matching bulk RNA-Seq data using DWLS and FARDEEP—these methods performed relatively well in our tests—and compared predicted cell type abundances to the gold standard in each of the scRNA-Seq datasets. Both DWLS and FARDEEP showed good performance when comparing observed cell type abundances to those in the fresh and methanol-fixed kidney tissues but both performed poorly when compared to the gold standard for cryopreserved scRNA-Seq dataset (Figure 3C). We note that while the overall deconvolution accuracy for the breast cancer dataset was lower than that of the kidney dataset, there remained a significant difference in performance between fresh suspensions and cryopreserved suspensions (Figure S2). Because of the variability in deconvolution accuracy, we included a comparison of all deconvolution and normalization methods for the kidney dataset (Figures 4C and S5). We note that while the normalization choice had a relatively small impact on deconvolution accuracy for the kidney dataset, normalization had a strong effect on the performance of each tested method for the breast cancer dataset (Figures 4A and S2).

Transformation of bulk RNA-Seq data with SQUID improves deconvolution accuracy

DWLS consistently outperformed other deconvolution methods in our tests. However, its accuracy was poor in several datasets, limiting its potential applications. Note that lower accuracy may be due to method-independent factors, including physically different cellular compositions between scnRNA-Seq and bulk RNA-Seq samples, and technical differences in sample processing that results in diverging estimates. Most importantly, deconvolution accuracy is dependent on accurate gene expression estimates, and—as is the case for our cell mixtures—scnRNA-Seq-derived gene

expression profiles may be imprecise. Indeed, we showed that OLS-based deconvolution using bulk RNA-Seq profiles of each cell type (Figure 2G) produced more accurate results than deconvolution using scRNA-Seq-derived profiles (Figure 2H) on our cell mixtures. Similarly, deconvolution with DWLS using bulk RNA-Seq profiles of each cell type was in excellent agreement with mixture composition as estimated by cell counts (Figure 5A), and its performance declined when using scRNA-Seq-derived profiles (Figure 5B). We note that the same was observed when estimating mixture abundances using either scRNA-Seq analysis or flow cytometry. However, in all cases, deconvolution with DWLS was more accurate than OLS. Relative to cell-count estimated mixture abundances, DWLS and OLS predictions had r=0.98 and RMSE=0.04 vs. r=0.95 and RMSE=0.06 when using bulk RNA-Seq profiles, and r=0.93 and RMSE=0.08 vs. r=0.78 and RMSE=0.12 when using scRNA-Seq-derived profiles, respectively. Based on these observations, we attempted to further improve DWLS performance by transforming bulk RNA-Seq profiles to scRNA-Seq vector spaces. This approach, which we coined 'SQUID', employed linear bulk RNA-Seq transformation followed by dampened weighted least squares and further improved deconvolution accuracy (r=0.95 and RMSE=0.06, Figure 5C).

To systematically test the benefit of bulk transformation and deconvolution with SQUID, we compared the performance of SQUID, DWLS, and OLS for our cell mixtures, as well as for pediatric AML, NB1, NB2, Synapse (ROSMAP brain), breast cancer, and kidney datasets using a leave-one-out cross-validation strategy. Namely, iteratively, concurrent RNA-Seq and scnRNA-Seq profiles of all but one of the samples were used to predict the composition of the remaining sample based on its bulk RNA-Seq profile (Figure 5D). Our results suggested consistently and significantly improved prediction accuracies with SQUID. Comparisons of SQUID accuracy with the other methods, including DWLS, CIBERSORT, FARDEEP, RLR, NNLS, and MuSiC, without cross validation—analogous to Figure 4 comparisons—are given in Figures S2-6.

Deconvolution of pediatric AML and neuroblastoma dataset with SQUID

To assess the utility of deconvolution on bulk RNA-Seq of clinical samples we focused on profiles of pediatric AML and neuroblastoma samples. Large-scale clinical and bulk RNA-Seq profiles are available for both these tumor types from the TARGET consortium, including the profiles of 181 pediatric AML[25] and 161 neuroblastoma[26] patient samples. We profiled paired diagnostic pre-treatment and relapse samples for 6 AML patients using concurrent RNA-Seq and scRNA-Seq assays, and we profiled the expression of 14 neuroblastoma samples using bulk RNA-Seq; note that we previously reported on the scRNA-Seq profiles of the 14 neuroblastomas[27] and used it here to evaluate deconvolution accuracy (the NB1 dataset). Pre-treatment AML samples were expected to be enriched for chemosensitive cancer cells, while relapse AML samples were expected to be enriched for chemoresistant cancer cells[28]. We used predicted cell types and expression profiles from these scRNA-Seq data to deconvolve RNA-Seq profiles of TARGET AML and neuroblastoma diagnostic samples.

Paired diagnostic-relapse pediatric AML samples were collected to identify chemoresistant tumor subclones. After integration and clustering (Figure 6A), we sought to identify AML subclones (clusters) that are present before treatment and expand at relapse. We found one AML cluster that included diagnostic and relapse cells from at least half of the patients and expanded at relapse. We refer to this subclone as the AML expanding subclone, or AML-X for short (Figure

6B). We used SQUID, DWLS, CIBERSORTx, and OLS to predict the composition of TARGET AML samples from chemotherapy trials AAML03P1 (40 patients), AAML0531 (171 patients), and CCG-2961 (24 patients). We then used the predicted abundance of AML-X cells in each diagnostic sample to predict patient outcomes by survival analysis. Note that AAML03P1 and AAML0531 patients were treated with a variety of chemotherapy and CD33-inhibitor combinations, and the earlier CCG-2961 patients were treated by combinations of chemotherapy and anthracyclines. Abundance estimates by DWLS, CIBERSORTx, and OLS were not predictive of outcomes, however, cell-type abundance estimates by SQUID suggested that diagnostic samples whose composition included at least 5% AML-X cells had significantly worse outcomes (p=1.90E-3, Kaplan–Meier estimator, Figure 6C). SQUID composition estimates were also the only ones that were predictive of survival by Cox regression (p=6E-4, compared to p=0.07, p=0.41, and p=0.68 using DWLS, CIBERSORTx, and OLS, respectively). Note AML-X cells accounted for ~10% of the AML cells in the three scRNA-Seq-profiled diagnostic samples with AML-X cells; this composition estimate was consistent with SQUID estimates in TARGET diagnostic samples after accounting for tumor purity. The most upregulated genes in AML-X were MALAT1, NEAT1, and ZEB2 (Figure 6D). The long non-coding RNAs NEAT1 and MALAT1 co-localize in Chr11Q13.1, are co-expressed in pediatric AML, and are predicted to transcriptionally co-inhibit hundreds of genes[29,30]. Their common targets were significantly downregulated in AML-X (Figure 6E). Moreover, NEAT1 has been previously implicated with chemoresistance in cancer[31], and both NEAT1 and MALAT1 have been associated with poor prognosis in childhood leukemia[32]. In addition, MALAT1 has been shown to post-transcriptionally upregulate ZEB2 in cancer[33,34], and ZEB2 was the third most upregulated gene in AML-X.

To identify neuroblastoma cell clusters that are associated with outcomes, we integrated scRNA-Seq data across the 14 neuroblastoma samples from the NB1 dataset and identified 15 cell clusters (Figure 6F). Each cluster was tested for patient outcomes prediction based on abundance estimates by SQUID, DWLS, CIBERSORTx, and OLS using TARGET RNA-Seq data. The target dataset includes profiles of 161 samples from 69 clinical trials where patients were treated by combinations of a variety of chemotherapies and other therapies including GD2 and thymidylate-synthase inhibitors. In total, two cell clusters were identified to be predictive of outcomes using SQUID abundance estimates (Figure 6G, Cluster NB-s1 at p=1.4E-3 and Cluster NB-s2 at p=1.0E-2, Kaplan–Meier estimator). No cluster was predictive of outcomes using estimates from other survival methods: DWLS, CIBERSORTx, and OLS abundance estimates for NB-s1 were predictive of survival at p=0.23, p=0.95, and p=0.99, and for NB-s2 at p=0.34, 0.93 and p=0.99, respectively. Notably, among the top 5 most upregulated genes in cluster NB-s1 were HRAS, SEMA3D, and H3F3B (Figure 6H). RAS pathway mutations have previously been identified in relapsed neuroblastomas[35]. More recently, upregulation of H3F3B has been associated with the alternative lengthening of telomeres (ALT) phenotype in neuroblastoma, which is associated with poor outcomes[36]. Moreover, tumors harboring RAS pathway mutations in combination with telomere maintenance mechanisms were shown to have extremely poor survival rates[37]. In cluster NB-s2, we observed the upregulation of 6 members of the semaphorin family, including SEMA3D (Figure 6I). SEMA3D upregulation has been documented in metastatic neuroblastomas and was shown to affect neuroblastoma cell migration[38].

DISCUSSION

Profiling technologies at single-cell resolutions are enabling efforts to characterize the cellular composition of complex tissues. Single-cell resolution RNA profiling technologies, including 10X Genomics platforms, are used to characterize the transcriptomes of individual cells, which, in turn, can be used to identify these cell types in past and future assays. Consequently, ongoing large-scale efforts, including The Human Cell Atlas[39], single-cell tumor immune atlas[40], and single-cell Atlas in Liver Cancer[41], are mapping out healthy and disease tissues and characterizing the transcriptomes of their composing cell types. These efforts are building resources that promise to improve our understanding of intercellular dependencies between healthy and diseased cells and to enable comparisons of tissues at high resolutions. Single-cell atlases promise to help interpret future single-cell assays and help maximize knowledge gained from RNA-Seq profiles. RNA-Seq profiles remain by far the most frequently used type of molecular data collected in the biological and health sciences, and they account for more publicly available molecular datasets than any other data type. Because of the technical and financial challenges associated with scnRNA-Seq, RNA-Seq is likely to remain the most frequently used assay for the foreseeable future. Consequently, computational deconvolution of bulk transcriptomes could serve as an alternative for scRNA-Seq to enumerate cell types in complex tissues, including tumor biopsies.

