nature biotechnology

SUPPLEMENTARY INFORMATION

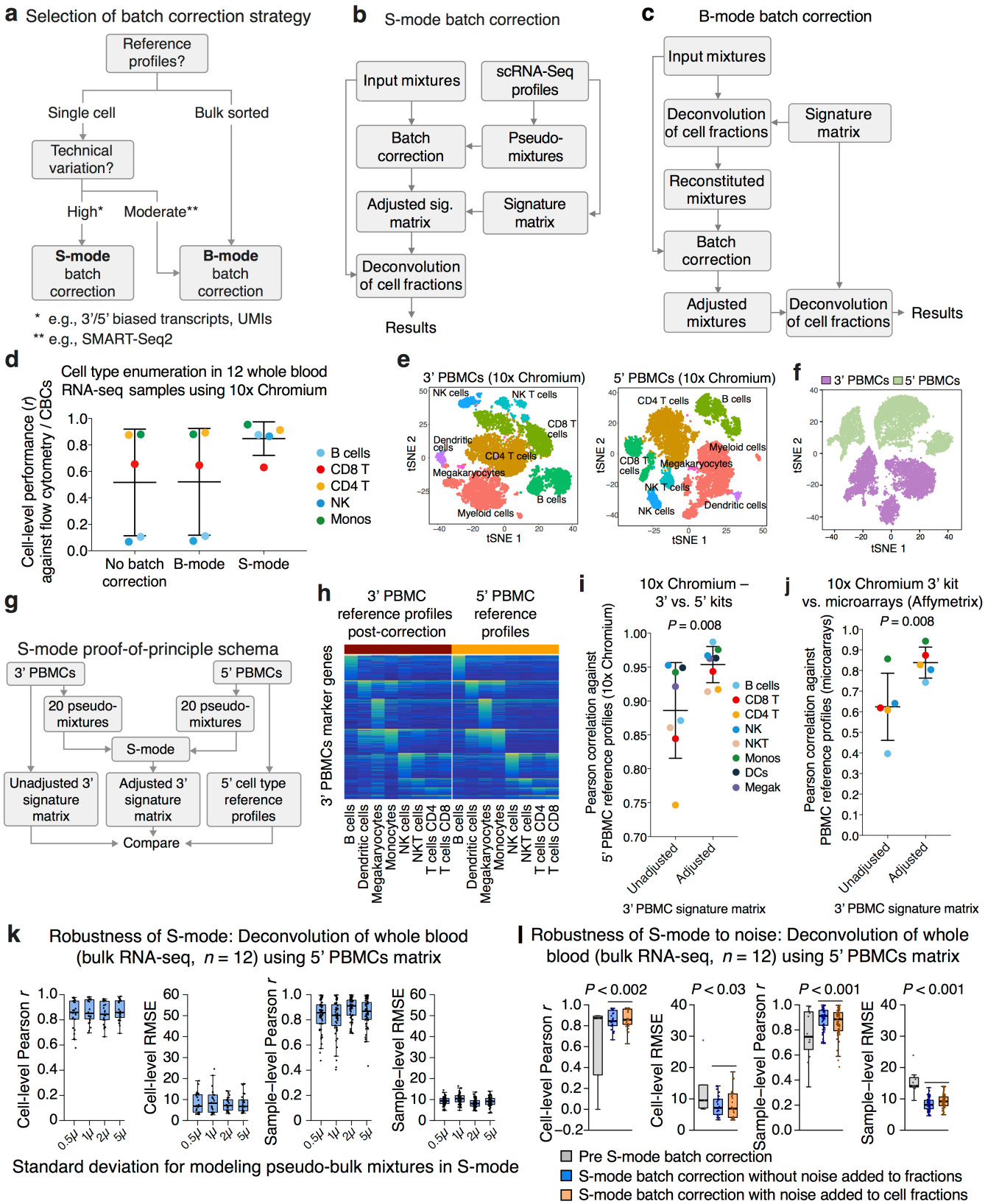
https://doi.org/10.1038/s41587-019-0114-2

In the format provided by the authors and unedited.

Determining cell type abundance and expression from bulk tissues with digital cytometry

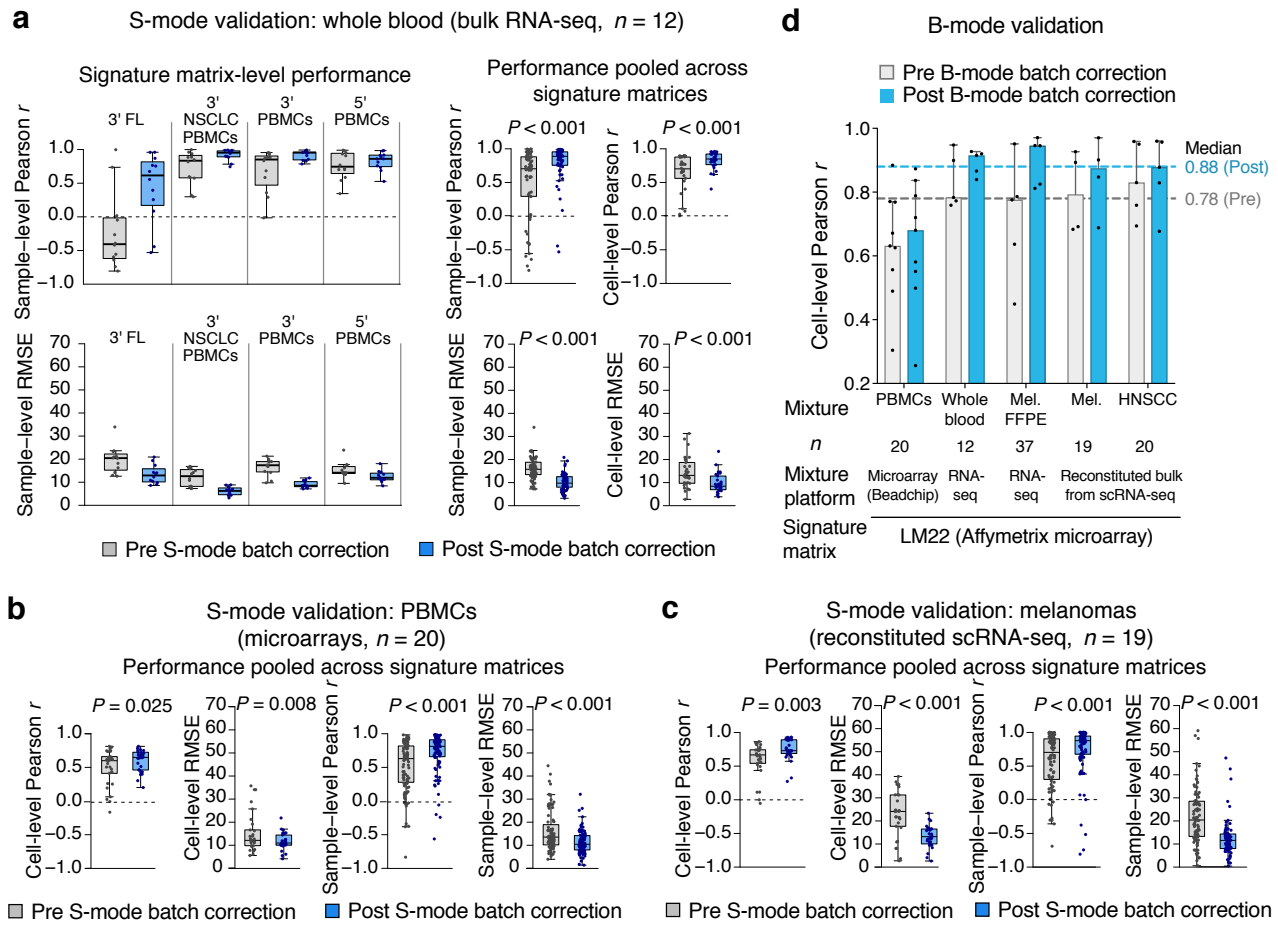
Aaron M. Newman[1,2*], Chloé B. Steen[3,4], Chih Long Liu[1,3], Andrew J. Gentles[2,3,5,6], Aadel A. Chaudhuri[7,8], Florian Scherer[3,9], Michael S. Khodadoust[3], Mohammad S. Esfahani[3,5,8], Bogdan A. Luca[6], David Steiner[3], Maximilian Diehn[1,6,8] and Ash A. Alizadeh[1,3,5,8,9*]

[1]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [3]Division of Oncology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA, USA. [4]Department of Informatics, University of Oslo, Oslo, Norway. [5]Center for Cancer Systems Biology, Stanford University, Stanford, CA, USA. [6]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA, USA. [7]Department of Radiation Oncology, Stanford University, Stanford, CA, USA. [8]Stanford Cancer Institute, Stanford University, Stanford, CA, USA. [9]Division of Hematology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA, USA. *e-mail: amnewman@stanford.edu; arasha@stanford.edu

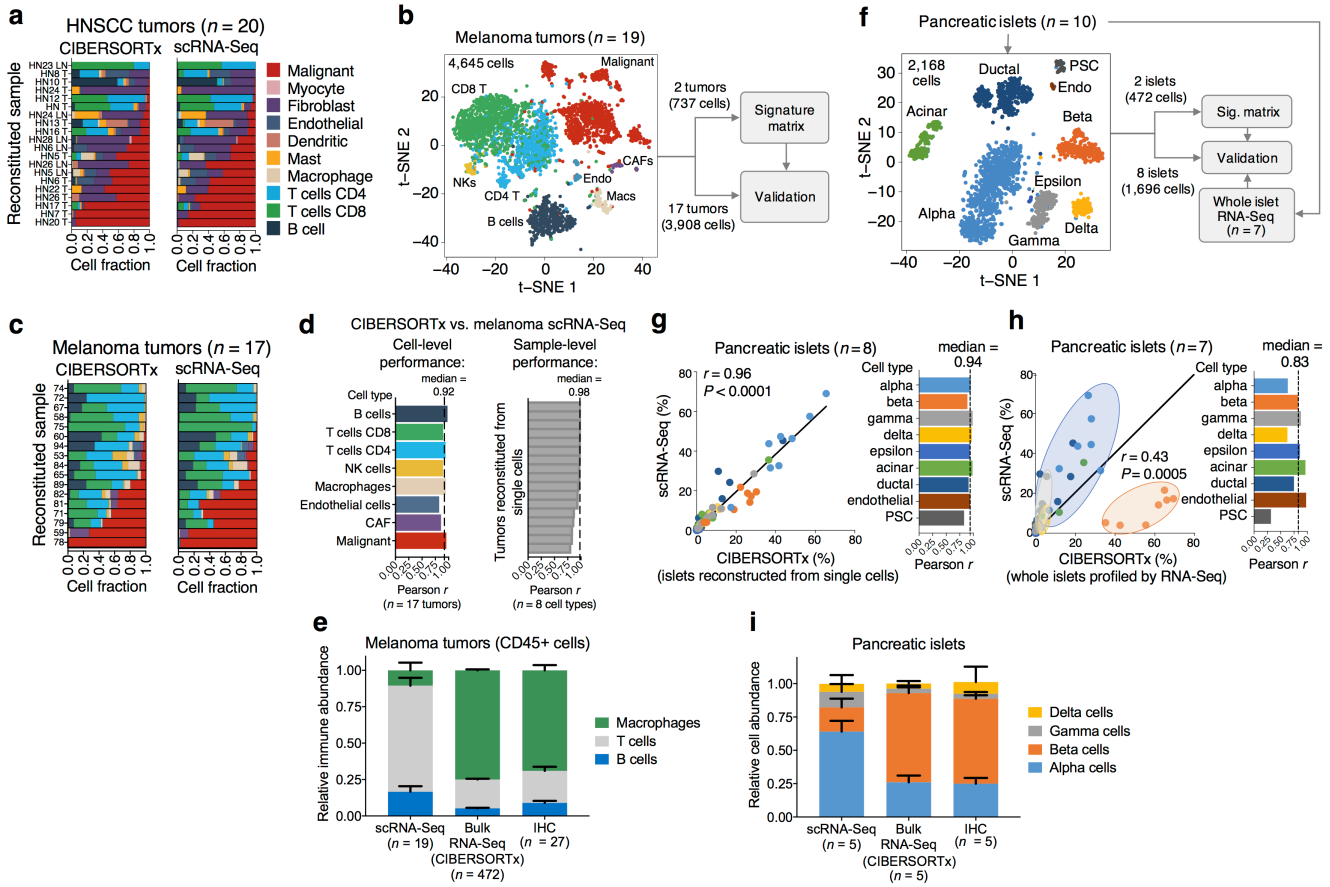


Supplementary Figure 1: Overview and robustness of cross-platform deconvolution. (**a**) Decision tree for selecting the most appropriate CIBERSORTx batch correction strategy for a given input signature matrix: (i) single-cell reference mode ('S-mode') or (ii) bulk reference mode ('B-mode'). (**b-c**) Schematics for S-mode (**b**) and B-mode (**c**) batch correction. (**d**) Analysis of deconvolution

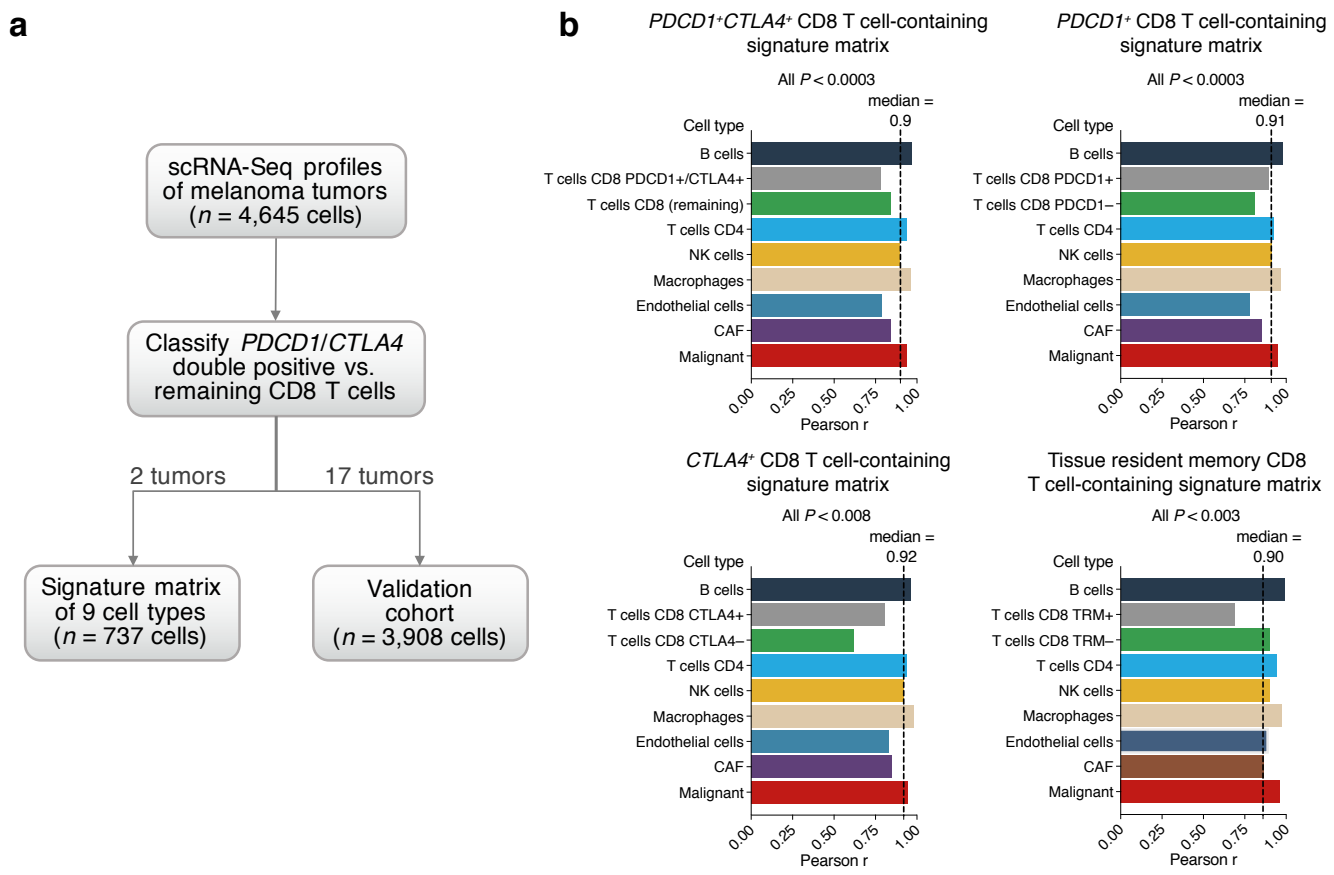
performance using the signature matrix and mixture dataset from Fig. 2b, but comparing the relative utility of B-mode and S-mode for overcoming technical variation between 10x Chromium v2 and bulk RNA-seq (*n* = 12 blood samples). (**e**) t-SNE projections of two publicly available single-cell transcriptome datasets of PBMCs profiled by Chromium v2 using 3' and 5' kits (*n* = 8,370 and 7,688 cells, respectively). Cell labels were determined manually using Seurat and canonical marker gene assessment, as described in **Methods**, and used to generate two signature matrices, 3' PBMCs and 5' PBMCs, respectively (**Supplementary Table 2**). (**f**) t-SNE plot showing platform-specific variation between the datasets from panel e (*n* = 16,058 cells). (**g-j**) Evaluation of S-mode for adjusting cell-type reference profiles derived from single-cell transcriptome data. (**g**) Schematic of the experimental approach, using datasets in panel e to illustrate the application of S-mode to minimize technical variation between reference profiles derived from distinct 10x Chromium kits (panel f). Here, a signature matrix from the 3' kit (3' PBMCs; **Supplementary Table 2**) was applied to 20 randomly drawn mixtures ("pseudo-mixtures") of PBMC samples derived from single cells in the 5' PBMCs dataset. Following S-mode, the adjusted 3' cell type reference profiles are more closely aligned with their corresponding reference profiles from the 5' kit, as demonstrated by heat maps (**h**) and Pearson correlation (**i**). (**j**) Same as **g**, but replacing the 5' PBMC single-cell dataset with 20 pseudo-mixtures of sorted leukocyte subsets profiled by microarray[1] (*n* = 5 cell types). Cell subsets are colored as in panel i. For panels g-j, only genes in the 3' PBMCs signature matrix were considered as reference profiles for each cell type. Group differences in i,j were evaluated by a two-sided Wilcoxon signed-rank test. Data in panels d, i, and j are expressed as means ± s.d. (**k,l**) Robustness of S-mode to variation in key parameters. Within S-mode, the mixing coefficients for each pseudo-bulk mixture are determined according to a Normal distribution N(μ, σ), where μ is set, by default, to the fractional abundance of each cell type within the source single-cell reference matrix $\{\mu_1, \mu_2, \ldots, \mu_c\}$ and σ is set, by default, to $\{2\mu_1, 2\mu_2, \ldots, 2\mu_c\}$, where *c* is equal to the number of cell types in the signature matrix (**Supplementary Note 1**). (**k**) Robustness of S-mode to variation in σ. In rank space, the top-performing value of σ was 2μ, though the benefit was marginal. (**l**) Robustness of S-mode to variation in $\{\mu_1, \mu_2, \ldots, \mu_c\}$, where each μ was multiplied by uniformly distributed noise and the results across all μ were normalized to 1. The mean R^2 comparing $\{\mu_1, \mu_2, \ldots, \mu_c\}$ before noise and $\{\mu_1, \mu_2, \ldots, \mu_c\}$ after noise was 0.43 [0.1–0.69]. P-values represent the statistical significance of the difference between no normalization and S-mode normalization. Group differences were evaluated by a two-sided Wilcoxon rank sum test. Results in k,l are presented as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers) and shown for 5 runs of S-mode applied to enumerate leukocyte subsets in bulk RNA-seq profiles of whole blood (*n* = 12 healthy adults) using a 10x Chromium-derived PBMC signature matrix (5' PBMCs; **Supplementary Tables 1, 2**). The same five cell types in panel d were evaluated here (*n* = 5). 'Cell-level' = performance for each cell type across samples; 'Sample-level' = performance for each sample across cell types.



