Supplementary Information

Table of Contents

1. TOTAL MRNA EXPRESSION IN SINGLE-CELL RNA SEQUENCING DATA

1.1. Datasets

Colorectal cancer single-cell RNA sequencing data

Two fresh colorectal adenocarcinoma samples of primary tumor were collected from patients who were receiving chemotherapies by surgical resection at the University of Texas MD Anderson Cancer Center (**Table S1**). Single-cell data preparation was achieved by using the Chromium Single Cell 3' Library, Gel Bead & Multiplex Kit, and Chip Kit (v3, 10x Genomics). Libraries were sequenced on an Illumina NovaSeq6000. The analysis of alignment, tagging, and gene and transcript counting were conducted by using the 10x Genomic Cell Ranger pipeline (version 3.0).

Liver cancer single-cell RNA sequencing data[1]

Three fresh hepatocellular carcinoma samples of primary tumor were collected at NIH Clinical Center for immune checkpoint inhibition studies (NCT01313442) (**Table S1**). Two of them (patient 1 and patient 2) were from patients who were receiving immunotherapies by needle biopsy, and the other one was from an untreated patient by surgical resection. Single-cell data preparation was conducted by using the Chromium Single Cell 3' Library, Gel Bead & Multiplex Kit, and Chip Kit (v2, 10x Genomics). Libraries were sequenced on an Illumina NextSeq500. The analysis of alignment, tagging, and gene and transcript counting were conducted by using the 10x Genomic Cell Ranger pipeline (version 2.0.2).

Lung cancer single-cell RNA sequencing data[2]

Two fresh lung adenocarcinoma samples of primary, non-metastatic lung tumor were collected from untreated patients by surgical resection at University Hospital Leuven (**Table S1**). Single-cell data preparation was conducted by using the Chromium Single Cell 3' Library, Gel Bead & Multiplex Kit, and Chip Kit (v1, 10x Genomics). Libraries were sequenced on Illumina HiSeq4000. The analysis of alignment, tagging, and gene and transcript counting were conducted by using the 10x Genomic Cell Ranger pipeline (version 2.0.0).

Pancreatic cancer single-cell sequencing data[3]

Two untreated patients with primary pancreatic cancer were recruited at the University of Texas MD Anderson Cancer Center and informed written consents following institutional review board approval were obtained (Lab00-396 and PA15-0014). Fresh biopsies were collected from the tumors by fine needle aspiration (**Table S1**). Single-cell data preparation was achieved by using the Chromium Single Cell 3' Library, Gel Bead & Multiplex Kit, and Chip Kit (v1, 10x Genomics). Libraries were sequenced on an Illumina NextSeq500. The analysis of alignment, tagging, and gene and transcript counting were conducted by using the 10x Genomic Cell Ranger pipeline (version 3.1).

1.2. Single-cell RNA sequencing data processing

In this section, we first introduce the preprocessing for the single-cell RNA sequencing (scRNAseq) datasets described above, including quality control, cell clustering, cell type annotation. It is followed by a method to group cell clusters within a cell type based on gene counts, i.e., the total number of expressed genes, to simplify the characterization of heterogeneity within the cell type, and a scale normalization method to correct for sequencing or experimental biases on total UMI counts. The cell cycle state for each cell in tumor cell clusters is scored based on canonical marker genes.

SI Table 1. Marker genes used to annotate cell types in scRNAseq patient samples from four cancer types.