Deconvolution methods that use scnRNA-Seq profiles to predict the composition of bulk-profiled samples are expected to play major roles in analyses based on single-cell atlases. However, their absolute accuracy remains unstudied, and their potential users face multiple unaddressed challenges. First and foremost, current deconvolution methods are heuristics that always produce composition estimates irrespective of accuracy. Most methods do not provide accuracy evaluations and efforts to evaluate accuracy will have limited success without assay-specific quality controls, which are not always available. Other challenges include the lack of guidance for choosing technical parameters in data analysis, including the choice of methods and parameters for data normalization, data harmonization, and clustering. These choices dictate the accuracy of scnRNA-Seq analysis and its use for deconvolution. In summary, the deconvolution of RNA-Seq profiles based on scnRNA-Seq data will benefit from reliable accuracy evaluation and guidance for selecting analysis parameters and methods.

Here, we produced comparative performance analyses of deconvolution methods based on constructed cell mixtures with known cell abundances and expression profiles, and based on concurrent scnRNA-Seq and bulk RNA-Seq data across a variety of tissue types. Our analyses of cell-mixtures samples suggested that current scRNA-Seq assays using the 10x Genomics platform can produce excellent sample-composition estimates, but these assays may produce relatively poor transcriptome characterizations for each identified cell type and particularly for rare cell types. Moreover, our results suggested that scRNA-Seq assays tend to under-sample adherent cells when non-adherent cells are present. We showed that when given accurate cell-type expression profiles, direct approaches like OLS for predicting cell-type abundances from bulk profiles produced excellent results (Figure 2G). However, deconvolution using scnRNA-Seq-derived profiles using the same approach produced poor cell-type abundance estimates (Figure 2H). We note that other deconvolution methods, including MuSiC, DWLS, and SQUID had substantially more accurate abundance estimates than OLS.

Our results suggested that accurate evaluations and performance of scnRNA-Seq-based deconvolution methods for any given context will greatly benefit from the collection of concurrent scnRNA-Seq and bulk RNA-Seq data. Namely, bulk RNA-Seq profiles allowed us to produce upper bounds on the accuracy of deconvolution methods that rely on the corresponding scnRNA-Seq assays, and the integration of concurrent bulk RNA-Seq in the deconvolution process with SQUID improved deconvolution accuracy for all datasets. In addition, we observed substantial and consistent performance differences that were associated with library preparation methods, as well as analysis and deconvolution methods. Namely, comparisons of related datasets—e.g., our two neuroblastoma datasets—suggested that datasets with few scnRNA-Seq profiles lead to worse deconvolution accuracy. We note that while the choice of scnRNA-Seq normalization methods influenced deconvolution methods performance in some datasets, the best choices varied across datasets and deconvolution methods. For example, while LogNormalize led to good performance for most deconvolution methods in our cell-mixture scRNA-Seq data, it was associated with reduced DWLS accuracy (Figure 3A). Overall, TPM normalization produced some of the most consistent results. However, the resolution of scnRNA-Seq clustering had a greater influence on deconvolution success. Namely, high clustering resolutions could lead to reduced deconvolution accuracy when multiple cell clusters share the same cell type and have highly similar transcriptomes, while low cluster resolutions could lead to heterogenous cell clusters that are not associated with specific cell types. In both cases, cell-type specific deconvolution marker genes were difficult to identify and had limited cell-type selectivity. To resolve this, we opted to either merge clusters with similar transcriptomes, or select resolutions to optimize the accuracy of OLS deconvolution of concurrent RNA-Seq profiles. Both approaches lead to dramatic improvements in deconvolution accuracy for all methods.

We developed a deconvolution strategy with substantially improved accuracy using concurrent scnRNA-Seq and bulk RNA-Seq profiles. Jew et al. suggested that the transformation of bulk RNA-Seq profiles to scnRNA-Seq space could improve the accuracy of RNA-Seq deconvolution, and their proposed method Bisque[20] combined RNA-Seq transformation and NNLS to predict the composition of RNA-Seq profiled samples based on scnRNA-Seq profiles. In our tests, Bisque was outperformed by other deconvolution methods, including DWLS, CIBERSORT, and FARDEEP. However, by combining RNA-Seq transformation with the dampened weighted least squares strategy employed by DWLS[14], we were able to dramatically improve deconvolution accuracy. Indeed, our proposed strategy (SQUID) outperformed all other strategies on all datasets with or without cross validation (Figure 5 and Figures S2-6, respectively), and when estimating cell abundances based on IHC, cell counts, flow cytometry, or scnRNA-Seq analyses.

To investigate the effects of tissue preservation methods on deconvolution accuracy, we evaluated deconvolution methods using scRNA-Seq profiles of matched suspensions derived from methanol fixed, cryopreserved, or fresh tissues. We showed that deconvolution based on scRNA-Seq profiles of fresh and methanol-fixed tissues can perform with good accuracy, but performance based on matched cryopreserved samples was markedly worse. These results are in line with observations made by Denisenko et al. that cryopreservation resulted in the loss of proximal tubule (epithelial) cell types, while methanol fixation maintained cellular composition[7]. Consequently, cryopreservation distorted abundance estimates, leading to a poor correlation between the predicted and expected cell type abundances. Interestingly, while not as accurate

as using fresh or methanol-fixed profiles, SQUID predictions based on profiles of cryopreserved samples were dramatically more accurate than other deconvolution methods. This was due, in part, to its employment of RNA-Seq transformation, which transformed bulk profiles to mirror cell-type depletions observed in scRNA-Seq profiles. Thus, while SQUID reduced the discrepancy between the concurrent profiles, it did not fully correct scRNA-Seq inaccuracies. We argued that because scRNA-Seq profiles can include inaccuracies, frameworks to evaluate deconvolution need to include datasets where both expression profiles and cell-type abundances are fully characterized, as in our mixture data.

Finally, we showed that improved deconvolution accuracy may be necessary for enabling its applications in cancer diagnostics. To evaluate this, we produced concurrent RNA-Seq and scRNA-Seq profiles for pediatric AML and neuroblastoma samples and analyzed RNA-Seq profiles and clinical annotations from TARGET-profiled samples. We identified a potentially chemoresistant pediatric AML subclone by comparing scRNA-Seq profiles of matching diagnosis and relapse samples, and we generated subclone characterizations for neuroblastoma. We showed that only SQUID-predicted tumor subclone abundances in diagnostic samples were predictive of patient outcomes. Interestingly, while composition estimates by other methods failed to associate subclone abundances and patient outcomes in these datasets, the significance of outcomes predictions based on abundance estimates by DWLS, CIBERSORTx, and OLS mirrored their estimated accuracy in our benchmark. Namely, DWLS-predicted abundance estimates for our candidate high-risk subclones were the most predictive of outcomes while OLS estimates were the least predictive.

In summary, we identified key prerequisites and provided guidance to produce accurate deconvolution of RNA-Seq profiled tissues based on a scnRNA-Seq dataset. We found that scnRNA-Seq-based composition estimates are often inaccurate for cryopreserved tissues, that expression normalization methods should be selected in context-specific manner, and that cell clustering resolution should be carefully calibrated. Our analyses suggested that, albeit at a marginally higher cost than scnRNA-Seq profiles alone, concurrent RNA-Seq and scnRNA-Seq profiles could be used to optimize normalization and clustering, evaluate the accuracy of deconvolution methods, and improve deconvolution accuracy. Taken together, our results suggested that RNA-Seq deconvolution using scnRNA-Seq data can produce accurate cell-type abundance estimates and that atlases of concurrent RNA-Seq and scnRNA-Seq profiles could be used to reevaluate the compositions of other RNA-Seq datasets.

FIGURES

Figure 1. We benchmarked data normalizations and deconvolution approaches in datasets with concurrent bulk RNA-Seq and scnRNA-Seq profiles (*). Cells were clustered in an unsupervised fashion (**). Gold standard abundance estimates (***) for each cell type were obtained by either aggregating cells or nuclei in each scnRNA-Seq cluster, immunohistochemistry, fluorescence-activated cell sorting, or cell counts. Deconvolution methods used either full scnRNA-Seq expression profiles or cluster-specific biomarkers to predict cell-type abundances based on bulk RNA-Seq profiles. Deconvolution accuracies in each sample were assessed by comparing predicted abundances from bulk RNA-Seq and gold standard estimates.

Figure 2. Cell mixture design, characterization, and analysis by OLS. (**A**) Breast cancer cells (BT474, T47D, and MCF7), leukemia cells (THP1 and Jurkat), and human mesenchymal stem cells (hMSCs) were used to generate six populations of mixed cells (cell mixtures). Each cell line was profiled individually by bulk RNA-Seq in triplicates, and each mixture was profiled by bulk RNA-Seq and flow cytometry in triplicates as well as by a 10x genomics Chromium controller. (**B**) Cell mixtures were composed of varying proportions of cancer cells, with leukemia cells accounting for 15% and hMSC accounting for 0.5% (M1 and M4) to 2% (M3 and M6) of each mixture. (**C**) The clusters derived from scRNA-Seq data corresponded to composing cell types, as identified by cell-type biomarkers. (**D**) The integration of scRNA-Seq profiles of our mixtures (in gray) and scRNA-Seq profiles of BT474, T47D, MCF7, BT483, AU565, HCC70, and DU4475 (Gambardella et. al., 2022) revealed a significant overlap between profiles of BT474, T47D, and MCF7 cells, while negative controls, including BT483, AU565, HCC70, and DU4475, clustered separately. (**E**) Cell counts at the time of mixture generation were significantly correlated with cellular composition estimates by flow cytometry (r=0.97) and (**F**) by scRNA-Seq analysis (r=0.96). However, the correlation between the estimates by flow cytometry and scRNA-Seq was significantly lower (r=0.92, $p<0.05$, Fisher's transformation). (**G**) Ordinary least squares regression (OLS) using bulk RNA-Seq profiles of composing cell types estimated the composition of our mixtures with high accuracy (r=0.95). (**H**) OLS deconvolution abundance estimates using cell-type expression profiles from scRNA-Seq analysis were also accurate (r=0.72, $p<1E-4$) but significantly worse ($p<1E-5$, Fisher's transformation).