Supplementary Figure 2: Validation of cross-platform deconvolution. (**a**) Evaluation of S-mode batch correction, showing deconvolution of whole blood from healthy donors (**Supplementary Table 1**) using four signature matrices of leukocyte subsets derived from 10x Chromium v2 (**Supplementary Table 2**). *Left*: Sample-level Pearson correlation (top) and RMSE (bottom) for each signature matrix before and after S-mode batch correction. *Right:* Sample-level and cell-level metrics for the results pooled across the four signature matrices ('sample-level' = performance for each sample across cell types; 'cell-level' = vice versa). Ground truth cell proportions were determined by flow cytometry and complete blood counts. (**b-c**) Same as panel a but applied to (**b**) microarray profiles of PBMCs with ground truth flow cytometry data[2] and (**c**) pseudo-bulk RNA-seq profiles of melanoma tumors[3] reconstituted from single-cell transcriptomes. Group differences were evaluated by a two-sided Wilcoxon signed-rank test in a-c, with data presented as box plots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers). (**d**) Impact of B-mode batch correction on cell subset enumeration with LM22[2] across 5 platforms, 3 tissue types, and fresh/frozen vs. fixed tissue preservation states: Illumina Beadchip microarrays of cryopreserved PBMCs vs. flow cytometry (GSE65133[2]), bulk RNA-seq of fresh whole blood vs. flow cytometry/CBCs (this work), bulk RNA-seq of FFPE melanoma (Mel.) tumors[4] vs. surrogate cell markers, and reconstituted bulk melanoma (Mel) and HNSCC tumors from scRNA-seq profiles (Smart-Seq2; TIL subset frequencies in reconstituted single-cell expression data from melanoma[3] and HNSCC[5] tumors, respectively). Points represent Pearson correlation coefficients between expected and inferred levels for each cell type, and bar plots show medians with 95% confidence intervals expressed as error bars. When applying LM22 to reconstituted tumors, only samples with at least 2 immune cells were considered. To evaluate the performance of B-mode normalization on FFPE melanomas, selected leukocyte subpopulations were compared to canonical marker genes in unadjusted RNA-seq data (*PTPRC* vs. inferred immune content, *CD19* vs. B cells, *CD3D* vs. T cells, *CD8A* vs. CD8 T cells, *CD68* vs. monocytes/macrophages), and CIBERSORTx results were scaled by total immune content using a previously described approach to calculate a 'normalized immune index' for each tumor sample[2].



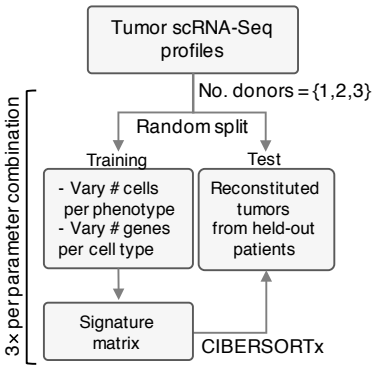
Supplementary Figure 3: Dissection of solid tumors and pancreatic islets using single-cell reference profiles. (**a**) Stacked bar plot depiction of the results in **Fig. 2d**, showing estimated (*Left*) and expected (*Right*) cell subset frequencies in 20 HNSCC tumor samples[5] reconstituted from single-cell transcriptomes. (**b-e**) Single-cell-guided deconvolution of melanoma tumors. (**b**) t-SNE visualization of scRNA-seq data from 19 melanoma tumors[3] (*Left*) and approach for testing single-cell deconvolution performance (*Right*). (**c-d**) Concordance between cell type proportions measured by CIBERSORTx deconvolution and scRNA-seq for 17 held-out melanoma tumors reconstructed from single-cell data, related to panel b. Shown are stacked bar plots (**c**) and Pearson correlations (**d**) of observed versus expected cell subset proportions. (**e**) Analysis of melanoma TIL proportions as measured by scRNA-seq[3], deconvolution of bulk RNA-seq profiles of skin cutaneous melanoma tumors[6] using the signature matrix from panel b, and IHC quantification[7]. (**f-i**) Single-cell-guided deconvolution of pancreatic islets. (**f**) t-SNE visualization of scRNA-seq data from 10 pancreatic islet specimens[8] (*Left*) and approach for testing single-cell deconvolution performance (*Right*). (**g**) Scatterplot comparing scRNA-seq and CIBERSORTx proportions for 9 pancreatic cell types in 8 held-out islets (*Left*), with corresponding Pearson correlation coefficients for individual cell types (*Right*). Data points in the scatterplot are color-coded by cell type to match the bar plot (*Right*). (**h**) Same as g, but for 7 islet samples profiled by bulk RNA-seq and scRNA-seq. The same single-cell-derived signature matrix used to analyze reconstructed tumors in g was applied to the same islet specimens profiled by bulk RNA-seq (*x*-axis). Prominent cell types are highlighted with ellipses. (**i**) Comparison of alpha, beta, gamma, and delta cell proportions in pancreatic islets, as measured by (1) scRNA-seq (using author-supplied cell label annotations), (2) CIBERSORTx deconvolution of bulk islets using a single-cell-derived signature matrix (panel f), and (3) IHC quantification in non-disaggregated islets[8]. Data in e,i are expressed as means ± s.e.m.



Supplementary Figure 4: Evaluation of digital gating for deconvolution of specific cell subsets in bulk tissue expression profiles. (**a**) Schema for creating and validating a single-cell-derived signature matrix to enumerate *PDCD1*+/*CTLA4*+ CD8 T cells along with eight other major melanoma tumor cell types (shown in panel b, upper left). (**b**) The framework in panel a was applied to create four signature matrices, one evaluating the performance of *PDCD1*$^{+}$*CTLA4*$^{+}$ CD8 T cells, one testing the performance of *PDCD1*$^{+}$ CD8 T cells, one assessing *CTLA4*$^{+}$ CD8 T cells, and one evaluating tissue resident memory CD8 T cells. Shown are bar plots depicting the deconvolution performance of each signature matrix applied to 17 reconstituted melanoma tumors held out from signature matrix construction. The Pearson correlation (and corresponding statistical significance) between imputed and known cell proportions is provided for each cell subset. See **Supplementary Note 1** for details of digital gating.

a

Parameter sweep schema



b

Performance in held-out tumors (median Pearson *r*)

HNSCC

Donors	Cells	Gene range 50-150	150-300	300-500
1	3	0.885	0.912	0.916
	5	0.857	0.909	0.925
	10	0.917	0.942	0.942
	20	0.906	0.910	0.925
2	3	0.891	0.920	0.921
	5	0.868	0.930	0.925
	10	0.874	0.923	0.929
	20	0.904	0.934	0.941
	40	0.914	0.944	0.950
3	3	0.897	0.923	0.932
	5	0.894	0.922	0.934
	10	0.919	0.925	0.931
	20	0.917	0.918	0.921
	40	0.925	0.952	0.953

Melanoma

Donors	Cells	Gene range 50-150	150-300	300-500
1	3	0.818	0.829	0.823
	5	0.815	0.853	0.846
	10	0.804	0.809	0.811
	20	0.869	0.887	0.890
2	3	0.780	0.747	0.746
	5	0.887	0.879	0.875
	10	0.876	0.899	0.900
	20	0.876	0.890	0.893
	40	0.857	0.880	0.895
3	3	0.814	0.779	0.797
	5	0.853	0.867	0.859
	10	0.898	0.903	0.897
	20	0.867	0.905	0.894
	40	0.875	0.878	0.889

Statistical signifance of signature matrices versus random genes

HNSCC

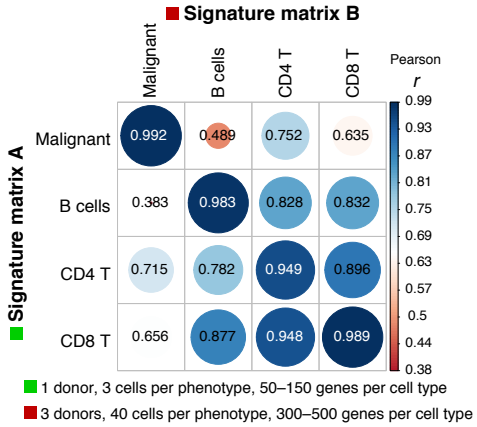
Donors	Cells	Gene range 50-150	150-300	300-500
1	3	<0.05	<0.05	<0.05
	5	<0.05	<0.05	<0.05
	10	<0.05	<0.05	<0.05
	20	<0.05	<0.05	<0.05
2	3	<0.05	<0.05	<0.05
	5	<0.05	<0.05	0.05
	10	0.05	<0.05	<0.05
	20	<0.05	<0.05	0.05
	40	<0.05	<0.05	<0.05
3	3	<0.05	<0.05	<0.05
	5	<0.05	<0.05	<0.05
	10	<0.05	<0.05	<0.05
	20	<0.05	<0.05	0.05
	40	<0.05	<0.05	<0.05

Melanoma

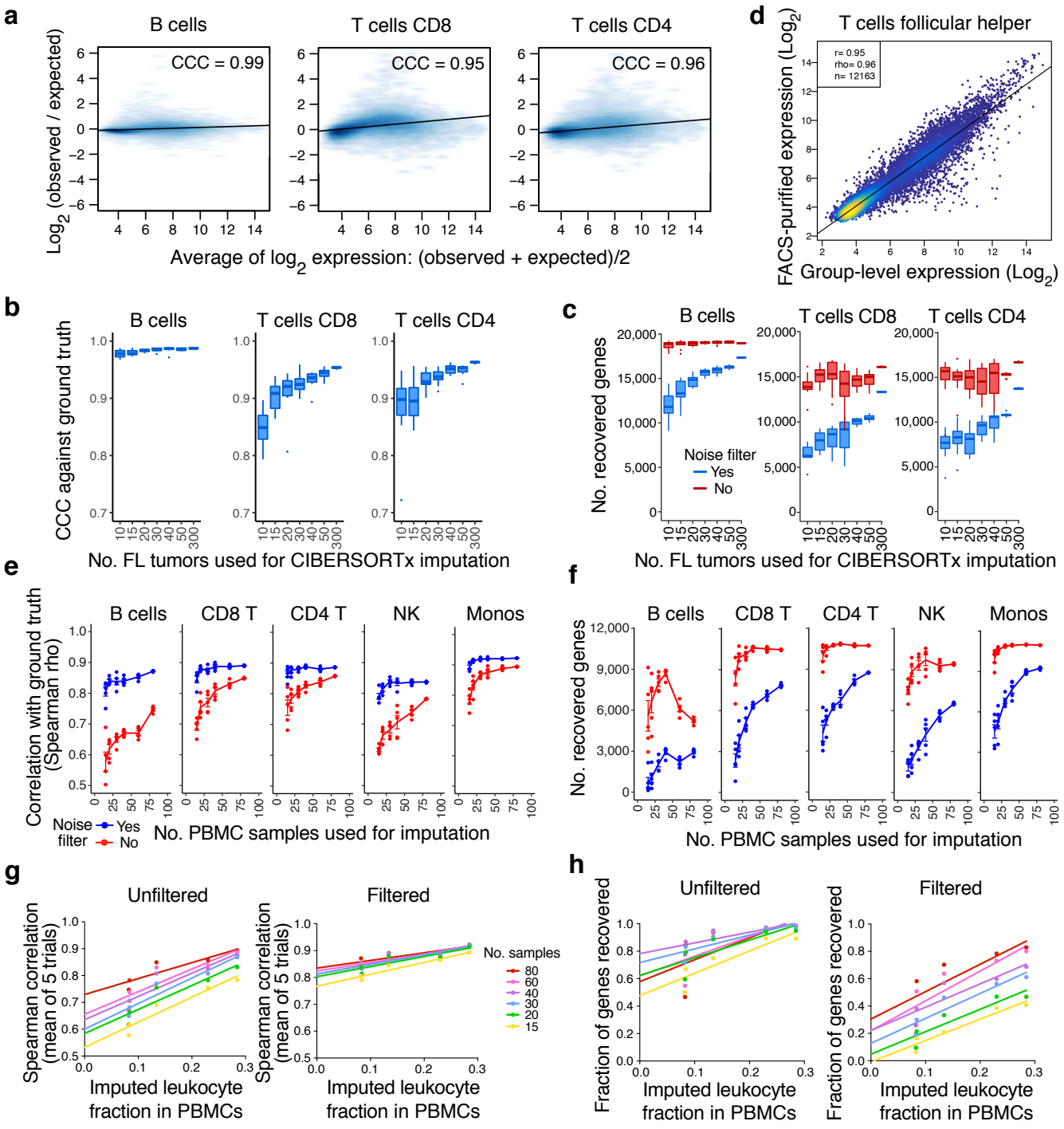
Donors	Cells	Gene range 50-150	150-300	300-500
1	3	<0.05	<0.05	<0.05
	5	<0.05	<0.05	<0.05
	10	<0.05	<0.05	<0.05
	20	<0.05	0.05	0.20
2	3	<0.05	0.05	0.10
	5	<0.05	<0.05	<0.05
	10	<0.05	<0.05	<0.05
	20	<0.05	<0.05	<0.05
	40	<0.05	<0.05	<0.05
3	3	0.15	0.15	0.50
	5	<0.05	<0.05	<0.05
	10	<0.05	<0.05	<0.05
	20	<0.05	<0.05	<0.05
	40	0.05	<0.05	<0.05

c

Concordance between group-level GEPs imputed from melanoma tumors (TCGA RNA-Seq)

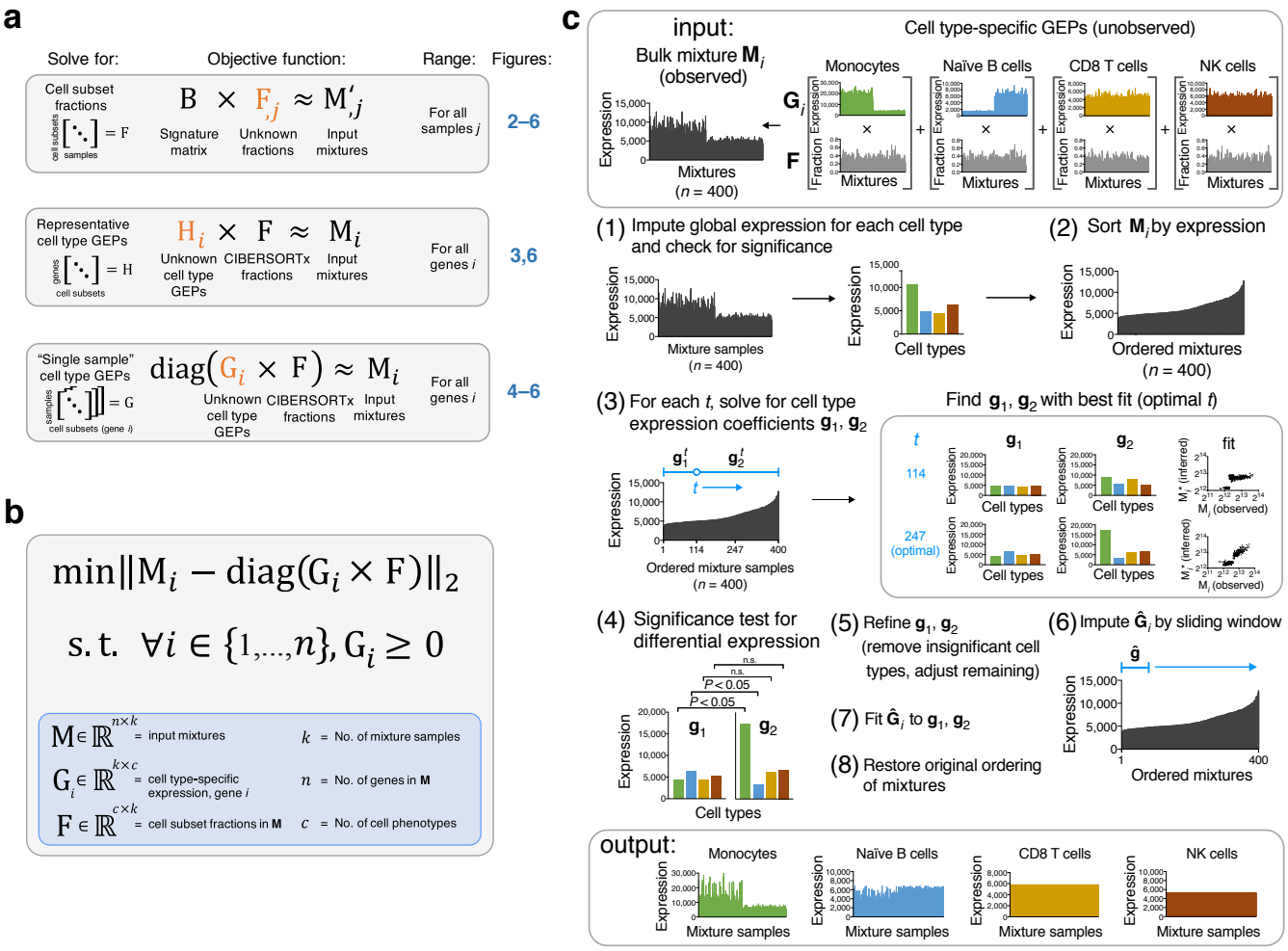


Supplementary Figure 5: Impact of key parameters on single-cell-guided deconvolution. (**a**) Schema of the parameter sweep. Deconvolution performance for estimating cell proportions was assessed in relation to the number of single cells per phenotype, the number of donors per signature matrix, and the range of marker genes per phenotype (**Supplementary Note 1**). (**b**) Results of the parameter sweep in panel a for two tumor types profiled by single-cell RNA-seq: head and neck squamous cell carcinoma (HNSCC) tumors[5] and melanoma tumors[3] (*n* = 126 signature matrices per tumor type; **Supplementary Note 1**). *Top:* The concordance between inferred and known cell proportions was assessed for each cell type by Pearson correlation, and the median correlation coefficient for each parameter combination is shown in tabular format. Gold and gray colors indicate higher and lower correlation coefficients, respectively. *Bottom:* Empirical p-values for obtaining correlation coefficients at least as high as those shown in the upper panel were calculated as described in **Supplementary Note 1**. (**c**) Consistency of signature matrices defined from opposite ends of the parameter range (denoted by green and red rectangles in panel b) for resolving group-mode GEPs (**Fig. 1**, step 3) from bulk melanoma tumors profiled by TCGA[9] (*n* = 472 tumors). Pearson correlations were performed on $\log_2$-adjusted expression values. Only genes with detectable expression after adaptive filtering were analyzed.

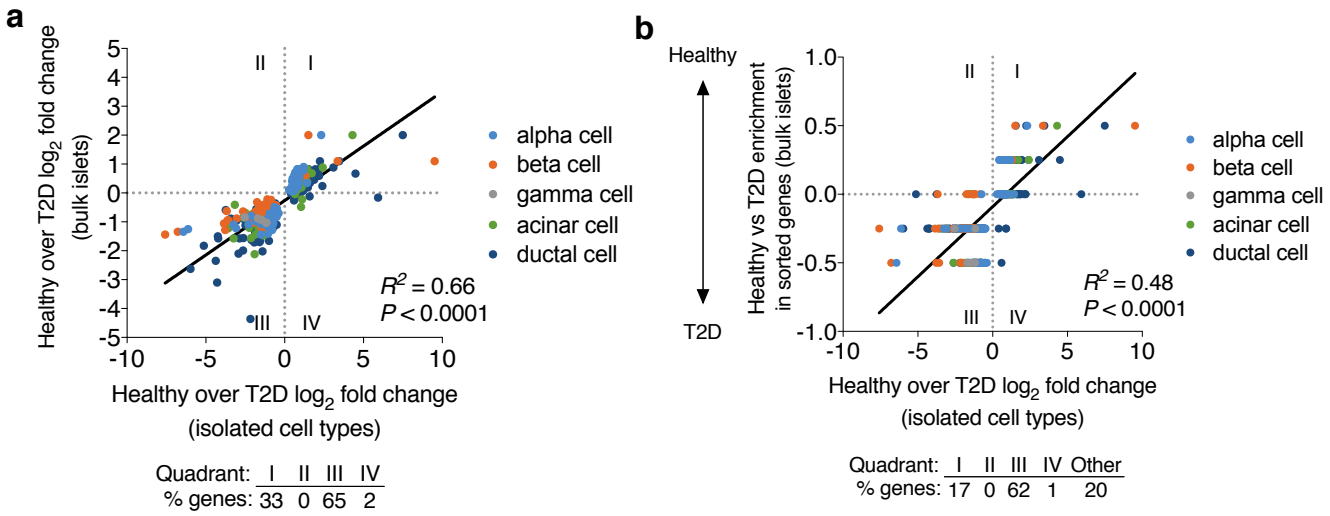


Supplementary Figure 6: Impact of key parameters on group-mode expression purification. (**a**) Bland-Altman plots[10] of B cell, CD8 T cell, and CD4 T cell GEPs imputed from FL tumor samples (*n* = 302), related to **Fig. 3d**. Concordance is captured by a linear regression line and Lin's concordance correlation coefficient[11] (CCC). (**b**) Same as **Fig. 3e**, but showing the distribution of CCC values as a function of the number of tumors used for group-mode GEP imputation, related to **Fig. 3e**. (**c**) Same as **Fig. 3e**, but showing the number of detectable genes (i.e., genes with nonzero expression) in relation to the number of FL tumors used for expression imputation. Data in b,c are expressed as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers). (**d**) Analysis of group-mode expression purification for reconstructing the transcriptome of follicular helper T cells from 302 FL GEPs, as compared to a ground truth profile obtained by FACS-purification of follicular helper T cells from FL lymph nodes[12]. The profile was generated by collapsing LM22 into B