	Colorectal adenocarcinoma[5]	Hepatocellular carcinoma[1]	Lung adenocarcinoma[2]	Pancreatic adenocarcinoma[3,7,8]
B cell	*CD79A*, *CD38*	*CD79A*, *SLAMF7*, *BLNK*	*CD79A*, *IGKC*, *IGLC3*	*CD79A*, *CD38*
T cell	*CD2*, *CD3E*, *CD3D*	*CD2*, *CD3E*, *CD3D*	*CD3D*, *TRBC1*, *TRBC2*	*CD2*, *CD3D*
NK cell				*NKG7*, *KLRF1*
Myeloid	*CD14*, *CD68*, *ITGAX*	*CD14*, *CD163*, *CD68*	*LYZ*, *MARCO*, *CD68*	*CD14*, *CD68*
Fibroblast	*COLA1A*, *COL1A2*, *COL3A1*	*COL1A2*, *FAP*, *PDPN*	*COLA1A*, *DCN*, *COL1A2*	*COLA1A*, *COL1A2*
Endothelial	*PECAM1*, *VWF*, *ENG*	*PECAM1*, *VWF*, *ENG*	*CLDN5*, *FLT1*, *CDH5*	
Alveolar			*FOLR1*, *AQP4*, *PEBP4*	
Epithelial	*EPCAM*, *KRT18*, *KRT20*		*CAPS*, *TEME190*, *PIFO*, *SNTN*	*EPCAM*, *KRT18*, *KRT20*
Tumor cell			*LCN2*, *CCL20*, *PTTG1*	

Quality control, clustering and cell type annotation.

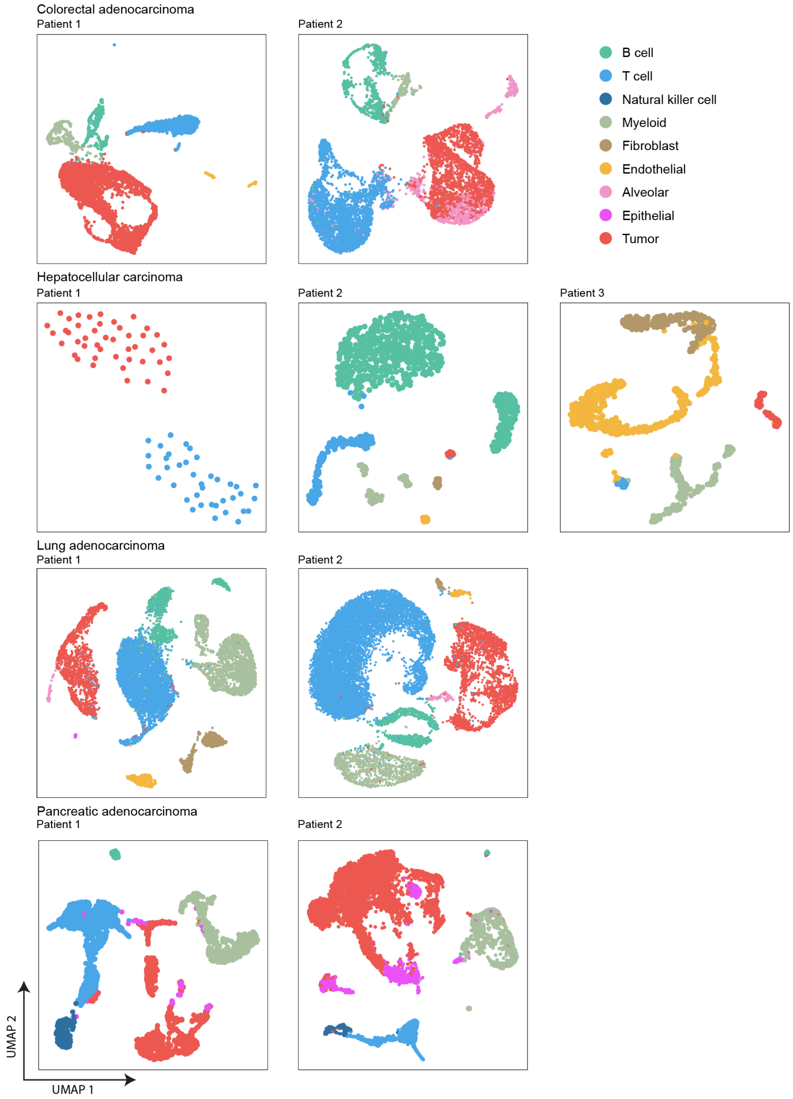
For each of the two colorectal adenocarcinoma scRNAseq samples generated at MD Anderson, genes expressed in less than three cells were removed. Cells were filtered out that have either fewer than 500 total UMIs, below 200 expressed genes, or more than 50% total UMI counts derived from mitochondrial genes. The total number of transcripts in each cell was normalized to 10,000, which was followed by a natural log transformation. The highly variable genes were detected and used for principal component analysis (PCA). Cells were then clustered with the Seurat package[4]. The cell type of each cell was annotated based on known marker genes[5] (**SI Figure 1**, **SI Table 1**). Initial somatic copy number variation (CNV) estimates were made using inferCNV[6], based on which CNV scores and CNV correlation scores[1] were calculated. The CNV score of a single cell was defined as the sum of the squared copy number variations across all gene positions. The CNV correlation score was calculated as the correlation between the copy number variations of a single cell and the average copy number variation of the top 2% cells ranked by CNV scores from the same sample. Tumor cells were identified as epithelial cells with an average CNV score greater than 0.0015. The two samples from patient 1 and patient 2 had 5,422 and 7,462 cells remaining, respectively, after data pre-processing.