Figure 3. The accuracy of cell-mixture deconvolution. (**A**) The impact of RNA-Seq and scRNA-Seq normalization strategies and the choice of deconvolution methods on deconvolution accuracy, as assessed by Pearson correlation and root mean square error (RMSE); darker and larger circles represent higher Pearson and lower RMSE values, respectively. (**B**) Deconvolution results for the normalization strategy with the lowest RMSE; axes are in $\log_{10}$ scales. Each scatterplot contains 36 data points corresponding to 6 cell lines in 6 mixtures, with gold standard abundance estimates based on cell counts and predicted abundances based on deconvolution.

Figure 4. Deconvolution accuracy on concurrently profiled tissues. (**A**) The impact of bulk RNA-Seq and scRNA-Seq normalization strategies on the accuracy of deconvolution methods, as assessed by Pearson correlation and root mean square error (RMSE); darker and larger circles represent stronger correlations and smaller errors, respectively. (**B**) Numerical visualizations of deconvolution accuracies in cell mixtures and tissues across normalization strategies, as assessed by Pearson correlation (top) and RMSE (bottom). (**C**) The effects of profiling cryopreserved, fresh, and methanol-preserved cold-dissociated kidney samples on the accuracies of deconvolution by DWLS and FARDEEP.

Figure 5. Mixture deconvolution with transformed RNA-Seq data. (**A**) DWLS deconvolved the composition of our mixtures with near-perfect accuracy when given the bulk RNA-Seq expression profiles of each cell type (r=0.98, RMSE=0.04), and (**B**) with high accuracy when using cell-type expression estimates from scRNA-cluster profiles (r=0.93, RMSE=0.08). (**C**) SQUID deconvolution accuracy, relative to cell counts, when using cell-type expression estimates from scRNA-cluster profiles (r=0.95, RMSE=0.06) was significantly better than DWLS ($p<2E-4$, Fisher's transformation). (**D**) Deconvolution accuracies of concurrent RNA-Seq and scnRNA-Seq

profiled tissues using SQUID, DWLS, and OLS, as assessed by Pearson correlation and root mean square error (RMSE).

Figure 6. SQUID-predicted cell-type abundances identified outcomes-predictive subclones in pediatric AML and neuroblastoma diagnostic biopsies. (**A**) Clusters identified in scRNA-Seq profiles of paired diagnostic and relapse samples from 6 pediatric AML patients included the cluster AML-X. (**B**) AML-X cells were present in the diagnostic biopsies of 3 patients and their abundance increased at relapse. (**C**) SQUID-predicted AML-X abundances in TARGET-profiled diagnostic AML biopsies were predictive of patient outcomes. (**D**) The cancer lncRNAs MALAT1 and NEAT1 and (**E**) their direct targets were upregulated in AML-X and were predicted to regulate chemoresistance in AML. (**F**) Clusters identified in scRNA-Seq profiles of neuroblastoma samples included NB-s1 and NB-s2. (**G**) SQUID-predicted abundances of both NB-s1 and NB-s2 in TARGET-profiled diagnostic biopsies were predictive of patient outcomes. (**H**) Upregulated genes in NB-s1 included SEMA3D and HRAS, and upregulated genes in NB-s2 (**I**) included SEMA3D and other semaphorin family members.

SUPPLEMENTARY FIGURES

Figure S1. (**A**) UMAP of merged scRNA-Seq profiles of the 6 cell mixtures with 21 clusters predicted using default parameters revealed groups of adjacent clusters. (**B**) The expression profiles of some clusters were significantly correlated, and supervised bi-clustering using a 2000-marker gene set identified cliques of correlated clusters that were also adjacent in the UMAP. Merging 3 cluster cliques, shown as regrouped clusters (top), reduced the total number of clusters to 6. (**C**) Deconvolution using the original 21 clusters, and (**D**) the reduced set of 6 clusters produced markedly different deconvolution accuracy evaluations. Here, the gold standard was estimated from scRNA-Seq analysis for fairness since there is no 1-to-1 mapping between 21 clusters and 6 cell types. Accuracy was evaluated using Pearson correlations (P) and RMSE (E).

Figure S2. Matched breast cancer samples were profiled fresh or after cryopreservation by scRNA-Seq. Confirming our observations from the analysis of kidney samples (Figure 4C), the accuracy of tested deconvolution methods was lower when using profiles of cryopreserved tissues. However, SQUID composition estimates were the most accurate based on fresh profiles and were nearly unaffected by tissue preservation.

Figure S3. Analogous to Figure 4A, deconvolution accuracy on concurrently profiled tissues suggested that DWLS outperforms other published methods irrespective of normalization. However, SQUID estimates were the most accurate on every dataset. Note that accuracy estimates reported in Figure 5 were based on cross-validation test errors, while estimates reported here and in Figure 4 did not use cross validation.

Figure S4. Analogous to Figure 4B, normalization strategies altered deconvolution accuracy estimates, but the top-performing methods outperformed other methods in nearly all tests, irrespective of normalization. In particular, the accuracy of SQUID estimates was nearly unaffected by the normalization strategy used.

Figure S5. Analogous to Figure 4C, deconvolution accuracies were estimated for fresh, methanol-fixed, and cryopreservation after warm and cold dissociation of kidney tissues.

Deconvolution accuracy was lower for profiles of cryopreserved kidney tissues than for profiles of fresh or methanol-preserved tissues. However, SQUID's accuracy was high irrespective of the tissue preservation technique used.

Figure S6. Analogous to Figure 3B, we present a detailed comparison between predicted abundance estimates by deconvolution and gold standard abundance estimates for the normalization strategy with the lowest RMSE for each dataset; axes are in $\log_{10}$ scale.

SUPPLEMENTARY TABLES

Table S1. The composition of each of the 6 cell mixtures, based on cell counts, compared to estimates based on flow cytometry and scRNA-Seq analysis.

Table S2. Cell-type specific biomarkers for cell lines used in the 6 cell mixtures.

Table S3. RNA-Seq and scRNA-Seq estimated expression profiles for cell mixtures and cell lines.

Table S4. Datasets used to evaluate deconvolution accuracy. Annotations include methods for setting the gold-standard composition, quality of deconvolution, sample types, cell or nuclei counts, UMI counts, and the number of samples in each dataset.

Table S5. Predicted abundance and clinical annotations of pediatric AML and neuroblastoma TARGET patients that were used to evaluate the outcomes-predictive value of subclone abundances in diagnostic samples.

ACKNOWLEDGEMENTS

The results published here are in part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments initiative. The work was supported by CPRIT award RP180674, European Union's Horizon 2020 research and innovation programme under grant agreement 826121, NCI award R21CA223140, and Special Research Fund postdoctoral scholarship from Ghent University (BOF21/PDO/007). Cell mixtures were profiled by the BCM Single Cell Genomics Core, which is supported by NIH shared instrument grants S10OD018033, S10OD023469, S10OD025240 and P30EY002520. We thank Elena Denisenko and Alistair Forrest (Harry Perkins Institute of Medical Research, Australia) for providing the necessary information to match bulk and scnRNA-Seq samples from the GSE141115 dataset. Alex Swarbrick, Kate Harvey, Sunny Wu and Dan Roden (Garvan Institute of Medical Research, Australia) for providing information regarding the state of tissue for scRNAseq captures and the matching tissue that was used for bulk RNAseq ("breast" dataset), and Andras Heczey for providing Jurkat (J32) cells.

DATA AVAILABILITY

Bulk RNA-Seq and scRNA-Seq data from patients in the NB1 and NB2 datasets are available from the European Genome Phenome Archive EGA (EGAS00001006723; EGAS00001006823; GSE218450; EGAS00001004349). Pediatric AML and cell mixture scRNA-Seq datasets are available at GEO GSE220608. Patient RNA-Seq data used for this study, phs000467 (neuroblastoma) and phs000465 (AML), are available at TARGET's GDC portal. See Table S4 for details.

CODE AVAILABILITY

All software and scripts used to manufacture the reported analyses, including R implementations of SQUID with and without cross validation, are freely available without restrictions at the Github repository https://github.com/favilaco/deconv_matching_bulk_scnRNA.

REFERENCES

1 Chen, S. *et al.* Intrinsic age-dependent changes and cell-cell contacts regulate nephron progenitor lifespan. *Developmental cell* **35**, 49-62 (2015).

2 Gregorieff, A., Liu, Y., Inanlou, M. R., Khomchuk, Y. & Wrana, J. L. Yap-dependent reprogramming of Lgr5+ stem cells drives intestinal regeneration and cancer. *Nature* **526**, 715-718 (2015).

3 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).

4 Pfister, D. *et al.* NASH limits anti-tumour surveillance in immunotherapy-treated HCC. *Nature* **592**, 450-456 (2021).

5 Jordan, N. V. *et al.* HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature* **537**, 102-106 (2016).

6 Maynard, A. *et al.* Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. *Cell* **182**, 1232-1251. e1222 (2020).

7 Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* **21**, 130, doi:10.1186/s13059-020-02048-6 (2020).

8 Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6** (2017).

9 Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature communications* **8**, 1-12 (2017).

10 Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications* **11**, 1-14 (2020).

11 Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. in *Cancer systems biology* 243-259 (Springer, 2018).

12 Monaco, G. *et al.* RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell reports* **26**, 1627-1640. e1627 (2019).