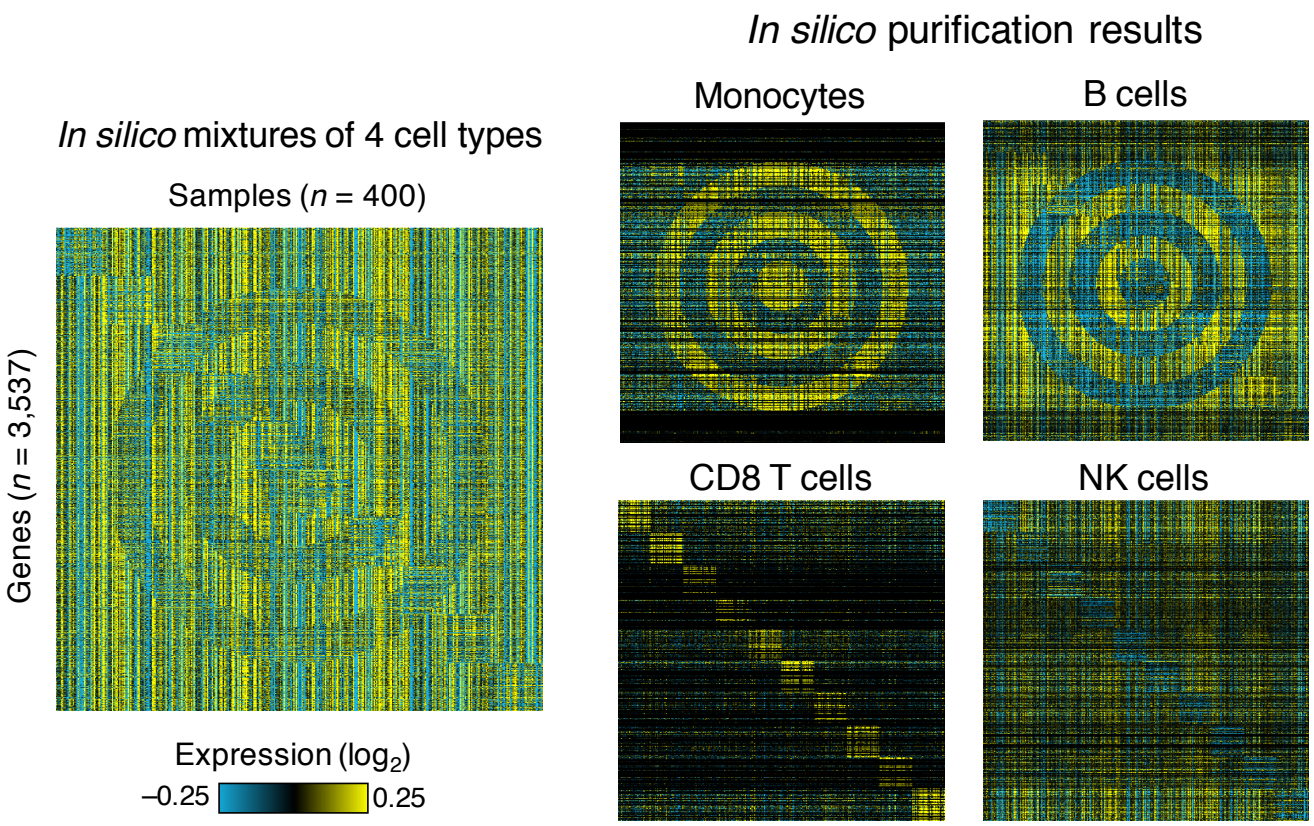
cells, CD8 T cells, follicular helper T cells, and other CD4 T cells, and by aggregating the remaining cell subsets into a single component (also see **Supplementary Table 2**). Imputed GEPs in panels a, b, and d were processed by adaptive noise filtration prior to analysis. Concordance was determined by Pearson correlation (*r*) and Spearman correlation (*rho*), and visualized by linear regression. (**e**) Same as **Fig. 3e**, but showing the relationship between the number of mixture samples analyzed and the performance of leukocyte transcriptome imputation, shown before and after adaptive noise filtration ('*Imputation of group-mode expression profiles*' in **Supplementary Note 1**), on a large cohort of PBMC samples[13]. Samples were subsampled without replacement from the entire PBMC cohort (*n* = 83) five times. The LM22 signature matrix was collapsed into 10 major leukocyte subsets after cell type enumeration and prior to transcriptome purification (**Supplementary Table 2**). Leukocyte reference profiles were derived from sorted peripheral blood populations obtained from healthy donors[1] and normalized as described in **Methods**. (**f**) Same as e, but showing the effect of the number of PBMC samples on the recovery of genes with nonzero expression. Lines in e,f are expressed as means with error bars indicating s.e.m. (**g,h**) Same as e,f, but illustrating the impact of inferred leukocyte abundance on both the accuracy of transcriptome purification (**g**) and the fraction of protein coding genes with nonzero expression (**h**), both before and after adaptive noise filtration ('Unfiltered' and 'Filtered', respectively) on the same PBMC dataset[13]. Linear regression lines in d,e were extrapolated to $x = 0$ in order to model the impact of lower cell fractions on performance.



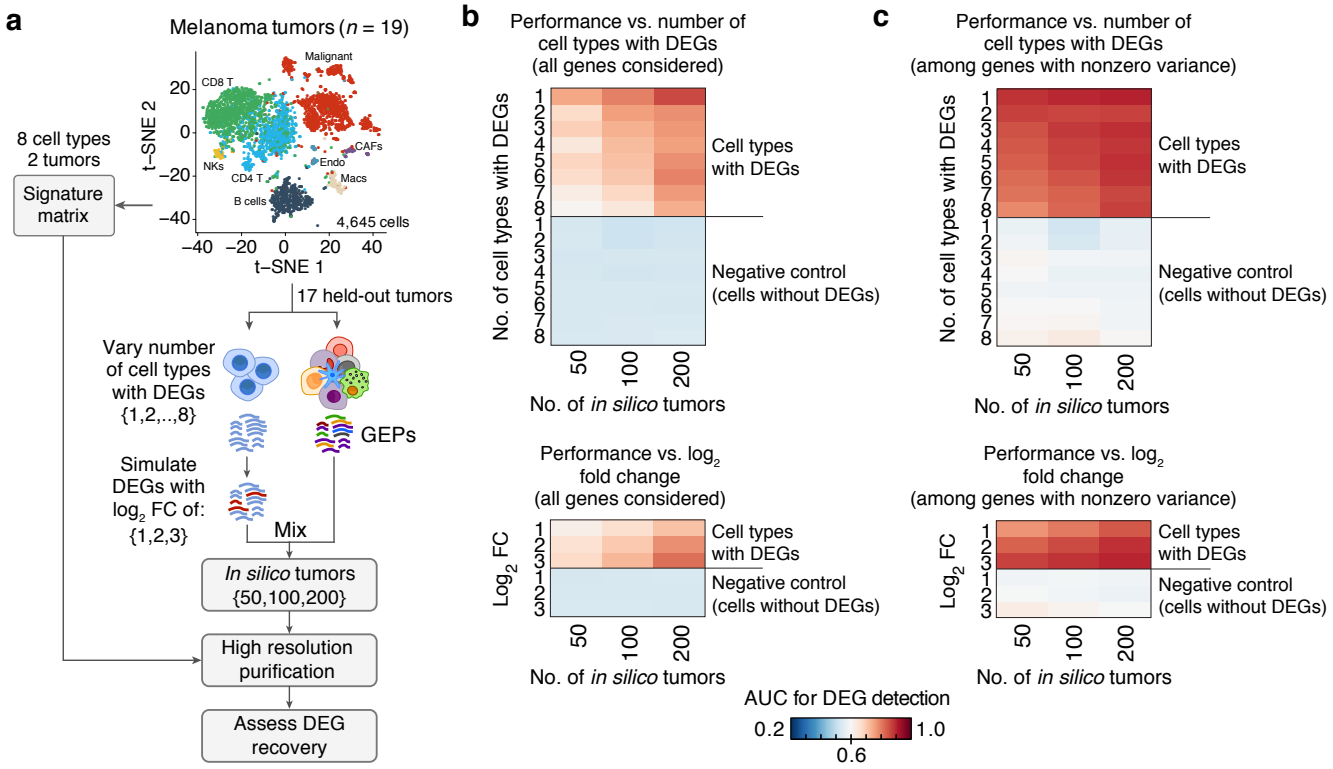
Supplementary Figure 7: CIBERSORTx framework and approach for high-resolution expression purification. (**a**) Key objective functions within the CIBERSORTx framework for enumeration of cell proportions (*Top*), imputation of group-mode cell type-specific transcriptomes, (*Middle*), and high-resolution transcriptome purification (*Bottom*) (**Methods, Supplementary Note 1**). (**b**) Objective function for high-resolution CIBERSORTx. Here, **M** is known, **F** is previously estimated (panel a, top), and **G**, the subject of estimation, is a three-dimensional matrix of n genes × k mixture samples × c cell types. (**c**) Overview of high-resolution CIBERSORTx, illustrated by a simple anecdotal example. Here, a single gene i is expressed by four immune subsets with mutually exclusive expression patterns in monocytes and naïve B cells and no differential expression in CD8 T cells and NK cells. The four profiles are randomly admixed into a bulk admixture GEP $\mathbf{M}_i$. The algorithm proceeds in eight major steps: (1) Determine cell type-specific coefficients for the entire mixture vector using NNLS and evaluate the significance of each coefficient. In this case, all cell types significantly express gene i. (2) Sort the bulk profile in ascending order. (3) Split the sorted vector by iterator t, and apply bootstrapped NNLS to either side of the ranked vector to impute representative cell type-specific expression coefficients ($\mathbf{g}^t_1$ and $\mathbf{g}^t_2$). Repeat this process for all possible values of t that satisfy regression constraints and window length w (**Methods**). For each value of t, estimate $\mathbf{M}_i$ using $\mathbf{g}^t_1$, $\mathbf{g}^t_2$, and **F** (**Supplementary Note 1**) and choose the $\mathbf{g}^t_1$, $\mathbf{g}^t_2$ pair that that results in the best approximation of $\mathbf{M}_i$. (4) Evaluate whether expression is significantly different between each cell type of the best $\mathbf{g}_1$ and $\mathbf{g}_2$ pair from step 3. In this case, only monocytes and naïve B cells show evidence of differential expression at the optimal t. (5) Refine $\mathbf{g}_1$, $\mathbf{g}_2$ estimates based on step 4, (6) Impute initial estimate of $\mathbf{G}_i$ (denoted $\hat{\mathbf{G}}_i$) using bootstrapped NNLS and a sliding window, (7) fit $\hat{\mathbf{G}}_i$ to anchor coefficients (adjusted $\mathbf{g}_1$, $\mathbf{g}_2$ from step 5), and (8) restore the original ordering to finalize $\hat{\mathbf{G}}_i$. For further details, see **Methods** and **Supplementary Note 1**.



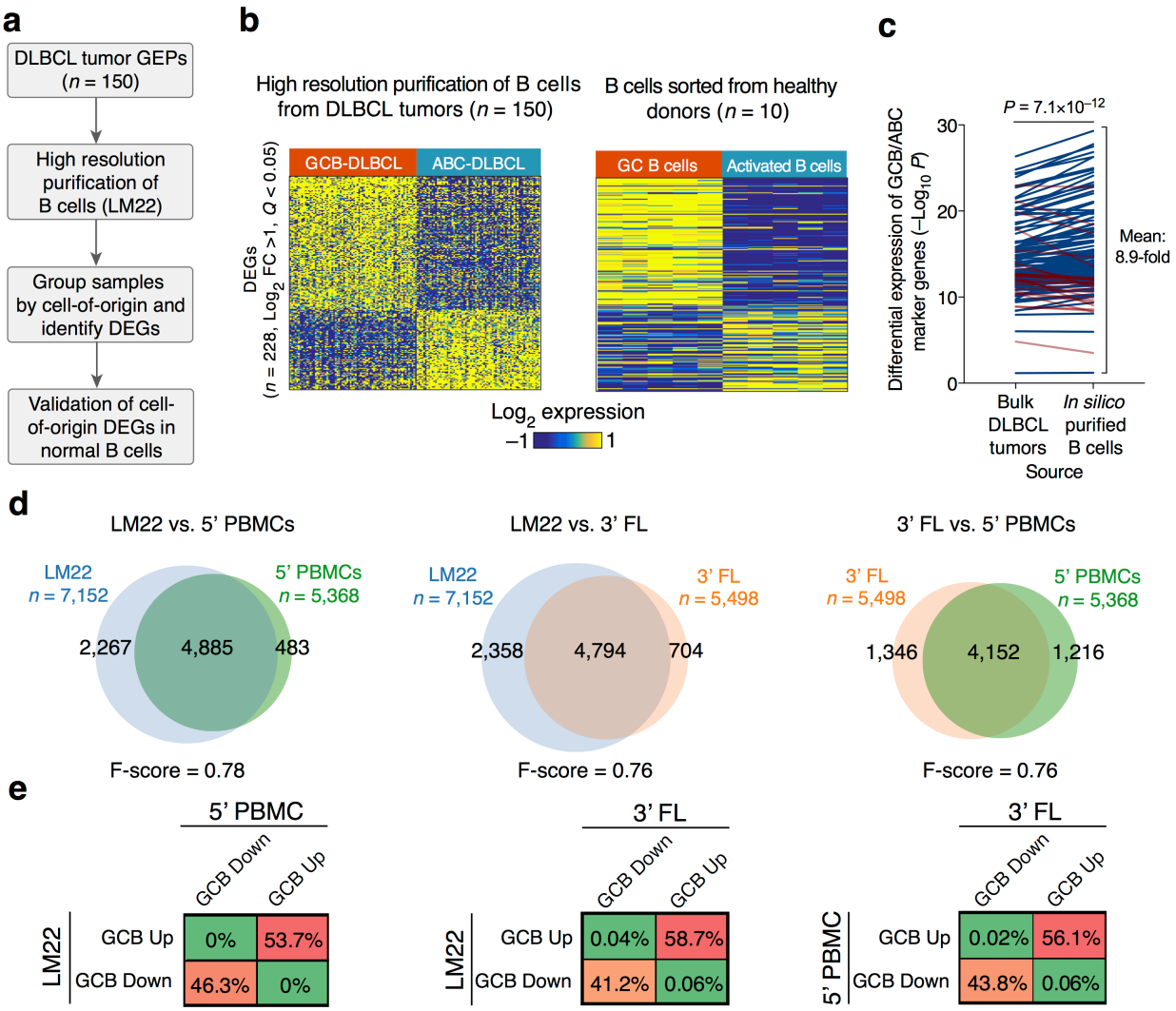
Supplementary Figure 8: Analysis of DEGs identified in single cells compared to their expression in bulk tissues. (**a**) Scatterplot showing fold changes of differentially expressed genes between healthy donors and patients with type II diabetes (T2D) in five major pancreatic islet cell types (*x*-axis) versus bulk islets reconstituted from single cells (*y*-axis). All DEGs and corresponding fold changes in isolated cell types were obtained from Supplemental Table 6 in Segerstolpe et al.[8] (*n* = 100 DEGs). Fold changes in bulk islets were obtained by summing all single-cell GEPs in each pancreatic islet specimen from Segerstolpe et al., normalizing the resulting transcriptomes to TPM, and calculating the $\log_2$ fold change between healthy and T2D specimens for each DEG. In all, 98% of cell type-specific DEGs showed a fold change in the same orientation, whether analyzed in bulk islets or individual cell types. (**b**) Same as a, but comparing the fold change of DEGs in isolated cell types (*x*-axis) to the relative skewing of T2D donors after sorting each gene in order of ascending expression (*y*-axis). To calculate enrichment, a binary matrix was defined such that patients with T2D were represented as 1 and healthy donors as 0. Then, for each gene (*n* = 100), the fraction of T2D patients in the top 50th percentile of expression was subtracted from 0.5. Consequently, genes with no skew received an enrichment of 0, genes skewed toward T2D received a negative value, and genes skewed toward health donors received a positive value. This allowed for an equitable comparison with the fold change orientation in the *x*-axis (i.e., Healthy/T2D). This analysis shows that sorting each gene by its expression in bulk admixtures can inform cell type-specific transcriptional heterogeneity. This concept is leveraged for the first step of high-resolution transcriptome purification with CIBERSORTx (**Supplementary Fig. 7c**, **Methods**, **Supplementary Note 1**). Concordance in a,b, was determined by linear regression and the significance of the result was assessed by a two-sided *t* test.



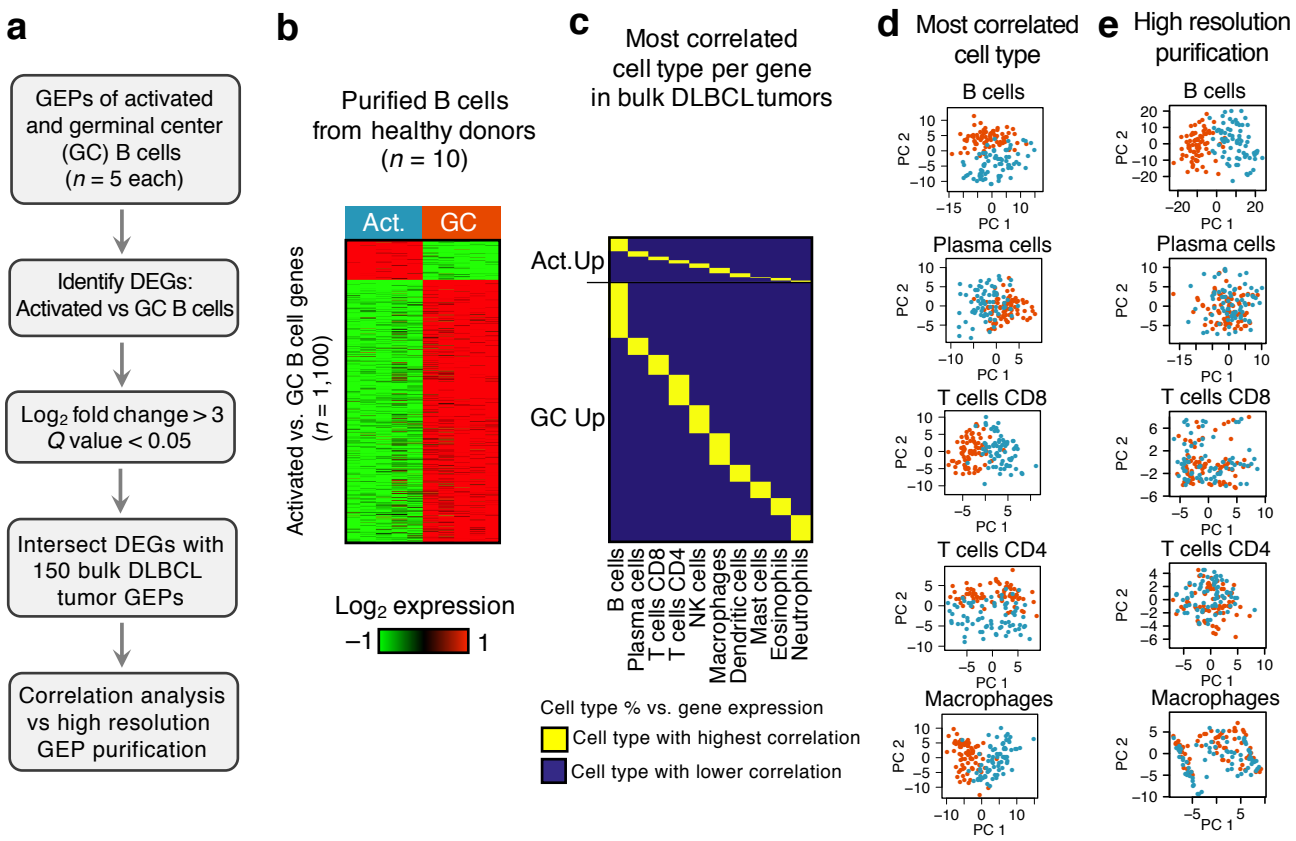
Supplementary Figure 9: High-resolution purification of DEGs with overlapping patterns. *Left:* Heat map showing 400 synthetic GEP admixtures of four immune subsets, each containing a unique cell type-specific DEG pattern: a set of concentric circles in monocytes ('bullseye'), a set of diagonal squares in CD8 T cells ('checkerboard'), and the inverse of these patterns in B cells and NK cells, respectively. *Right:* Heat maps showing high-resolution CIBERSORTx results (for details, see **Supplementary Note 1**). In contrast, Figure 4d shows the application of high-resolution purification to a toy example consisting of a set of concentric circles that are more highly "expressed" in monocytes than other cell types.