The quality control of the three hepatocellular carcinoma scRNAseq patient samples was conducted consistently with the study[1]. For each sample, genes expressed in less than 0.1% of the cells were removed. Cells with fewer than 700 total UMIs, fewer than 500 expressed genes, or more than 20% total UMI counts derived from mitochondrial genes were excluded. An additional quality control step of doublet removal was performed based on the number of cells loaded and recovered. The total number of transcripts in each cell was normalized to 10,000, followed by a natural log transformation. The highly variable genes were detected and used for the PCA. Cells were then clustered with the Seurat package[4]. The cell type of each cell was annotated based on known marker genes[1] (**SI Figure 1**, **SI Table 1**). Tumor cells were identified as epithelial cells with CNV scores above the 80th percentile and CNV correlation scores above 0.4. The three samples of patient 1, patient 2 and patient 3) had 83, 761 and 796 cells remaining, respectively, after data pre-processing.

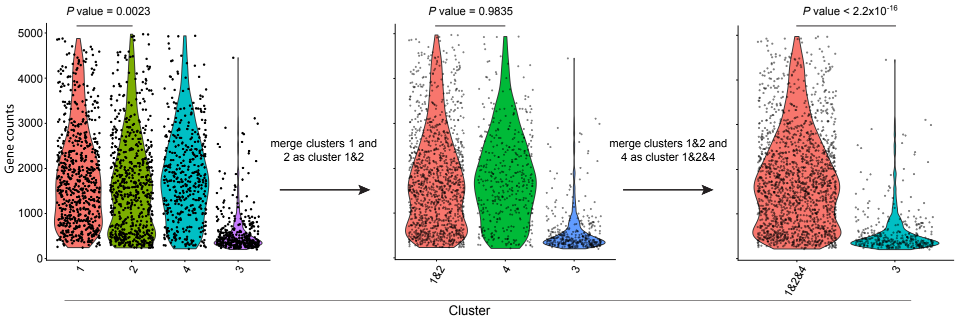
The quality control of the two lung adenocarcinoma scRNAseq patient samples was conducted consistently with the study[2]. For each sample, genes expressed in less than 0.5% of the cells were removed. Any cell with either fewer than 201 total UMI counts, below 101 or over 6,000 expressed genes, or more than 10% total UMI counts derived from mitochondrial genes were filtered out from downstream analysis. The total number of transcripts in each cell was normalized to 10,000, followed by a natural log transformation. The highly variable genes were detected and used for the principal component analysis (PCA). Cells were then clustered with the Seurat package[4]. The cell type (including tumor cell) of each cell was annotated based on known marker genes[2] (**SI Figure 1**, **SI Table 1**). The two samples of patient 1 and patient 2 had 8,845 and 13,658 cells remaining, respectively, after data pre-processing.