13 Steen, C. B., Liu, C. L., Alizadeh, A. A. & Newman, A. M. in *Stem Cell Transcriptional Networks* 135-157 (Springer, 2020).

14 Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nature communications* **10**, 1-9 (2019).

15 Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications* **10**, 1-9 (2019).

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

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

18 Ripley, B. *et al.* Package 'mass'. *Cran r* **538**, 113-120 (2013).

19 Mullen, K. M. & Van Stokkum, I. H. (R package version, 2007).

20 Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications* **11**, 1-11 (2020).

21 Gambardella, G. *et al.* A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. *Nature communications* **13**, 1-12 (2022).
22 Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758-763 (2018).
23 Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics* **53**, 1334-1347 (2021).
24 Papanicolaou, M. *et al.* Temporal profiling of the breast tumour microenvironment reveals collagen XII as a driver of metastasis. *Nature communications* **13**, 1-21 (2022).
25 Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nature medicine* **24**, 103-112 (2018).
26 Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* **45**, 279-284, doi:10.1038/ng.2529 (2013).
27 Jansky, S. *et al.* Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma. *Nature genetics* **53**, 683-693 (2021).
28 Rasche, M. *et al.* Successes and challenges in the treatment of pediatric acute myeloid leukemia: a retrospective analysis of the AML-BFM trials from 1987 to 2012. *Leukemia* **32**, 2167-2177 (2018).
29 Chiu, H.-S. *et al.* Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. *Cell reports* **23**, 297-312 (2018).
30 Lorenzi, L. *et al.* The RNA Atlas expands the catalog of human non-coding RNAs. *Nature biotechnology* **39**, 1453-1465 (2021).
31 Adriaens, C. *et al.* p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nature medicine* **22**, 861-868 (2016).
32 Pouyanrad, S., Rahgozar, S. & Ghodousi, E. S. Dysregulation of miR-335-3p, targeted by NEAT1 and MALAT1 long non-coding RNAs, is associated with poor prognosis in childhood acute lymphoblastic leukemia. *Gene* **692**, 35-43 (2019).
33 Xiao, H. *et al.* LncRNA MALAT1 functions as a competing endogenous RNA to regulate ZEB2 expression by sponging miR-200s in clear cell kidney carcinoma. *Oncotarget* **6**, 38005 (2015).
34 Cheng, H. *et al.* Long non-coding RNA MALAT1 upregulates ZEB2 expression to promote malignant progression of glioma by attenuating miR-124. *Molecular Neurobiology* **58**, 1006-1016 (2021).
35 Eleveld, T. F. *et al.* Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nature genetics* **47**, 864-871 (2015).
36 Burkert, M. *et al.* Copy-number dosage regulates telomere maintenance and disease-associated pathways in neuroblastoma. *bioRxiv* (2022).
37 Ackermann, S. *et al.* A mechanistic classification of clinical phenotypes in neuroblastoma. *Science* **362**, 1165-1170 (2018).
38 Delloye-Bourgeois, C. *et al.* Microenvironment-driven shift of cohesion/detachment balance within tumors induces a switch toward metastasis in neuroblastoma. *Cancer Cell* **32**, 427-443. e428 (2017).
39 Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451-453 (2017).
40 Nieto, P. *et al.* A single-cell tumor immune atlas for precision oncology. *Genome research* **31**, 1913-1926 (2021).
41 Ma, L. *et al.* Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *Journal of hepatology* **75**, 1397-1408 (2021).