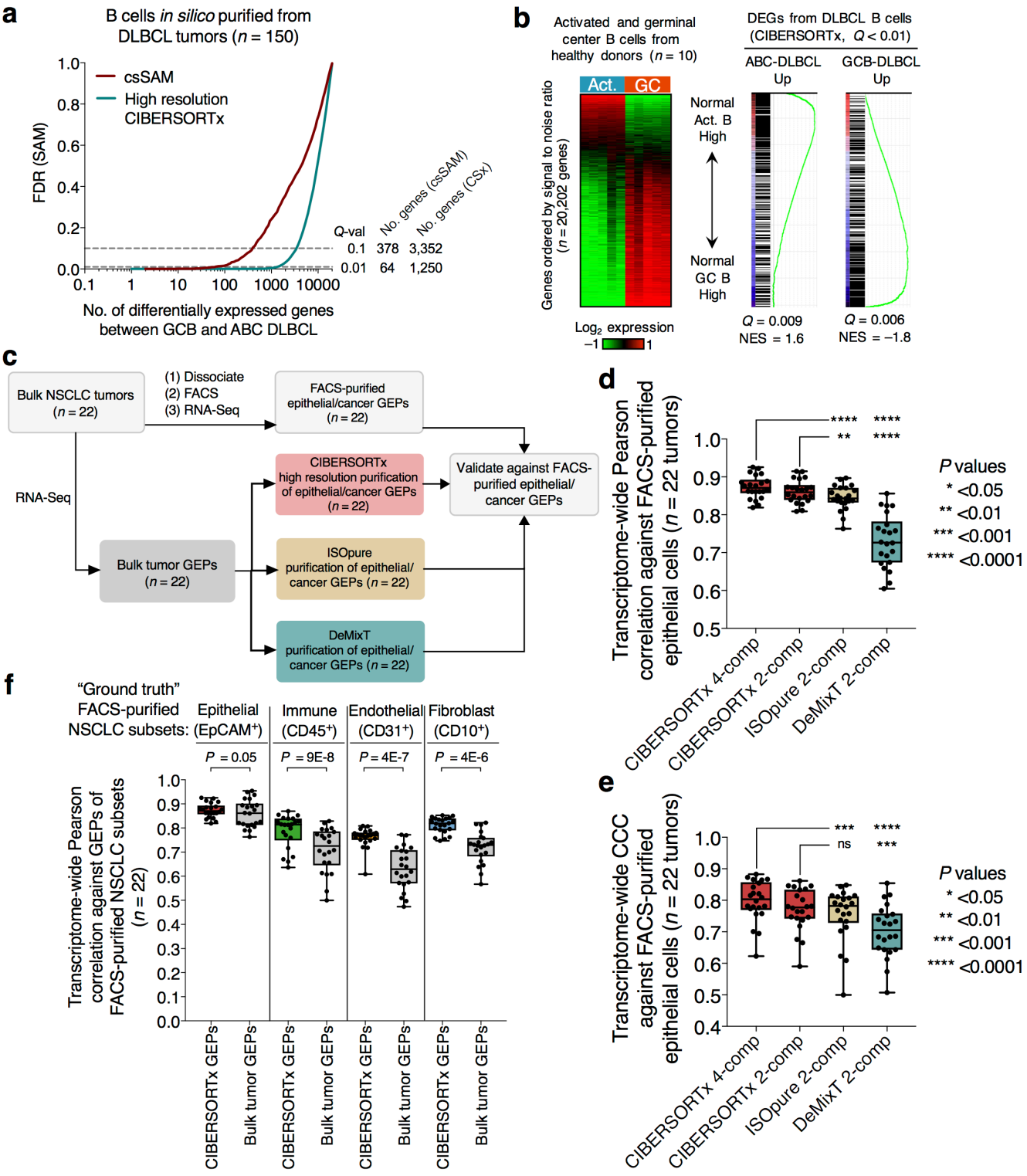
Supplementary Figure 10: Parameter sweep of high-resolution expression purification. (**a**) Schema depicting a parameter sweep to evaluate DEG recovery in reconstituted melanoma tumors containing 8 major cell types. Performance was assessed by systematically varying: (1) the total number of cell types with DEGs (1, 2, 3, …, 8), (ii) the total number of tumor samples (50, 100, 200), and (iii) the $\log_2$ fold change of DEGs (1, 2, and 3). DEG recovery was evaluated as in **Fig. 4e**, with additional details provided in **Supplementary Note 1**. Cell fractions were sampled uniformly according to their frequencies in the original dataset. (**b**) Heat map showing the mean area under the curve (AUC) for the recovery of known cell type-specific DEGs in different numbers of *in silico* tumors as a function of (*Top*) the number of cell types with artificial DEGs and (*Bottom*) the $\log_2$ fold change of DEGs. (**c**) Same as **b**, except the AUC is calculated with respect to imputed genes with non-zero variance. Thus, genes with insufficient evidence of expression are excluded from consideration. Vertical lines in the color bar show AUC intervals of 0.1. As shown in panel b, DEG recovery is highly specific; when known DEGs from one cell type are assessed in another, the false detection rate is no greater than random chance (AUCs of ~0.5). DEG detection rates increase as more tumor samples are profiled and as higher fold changes were evaluated. A modest reduction in performance can be seen as a function of the number of cell types with DEGs. However, these results reflect all genes, including those with insignificant evidence of cell type-specific expression. As shown in panel c, when considering performance among genes with variable expression (i.e., imputed with nonzero variance), the AUC for DEG recovery is consistently high (mean = 0.93, s.d. = 0.04).



Supplementary Figure 11: High-resolution purification of DLBCL molecular subtypes. (**a**) Schema for dissecting diffuse large B-cell lymphoma (DLBCL) tumors (GCB/ABC DLBCL, CHOP cohort, Lenz et al.[14], **Supplementary Table 1**) by high-resolution purification using LM22 (**Supplementary Table 2**). (**b**) *Left*: Heat map showing DEGs identified between ABC and GCB DLBCL within B cell transcriptomes digitally purified from DLBCL tumors. Samples are grouped by previously annotated subtypes[14]. *Right:* Heat map depicting the same genes in the same ordering as the left panel, but shown for germinal center and activated B cells derived from healthy donors[15,16]. Expression data in both heat maps were $\log_2$ adjusted and median-centered for each gene prior to rendering the heat maps. (**c**) Significance of differential expression between GCB and ABC DLBCL for previously published marker genes[14] ($-\log_{10}$ p-value from a two-sided unpaired *t* test with unequal variance), comparing bulk DLBCL tumors and digitally purified B cells. Only marker genes with non-zero expression in digitally purified B cells were evaluated (n = 141 of 144 genes; *SPINK2*, *TOX2*, and *FAM46C* were assigned to other cell types). (**d,e**) Comparative analysis of LM22 versus two 10x Chromium-derived signature matrices, one from peripheral blood and one from follicular lymphoma tumor cells[17] (**Supplementary Table 2**), for identifying DEGs according to the schema in panel a. S-mode was applied to both 10x-derived signature matrices to calculate cell type frequencies. DEGs were filtered as described for NSCLC and melanoma in **Supplementary Note 1** (absolute $\log_2$ fold change >0.1, Q < 0.5). (**d**) Overlap between DEGs, as quantified by the F-score. (**e**) Concordance in the direction of DEGs detected by each pair of signature matrices (e.g., GCB up in both), as expressed by the percentage of DEGs detected with both matrices (i.e., the intersection of the corresponding Venn diagrams in d).

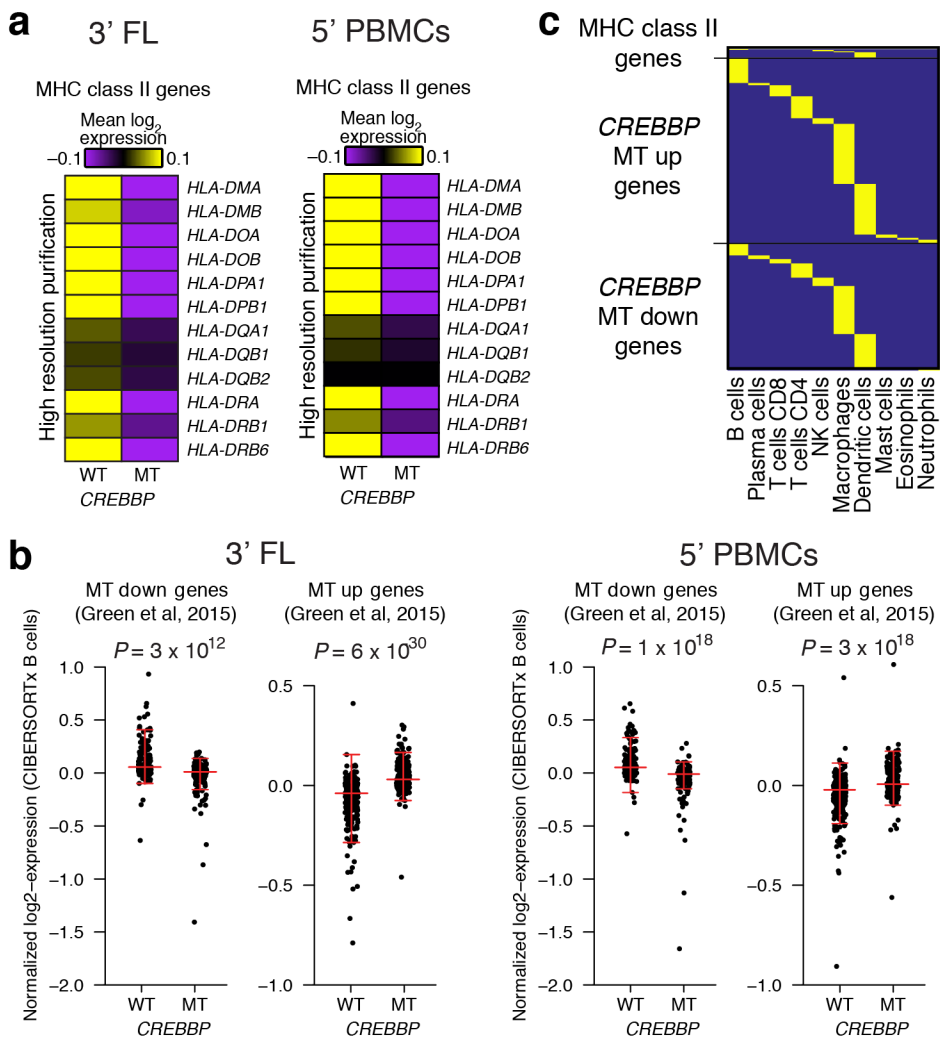


Supplementary Figure 12: Relative utility of high-resolution purification versus a correlation-based approach. (**a**) Schema for the identification and assessment of DEGs that distinguish activated and germinal center B cells profiled by microarray[15,16]. (**b**) Heat map showing DEGs identified as described in panel a. (**c**) Heat map showing the most correlated cell type (of n = 10 evaluated immune subsets; columns) with each gene from **b** (rows, n = 1,100) in bulk DLBCL tumors (n = 150; GCB/ABC DLBCL, CHOP cohort, Lenz et al.[14]). Correlations were calculated as the Pearson correlation between the non-log expression vector of each gene and the imputed fractional abundance of each cell type across all bulk tumors, using LM22[2] fractions collapsed into the indicated lineages. Genes are ordered identically in panels b and c. (**d,e**) PCA plots showing transcriptional variation ($\log_2$-adjusted) among the 5 most abundant cell types in DLBCL tumors (n = 150), color-coded by previously annotated molecular subtypes[14] (ABC DLBCL, blue; GCB DLBCL, orange). Each PCA plot in panel d was generated by uniquely assigning each gene in panel c to its most correlated cell type (e.g., only genes highlighted yellow in the leftmost column of c were used for B cells). Each plot in panel e was generated by applying PCA to the output of high-resolution expression purification, which was run on the same 150 bulk DLBCL tumor GEPs as the correlation-based strategy, and filtered to restrict the analysis to the same 1,100 DEGs (identified as described in panel a).

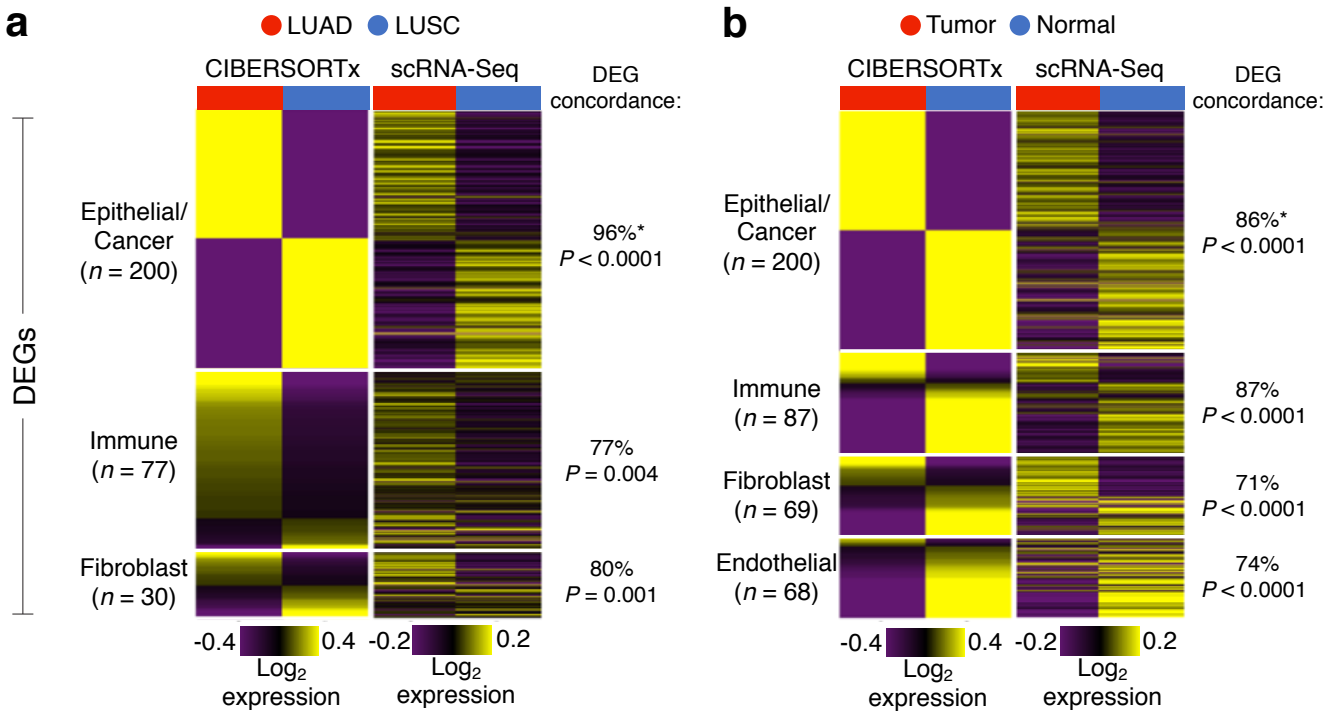


Supplementary Figure 13: Benchmarking of high-resolution gene expression purification against previous methods. (**a**) Analysis of DEGs identified in B cells between germinal center B cell (GCB) and activated B cell (ABC) DLBCL (*n* = 150 tumors; GCB/ABC DLBCL, CHOP cohort, Lenz et al.[14]). Performance is shown for high-resolution purification and csSAM[18] in relation to the false discovery rate (FDR) (*y*-axis) and number (*x*-axis) of DEGs. Cell proportion estimates were obtained using LM22[18] and the same estimates were used as input for both methods. While GCB and ABC labels were provided as input to csSAM, CIBERSORTx was run without prior knowledge of class labels. Significance analysis of microarrays (SAM) was applied to CIBERSORTx results in order to identify

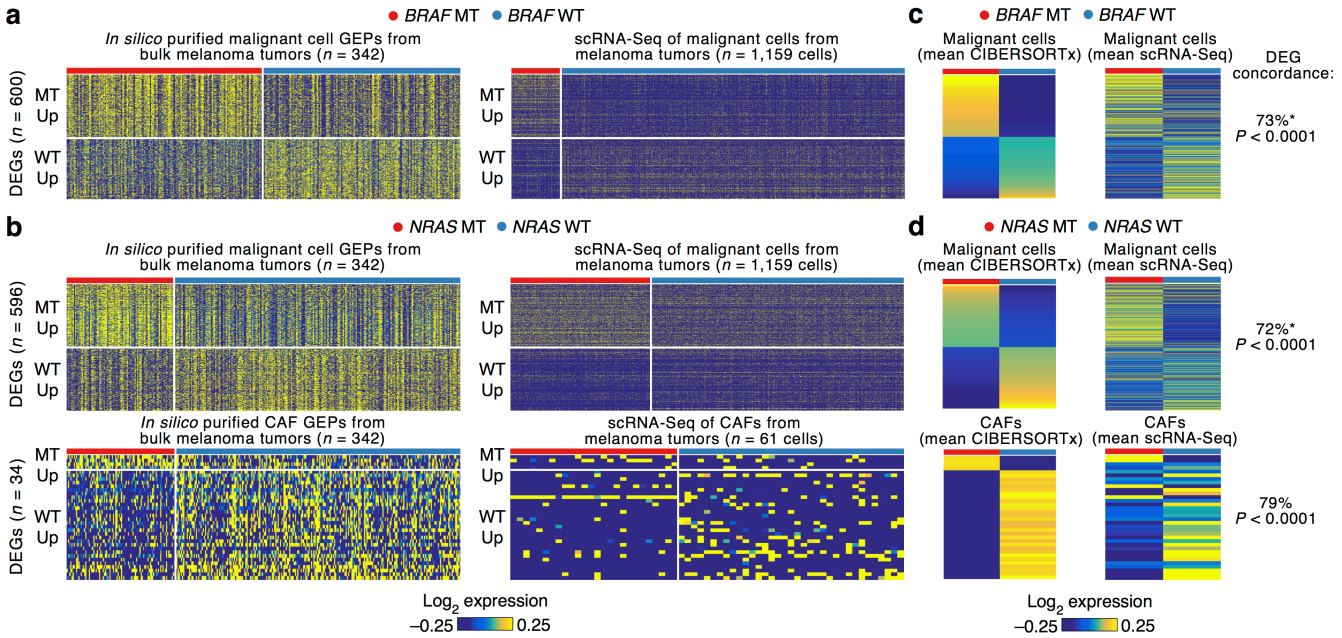
DEGs between GCB and ABC DLBCL. SAM was run with 100 permutations for both methods. (**b**) Gene set enrichment analysis[19] of CIBERSORTx-imputed DEGs (from panel a) for their enrichment in activated or germinal center B cell GEPs derived from healthy donors (n = 10)[15,16]. NES, normalized enrichment score. (**c**) Schema for benchmarking CIBERSORTx against ISOpure[20] and DeMixT[21] for sample-level GEP estimation of epithelial cells from 22 bulk RNA-seq profiles of NSCLC tumors. (**d,e**) Each imputed epithelial transcriptome from panel c was compared against its corresponding FACS-purified GEP from the same patient by (**d**) Pearson correlation and (**e**) Lin's concordance correlation coefficient[11] (CCC). CIBERSORTx high-resolution purification was run twice, once with four components (i.e., all four cell types sorted by FACS; **Fig. 5d**), and once with two components. For the latter, we applied high-resolution-mode to distinguish epithelial/cancer cells from the remaining cell types, whose imputed proportions were aggregated by summation. We ran DeMixT with two components (epithelial vs. non-epithelial) in order to compare all three methods using the same components. 'comp', component. ns, not significant. (**f**) Accuracy of high-resolution CIBERSORTx for simultaneously inferring the transcriptomes of four major cell subsets purified from bulk RNA-seq profiles of 22 bulk NSCLC tumors. To assess performance, each $\log_2$-adusted sample-level transcriptome was compared to RNA-seq profiles of (1) corresponding unsorted bulk tumors and (2) corresponding FACS-purified cell subsets from the same patients. The cell sorting scheme is provided in **Fig. 5d.** For the analyses presented in d-f, concordance and group comparisons were determined for genes with detectable expression, and the latter was evaluated by a two-sided paired t test. Data in d-f are expressed as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, maximum and minimum values). Additional details are provided in **Supplementary Note 1**.