For each of the two pancreatic adenocarcinoma scRNAseq samples, genes expressed in less than three cells were removed. Cells were filtered out that have either fewer than 500 total UMIs, below 200 expressed genes, or more than 50% total UMI counts derived from mitochondrial genes. The total number of transcripts in each cell was normalized to 10,000, followed by a natural log transformation. The highly variable genes were detected and used for the PCA. Cells were then clustered with the Seurat package[4]. The cell type of each cell was annotated based on known marker genes[7,8] (**SI Figure 1**, **SI Table 1**). Tumor cells were identified as epithelial cells with CNV score above 0.015 and CNV correlation above 0.4. The two samples of patient 1 and patient 2 had 2,404 and 7,037 cells remaining after QC, respectively, after data pre-processing.

Within each cell type, we further merged clusters that are not significantly different in gene counts (Wilcoxon rank-sum test, α=0.001) (**SI Figure 2**).



SI Figure 1. UMAPs of scRNAseq data from four cancer types.



SI Figure 2. **An example of merging cell clusters by gene counts.** Tumor cells in patient 2 of colorectal adenocarcinoma are used. The initial 4 clusters were determined by Seurat clustering (resolution=0.5). Wilcoxon rank-sum tests comparing gene counts were performed between clusters and those that did not pass the significance level of 0.001 were merged. The resulting two tumor cell clusters had 1,696 cells (low UMI cluster, e.g. 1&2&4) and 359 cells (high UMI cluster, e.g. 3), respectively. We repeated this process based on the initial Seurat clustering with resolution=1.0. There were still two tumor cell clusters after merging. The differences of tumor cells in the high UMI cluster and in the low UMI cluster based on the two resolutions were only 12 cells and 13 cells, respectively.

Normalized total UMI counts

We performed scale normalization on the raw count data to ensure the total unique molecular identifiers (UMI) count per cell across all cells are comparable for different samples. Specifically, let UMI_i = $\{UMI_{igc}\}_{GxCi}$ be a matrix of raw UMI counts for the scRNAseq data for sample i being investigated, with genes g on the rows and cells c on the columns. G denotes the total number of genes, C_i is the number of cells in sample i. Then, the normalized UMI matrix UMI_i, denoted as UMI_i^{norm}, is calculated as $UMI_i^{norm}=UMI_i/r_i$, where, $r_i = \frac{UMI_i^{sum}/C_i}{baseline}$, $baseline = median\{UMI_1^{sum}/C_1, UMI_2^{sum}/C_2, \ldots, UMI_n^{sum}/C_n\}$, $UMI_i^{sum}=\sum_{c=1}^{C_i}\sum_{g=1}^{G} UMI_{igc}$.

Given a cell cluster, we let u_{gc} denote the amount of mRNA of gene g in cell c. The average total mRNA amount per cell is $\sum_{c=1}^{C}(\sum_{g=1}^{G} u_{gc})/C$. For scRNAseq data, we assume the UMI_{gc} from gene g, cell c is proportional to the total mRNA u_{gc} of gene g in that cell, with a constant k_g that represents technical effects: $UMI_{gc}=k_g * u_{gc}$. The constant k_g is introduced because every single-cell sequencing platform presents a <100% capture efficiency for mRNA, and such efficiency varies across different platforms[9]. Under the assumption that the technical effect k_g remains constant across cells and is often evaluated as an average effect across genes within the same platform, we can evaluate total mRNA expression in the scRNAseq data using the average total UMI counts, which is $\sum_{c=1}^{C}(\sum_{g=1}^{G} UMI_{gc})/C$. Notably, we observed strong correlations between gene counts and total UMI across cells in each cell cluster across all cell types and cancer types (**Fig. S1B**). This observation supports our assumption of a stable technical effect

k_g within each study, and that the average total UMI counts serve as a reasonable surrogate to compare total mRNA expression across cells that are generated from the same experiment.

The average gene counts and average total UMI counts for both individual cell clusters and all the clusters pooled within a cell type are summarized in **SI Table 2**.

SI Table 2. The average gene counts and average total UMI counts for both individual cell clusters and all the clusters pooled. The 95% CI is estimated using bootstrapping with 1,000 iterations.