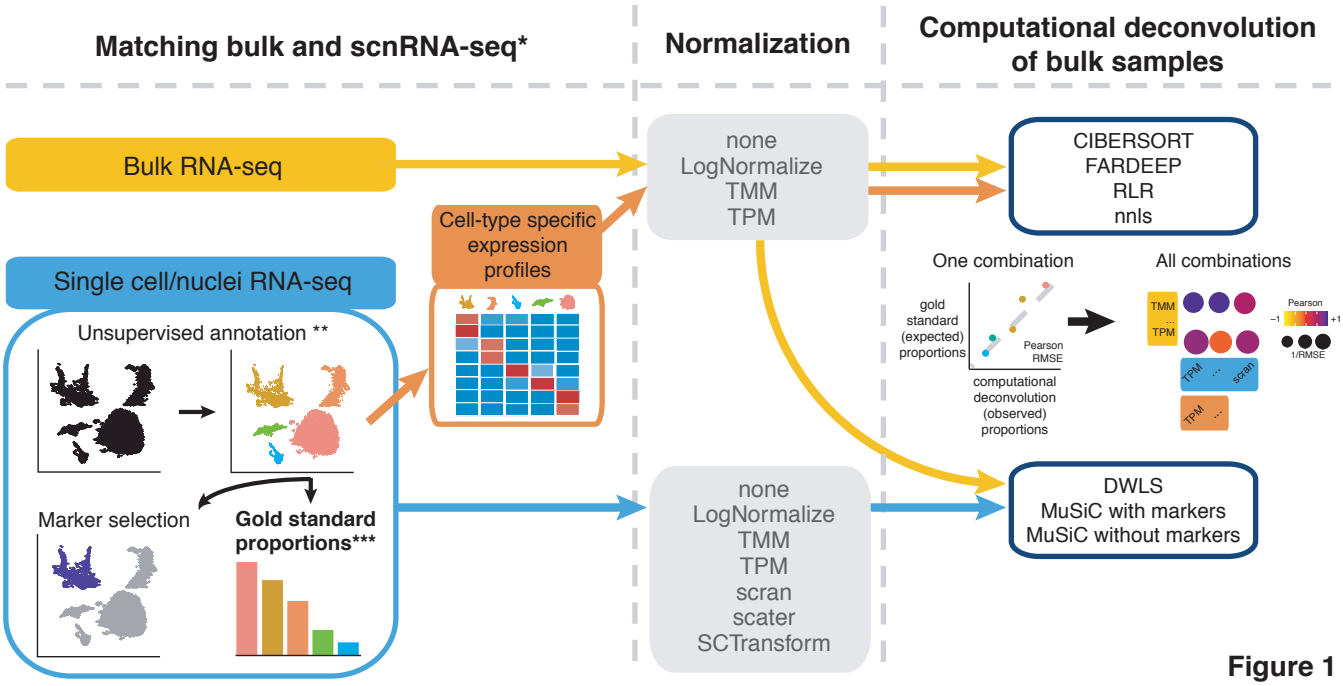

Figure 1

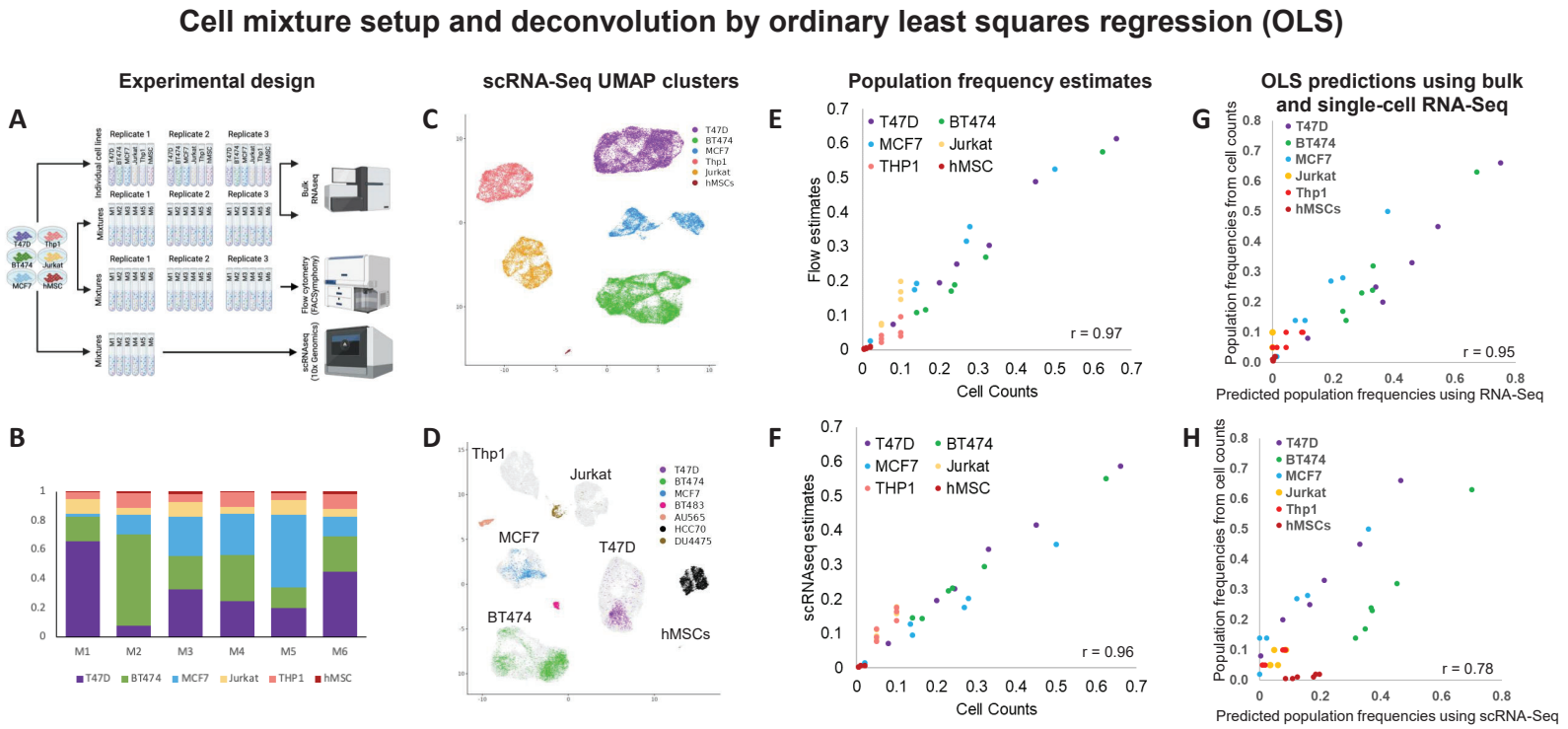

Figure 2

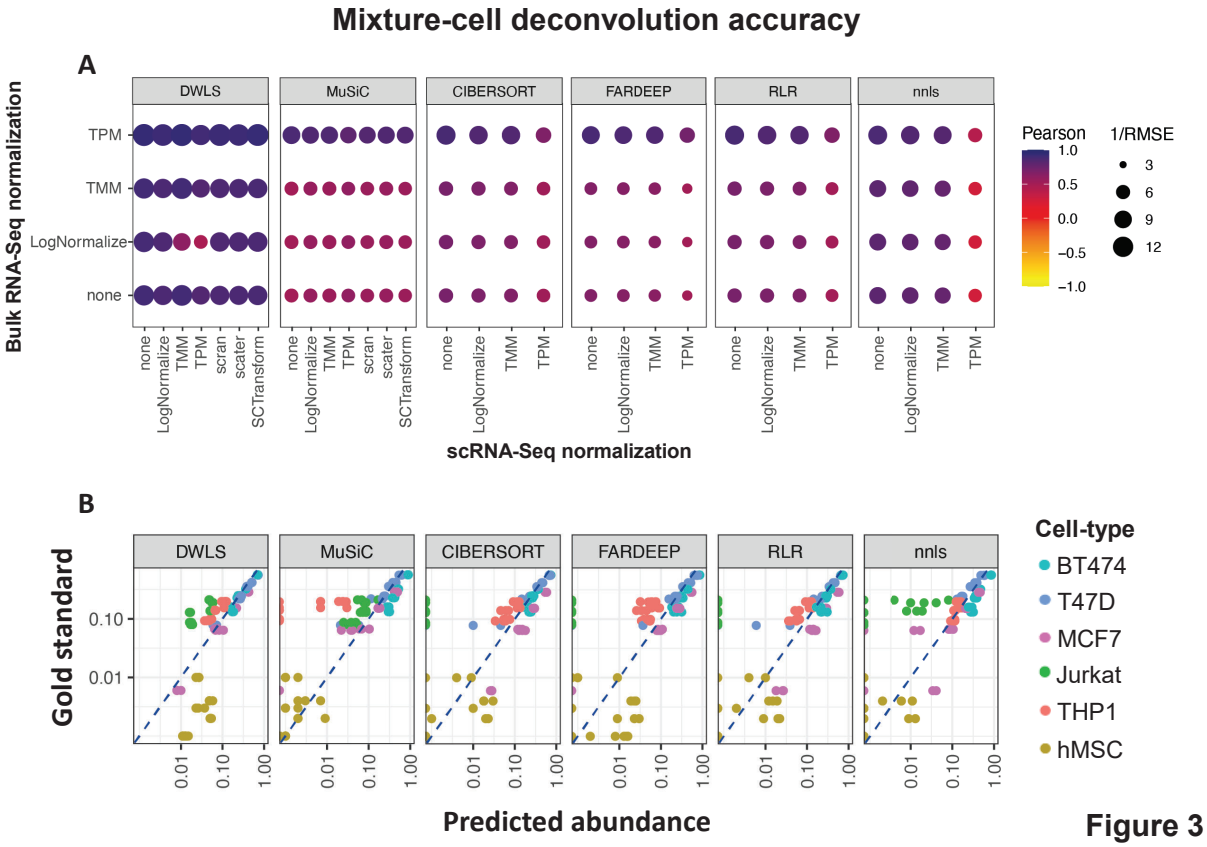

Figure 3

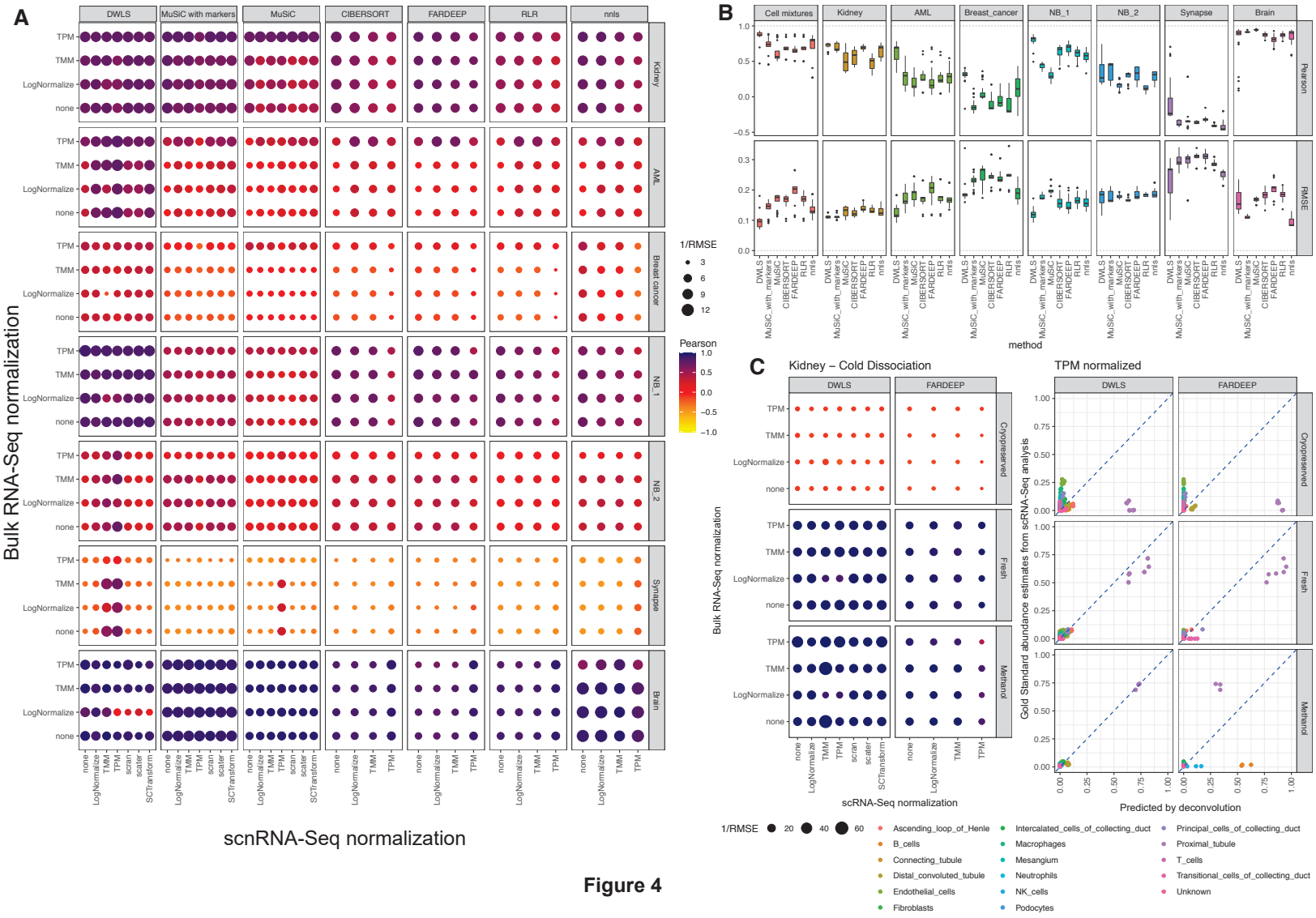