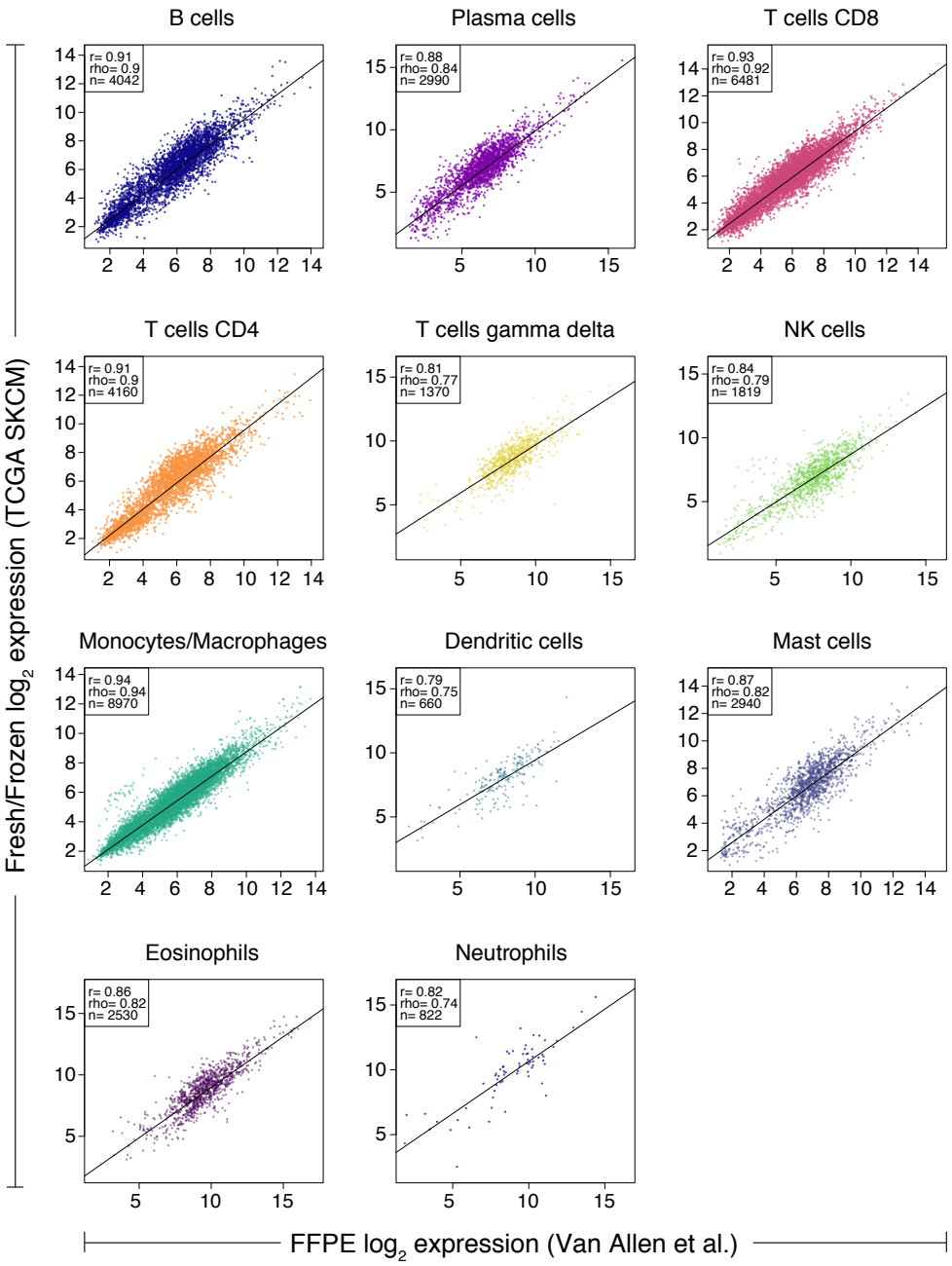
Supplementary Figure 14: Performance of 10x Chromium-derived signature matrices and bulk tumor GEPs for resolving *CREBBP* mutation-associated DEGs in FL B cells. (**a-b**) Same analyses as in **Fig. 5a-c**, but using two 10x Chromium-derived signature matrices, one from peripheral blood (5' PBMCs) and one from follicular lymphoma tumor cells[17] (3' FL) (**Supplementary Table 2**). S-mode batch correction was applied to calculate cell-type frequencies. In b, group comparisons were assessed by a two-sided Wilcoxon signed-rank test, and data are presented as means ± s.d. WT, wildtype (n = 10); MT, mutant (n = 14). (**c**) Same analysis as in **Supplementary Fig. 12c**, but using the gene sets, FL bulk GEP dataset, and inferred cell type frequencies from Fig. 5a-c.



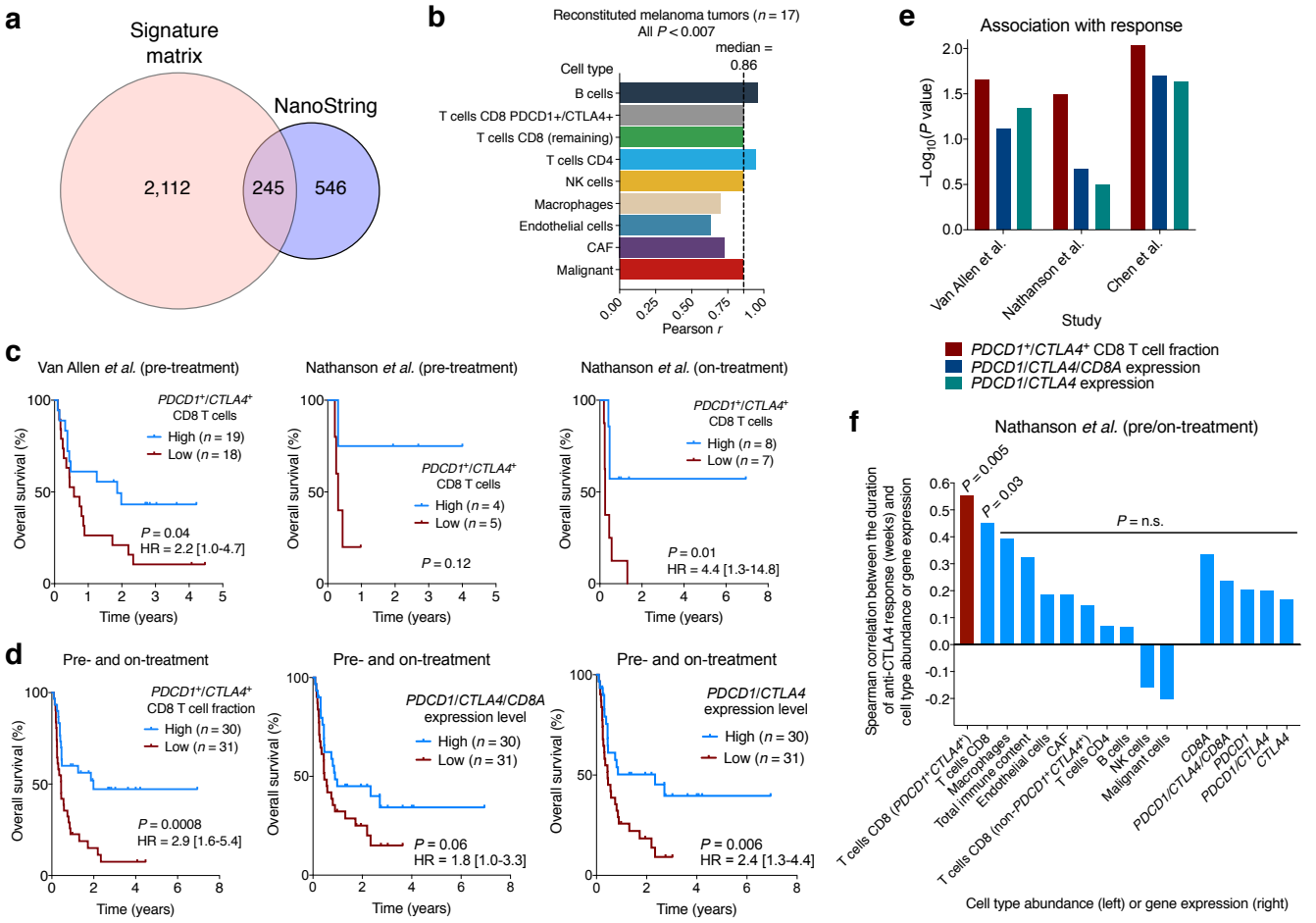
Supplementary Figure 15: Single-cell validation of histology-specific DEGs and tumor/normal DEGs in NSCLC tumors. (**a**) Analysis of the concordance between histology-specific DEGs identified by high-resolution expression purification in bulk NSCLC tumors (related to **Fig. 5d**, left; **Supplementary Table 4**) versus expression patterns in the same cell types from a scRNA-seq atlas of NSCLC (10x Chromium platform, 3' kit)[22]. Shown are the mean-centered $\log_2$-adjusted expression values for each DEG identified by CIBERSORTx (**Supplementary Table 4**), averaged by known phenotype (LUAD or LUSC), and filtered for genes with a mean absolute $\log_2$ fold change in the scRNA-seq data of at least 0.1. The filtration step was performed to overcome sparsity in the 10x Chromium data (median of 775 detectably expressed genes per single cell). Furthermore, given the large number of DEGs detected within epithelial/cancer cells (**Supplementary Table 4**), only the top 100 genes enriched in LUAD and LUSC are shown in the heat map for the sake of clarity. 'DEG concordance' denotes the fraction of genes with the same direction of differential expression between CIBERSORTx and scRNA-seq (higher in both or lower in both), and the p-value was calculated by a Monte Carlo strategy described in **Supplementary Note 1**. (**b**) Same as panel a, but showing DEGs identified between tumor and adjacent normal tissues for each of the four cell types in the signature matrix. The asterisks denote DEG concordance for the epithelial/cancer DEGs depicted in the two heat maps (top 100 DEGs in each direction). The asterisks denote DEG concordance for the epithelial/cancer cell DEGs depicted in a,b (top 100 DEGs in each direction). The concordance values for all detected epithelial/cancer DEGs are 75% (**a**) and 81% (**b**), both of which have Monte Carlo p-values < 0.0001. Of note, this analysis was restricted to cell populations that showed evidence of clustering by LUAD/LUSC or tumor/adjacent normal tissues in **Fig. 5e**. Further details are provided in **Supplementary Note 1**.



Supplementary Figure 16: Prediction and validation of melanoma driver mutation-associated DEGs in tumor cell subpopulations. (**a**,**b**) Heat maps depicting cell type-specific DEGs associated with *BRAF* (a) or *NRAS* (b) mutation status, as identified by high-resolution purification of bulk melanoma tumors (TCGA[9]), related to **Fig. 6b** (see **Supplementary Note 1** for details). Results are shown for *in silico* purified transcriptomes from TCGA[9] (*Left*) and scRNA-seq profiles in a validation cohort[3] (*Right*). Genes are organized top to bottom by their $\log_2$ fold change between mutant and wildtype groups in digitally purified transcriptomes. For visual clarity, only the top 300 over- or under-expressed genes are depicted (all detected DEGs are provided in **Supplementary Table 4**). (**c**,**d**) Same as a,b but averaged by mutation status. 'DEG concordance' denotes the fraction of genes with the same direction of differential expression between CIBERSORTx and scRNA-seq (higher in both or lower in both), and the p-value was calculated by a Monte Carlo sampling approach described in **Supplementary Note 1.** The asterisks denote DEG concordance for the malignant cell DEGs depicted in c,d (top 300 DEGs in each direction). The concordance values for all detected malignant cell DEGs are 64% (**c**) and 73% (**d**), both of which have Monte Carlo p-values < 0.0001.



Supplementary Figure 17: Comparison of TIL GEPs imputed from fresh/frozen versus fixed melanoma tumors. Leukocyte transcriptomes were imputed with CIBERSORTx from independent cohorts of fresh/frozen (TCGA[9,23]; *n* = 473 tumors) and fixed (Van Allen et al.[4]; *n* = 42 tumors) bulk melanoma tumors profiled by RNA-seq. Only genes with nonzero expression in both datasets are shown. The LM22 signature matrix was collapsed into 11 major leukocyte subsets after cell type enumeration and prior to transcriptome purification (**Supplementary Table 2**). Adaptive noise filtration was applied in both datasets (**Supplementary Note 1**). SKCM, skin cutaneous melanoma. Concordance was determined by Pearson correlation (*r*) and Spearman correlation (*rho*), and visualized by linear regression.



Supplementary Figure 18: Associations between *PDCD1*$^+$/*CTLA4*$^+$ CD8 T cells and survival in melanoma patients receiving immune checkpoint blockade. (**a**) Venn diagram depicting the overlap between a single-cell-derived melanoma signature matrix distinguishing 9 tumor cell subsets (**Supplementary Fig. 3b**, top left; **Supplementary Table 1**; **Supplementary Note 1**) and the NanoString nCounter platform employed by Chen and colleagues[7]. (**b**) Same analysis as **Supplementary Fig. 4b** (top left panel), but using the 245 signature matrix genes that overlap the NanoString nCounter assay, related to panel a. All tumors were reconstructed from single-cell GEPs held-out from signature matrix construction. (**c-f**) Gene expression signatures versus deconvolution for predicting response to immune checkpoint blockade. (**c**) Kaplan Meier survival plots showing the relationship between estimated levels of *PCDC1*$^+$/*CLTA4*$^+$ CD8 T cells and overall survival in datasets from Van Allen and colleagues[4] and Nathanson and colleagues[4]. *PCDC1*$^+$/*CLTA4*$^+$ CD8 T cell levels were stratified by their median value in each dataset. (**d**) Kaplan Meier plots comparing deconvolution of *PDCD1+CTLA4+* CD8 T cells to surrogate gene expression signatures for predicting overall survival in melanoma patients treated with anti-CTLA4 therapy. Both gene expression and imputed fractions were stratified by median split. Gene expression signatures were calculated as the geometric mean of *PDCD1*, *CTLA4*, and *CD8A* (*Center*) or *PDCD1* and *CTLA4* (*Right*). Statistical significance in c,d was determined using a two-sided log rank test. HR, hazard ratio. 95% HR confidence intervals are shown in brackets. (**e**) Binary association between response to immunotherapy and *PDCD1+CTLA4+* CD8 T cells levels or expression signatures (same as panel d). Significance was evaluated with a two-sided Wilcoxon rank-sum test and is expressed as –log10 p-values. Sample sizes (*n*) from left to right: 37, 24, 26. (**f**) Spearman correlation between the duration of response to anti-CTLA4 therapy and CIBERSORTx estimates of cell type abundance (*Left*) or average expression of selected genes in $\log_2$ FPKM space (*Right*) in the dataset of Nathanson and colleagues[4] (*n* = 24 tumors).

Supplementary Note 1 – Extended Methods

Table of Contents

For completeness, the description of CIBERSORTx from Online Methods is repeated below.

1. Processing and flow cytometry of human tissue specimens.

1.1. Whole blood.

Whole blood samples were collected from 12 healthy adult subjects and immediately processed to enumerate leukocyte composition by FACS using an FDA-approved in vitro diagnostic test (IVD Multitest 6-color TBNK, Becton Dickinson) and automated hematology analyzer for blood leukocyte differential counts (Sysmex XE-2100) in a CLIA hematology lab setting (Stanford Clinical Laboratories). Aliquots from the same whole blood samples were stored in PAXgene blood RNA tubes (Qiagen) for subsequent RNA extraction and RNA-seq library preparation.

1.2. Primary NSCLC tumors.

Freshly resected surgical tumor samples from patients with NSCLC were dissociated and sorted as previously described[24,25] using A700 anti-human CD45 clone HI30 (pan-leukocyte cell marker), PE anti-human CD31 clone XWM59 (endothelial cell marker), APC anti-human EpCAM clone X9C4 (epithelial cell marker), and PE-Cy7 anti-human CD10 clone XHI10a (fibroblast marker). All antibodies were obtained from BioLegend.

1.3. Primary FL tumors.

Cryopreserved FL tumor cell suspensions were thawed, processed, and sorted to isolate tumor B cells as previously described[26]. GEPs of FACS-purified CD8, CD4, and follicular helper T cells isolated from FL lymph nodes were obtained from previous studies[12,27] (**Supplementary Table 1**).

2. Gene expression profiling.

2.1. Single-cell RNA-seq.

A single-cell library was prepared from a cryopreserved NSCLC PBMC single-cell suspension using Chromium with v2 chemistry (10x Genomics), and was sequenced on a NextSeq 500 (Illumina). Reads were aligned, filtered, de-duplicated, and converted into a digital count matrix using Cell Ranger 1.2 (10x Genomics). Additional quality control (QC) was performed by visual inspection of QC plots using Seurat v1.4.0.16[28]. Outlier cells were identified and filtered based on the following criteria: (1) anomalously high/low mitochondrial gene expression (cells with >10% or <1% mitochondrial content were removed), or (2) potential doublets/multiplets, as identified by comparing the number of expressed genes detected by per cell versus the number of unique molecular identifiers (UMIs) detected per cell (cells with greater than 3,500 and less than 500 expressed genes were removed). Using Seurat[28], clusters were identified by (1) regressing out the dependence of gene expression on the number of UMIs and the percentage of mitochondrial content, and (2) by running “FindClusters” on the first 10 principal components of the data with the resolution parameter set to 0.8. Cell labels were assigned according to the expression of canonical leukocyte marker genes (*MS4A1* high = B cells; *CD8A* high and *GNLY* low = CD8 T cells; *CD3E* high, *CD8A* low, and *GNLY* low = CD4 T cells; *GNLY* high and *CD3E* low = NK cells; *GNLY* high and *CD3E* high = NKT cells; *CD14* high = monocytes).

2.2. Bulk RNA-seq.

Total RNA was isolated from whole blood samples of 12 healthy adults stored in PAXgene tubes. RNA was isolated using the PAXgene Blood RNA Kit (Qiagen) according to the manufacturer’s recommendations. then quantitated and quality assessed using a 2100 Bioanalyzer (Agilent). Library preparation was performed with the TruSeq RNA exome kit (Illumina) per the manufacturer’s recommendations. RNA-seq libraries were multiplexed together and sequenced on a single HiSeq 4000 lane (Illumina) using 2 x 150bp reads.

Total RNA was extracted from bulk NSCLC tumors and sorted cell populations (range: 500 to 25,000 cells) using AllPrep DNA/RNA Micro kit (Qiagen). A median of ~30ng total RNA was amplified using the Ovation RNA-seq System V2 (NuGEN). The resulting cDNA was sheared by sonication (Covaris S2 System) to an average size of 150-200bp and used to construct DNA libraries using the NEBNext DNA Library Prep Master Mix (New England Biolabs). Libraries were sequenced on a HiSeq 2000 (Illumina) to generate 100bp paired end reads with an average of 100M reads per sample.

To maximize linearity in the context of deconvolution analyses[29], raw FASTQ reads were processed with Salmon v0.8.2[30] using GENCODE v23 reference transcripts, the –biasCorrect flag, and otherwise default parameters. RNA-seq quantification results were merged into a single gene-level TPM matrix using the R package, *tximport*[31].

2.3. Microarrays.

Total RNA was extracted from bulk FL specimens and sorted B cells, and assessed for yield and quality as previously described[2,26]. cRNA was prepared from 100ng of total RNA following linear amplification (3' IVT Express, Affymetrix) and then hybridized to HGU133 Plus 2.0 microarrays (Affymetrix) according to the manufacturer's protocol. All CEL files were pooled with a publicly available Affymetrix dataset containing CD4 and CD8 TILs FACS-sorted from FL lymph nodes (GSE27928[27]). The resulting dataset was then RMA normalized using the "affy" package in Bioconductor, mapped to NCBI Entrez gene identifiers using a custom chip definition file (Brainarray version 21.0; http://brainarray.mbni.med.umich.edu/Brainarray/), and converted to HUGO gene symbols. Replicates of sorted cell subsets were combined to create ground truth reference profiles (e.g., as in **Fig. 3b**) using the geometric mean of expression values.

3. Tumor genotyping.

Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq) was applied to 12 cryopreserved FL tumor biopsies using a dedicated lymphoma targeted sequencing panel and sample processing protocol, as previously described[32,33]. Variant calling was performed as previously described[32], with matching peripheral blood leukocytes employed as germline controls. High-throughput sequencing was performed using 2x100bp reads on an Illumina HiSeq 2500 instrument. Since CREBBP hotspot mutations occur within the lysine acetyltransferase (KAT) domain[26], only non-silent variants (single nucleotide variants and insertions/deletions) within the KAT domain were considered "mutant" in this work. To increase the sample size, we also analyzed the *CREBBP* mutation status of 12 previously genotyped patients from our FL cohort[26]. Variants and patient identifiers are provided in **Supplementary Table 3**.

4. Single-cell signature matrix construction.

4.1. Additional details.

See **Methods** for the general approach. Details regarding specific datasets are provided below.

HNSCC: The signature matrix derived from the HNSCC scRNA-seq data generated by Puram and colleagues[5] was trained on tumor ID MEEI18 (primary tumor) and MEEI25 (primary tumor and paired lymph node metastasis) (**Fig. 2c**).

Melanoma: All signature matrices derived from the scRNA-seq data generated by Tirosh and colleagues[3] were trained on tumor IDs 80 and 88 (**Fig. 6a**, **Supplementary Fig3. 2b-e, 4, 10, 18**). Previously annotated cell phenotype labels from the authors were employed in both cases, with the exception of CD8 T cells, which were not provided by the authors and which we defined as T lymphocytes expressing *CD8A* or *CD8B* (TPM>0).

Pancreas: The pancreatic islet signature matrix (**Fig. 2f**, **Supplementary Fig. 3f-i**) was trained on sample IDs HP1508501T2D and HP1504901 using the previously assigned cell phenotypes[8]. We omitted mast cells and MHC II cells from the islet signature matrix since they had insufficient single cells for adequate learning in the training cohort (i.e., <3 cells; see '*Single-cell signature matrix construction*'). Glucagon, the single most highly expressed marker gene in alpha cells, was found to adversely affect deconvolution performance on pancreatic islet training data despite normalization, and was therefore omitted.

All single-cell-derived signature matrices are available in **Supplementary Table 2**.

4.2. Digital gating.

Related to **Supplementary Figure 4** and **Figure 6d**, single cells were pre-labeled by the original investigators according to 7 previously annotated phenotypes[3]: B cells, T cells, NK cells, macrophages, endothelial cells, cancer-associated fibroblasts, and malignant melanocytes. As described above, we further divided T lymphocytes into CD4 and CD8 T cells by the expression of *CD8A* or *CD8B* (TPM>0). CD8 T cells were further subdivided into four distinct subpopulations as follows. To isolate candidate CD8 T cells positive for *PDCD1* and/or *CTLA4* expression, CD8 T cells were split by the expression of *PDCD1* (non-log TPM >5th percentile of *PDCD1* expression values among CD8 T cells) and/or *CTLA4* (non-log TPM >5th percentile of *CTLA4* expression values among CD8 T cells). To isolate candidate tissue resident memory CD8 T cells, we selected CD8 T cells that expressed *ITGAE* (TPM>0) and *CD69* (TPM>0), but not *CCR7* (TPM=0)[34].