Cancer type	Patient id	Cell cluster	Cell type									
			Tumor		Epithelial		Alveolar		Endothelial		Fibroblast	
			Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)
Colorectal adenocarcinoma	Patient 1 (Stage IVA, PFS = 4 months)	Cluster 3	5,649 (5,526, 5,779)	48,706 (46,811, 50,691)	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 2	2,438 (2,325, 2,561)	13,787 (12,661, 15,063)	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 1	646 (620, 675)	1,929 (1,805, 2,058)	NA	NA	NA	NA	NA	NA	NA	NA
		Pooled	1,926 (1,785, 2,074)	13,626 (11,990, 15,228)	NA	NA	NA	NA	2,135 (2,041, 2,239)	7,378 (6,845, 7,899)	NA	NA
	Patient 2 (Stage IVA, PFS ≥ 8 months)	Cluster 2	1,782 (1,710, 1,848)	7,307 (6,910, 7,700)	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 1	604 (573, 638)	1,964 (1,800, 2,142)	NA	NA	NA	NA	NA	NA	NA	NA
		Pooled	1,576 (1,506, 1,642)	6373 (5988, 6781)	1,272 (1,225, 1,326)	4,455 (4,199, 4,721)	NA	NA	NA	NA	NA	NA
Hepatocellular carcinoma	Patient 1 (Stage IV, PFS < 5 months)	Cluster 2	5,338 (5,252, 5,425)	48,457 (47,098, 49,861)	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 1	1,364 (1,325, 1,405)	4,670 (4,402, 4,954)	NA	NA	NA	NA	NA	NA	NA	NA
		Pooled	3,660 (3,524, 3,800)	29,969 (28,109, 31,625)	NA	NA	NA	NA	NA	NA	NA	NA
	Patient 2* (Stage IV, PFS ≥ 18 months)	Pooled	1,871 (1,787, 1,949)	7,961 (7,471, 8,460)	NA	NA	NA	NA	1,760 (1,734, 1,787)	4,150 (4,037, 4,259)	1,947 (1,906, 1,986)	5,255 (5,089, 5,415)
	Patient 3 (Stage I, PFS ≥ 18 months)	Cluster 2	NA	NA	NA	NA	NA	NA	4,149 (4,064, 4,241)	16,410 (15,796, 17,043)	NA	NA
		Cluster 1	NA	NA	NA	NA	NA	NA	2,368 (2,306, 2,429)	7,131 (6,813, 7,462)	NA	NA
		Pooled	2,921 (2,876, 2,966)	13,289 (12,897, 13,647)	NA	NA	NA	NA	2,708 (2,634, 2,788)	8,904 (8,447, 9,380)	1,961 (1,918, 2,002)	5,699 (5,491, 5,922)
Lung adenocarcinoma	Patient 1 {Stage IIIB}	Cluster 2	3,999 (3,921, 4,073)	15,664 (15,180, 16,128)	NA	NA	NA	NA	NA	NA	1,663 (1,612, 1,713)	4,179 (3,995, 4,375)
		Cluster 1	649 (616, 682)	1,230 (1,132, 1,341)	NA	NA	NA	NA	NA	NA	717 (675, 767)	1,680 (1,498, 1,869)
		Pooled	1,869 (1,761, 1,979)	6,489 (5,952, 7,050)	3,233 (3,156, 3,314)	9,846 (9,502, 10,176)	724 (692, 754)	1,479 (1,394, 1,569)	1,371 (1,314, 1,432)	3,200 (2,989, 3,417)	1,520 (1,464, 1,574)	3,801 (3,612, 3,993)
	Patient 2 {Stage IIB}	Cluster 2	2,778 (2,680, 2,871)	8,458 (8,041, 8,868)	NA	NA	2,097 (2,039, 2,160)	6,123 (5,856, 6,411)	NA	NA	1,898 (1,836, 1,968)	5,179 (4,899, 5,486)
		Cluster 1	831 (784, 880)	1,703 (1,535, 1,897)	NA	NA	586 (561, 612)	1,091 (1,024, 1,168)	NA	NA	649 (623, 676)	1,255 (1,190, 1,321)
		Pooled	1612 (1515, 1703)	4,411 (4,058, 4,763)	NA	NA	1200 (1134, 1266)	3,135 (2,917, 3,360)	684 (652, 715)	1,316 (1,232, 1,401)	1,128 (1,071, 1,188)	2,760 (2,549, 2,985)
Pancreatic adenocarcinoma	Patient 1 (Stage IV, PFS = 11 months)	Cluster 2	4,315 (4,205, 4,421)	21,718 (20,860, 22,550)	2,549 (2,437, 2,654)	11,066 (10,341, 11,818)	NA	NA	NA	NA	NA	NA
		Cluster 1	1,510 (1,439, 1,578)	4,631 (4,314, 4,938)	616 (588, 645)	1,491 (1,393, 1,599)	NA	NA	NA	NA	NA	NA
		Pooled	3,323 (3,193, 3,458)	15,675 (14,823, 16,549)	1,423 (1,334, 1,527)	5,489 (4,933, 6,075)	NA	NA	NA	NA	NA	NA
	Patient 2 (Stage IIB, PFS ≥ 26 months)	Cluster 2	2,235 (2,160, 2,306)	8,896 (8,437, 9,317)	1,382 (1,324, 1,442)	6,017 (5,635, 6,386)	NA	NA	NA	NA	NA	NA
		Cluster 1	997 (959, 1,039)	3,381 (3,173, 3,581)	779 (728, 831)	2,496 (2,246, 2,765)	NA	NA	NA	NA	NA	NA
		Pooled	1,614 (1,542, 1,682)	6,129 (5,715, 6,486)	1,086 (1,028, 1,142)	4,286 (3,965, 4,644)	NA	NA	NA	NA	NA	NA