Figure 4

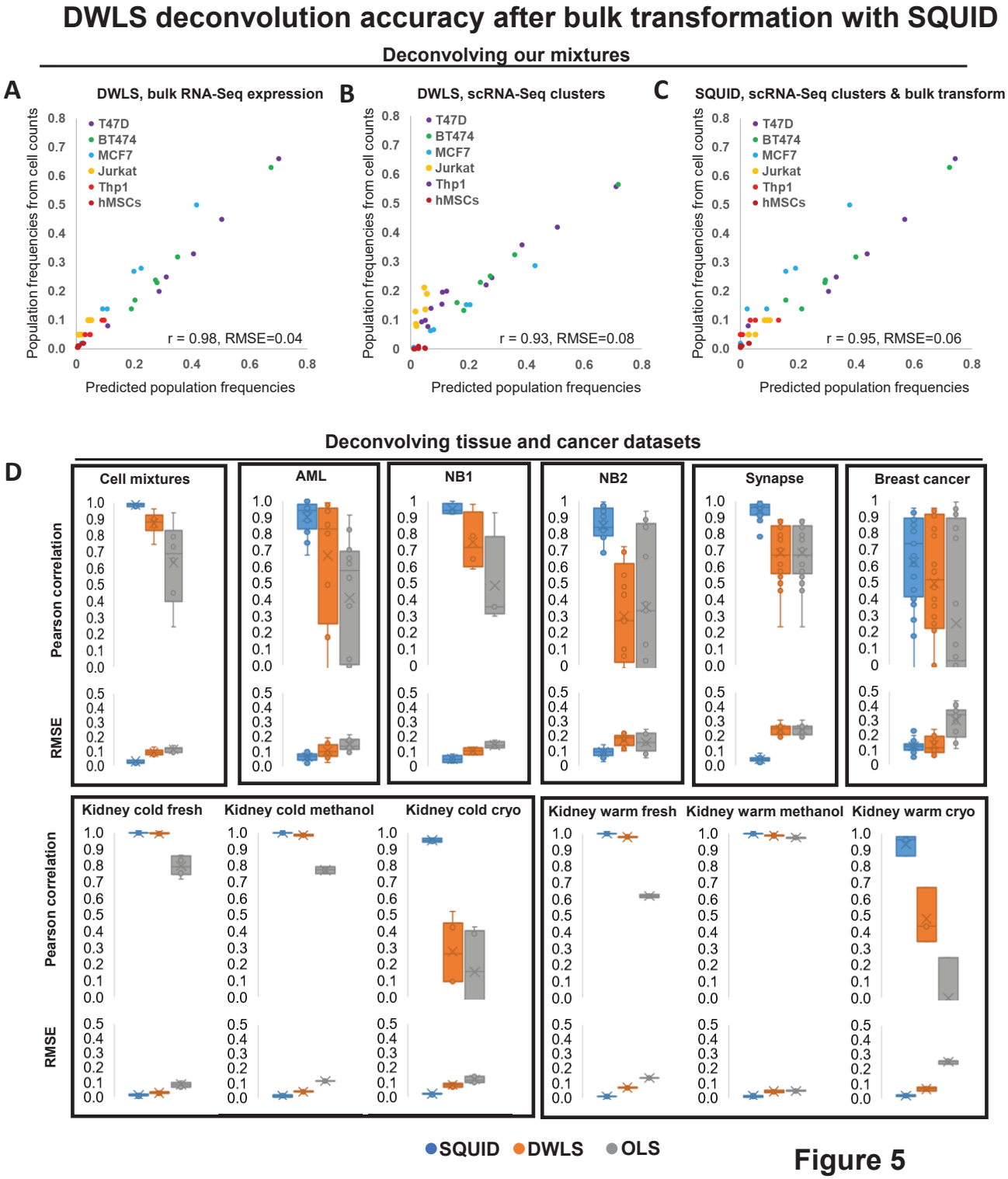

Figure 5

Deconvolution of large-scale pediatric AML and NB RNA-Seq datasets

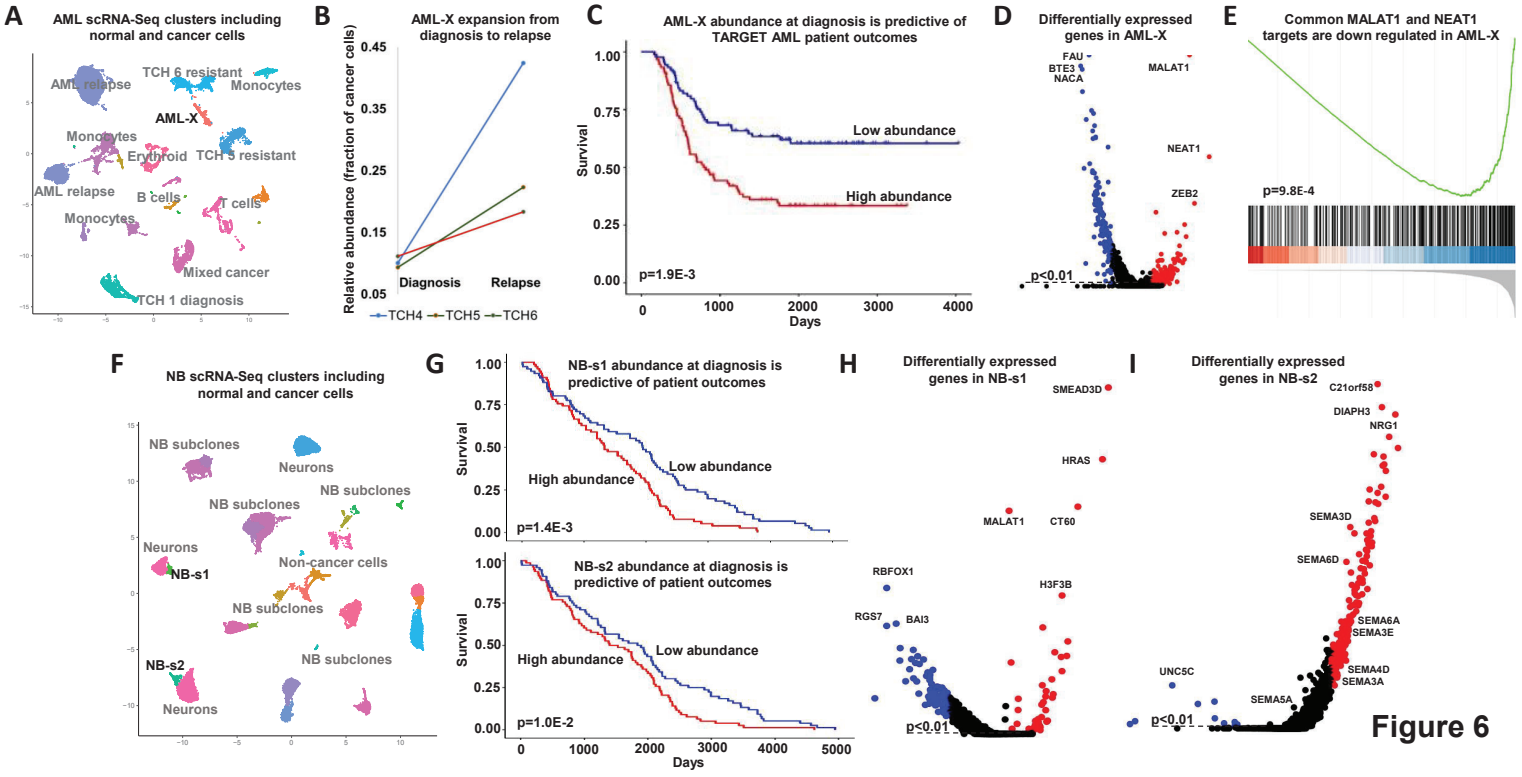