4.3. t-SNE.

Visualization of scRNA-seq data by t-distributed stochastic neighbor embedding (t-SNE) was performed as previously described[35] using the *Rtsne* package in R with default parameters unless stated otherwise. For **Figure 2f** and **Supplementary Figure 3f**, we applied ComBat[36] to eliminate subject-specific variation prior to t-SNE projection, as previously described by the authors[8].

5. Parameter sweep and cross-platform deconvolution analysis.

To systematically evaluate the impact of key signature matrix parameters on deconvolution performance, we assessed: (1) the number of cells per phenotype (3, 5, 10, 20 and 40), (2) the number of signature genes per cell type (ranges of 50-150, 150-300, and 300-500), and (3) the number of donor/tissue samples used for training (1, 2, or 3) (**Supplementary Fig. 5a**). We created separate single-cell-derived signature matrices for melanoma[3] and HNSCC[5] tumors within a training/validation framework (**Supplementary Fig. 5a**). To ensure the robustness of the results, each parameter combination was tested in triplicate by random sampling (**Supplementary Fig. 5a**). In addition, to ensure sufficient representation in both the training and validation datasets across all evaluated parameter settings, the analysis was restricted to the following cell phenotypes: malignant cells (HNSCC, melanoma), CD8 and CD4 T cells (HNSCC, melanoma), B cells (melanoma), and fibroblasts (HNSCC). Human tissue samples were selected from a set of samples that had at least 20 cells per cell phenotype (subject IDs HN16_T, HN17_T, HN25_LN, HN25_T, HN18_T for HNSCC tumors; subject IDs 79, 80, 88, 89 for melanoma tumors). Four HNSCC samples (HN10_T, HN13_T, HN24_LN, and HN7_T) were excluded since they each contained <50 cells total across the evaluated phenotypes. When sampling only one patient, melanoma subject 88 was also excluded owing to insufficient cells (<50). In all, 126 signature matrices were tested per dataset. Each signature matrix was created using the approach described in '*Single-cell signature matrix construction.*' Performance was quantified as the median cell-level correlation against ground truth cell proportions from reconstituted tumor samples held out from signature matrix construction (**Supplementary Fig. 5b**). To evaluate the statistical significance of the resulting correlation coefficients, we ran CIBERSORTx using signature matrices comprised of *n* genes randomly sampled from the transcriptome, where *n* was equivalent to the number of genes used in the corresponding 126 signature matrices. This process was repeated 20 times to calculate empirical p-values (**Supplementary Fig. 5b, bottom**).

To assess deconvolution across platforms and tissues, we evaluated single-cell-derived signature matrices across three scRNA-seq protocols (SmartSeq2, 3'- and 5'-biased 10x Chromium expression profiling kits) and three tissue types (reconstituted melanoma tumors, reconstituted HNSCC tumors, and peripheral blood). In all cases, imputed proportions were compared against ground truth frequencies quantified by flow cytometry / automated hematology analyzer (blood samples) or scRNA-seq (melanoma[3], HNSCC[5]). As a comparator, LM22[2], a microarray-derived signature matrix for distinguishing 22 hematopoietic subsets derived from healthy adult subjects, was also tested (**Fig. 2e**, **Supplementary Fig. 2d**). Only cell types present in both the signature matrix and tissue samples were considered, and all cell fractions were normalized to one prior to comparison (**Fig. 2e, Supplementary Fig. 2**). Of note, NKT and CD8 T cell fractions were pooled when comparing inferred CD8 T cell levels against flow cytometry since (1) NKT cells were not enumerated by the FDA-approved *in vitro* diagnostic test that we employed to enumerate leukocyte composition, and (2) NKT cells inferred by 10x profiling were found to express *CD8A* and *CD3D*. For SMART-Seq2 datasets, only previously annotated cell phenotypes were assessed, with the exception of CD8/CD4 T cells in HNSCC and melanoma tumors, which were defined as described in '*Single-cell signature matrix construction'* above and evaluated in HNSCC tumors using a melanoma signature matrix, and vice versa. In all cases, deconvolution was applied with S-mode normalization for 10x-derived signature matrices and B-mode normalization for the remaining signature matrices (see '*Cross-platform normalization schemes for deconvolution*' below).

6. Group-mode purification of NSCLC cell subsets.

To generate an NSCLC signature matrix for expression purification (**Fig. 3g**), we analyzed RNA-seq GEPs of epithelial ($EpCAM^+$), hematopoietic ($CD45^+$), endothelial ($CD31^+$), and stromal ($CD10^+$) subpopulations, each of which was FACS-purified from primary NSCLC tumor biopsies obtained from four subjects with newly diagnosed cancer. To eliminate the confounding effect of variably efficient rRNA depletion during RNA-seq library preparations, we removed mitochondrial rRNA transcripts (*MT-RNR1* and *MT-RNR2*) and renormalized the data matrix in TPM space. Signature matrix parameters were identical to those described in '*Single-cell signature matrix construction*', but without the preprocessing steps employed for scRNA-seq. The resulting signature matrix, comprised of 933 genes, was applied to RNA-seq profiles of bulk NSCLC tumor samples obtained from another set of human subjects, preprocessed as described above (n = 26; **Supplementary Table 1**), and the results were analyzed to impute group-mode cell type-specific GEPs using the method described in '*Imputation of group-mode expression profiles*' (below). Ground truth cell type GEPs were derived by taking the geometric mean of replicate RNA-seq profiles of $EpCAM^+$, $CD45^+$, $CD31^+$, and $CD10^+$ subpopulations that were FACS-sorted from the 2nd set of subjects.

7. Generation and analysis of synthetic GEP admixtures.

Simulated GEP mixtures with overlapping, block-like patterns in **Figure 4a** were created using reference profiles of c = 4 cell types (naïve B cells, CD8 T cells, NK cells, and monocytes) obtained from the source data underlying LM22[2]. As an initial proof of concept and for efficiency, we restricted the analysis to 1,000 randomly selected genes from the transcriptome, along with all 547 genes from the LM22 signature matrix. We then $\log_2$-adjusted the four cell type expression vectors and replicated each of them m = 400 times. Variability was introduced into each replicate by calculating the standard deviation sd_j of each gene j across the four cell types, and by replacing each gene's expression value v by drawing random values from the distribution $N(v, sd_j/5)$. To simulate biological groupings, synthetic DEGs were added to each cell type, such that the first 200 and second 200 samples were stratified into two classes. Overlapping DEG patterns were inserted into monocyte, B cell, CD8 T cell, and NK cell GEPs, as depicted by orange cells, green cells, blue cells, and purple cells, respectively, in **Figure 4a** (left panel). All DEGs were created to reflect a ~5x higher level of expression. Next, we generated a $m \times c$ matrix of mixture coefficients randomly drawn from a normal distribution with $N(1/k, 1/(2k))$. Coefficients for each mixture were normalized to sum to 1. Prior to mixing, all synthetic cell type GEPs

were converted to non-log linear space. We then combined them according to their corresponding mixing coefficients, yielding 400 simulated tissue GEPs. High-resolution CIBERSORTx was run on the resulting admixture dataset using a window size of 20 (=5*c*) and LM4, a subset of the LM22 signature matrix containing barcode genes for monocytes, B cells, CD8 T cells, and NK cells.

For the synthetic 'bullseye' dataset in **Figure 4d**, we repeated the above approach with *m* = 11,845 (the full list of genes in the source dataset from which LM22 was derived), and the following changes. First, we embedded up-regulated DEGs in the shape of a 'bullseye' target (three concentric circles) into monocyte GEPs. To mask the DEG image in bulk admixtures, DEGs were added every 22 genes into monocyte GEPs; the inverse (downregulated DEGs of equal fold change) was added into the remaining cell type GEPs (*n* = 3) in alternating order (i.e., every 66 genes). High-resolution CIBERSORTx was then applied as described above to recover the embedded bullseye.

For the synthetic 'bullseye' and diagonal squares dataset in **Supplementary Figure 9,** we repeated the approach used for **Figure 4d**, with the exception that we restricted our analysis to a random selection of 3,000 genes along with all genes in LM22. We then embedded two upregulated DEG patterns, a target into monocyte GEPs (as in **Fig. 4d**, but inserted every other gene), and a set of 10 diagonally aligned squares into CD8 T cell GEPs (added to every gene). To obfuscate these DEGs in admixture profiles, we inserted downregulated DEGs into B cells and NK cells in a reciprocal manner that complemented the respective patterns. High-resolution CIBERSORTx was again applied as described above.

8. Analysis of cell type-specific DEG detection with high-resolution CIBERSORTx.

8.1. Synthetic tumors with five cell types.

Simulated mixtures were created as described for 'Generation and analysis of synthetic GEP admixtures', except in this experiment, *k* was set to 50 mixtures, *m* was set to 3,000 genes (containing LM22 genes), and DEGs were exclusively inserted into CD8 T cell profiles (samples 26-50; all DEGs were over-expressed relative to baseline levels). To assess performance across a range of realistic settings, we evaluated a distinct set of 50 mixtures for different combinations of (1) DEG fold changes (1.5x–5x) and (2) average fractional abundances of CD8 T cells (1% to 25%) (**Fig. 4e**). High-resolution CIBERSORTx was run on all mixture samples using the LM4 signature matrix (described above) using a window of length 16 (=4*c*). To evaluate the presence of unknown mixture content (e.g., malignant cells in a tumor that are not present in the signature matrix), we repeated the analysis by adding an average of 50% HCT116, a colon cancer cell line, to each mixture dataset and by applying high-resolution CIBERSORTx with cross-platform normalization. We used the same strategy for simulating mixing coefficients as described for '*Generation and analysis of synthetic GEP admixtures*', except the mean proportions of CD8 T cells and HCT116 were set to their target abundances. We used the same HCT116 GEP dataset and expression normalization steps as described previously[2]. The area under the curve was assessed by calculating gene-level fold changes in $\log_2$ space between known DEG classes (25 samples each), and by comparing the resulting quantities between known DE genes (positives) and remaining genes (negatives). To evaluate specificity, this process was repeated for each imputed cell type GEP.

8.2. Simulated melanoma tumors with eight cell types.

Single-cell RNA-seq profiles from 19 melanoma patients were obtained as described in **Methods** ('*External datasets'*), and labeled according to seven previously annotated cell phenotypes: B cells, T cells, NK cells, macrophages, endothelial cells, cancer-associated fibroblasts, and malignant cells. As described above, T lymphocytes were subsequently divided into CD8 and CD4 T cells by the expression of *CD8A* or *CD8B* (TPM>0), for a total of *c* = 8 evaluable cell types. In order to simulate melanoma tumors, we used an approach that mirrors the distribution of cell type frequencies across dissociated tumors. Specifically, to build each synthetic tumor, we iterated through each cell type, one

at time, and selected all single cells corresponding to that cell type from a randomly chosen tumor in the original cohort. We iterated this process to create cohorts of k = 50, 100, or 200 tumors. For each synthetic tumor, single-cell GEPs in non-log linear space were aggregated by summation and TPM normalized. For each cohort, we varied: (1) the total number of cell types with artificial DEGs (d = 1, 2, 3, …, 8), and (2) the $\log_2$ fold change of DEGs (fc = 1, 2, and 3) (**Supplementary Fig. 10a**). For each value of d and fc, we created artificial DEGs by adding expression values drawn from a normal distribution to a set of 2,000 randomly selected genes, with a mean of fc and s.d. of 0.5.

Differential expression was simulated differently depending on d, the number of cell types with DEGs. For d = 1, half of the cell type GEPs were assigned to a class with artificial DEGs and half retained their baseline values. For d > 1, we sought to simulate DEGs in a manner that would make their recovery in bulk tumor transcriptomes more difficult. To do this, we set the class size equal to 25% greater than the number of tumors divided by d, and tiled the resulting DEG classes evenly across cell types. To enumerate cell proportions, we used a melanoma signature matrix derived from two tumors capturing the same eight cell types (**Supplementary Fig. 3b**). All synthetic tumors were derived from the remaining 17 held-out tumors. A window of length 32 (=4c) was used for high-resolution transcriptome purification. Accuracy was evaluated as described above for '*Synthetic tumors with five cell types*'. The $\log_2$ fold change that maximized DEG recovery had a mean of 0.11 with a 95% confidence interval of +/- 0.009.

9. Analysis of primary tumor specimens with high-resolution CIBERSORTx.

9.1. Lymphoma.

For **Figure 5a-c** (FL) and **Supplementary Figure 11** (DLBCL), we collapsed LM22 into c = 10 major leukocyte subsets following cell type enumeration and preceding transcriptome purification (**Supplementary Table 2**), and used a window of length 40 for high resolution purification (=4c). 10x-derived signature matrices (3' FL and 5' PBMCs, **Supplementary Table 2**) were run with window lengths of 12 (=4c) and 32 (=4c), respectively (**Supplementary Figs. 11d, 14a,b**).

9.2. NSCLC.

For **Figure 5d-f** and **Supplementary Figure 13c-f**, we used a window of length 20 to enumerate c = 4 cell types (**Fig. 5d**) with the signature matrix described in '*Group-mode purification of NSCLC cell subsets'* (**Supplementary Table 2**). For high-resolution purification of TCGA samples, RNA-seq profiles of LUAD and LUSC tumors (and their adjacent normal tissues) were uniformly processed and normalized[23], merged into a single expression matrix, and visualized by t-SNE for evidence of possible outliers and batch effects. Among 1,156 total specimens, 24 putative outlier samples (2%) were identified using t-SNE plots and omitted from further analysis (TCGA IDs provided in **Supplementary Table 1**). Following high-resolution purification, cell type-specific DEGs were identified between LUAD and LUSC tumor samples by a two-sided unequal variance t test applied to $\log_2$ adjusted expression data. To filter DEGs for significance, we used an empirically derived threshold ($\log_2$ fold change >0.1) that maximized the sensitivity and specificity of DEG recovery in simulated melanoma tumors (see '*Simulated melanoma tumors with eight cell types*' and **Supplementary Fig. 10**) coupled with a Q value threshold (Benjamini-Hochberg) of <0.5 for additional stringency. To confirm these results, predicted cell type-specific DEGs were cross-referenced with a validation cohort of GEPs from FACS-purified cell subsets obtained from 21 NSCLC tumors (**Fig. 5d, Supplementary Tables 1** and **4**). Only DEGs with detectable (i.e., nonzero) expression in the validation cohort are shown in Figure 5f, filtered to show only the top 300 over- or under-expressed DEGs per cell type, per condition. Additionally, to increase stringency and guard against the possibility of cellular contamination in the validation cohort, we identified highly specific expression markers of epithelial and immune subsets and censored these genes in less abundant cell types. More specifically, we identified epithelial-specific genes in the validation cohort (i.e., genes with high expression in EpCAM+ cells and low expression in the remaining cell subsets), and removed DEGs with Q<0.25 (Benjamini-Hochberg) in GEPs from immune and

stromal subsets by setting their expression to 0. We repeated this process by identifying immune-specific genes and eliminating all DEGs with $Q<0.25$ from stromal subsets. DEGs were identified using an unpaired two-sided *t* test with unequal variance applied to $\log_2$-adjusted expression data.

To further validate cell type-specific DEGs identified by CIBERSORTx, we cross-referenced them with a recently published scRNA-seq atlas of NSCLC tumors (10x Chromium platform, 3' kit)[22]. In total, we analyzed 31,984 single cells from 2 LUAD tumors, 2 LUSC tumors, and 3 adjacent normal tissues to validate histology-specific DEGs and tumor/adjacent normal DEGs (**Supplementary Fig. 15**). Normalized and quality-controlled expression data were downloaded (**Supplementary Table 1**) and analyzed without further normalization. Of the four cell types that we analyzed in our signature matrix, immune cells are unique in being comprised of many well-defined cell subsets. We therefore pooled single-cell leukocyte transcriptomes into 100 pseudo-bulk transcriptomes within each phenotypic group (LUAD or LUSC; tumor or adjacent normal) to better approximate population-level leukocyte GEPs prior to DEG validation. In each mixture, for each cell type ∂, we sampled its fraction from a Gaussian distribution with μ = the fraction of ∂ from our prior characterization of TIL composition in NSCLC[25] and $\sigma = 0.1$. Negative fractions were subsequently replaced with 0 and all fractions were rescaled within each mixture to sum to 1. Epithelial/cancer, fibroblast, and endothelial cells were analyzed directly at the single-cell level.

For each validation dataset and evaluated cell type, we calculated 'DEG concordance' as the fraction *f* of genes with the same direction of differential expression between CIBERSORTx and the validation data (higher in both or lower in both) after mean-aggregating the expression data by known phenotypic groupings (**Fig. 5f, Supplementary Figs. 15**). We then calculated an empirically defined p-value for each cell type as the number of times the DEG concordance was $\geq f$ when randomly sampling the same number of genes from the same cell type in the validation data (n = 10,000 iterations). Only genes with non-zero expression were considered for validation in the bulk RNA-seq cohort (**Fig. 5f**). Given the high sparsity of 10x Chromium v2 expression data, we pre-filtered genes in the scRNA-seq validation cohort as described in **Supplementary Fig. 15**.

To benchmark high-resolution CIBERSORTx against ISOpure[37] and DeMixT[21] in **Supplementary Fig. 13**, we applied all three methods to 22 bulk NSCLC tumor expression profiles, all of which had corresponding GEPs from FACs-purified epithelial, immune, and stromal subpopulations (**Supplementary Table 1**). To supply ISOpure and DeMixT with reference profiles of non-cancer cells (i.e., "normal cells"), we reconstituted the tumor microenvironment for each of the 22 tumor samples by combining patient-matched immune, fibroblast, and endothelial subpopulation GEPs according to their CIBERSORTx-inferred proportions in bulk tumors. We then filtered the input to only consider genes with a minimum $\log_2$ TPM value >0 within normal reference profiles. ISOpure and DeMixT were run using default parameter values from the R packages, *ISOpureR*[37] v1.1.1 and DeMixT v0.2.1 (https://github.com/wwylab/DeMixT), respectively. High resolution CIBERSORTx was run using the same 4-component signature matrix described above with a window size of 12 (=3c) for 4-component deconvolution and 8 (=4c) for 2-component deconvolution (epithelial/cancer vs. remaining cell types), as described in **Supplementary Fig. 15**. Window lengths were selected based on the criteria outlined in '*Pseudocode for high-resolution expression purification*'.

9.3. Melanoma.

For **Figure 6b**, we employed a single-cell-derived melanoma signature matrix targeting c = 8 major cell types (**Supplementary Fig. 3b**, **Supplementary Table 2**), and used a window of length 32 (=4c) for high-resolution purification. DEGs were identified and filtered as described above for NSCLC, however since we validated predicted DEGs using scRNA-seq data[3], the step to remove potential contaminants was omitted. To statistically quantify overlapping DEGs between CIBERSORTx and the scRNA-seq validation cohort, we used the same strategy as described above for NSCLC (**Supplementary Fig. 16c,d**). However, since the scRNA-seq dataset was profiled using SMART-Seq2, which captures

substantially more genes per cell than droplet-based techniques[38], all genes with non-zero expression in the scRNA-seq dataset were considered for validation (**Supplementary Fig. 16**).

10. Analysis of *PDCD1*$^+$/*CTLA4*$^+$ CD8 T cells in bulk melanomas.

CIBERSORTx was applied with cross-platform normalization to enumerate cell composition in bulk melanoma tumors from three datasets, including two profiled by RNA-seq[4,39] and one by NanoString[7]. For all deconvolutions, we applied a melanoma signature matrix designed to enumerate *PDCD1*$^+$/*CTLA4*$^+$ CD8 T cells along with eight other major cell types (**Supplementary Fig. 4**). Given the small number of genes interrogated by NanoString nCounter, overlapping genes within the signature matrix were benchmarked on melanoma tumors reconstituted from single cells and found to exhibit favorable performance (**Supplementary Fig. 16a,b**).

11. Overview of CIBERSORTx framework (same as Online Methods).

11.1. Introduction.

A number of computational methods have been proposed to infer cell type abundance, cell type-specific GEPs, or both from bulk tissue expression profiles[20,40-45]. These methods generally assume that biological mixture samples can be modeled as a system of linear equations, where a single mixture transcriptome **m** with n genes is represented as the product of **H**, a $n \times c$ cell type expression matrix consisting of expression profiles for the same n genes across c distinct cell types, and a vector **f** of size c, consisting of cell type mixing proportions.