SI Table 2. (Continued)

Cancer type	Patient id	Cell cluster	Cell type							
			Myeloid		T cell		Natural killer cell		B cell	
			Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)	Average gene counts (95% CI)	Average total UMI counts (95% CI)
Colorectal adenocarcinoma	Patient 1 (Stage IVA, PFS = 4 months)	Cluster 3	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 2	2,984 (2,920, 3,054)	13,050 (12,518, 13,573)	NA	NA	NA	NA	NA	NA
		Cluster 1	535 (523, 548)	1,415 (1,372, 1,457)	NA	NA	NA	NA	NA	NA
		Pooled	1,787 (1,691, 1,877)	7,365 (6,853, 7,920)	1,211 (1,177, 1,243)	3,600 (3,473, 3,743)	NA	NA	1,102 (1,058, 1,149)	5,422 (5,012, 5,873)
	Patient 2 (Stage IVA, PFS ≥ 8 months)	Cluster 2	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 1	NA	NA	NA	NA	NA	NA	NA	NA
		Pooled	662 (627, 700)	2,054 (1,881, 2,231)	1,203 (1,177, 1,230)	4,377 (4,250, 4,505)	NA	NA	963 (911, 1,011)	3,633 (3,330, 3,936)
Hepatocellular carcinoma	Patient 1 (Stage IV, PFS < 5 months)	Cluster 2	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 1	NA	NA	NA	NA	NA	NA	NA	NA
		Pooled	NA	NA	1,318 (1,285, 1,350)	3,720 (3,561, 3,879)	NA	NA	NA	NA
	Patient 2 (Stage IV, PFS ≥ 18 months)	Pooled	1,406 (1,376, 1,434)	4,228 (4,085, 4,368)	1,303 (1,272, 1,340)	3,221 (3,079, 3,366)	NA	NA	1,105 (1,088, 1,122)	11,686 (11,426, 11,965)
	Patient 3 (Stage I, PFS ≥ 18 months)	Cluster 2	NA	NA	NA	NA	NA	NA	NA	NA
		Cluster 1	NA	NA	NA	NA	NA	NA	NA	NA
		Pooled	1,602 (1,571, 1,634)	5,879 (5,711, 6,057)	1,410 (1,373, 1,448)	4,590 (4,412, 4,782)	NA	NA	NA	NA
Lung adenocarcinoma	Patient 1 {Stage IIIB}	Cluster 2	1,293 (1,253, 1,329)	3,936 (3,776, 4,102)	NA	NA	NA	NA	NA	NA
		Cluster 1	393 (381, 407)	876 (838, 917)	NA	NA	NA	NA	NA	NA
		Pooled	1,233 (1,191, 1,271)	3,732 (3,573, 3,889)	584 (566, 602)	1,050 (1,011, 1,090)	NA	NA	544 (520, 568)	2,569 (2,278, 2,858)
	Patient 2 {Stage IIB}	Cluster 2	1,207 (1,164, 1,251)	3,504 (3,302, 3,721)	NA	NA	NA	NA	NA	NA
		Cluster 1	361 (352, 370)	728 (699, 754)	NA	NA	NA	NA	NA	NA
		Pooled	1,137 (1,089, 1,183)	3,275 (3,080, 3,482)	765 (745, 789)	1362 (1310, 1417)	NA	NA	762 (732, 796)	4,396 (4,083, 4,746)