Figure 6

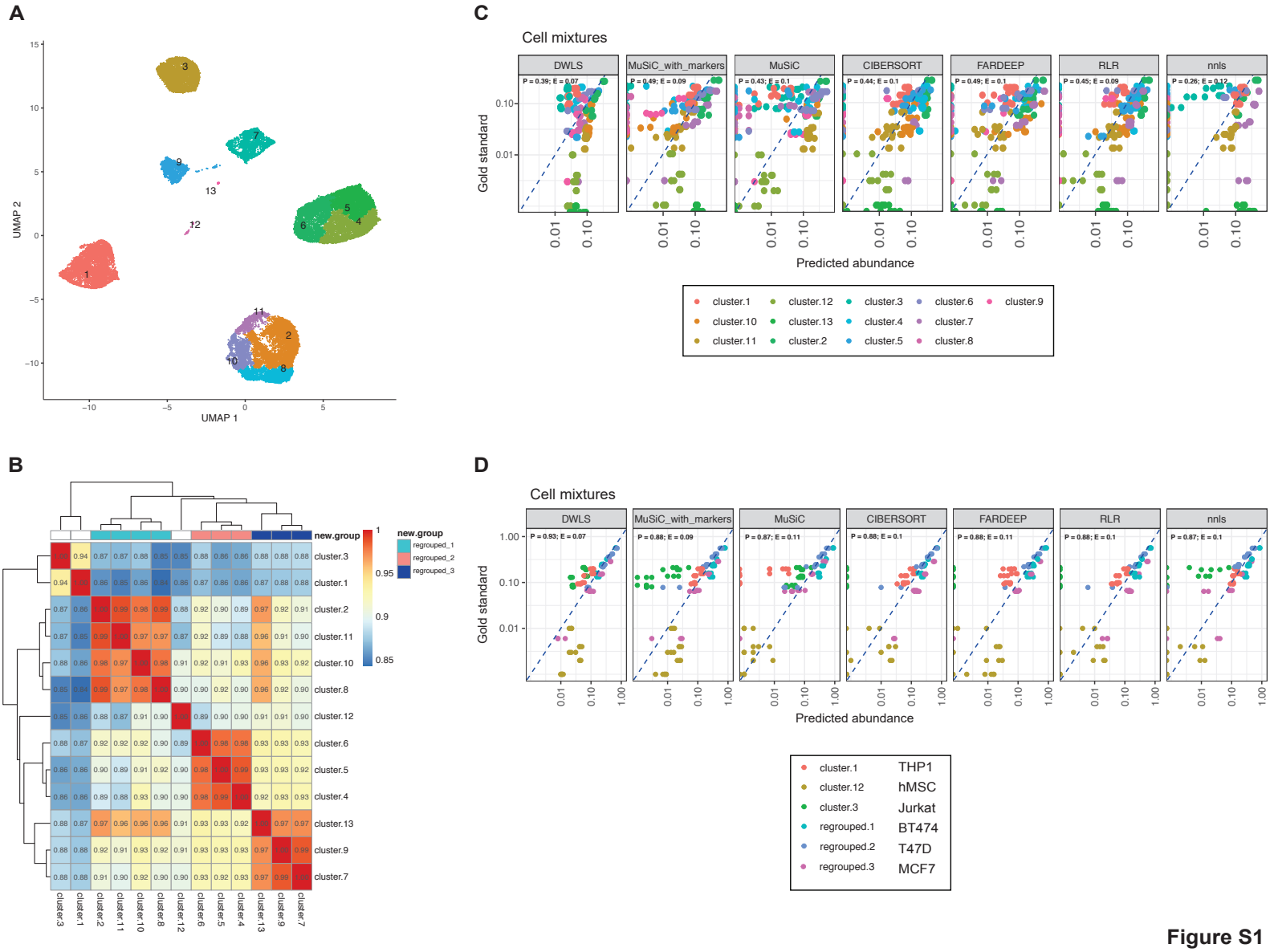

Figure S1

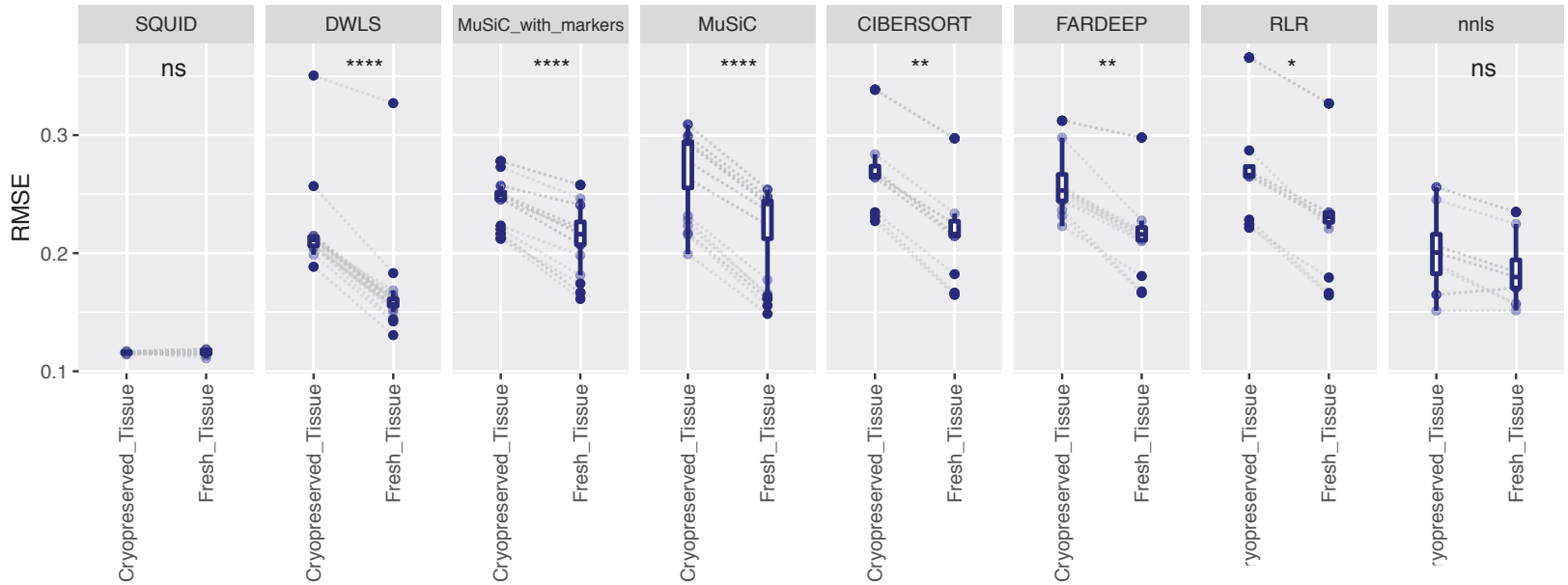

Figure S2

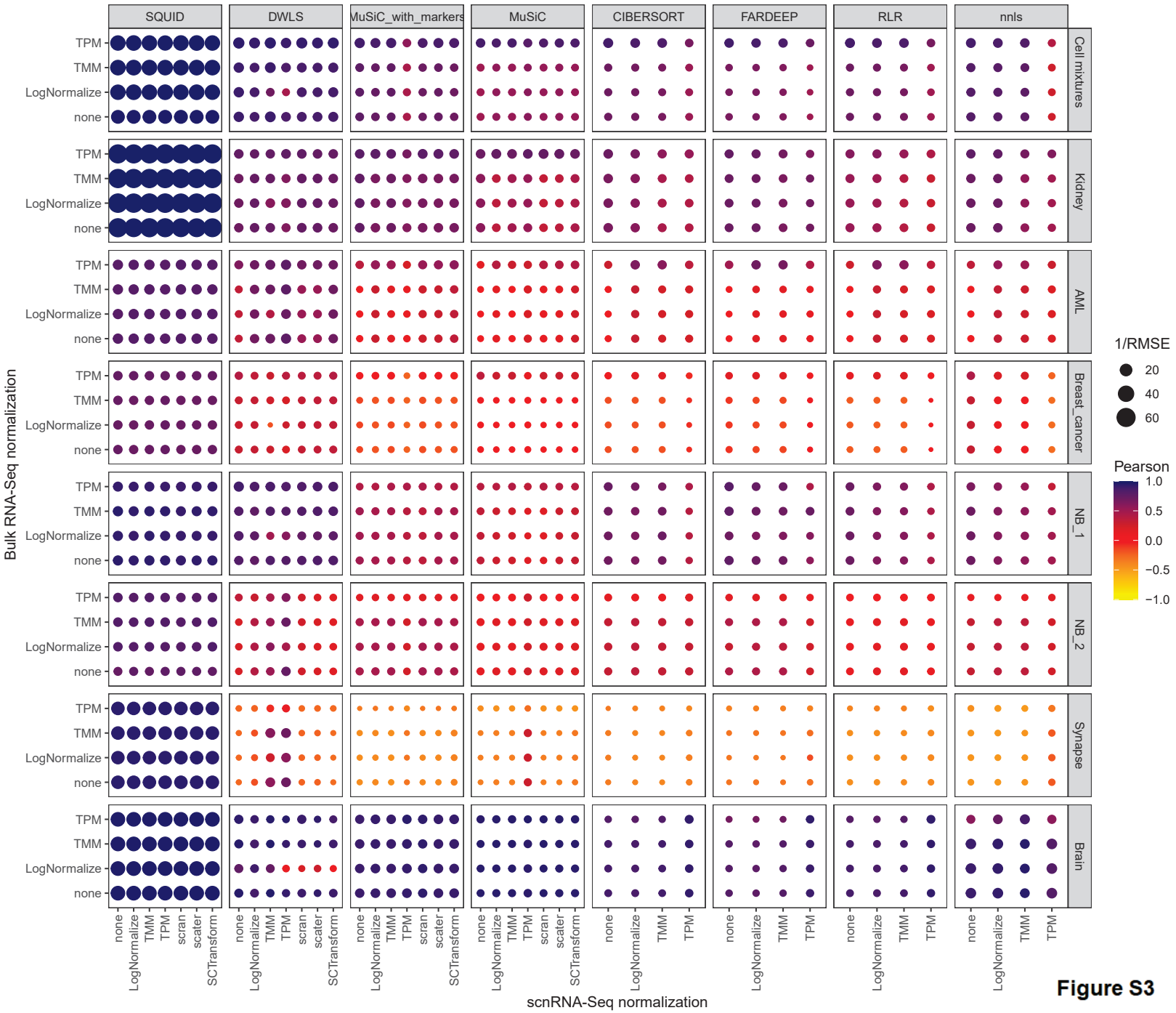

Figure S3

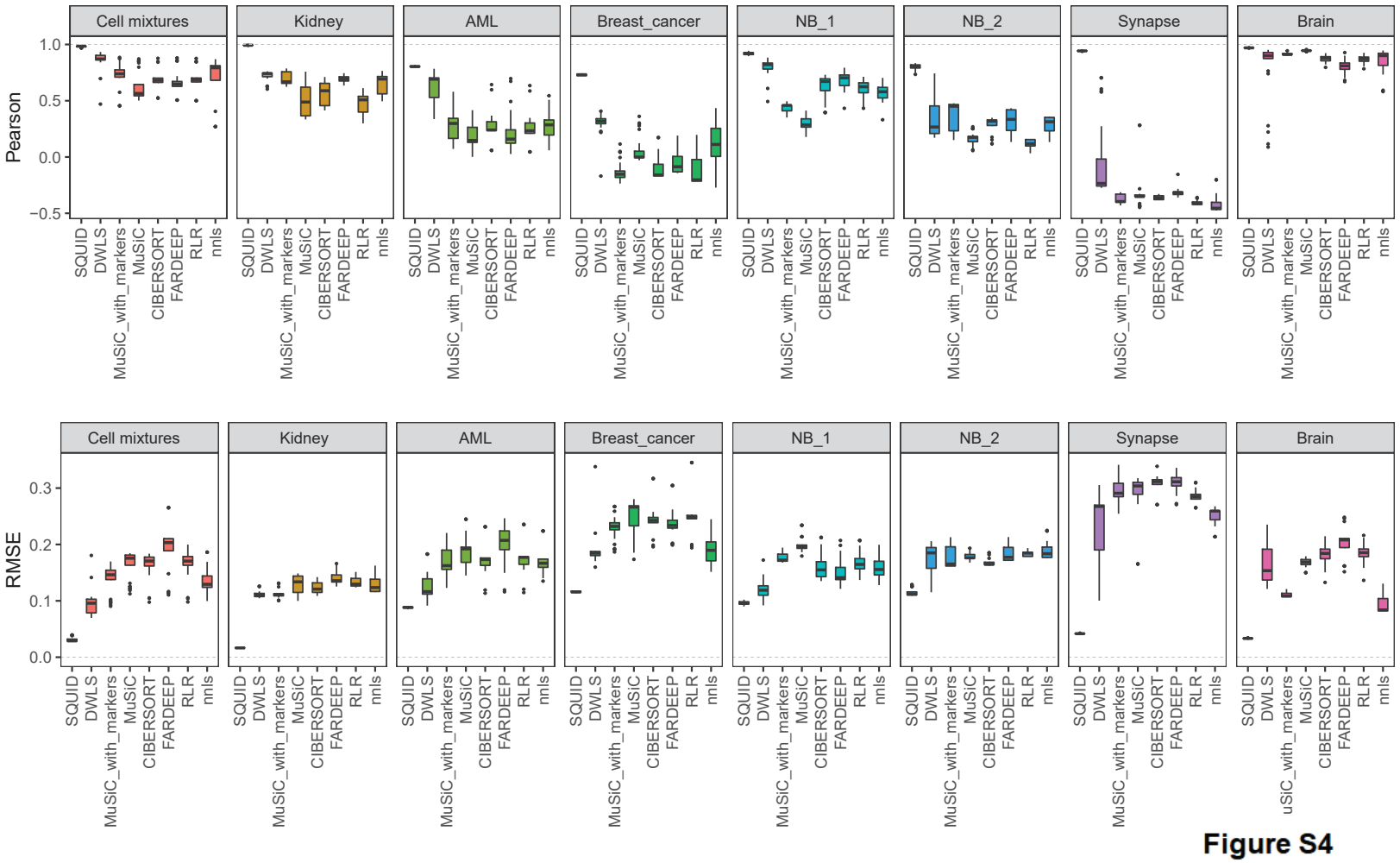