To infer cell type abundance using this linear model within CIBERSORTx, let **M** be an $n \times k$ matrix with n genes and k mixture GEPs, let matrix **B** be a subset of **H** containing discriminatory marker genes for each of the c cell subsets (i.e., signature or basis matrix[2,46,47]), and let **M'** be the subset of **M** that contains the same marker genes as **B**. Given **M'** and **B**, the following equation can then be used to impute **F**, a $c \times k$ fractional abundance matrix with columns [$\mathbf{f}_1$,$\mathbf{f}_2$,…,$\mathbf{f}_k$]:

$$\mathbf{B} \times \mathbf{F}_{\bullet,j} = \mathbf{M}'_{\bullet \mathbf{j}}, 1 \leq j \leq k \qquad (1)$$

where $\mathbf{F}_{i,j} \geq 0$ for all i, j, the system is overdetermined (i.e., $n > c$), and expression data in **M'** and **B** are represented in non-log linear space[48]. (Note that $\mathbf{M}_{i,\bullet}$ and $\mathbf{M}_{\bullet,j}$ denote row i and column j of matrix **M**, respectively). Many methods either normalize **F** or impose an additional constraint on **F** such that for each mixture sample, the inferred mixing coefficients sum to one, allowing **F** to be directly interpreted as cell type proportions (with respect to the cell subsets in **B**)[41]. We previously introduced CIBERSORT as a method to estimate **F** using an implementation of ν-support vector regression, a machine learning technique that is robust to noise, unknown mixture content, and collinearity among cell type reference profiles[2]. CIBERSORT was used to impute **F** in this work; the normalization scheme described below in Section 12.4 was used for all cross-platform analyses, unless stated otherwise.

11.2. Group-mode expression purification.

The fractional abundance matrix **F** can be determined for the bulk GEP matrix **M**, either through expression deconvolution as described above[40,41] or with prior knowledge (e.g., by an automated hematology analyzer, or by flow cytometry)[18]. Once **F** is determined for a given **M**, a representative imputed GEP for each cell type in **F** can be estimated by solving the following system of linear equations:

$$\mathbf{H}_{i,\bullet} \times \mathbf{F} = \mathbf{M}_{i,\bullet}, 1 \leq i \leq n \qquad (2)$$

where **H** is a $n \times c$ expression matrix of n genes and c cell types, $\mathbf{H}_{i,j} \geq 0$ for all i, j, and **F** is defined as above with the constraint that relative cell fractions sum to one for each mixture sample. Like Equation 1 above, the system should be overdetermined ($k > c$), with a greater difference between k and c generally leading to improved GEP estimation (**Fig. 3e, Supplementary Fig. 6**). To ensure biologically realistic estimates of gene expression, we employ non-negative least squares regression (NNLS), an

optimization framework to solve the least squares problem with non-negativity constraints. Although NNLS is robust on simple mixtures and toy examples, its performance on more complex mixtures inherent within real tissue samples can be affected by noise, imprecision, and missing data in the linear system[2]. We therefore developed a series of novel data normalization and filtering techniques to help mitigate these issues (**Fig. 3b-d**; see Section 12.2 for additional details).

11.3. High resolution expression purification.

Despite the utility of Equation 2 for imputing cell type-specific gene expression from bulk tissues (e.g., *in silico* purifications in **Fig. 3**), it can only estimate a *single* representative GEP for each cell type. Therefore, to explore cell type-specific differentially expressed genes (DEGs) between more than one condition of interest, cell type transcriptomes must be re-generated for each condition (e.g., responders versus non-responders to a given therapy, early versus advanced stage disease, etc.). To more broadly address this issue, we extended the model in Equation 2 within a novel method for "high-resolution" *in silico* cell purification.

Our approach decomposes a matrix of bulk tissue GEPs (i.e., **M**) into *c* gene expression matrices of equal size, one for each cell subset in the cell fraction matrix **F**. Unlike previous approaches[18,49-53], this method is entirely agnostic to phenotypic class structure and can be formulated as a non-negative matrix factorization (NMF) problem with partial observations. Let **M** and **F** be defined as above ($n \times k$ and $c \times k$ matrices, respectively), and assume the latter is estimated by CIBERSORT[2]. Then, for each gene *i,* we seek to determine an expression matrix $\mathbf{G}_i$, defined here as a $k \times c$ expression matrix of *k* mixture samples by *c* cell types. Fixing **M** and **F** to solve for **G** yields the following constrained matrix decomposition problem:

$$\mathrm{diag}\left(\mathbf{G}_{i,\bullet,\bullet} \times \mathbf{F}\right) = \mathbf{M}_{i,\bullet}, 1 \leq i \leq n \qquad (3)$$

where $\mathbf{G}_{i,j,k} \geq 0$ for all *i*, *j*, *k*. Unlike the linear systems in Equations 1 and 2, there are no closed-form solutions for **G**, which is a 3D $n \times k \times c$ matrix, and existing numerical techniques are unlikely to yield biologically plausible estimates without additional constraints (e.g., regularization). We therefore developed a novel heuristic algorithm to estimate **G**, depicted schematically in **Supplementary Fig. 7** and described below.

11.3.1. Rationale for high-resolution expression purification.

Our approach for inferring **G** makes two distinct assumptions that improve the tractability of the problem while generating biologically plausible solutions. First, we assume each gene can be analyzed independently. Although ignoring gene-gene covariance relationships will likely impact the resolving power for some genes, we found this assumption to be effective in practice (e.g., **Figs. 4, 5, 6a-b, Supplementary Fig. 9**). Second, for a given gene, we assume that at least some evidence of cell type-specific differential expression is detectable in bulk tissue samples, even if statistically insignificant.

We evaluated this hypothesis using a previously published dataset of cell type-specific DEGs in the human pancreas[8]. In this dataset, the authors analyzed 5 pancreatic endocrine islet cell types (ranging from ~5% to 44% median fractional abundance within whole islets) for DEGs exhibiting >1.2 fold-change between non-diabetic normal subjects (n = 6) and patients with type 2 diabetes mellitus (T2D; n = 4)[8]. Importantly, these DEGs were identified by scRNA-seq profiling[8], allowing us to test for evidence of differential expression in bulk islets reconstructed *in silico*.

Strikingly, when grouped by known phenotypic classes (non-diabetic versus T2D), 98% of DEGs associated with T2D showed at least some fold change in the correct orientation in reconstructed islets (**Supplementary Fig. 8a**). This suggested that cell type-specific DEGs might be discernible in bulk tissues without prior knowledge of phenotypic class labels. Indeed, when we independently ordered each gene in reconstructed islets by expression levels, and split each vector evenly into high and low expression groups (i.e., 50th percentile split), diabetic patients were skewed to low or high expression for 80% of previously defined cell type-specific DEGs. Among these genes, the enrichment direction

matched the orientation of the known fold change in 99% of cases (**Supplementary Fig. 8b**). Thus, variation in bulk gene expression data can be leveraged to infer latent phenotypic class structure in the underlying cell subpopulations.

11.3.2. Pseudocode for high resolution expression purification.

Given the foundational assumptions in Section 11.3.1 along with a mixture GEP matrix **M** and cell type fractional abundance matrix **F**, we developed a heuristic algorithm for high-resolution purification. An overview of this heuristic is outlined below, with a corresponding graphical summary in **Supplementary Figure 6c** and additional algorithmic details provided in Section 12.3:

1. For a given gene *i*, estimate whether the gene is significantly expressed by at least one cell type (**Supplementary Fig. 7c**, step 1). If so, proceed to step 2. If not, proceed to the next gene.
2. Sort the gene's corresponding bulk mixture vector $\mathbf{M}_{i,\bullet}$ into ascending order (**Supplementary Fig. 7c**, step 2). Any cell type-specific DEGs with detectable signal in $\mathbf{M}_{i,\bullet}$ will influence this ordering, and the most prominent DEGs are likely to skew to one side of the distribution (e.g., **Supplementary Fig. 8b**).
3. Split the vector into two groups using a sliding window strategy (**Supplementary Figure 7c**, step 3). For each group, impute cell type-specific gene expression coefficients for *c* cell types (denoted $\mathbf{g}_1$ for group 1 and $\mathbf{g}_2$ for group 2) (**Supplementary Figure 7c**, step 3, left). Find the $\mathbf{g}_1$, $\mathbf{g}_2$ pair that best explains $\mathbf{M}_{i,\bullet}$ (**Supplementary Figure 7c**, step 3, right).
4. Perform statistical tests to determine whether the gene is significantly different between $\mathbf{g}_1$ and $\mathbf{g}_2$ for each cell type (**Supplementary Figure 7c**, step 4).
5. Refine estimates of $\mathbf{g}_1$ and $\mathbf{g}_2$ based on the significance of their differential expression (**Supplementary Figure 7c**, step 5).
6. Impute smooth expression values for each cell type using a sliding window strategy, and store as matrix $\hat{\mathbf{G}}$ (**Supplementary Figure 7c**, step 6).
7. Adjust smooth expression vectors in $\hat{\mathbf{G}}$ to maximize their agreement with $\mathbf{g}_1$ and $\mathbf{g}_2$ (**Supplementary Figure 7c**, step 7).
8. Restore the original sample ordering of $\hat{\mathbf{G}}$ (based on sample ordering in step 2; **Supplementary Figure 7c**, step 8).
9. Repeat steps 1 through 8 for all genes.

Within this framework, cell type-specific gene expression vectors are imputed using NNLS (detailed in Section 12.3 below). In order to capture variation in expression across $\mathbf{M}_{i,\bullet}$, NNLS is iteratively applied on subsets of mixture samples grouped by similar expression values. The size of each subset is governed by a sliding window of length *w*. In order to satisfy NNLS constraints and to avoid overlapping phenotypic classes, *w* is bounded by the interval $c < w \leq (k/2)$, where *c* denotes the number of cell types and *k* denotes the number of samples. We observed favorable performance of this approach across a broad range of *w* values. Nevertheless, given the marginal gains observed with increasingly large *w* values within our saturation analysis of group-mode expression purification (**Fig. 3e, Supplementary Fig. 6**), we set *w* to 4–5 fold greater than *c*, which balances performance with practical considerations based on our empirical observations. To address potential instability in the linear system and to infer expression coefficients with robust standard errors and confidence intervals, NNLS is run using bootstrapping. Additional details are provided in Section 12.3 below.

12. Details of CIBERSORTx framework.

12.1. Equations and goals.

This section extends the analytical description of CIBERSORTx above (Section 11). Equations 1–3 are copied here for convenience (also see **Supplementary Fig. 7** for a schematic depiction of these methods):

$$\mathbf{B} \times \mathbf{F}_{\bullet,j} = \mathbf{M}'_{\bullet,j}, 1 \leq j \leq k \qquad (1)$$

$$\mathbf{H}_{i,\bullet} \times \mathbf{F} = \mathbf{M}_{i,\bullet}, 1 \leq i \leq n \qquad (2)$$

$$\mathrm{diag}\left(\mathbf{G}_{i,\bullet,\bullet} \times \mathbf{F}\right) = \mathbf{M}_{i,\bullet}, 1 \leq i \leq n \qquad (3)$$

Given **B**, an $m \times c$ signature matrix (e.g., LM22[41]) consisting of m marker genes by c distinct cell types, and **M'**, an $m \times k$ mixture of GEPs consisting of the same m genes by k samples, the goal of Equation 1 is to impute **F**, a $c \times k$ matrix consisting of the fractional abundances of c cell types for each sample in **M'**.

The goal of Equation 2 is to impute **H**, an $n \times c$ matrix of representative cell type-specific expression signatures given **F** and **M,** an $n \times k$ mixture of GEPs consisting of n genes by k samples (see Section 12.2).

The goal of Equation 3 is to impute **G**, an $n \times k \times c$ matrix consisting of n genes, k samples, and c cell types, given **F** and **M** (see Section 12.3).

12.2. Imputation of group-mode expression profiles.

We employed non-negative least squares regression (NNLS) using the *nnls* package in R to solve Equation 2 for **H** given **F** and **M**. We selected NNLS since it can produce biologically plausible solutions by enforcing non-negativity constraints. Bootstrapping was incorporated to improve the reliability of the results and to compute standard errors and confidence intervals. To estimate the significance of the regression coefficients for each gene, we approximated the analytical p-value of the regression coefficients using a *t*-statistic (as performed in ordinary least squares regression)[54]. For each cell type, p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method. In cases where adjusted p-values (i.e., q-values) were highly significant (q-value < 10^{-5}, by default) for a given gene i and cell type j, we conservatively considered all other cell types with a q-value > 0.25 in gene i as insignificantly detected (expression set to 0). To further reduce confounding noise, we filtered genes based on their geometric coefficient of variation (geometric c.v.[55]), calculated using the natural logarithm of subsampled regression coefficients. We defined geometric c.v. thresholds using an adaptive approach. Specifically, for each cell type, we ranked each geometric c.v. in ascending order, then mapped each c.v. onto a two-dimensional coordinate system such that the lowest c.v. ('cv_1') became coordinate (1, cv_1), the second lowest became (2, cv_2), and so on, until the maximum c.v. ('cv_{max}') was reached, defined as the maximum of either 1 or the 75th percentile of all c.v. values. We then identified an inflection point in the c.v. curve, defined as the point farthest from (0, cv_{max}) in Euclidean space. The geometric c.v. corresponding to this threshold was used as an adaptive cell type-specific noise threshold (**Fig. 3b-e, Supplementary Fig. 6**).

12.3. Imputation of high-resolution cell type-specific expression profiles.

The algorithm proceeds in the following major steps, which parallel the corresponding schema presented in **Supplementary Figure 6c**.

12.3.1. Estimate cell type expression coefficients across all samples

For each gene i in mixture matrix **M**, our approach starts by testing whether gene i shows evidence of expression in each cell type across all samples (**Supplementary Fig. 7c**, step 1). Toward this end, we apply NNLS by bootstrapping the mixture samples with 100 iterations (by default). This yields a vector of length c consisting of the median expression value for each cell type (denoted $\mathbf{g}^{global}$). The empirical p-value for each gene/cell type pair is calculated as the number of times a coefficient is equal to zero, divided by 100. Despite the well-established advantages of this approach, it is susceptible to inaccuracies caused by inadequate sample size or insufficient bootstrap draws. For example, the latter imposes a lower bound on empirical p-values, which may be considerably underestimated for genes with otherwise exceptional statistical support (e.g., cell type marker genes). Thus, we also consider analytically derived p-values adjusted for multiple hypothesis testing, using the same approach described for Section 12.2 above. In cases where the latter is highly significant (Benjamini-Hochberg q-value < 10^{-5}, by default) for a given gene i and cell type j, all other cell types with q-value > 0.01 in gene i are considered insignificant (q-value set to 1), whereas all other q-values are left unchanged. To

combine the output of both tests, we use Fisher's method to boost or penalize p-values based on their concordance. The output of Fisher's method is a meta-score, denoted $\mathbf{p}^{global}$. Gene/cell type pairs with $\mathbf{p}^{global} < \alpha$ (=0.05, by default) are considered significantly detectable and subjected to high-resolution purification. Genes that are not significantly detectable in any cell type are omitted from further analysis.

As all remaining steps are performed on each gene independently, we focus our description below on a single gene.

12.3.2. Sorting bulk gene expression

Given a gene *i* with evidence of significant expression in at least one cell type (Section 12.3.1), we sort the gene's bulk expression vector $\mathbf{M}_{i,\bullet}$ (hereafter, denoted **m** for simplicity) in ascending order, resulting in vector $\mathbf{m}^*$ (**Supplementary Fig. 7c**, step 2). Any cell type-specific DEGs with detectable signal in the bulk mixture will influence this ordering, and the most prominent signals are likely to skew to one side of the distribution (e.g., **Supplementary Fig. 8b**).

12.3.3. Cell type expression coefficients that best explain the bulk GEP

Next, we reorder the mixture samples in **F** (denoted $\mathbf{F}^*$) to match the ordering of samples in $\mathbf{m}^*$, and define iterator *t,* which we set equal to *w*, a parameter specifying the number of mixture samples to use for imputing cell type-specific expression coefficients using NNLS. For each integer value of *t* in [*w*, (1 + *k* – *w*)], we apply NNLS to solve for cell-type coefficients twice, once for mixture samples in $\mathbf{m}^*$ spanning indices 1 to *t* and once for mixture samples in $\mathbf{m}^*$ with indices (*t* + 1) to *k*. Thus, for each value of *t,* this process yields two vectors of representative cell type-specific gene expression coefficients of length *c*, $\mathbf{g}_1$ and $\mathbf{g}_2$, respectively (**Supplementary Fig. 7c**, step 3).

Let $\hat{\mathbf{G}}^t$ be a $k \times c$ matrix consisting of *t* rows of $\mathbf{g}_1$ followed by *k*–*t* rows of $\mathbf{g}_2$.

$$\hat{\mathbf{G}}^t = \begin{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \dots \\ \mathbf{g}_1 \end{bmatrix}_{t \times c} \\ \begin{bmatrix} \mathbf{g}_2 \\ \dots \\ \mathbf{g}_2 \end{bmatrix}_{(k-t) \times c} \end{bmatrix}$$

Since $\hat{\mathbf{G}}^t$ captures cell-type-specific expression estimates with two degrees of freedom (two possible values per gene), it is an approximation of $\mathbf{G}_{i,\bullet,\bullet}$ (see Equation 3). To assess the goodness of fit between the reconstituted mixture derived from $\hat{\mathbf{G}}^t$ and $\mathbf{F}^*$ and the original sorted mixture $\mathbf{m}^*$, we calculate the residual sum of squares and determine the value of *t* that minimizes it:

$$\min_{\forall t} \left\| \mathbf{m}^* - \mathrm{diag}\left(\mathbf{F}^* \times \hat{\mathbf{G}}^t\right) \right\|_2 \qquad (4)$$

The value of *t* that minimizes Equation 4, denoted $t^{\min}$, defines the matrix $\hat{\mathbf{G}}$ and the pair of underlying expression coefficient vectors, $\mathbf{g}_1$ and $\mathbf{g}_2$, that best reconstitute $\mathbf{m}^*$ given $\mathbf{F}^*$ (**Supplementary Fig. 7c**, step 3).

12.3.4. Testing for significant differences between cell type expression coefficients

Next, each gene/cell type pair is subjected to a series of statistical tests to evaluate the null hypothesis of no differential expression between cell type-specific coefficients in $\mathbf{g}_1$ and $\mathbf{g}_2$. (**Supplementary Fig. 7c**, step 4). This is accomplished using the following two strategies:

First, we make a simplifying assumption that the distributions underlying the standard deviations of $\mathbf{g}_1$ and $\mathbf{g}_2$ (denoted sd_1 and sd_2, respectively), which are obtained via bootstrapping (as described in Section 12.3.1), are Gaussian. It is then trivial to calculate *Z*-scores for differential expression as the difference between $\mathbf{g}_1$ and $\mathbf{g}_2$ divided by the square root of the sum of the squares of sd_1 and sd_2.

Resulting *Z*-scores are converted to two-sided p-values, one for each cell type, and stored as a vector $\mathbf{p}_1$.

Second, we test for differential expression using a permutation analysis. Here, we shuffle the columns (i.e., samples) of **F*** (denoted $\mathbf{F}^{rand}$), and then solve for cell type-specific coefficients using NNLS applied to $\mathbf{F}^{rand}$ and **m*** on sample indices from (t^{min} + 1) to *k* (these are the same indices used to define $\mathbf{g}_2$). This yields a randomized version of $\mathbf{g}_2$ with length *c*, denoted $\mathbf{g}^r_2$. We save the difference vector, $|\mathbf{g}^r_2 - \mathbf{g}_1|$, and repeat the process 50 times (by default). The null distribution of difference vectors is then compared to the test statistic, $|\mathbf{g}_2 - \mathbf{g}_1|$, and empirical p-values are calculated as the fraction of randomized difference vectors exceeding this test statistic. The resulting p-values are stored as vector $\mathbf{p}_2$.

Since the above tests evaluate differential expression in different ways, we combine $\mathbf{p}_1$ and $\mathbf{p}_2$ into a meta-score $\mathbf{p}^{deg}$ using Fisher's method.

12.3.5. Refinement of cell type-specific expression coefficients: part I

Because some gene/cell type pairs may not be detectably expressed ($\min(\mathbf{p}^{global}) \geq \alpha$) or variably expressed ($\min(\mathbf{p}^{deg}) \geq \alpha$), we update our estimates of the global coefficients $\mathbf{g}^{global}$ and differential cell type expression vectors ($\mathbf{g}_1$, $\mathbf{g}_2$) accordingly (**Supplementary Fig. 7c**, step 5). First, we define **F**** ← **F*** and set fractional abundance vectors in **F**** to zero for cell types with insignificantly detectable expression across all samples ($\mathbf{p}^{global} \geq \alpha$). We then recalculate $\mathbf{g}^{global}$ with **F**** using the same procedure described in Section 12.3.1. To update $\mathbf{g}_1$ and $\mathbf{g}_2$, we use **F**** to identify a new optimal cut point t^{min} and corresponding matrix $\hat{\mathbf{G}}$ using Equation 4 above (Section 12.3.3), and then set **F*** ← **F****.

12.3.6. Refinement of cell type-specific expression coefficients: part II

Next, we seek to adaptively reduce the difference between paired expression coefficients in $\mathbf{g}_1$, $\mathbf{g}_2$ in a manner that reflects our confidence in their differential expression (**Supplementary Fig. 7c**, step 5). To accomplish this, we first set $\mathbf{v} = -\log_{10}(\mathbf{p}^{diff})$, where $\mathbf{p}^{diff}$ is equal to $\mathbf{p}^{deg}$ (Section 12.3.4). Since **v** is in logarithm space, with higher values reflecting higher significance in $\mathbf{p}^{diff}$, we divide **v** by max(**v**) (if $\max(\mathbf{v}) > -\log_{10}(\alpha)$), where the maximum value of **v** is calculated separately for cell types with higher expression in $\mathbf{g}_1$ than in $\mathbf{g}_2$, and vice versa. Values greater than 1 are set to 1. This yields a new vector **v***, which is used to reduce the fold change between paired expression coefficients in $\mathbf{g}_1$, $\mathbf{g}_2$ by adding (or subtracting) the following quantity from $\mathbf{g}_1$, $\mathbf{g}_2$:

$$\Delta = (1 - (\mathbf{v}^*)^2) \frac{|\mathbf{g_1} - \mathbf{g_2}|}{2}$$

In this way, paired expression coefficients in $\mathbf{g}_1$, $\mathbf{g}_2$ will become more similar when our confidence in their differential expression is lower.

If no cell types show evidence of variable expression (i.e., $\min(\mathbf{p}^{deg}) \geq \alpha$), we use an alternative metric to calculate **v**. In this case, $\mathbf{p}^{diff}$ is a vector of length *c* that consists of the statistical significance (p-value) of the Pearson correlation between each cell type's inferred proportions in **F*** and an expression vector **m**** of length *k*. For each cell type *j*, we calculate **m**** using the following scalar-vector multiplication:

$$\mathbf{m}^{**} = \mathbf{m}^* - \sum_{\forall i \neq j} \left(\mathbf{g}_i^{global} \times \mathbf{F}_{\bullet,i}\right)$$

This is akin to a partial correlation and allows for the estimation of expression variability in cell type *j* while controlling for the effect of the remaining cell types on **m***.

After adjustment, paired expression coefficients in $\mathbf{g}_1$ and $\mathbf{g}_2$ with a fold change < 1.1 (by default) are either set to their corresponding coefficients in $\mathbf{g}^{global}$ (by default) or NA. Although we empirically determined that a cutoff of 1.1 delivers a reasonable tradeoff between sensitivity and specificity (e.g., **Supplementary Fig. 10**), the fold change threshold can be modified for greater sensitivity at the

expense of specificity (cutoff <1.1) or vice versa (cutoff >1.1). Following this adjustment step, we re-estimate $\hat{\mathbf{G}}$ using Equation 4.

12.3.7. Smoothing of cell type-specific expression profiles

Despite the improvements to $\hat{\mathbf{G}}$ thus far, it is still restricted to at most two distinct values per gene/cell type pair. To impute smooth expression coefficients with multiple degrees of freedom, we apply NNLS to impute a set of cell type-specific coefficients (denoted **g**') using **F*** and **m***, where sample indices are selected using a sliding window of length *w* for all possible window positions in **m*** (**Supplementary Fig. 7c**, step 6).

Let **Q** be a $(1+k-w) \times c$ matrix consisting of **g'** vectors for all possible windows:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{g}'_1 \\ \mathbf{g}'_2 \\ \dots \\ \mathbf{g}'_{1+k-w} \end{bmatrix}_{(1+k-w)\times c}$$

Because **Q** has *w*–1 fewer rows than $\hat{\mathbf{G}}$, we create two versions of **Q**, one with *w*–1 copies of row 1 added to the top of **Q**, denoted **Q'**, and one with *w*–1 copies of row *c* added to the bottom of **Q**, denoted **Q''**. We then average these two matrices, setting $\mathbf{Q} \leftarrow (\mathbf{Q}' + \mathbf{Q}'')/2$. Although this procedure results in smooth expression vectors harboring *k* degrees of freedom, it relies on the assumption that the mixture samples within each window are sufficiently similar with minimal cell type-specific expression variation. In other words, sample-specific expression differences for a given cell type are averaged out within a given window, despite being captured across windows. This issue can be partially addressed by decreasing the size of *w*, albeit at the risk of numerical instability within the regression framework.

Additionally, since bulk expression in **m*** is sorted, windows that cover regions with low expression may lead to imprecise estimates owing to inadequate information content. The same issue can arise from windows of insufficient size.

To help address these potential problems, we adjust expression coefficients in **Q** based on the final values of $\mathbf{g}_1$ and $\mathbf{g}_2$ within $\hat{\mathbf{G}}$ (**Supplementary Fig. 7c**, step 7) Specifically, for each cell type *j*, we individually fit the expression coefficients in **Q** to $\hat{\mathbf{G}}$ using ordinary least squares regression. This allows $\mathbf{g}_1$ and $\mathbf{g}_2$ to serve as anchor points and to ensure that the directionality of cell type-specific expression differences (i.e., high to low or low to high) across the ordered mixture is maintained. We employ linear least squares regression for this step, though other approaches could be used. The resulting matrix is denoted **Q'**.

12.3.8. Final steps

Finally, for each gene, we restore the original ordering of mixture samples in **Q'**, which we save as $\hat{\mathbf{G}}$ (**Supplementary Fig. 7c**, step 8). This process is repeated for all genes in **M**.

12.4. Cross-platform normalization schemes for deconvolution.

12.4.1. Background

Although several cross-platform normalization techniques have been developed to minimize technical variation in expression profiling experiments, to our knowledge, only two studies have applied this idea to gene expression deconvolution[56,57]. One study applied cubic splines with four degrees of freedom to model gene-level technical variation between paired microarray and RNA-seq profiles from TCGA tumor samples[56]. Although useful when paired samples are available, both new and archival expression datasets generally lack such pairs, underscoring the need for a more general approach. A second study[57] applied ComBat[36], an empirical Bayesian method, to eliminate technical variation between cell reference profiles and bulk tissue GEPs. Although ComBat was developed specifically for batch effect removal, the batches should ideally represent GEPs from the same tissue type; otherwise, the

approach can introduce major distortions in cell abundance estimation (as in Figure 2f-h in Newman et al.[58]).

There is no previous approach that can be applied to eliminate technical variation between mixture sample expression profiles (denoted **M**) and a signature matrix comprised of cell type reference profiles (denoted **B**) while preserving biological signal. This is because, unlike technical batches of the same sample type, both biological differences and technical batches are inherently conflated in the setting of deconvolution. Since previous techniques for addressing technical dropout and cross-platform normalization in bulk GEP and scRNA-seq data are not directly applicable to this problem[36,59-61], and since technical variation between the signature matrix and mixture samples can variably and profoundly confound deconvolution results, we developed a novel approach for cross-platform deconvolution (**Supplementary Fig. 1**). By exploiting the fact that **M** can be modeled as a linear combination of **B**, and by estimating **M** from **B** (denoted **M***), batch correction can be directly applied to **M** and **M***. This helps to ensure that sample types are comparable between batches, thus avoiding estimation artifacts and the need for paired samples profiled on both platforms.

Our approach is comprised of two distinct normalization strategies (B-mode and S-mode) to more reliably apply gene expression deconvolution across platforms and tissue storage types (e.g., fresh/frozen versus FFPE). The former adjusts the mixture samples while the latter adjusts the signature matrix. A decision tree to guide users in selecting the most appropriate strategy is provided in **Supplementary Figure 1a**.

12.4.2. B-mode.

Bulk-mode batch correction (B-mode) removes technical differences between a signature matrix derived from bulk sorted reference profiles (e.g., bulk RNA-seq or microarrays) and an input set of mixture samples (**Supplementary Fig. 1c**). The technique can also be applied to signature matrices derived from scRNA-seq platforms, provided that transcripts are measured analogously to bulk mixture GEPs (e.g., full-length transcripts without UMIs profiled by SMART-Seq2).

Given a set of mixture samples **M**, the approach creates a series of estimated mixture samples $\mathbf{M}^*$, where the latter consists of a linear combination of imputed cell type proportions in **M** along with the corresponding signature matrix profiles (in non-log linear space). Although the strategy is general and can flexibly accommodate different deconvolution and batch correction methods, we used ComBat[36] (an empirical Bayesian method) to minimize technical variation between **M** and $\mathbf{M}^*$ after $\log_2$-adjustment. Following this step, cell proportions are re-estimated using the adjusted mixture samples in non-log linear space. In addition to cell type enumeration, this approach can also be applied to cell type-specific gene expression purification, allowing for down-weighting of genes whose expression levels are low or absent from the collection of cell types in the signature matrix. An absolute minimum of 3 mixture samples is required to perform the batch correction procedure, though at least 10 is recommended. We found that application of this strategy to LM22, a microarray-derived signature matrix, results in improved deconvolution performance across multiple datasets and platforms (**Supplementary Fig. 2d**).

12.4.3. S-mode.

Despite the benefits of the above technique, deconvolution may fail when excessive technical variation is present (e.g., when cell types that are expected to be present are observed to "drop out" in bulk GEP samples after deconvolution). In this study, we observed this phenomenon when using signature matrices derived from 10x Chromium without proper normalization (**Fig. 2b**, **Supplementary Fig. 1d,l, 2a-c**). This discrepancy was not surprising given the major differences in transcriptome representation between UMI-based and 3'/5'-biased methods, such as 10x, and those that capture full transcripts without UMIs, such as SMART-Seq2 (Figure S3E in Ziegenhain et al.[38]). Because B-mode requires an initial round of fractional abundance estimates prior to normalization, it cannot address cellular dropout and is insufficient to overcome excessive variation (**Supplementary Fig. 1d**). We therefore developed a second strategy for single-cell batch correction (S-mode), tailored for signature matrices derived from

droplet-based or UMI-based scRNA-seq techniques, including 10x Chromium, and other challenging datasets (**Supplementary Fig. 1b**).

Like B-mode, the primary objective of S-mode is to obtain a cell frequency matrix **F** from a set of mixture GEPs **M** containing *k* samples, while minimizing technical variation. However, unlike B-mode, S-mode adjusts the signature matrix rather than the mixture samples.

To apply S-mode, we first seek to obtain initial estimates of cell frequencies $\mathbf{F}^*$ within a set of mixture samples **M** given (1) a signature matrix **B** containing *c* cell types, and (2) the set of single-cell reference profiles **R** (*n* genes $\times$ *r* single cells) from which **B** was derived. To accomplish this, we create a series of *k* artificial mixture profiles $\mathbf{M}^*$ reconstituted from single-cell GEPs within **R**. Mixing coefficients are determined according to a Normal distribution N(μ, σ) for each cell type separately, where μ is set to the fractional abundance of each cell type from **B** in **R** $\{\mu_1, \mu_2, \ldots, \mu_c\}$ and σ is set to $\{2\mu_1, 2\mu_2, \ldots, 2\mu_c\}$. Although these parameters were found to be favorable in empirical analyses, S-mode is robust to moderate variation in their values (**Supplementary Fig. 1k,l**). Once specified, cell type-specific mixing coefficients are randomly drawn from N(μ, σ) to create a matrix $\mathbf{F}^*$ consisting of *c* cell types $\times$ *k* samples. After setting negative values in $\mathbf{F}^*$ to 0, the resulting matrix is normalized across cell types such that the sum of coefficients for each mixture is equal to 1. Next, single-cell transcriptomes from each cell type are sampled according to their proportions in $\mathbf{F}^*$ and aggregated into *k* bulk expression profiles in TPM space, yielding $\mathbf{M}^*$. ComBat is subsequently applied to **M** and $\mathbf{M}^*$ in $\log_2$ space, yielding adjusted mixtures $\mathbf{M}^{adj}$ and $\mathbf{M}^{adj*}$, respectively.

After converting $\mathbf{M}^{adj*}$ to non-log linear space, both $\mathbf{M}^{adj*}$ and $\mathbf{F}^*$ are used as inputs to NNLS in order to reconstruct cell type expression coefficients for each gene in **B** (see Equation 2 above). The result is an adjusted signature matrix, $\mathbf{B}^{adj}$. Notably, unlike group-mode expression purification (**Fig. 1**), our use of NNLS in S-mode does not require adaptive noise filtration. This is because (1) the cell mixing proportions of each artificial mixture in $\mathbf{M}^{adj*}$ are derived from $\mathbf{F}^*$ and thus known with 100% certainty, and (2) ComBat only applies a linear transformation to each gene's $\log_2$ expression profile, thereby preserving the relative ordering of bulk expression values. Thus, the conditions for applying NNLS to propagate the adjustments in $\mathbf{M}^{adj*}$ to **B** are mathematically ideal. Indeed, we found that $\mathbf{B}^{adj}$ can significantly improve concordance between a given signature matrix and the cell type reference profiles underlying a set of input mixtures **M**, regardless of whether S-mode is applied to mixtures profiled by RNA sequencing or microarrays (**Supplementary Fig. 1e-j**). Once $\mathbf{B}^{adj}$ is obtained, the cell frequency matrix **F** is estimated from **M** using Equation 1 (above) and CIBERSORTx. We found that application of this strategy to four 10x-derived signature matrices and three bulk tissue GEP datasets resulted in significantly improved deconvolution performance (**Supplementary Fig. 2a-c**).

Of note, although NNLS requires a greater number of mixture samples than cell types (see '*Details of CIBERSORTx framework'* and Equation 2 above), for datasets that fail to directly meet this requirement, we create additional pseudo-mixtures from **R** by repeating the random sampling process above until $k > c$. Each additional pseudo-mixture is then paired with a randomly selected mixture sample in **M** prior to S-mode normalization.

Supplementary Note 2

Tissue dissociation-related considerations

There are two major factors underlying dissociation-related artifacts that can produce substantial distortions in the cellular representation of a bulk tissue: (1) the differential susceptibility of single cells to mechanical dissociation or enzymatic tissue digestion, and (2) biases in single-cell recovery rates due to variation in cell sorting parameters (e.g., sheath fluid or core stream flow rate, nozzle size, etc.). In this work, we characterized dissociation-induced skewing in melanomas and pancreatic islets by comparing cell type enumeration by scRNA-seq, bulk RNA-seq deconvolution, and IHC. We identified a major depletion of macrophages (melanomas) and pancreatic beta cells (islets) in single-cell data, with reciprocal changes in other cell types. As observed for melanomas, we would expect similar dissociation-related artifacts in single cell analyses of other solid tumor types, including HNSCC (**Fig. 2c,d**) and NSCLC (**Fig. 5d-f**). In fact, we observed a significant loss of plasma cells in NSCLC tumor cell suspensions (flow cytometry) as compared to CIBERSORT deconvolution of bulk NSCLC tumors and ISH in a prior study[25]. Such results are consistent with the widely recognized fragility of certain cell types including plasma cells[62].

The availability of paired scRNA-seq and bulk RNA-seq data from the same pancreatic islet specimens provided an opportunity to address the impact of tissue dissociation in a single experimental system (**Supplementary Fig 3h**). We found that deconvolution results of bulk RNA-seq data for individual islet cell types (right) were remarkably well-correlated with fractions determined by scRNA-seq (median cell-level r = 0.83, **Supplementary Fig 3h**). Although this suggests that the linearity of cell proportions may only be modestly impacted by tissue dissociation, this observation is anecdotal and will need to be rigorously evaluated across tissues, dissociation conditions, and sorting parameters.

Advantages of scRNA-seq for deconvolution

Single-cell RNA-seq data provides several advantages for CIBERSORTx deconvolution. These include: (1) the ability to create signature matrices that capture major cell types in a tissue without having to design and optimize antibody panels to sort each cell type, (2) the ability to study poorly understood or unknown transcriptional states, and (3) the ability to internally validate cell type reference profiles on simulated tissues reconstituted from single-cell GEPs. Although we found that reference signatures can generalize across tissue types that share common cell lineages (**Fig. 2e, Supplementary Fig. 2**), we recommend that users tailor the signature matrix to the tissue type of interest when tissue-specific expression data are available.

Supplementary References

1. Allantaz, F.*, et al.* Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. *PLoS One* **7**, e29979 (2012).
2. Newman, A.M.*, et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).
3. Tirosh, I.*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
4. Van Allen, E.M.*, et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-211 (2015).
5. Puram, S.V.*, et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624 (2017).
6. Akbani, R.*, et al.* Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696.
7. Chen, P.-L.*, et al.* Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discovery* (2016).
8. Segerstolpe, A.*, et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab* **24**, 593-607 (2016).
9. Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696 (2015).
10. Bland, J.M. & Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307-310 (1986).
11. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255-268 (1989).
12. Ame-Thomas, P.*, et al.* CD10 delineates a subset of human IL-4 producing follicular helper T cells involved in the survival of follicular lymphoma B cells. *Blood* **125**, 2381-2385 (2015).
13. Nakaya, H.I.*, et al.* Systems biology of vaccination for seasonal influenza in humans. *Nat Immunol* **12**, 786-795 (2011).
14. Lenz , G.*, et al.* Stromal Gene Signatures in Large-B-Cell Lymphomas. *New England Journal of Medicine* **359**, 2313-2323 (2008).
15. Compagno, M.*, et al.* Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459**, 717-721 (2009).
16. Jourdan, M.*, et al.* An in vitro model of differentiation of memory B cells into plasmablasts and plasma cells including detailed phenotypic and molecular characterization. *Blood* **114**, 5173-5181 (2009).
17. Milpied, P.*, et al.* Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. *Nature Immunology* **19**, 1013-1024 (2018).
18. Shen-Orr, S.S.*, et al.* Cell type-specific gene expression differences in complex tissues. *Nat Methods* **7**, 287-289 (2010).
19. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
20. Quon, G.*, et al.* Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine* **5**, 29 (2013).
21. Wang, Z.*, et al.* Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, 451-460 (2018).
22. Lambrechts, D.*, et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med* (2018).
23. Tatlow, P.J. & Piccolo, S.R. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *Sci Rep* **6**, 39259 (2016).

24. Gentles, A.J.*, et al.* Integrating Tumor and Stromal Gene Expression Signatures With Clinical Indices for Survival Stratification of Early-Stage Non-Small Cell Lung Cancer. *J Natl Cancer Inst* **107**(2015).
25. Gentles, A.J.*, et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* **21**, 938-945 (2015).
26. Green, M.R.*, et al.* Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proceedings of the National Academy of Sciences* **112**, E1116-E1125 (2015).
27. Kiaii, S.*, et al.* Follicular lymphoma cells induce changes in T-cell gene expression and function: potential impact on survival and risk of transformation. *J Clin Oncol* **31**, 2654-2661 (2013).
28. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotech* **33**, 495-502 (2015).
29. Jin, H., Wan, Y.-W. & Liu, Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics* **18**, 117 (2017).
30. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
31. Soneson, C., Love, M. & Robinson, M. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research* **4**, 1521 (2016).
32. Newman, A.M.*, et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine* **20**, 548 (2014).
33. Scherer, F.*, et al.* Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. *Science Translational Medicine* **8**, 364ra155-364ra155 (2016).
34. Boddupalli, C.S.*, et al.* Interlesional diversity of T cell receptors in melanoma with immune checkpoints enriched in tissue-resident memory T cells. *JCI Insight* **1**(2016).
35. Zheng, G.X.*, et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
36. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
37. Anghel, C.V.*, et al.* ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics* **16**, 156 (2015).
38. Ziegenhain, C.*, et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631-643.e634 (2017).
39. Nathanson, T.*, et al.* Somatic Mutations and Neoepitope Homology in Melanomas Treated with CTLA-4 Blockade. *Cancer Immunology Research* (2016).
40. Shen-Orr, S.S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr Opin Immunol* **25**, 571-578 (2013).
41. Newman, A.M. & Alizadeh, A.A. High-throughput genomic profiling of tumor-infiltrating leukocytes. *Curr Opin Immunol* **41**, 77-84 (2016).
42. Aran, D., Hu, Z. & Butte, A.J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 220 (2017).
43. 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**, e26476 (2017).
44. Angelova, M.*, et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biology* **16**, 64 (2015).
45. Becht, E.*, et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17**, 218 (2016).
46. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics* **17 Suppl 1**, S279-287 (2001).

47. Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H.F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* **4**, e6098 (2009).
48. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat Methods* **9**, 8-9; author reply 9 (2012).
49. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**, 2211-2212 (2013).
50. Liebner, D.A., Huang, K. & Parvin, J.D. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics* **30**, 682-689 (2014).
51. Zhong, Y., Wan, Y.W., Pang, K., Chow, L.M. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).
52. Zuckerman, N.S., Noam, Y., Goldsmith, A.J. & Lee, P.P. A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS Comput Biol* **9**, e1003189 (2013).
53. Onuchic, V.*, et al.* Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Reports* **17**, 2075-2086 (2016).
54. Sheather, S. *A Modern Approach to Regression with R*, (Springer-Verlag New York, 2009).
55. Koopmans, L.H., Owen, D.B. & Rosenblatt, J.I. Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika* **51**, 25-32 (1964).
56. Charoentong, P.*, et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Reports* **18**, 248-262 (2017).
57. Li, B.*, et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol* **17**, 174 (2016).
58. Newman, A.M., Gentles, A.J., Liu, C.L., Diehn, M. & Alizadeh, A.A. Data normalization considerations for digital tumor dissection. *Genome Biology* **18**, 128 (2017).
59. Bacher, R.*, et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods* **14**, 584 (2017).
60. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411 (2018).
61. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* **36**, 421 (2018).
62. Cogbill, C.H.*, et al.* Morphologic and cytogenetic variables affect the flow cytometric recovery of plasma cell myeloma cells in bone marrow aspirates. *International Journal of Laboratory Hematology* **37**, 797-808 (2015).