Figure S4

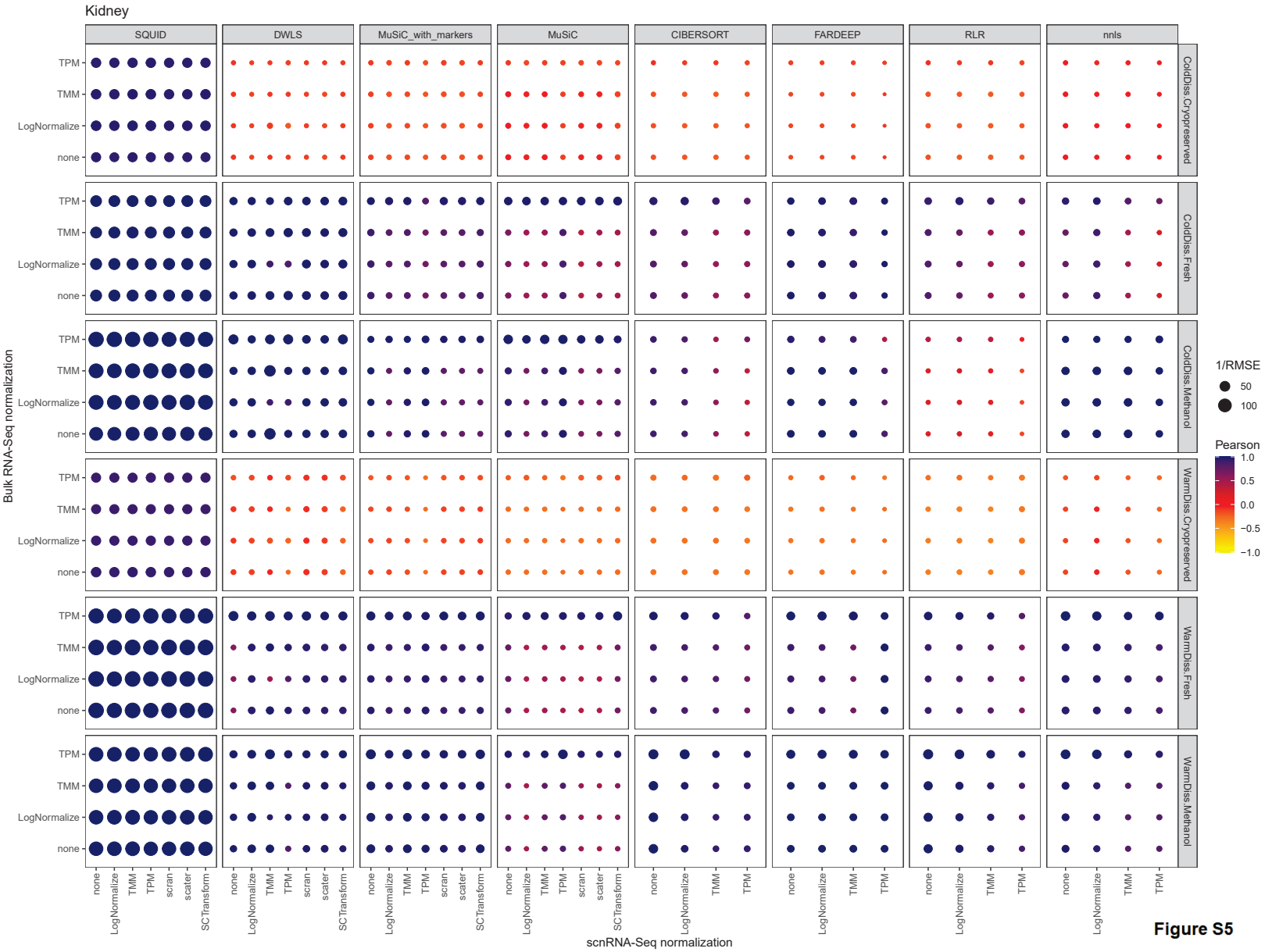

Figure S5

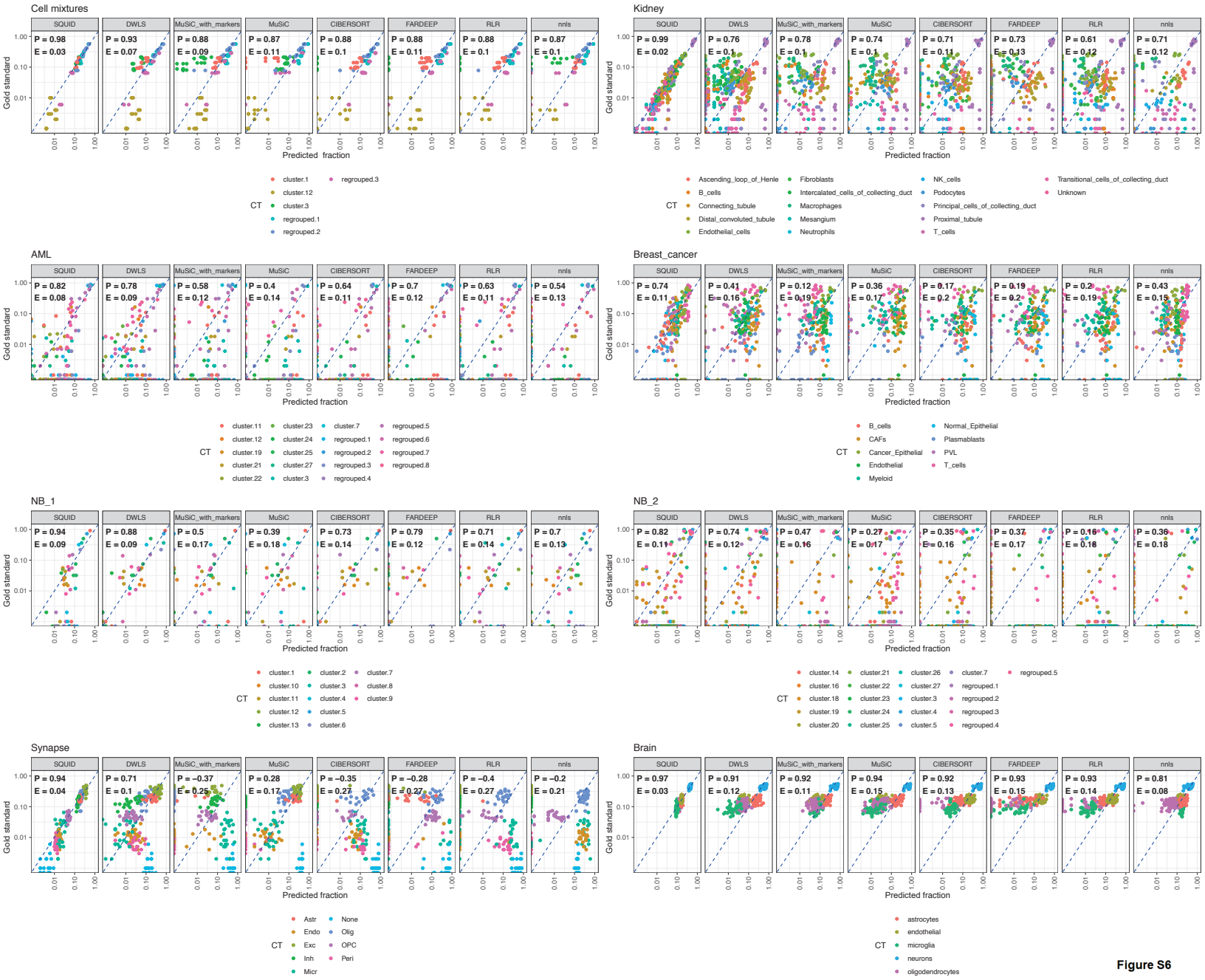

Figure S6

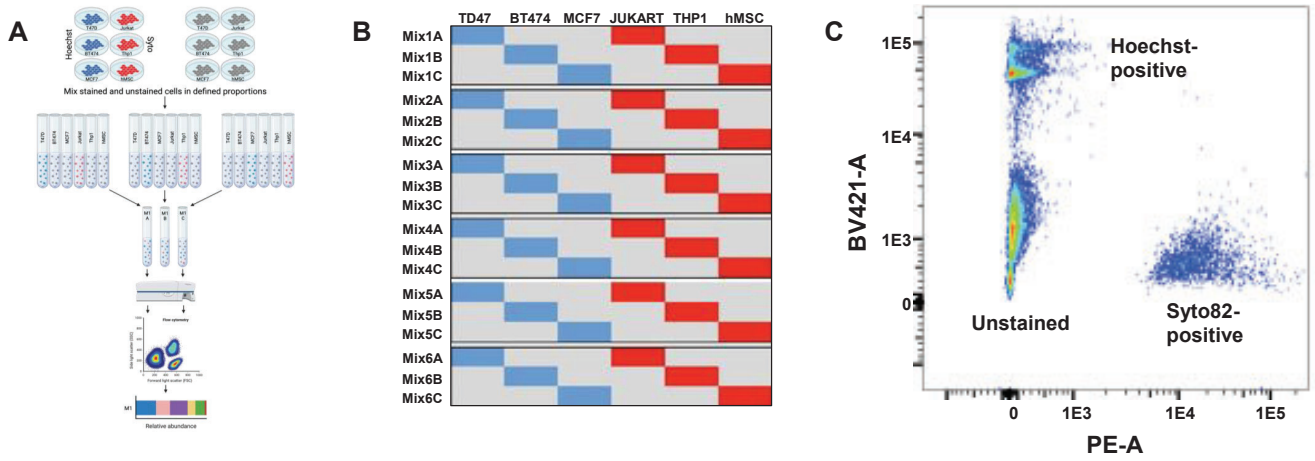

Figure S7