Pancreatic adenocarcinoma	Patient 1 (Stage IV, PFS = 11 months)	Cluster 2	3,213 (3,131, 3,292)	16,221 (15,618, 16,851)	NA	NA	NA	NA	NA	NA
		Cluster 1	1,460 (1,420, 1,504)	3,840 (3,688, 3,980)	NA	NA	NA	NA	NA	NA
		Pooled	1,788 (1,726, 1,851)	6,158 (5,746, 6,568)	1,531 (1,508, 1,554)	4,814 (4,716, 4,922)	1,651 (1,630, 1,671)	4,064 (4,001, 4,126)	1,352 (1,327, 1,378)	4,119 (4,022, 4,213)
	Patient 2 (Stage IIB, PFS ≥ 26 months)	Cluster 2	2,210 (2,155, 2,271)	10,285 (9,860, 10,748)	NA	NA	NA	NA	NA	NA
		Cluster 1	890 (857, 926)	2,523 (2,344, 2,711)	NA	NA	NA	NA	NA	NA
		Pooled	1,960 (1,891, 2,030)	8,818 (8,327, 9,245)	936 (919, 953)	2,526 (2,467, 2,586)	1,169 (1,148, 1,191)	2,855 (2,763, 2,950)	927 (904, 950)	2,684 (2,586, 2,775)

PFS: progression free survival.

*: for a patient, if all cell types have one cluster each, only the results from the pooled cells of each cell type are shown.

NA: due to no cells or only one cell cluster in the corresponding cell type; for the latter, the results of gene counts and total UMI counts are shown in the "Pooled" position.

The observed fold changes in total UMI counts between tumor cell clusters were significantly higher than those expected from expression dosage response from genome ploidy changes alone (at 2-3 fold[10,11]) among tumor cells (**SI Table 3**). For the two tumor cell clusters in each patient across four cancer types, our null hypothesis is that there is no difference between the distribution of the total UMI counts from the tumor cell high-UMI cluster and the distribution of the total UMI counts from the tumor cell low-UMI cluster multiplied by three. For each patient, the *P* value was obtained with a Wilcoxon rank-sum test and adjusted by the Benjamini-Hochberg method[12]; the 95% confidence interval for the ratio of means of total UMI counts from the respective tumor cell cluster was estimated using bootstrapping with 1,000 iterations.

SI Table 3. T-tests of total UMI counts between two tumor cell clusters within each patient across four cancer types.

Cancer type	Patient 1		Patient 2	
	P value	μ_2/μ_1(95% CI)*	*P* value	μ_2/μ_1(95% CI)
Colorectal adenocarcinoma	$1\text{x}10^{-11}$	25 (23, 27)	$6\text{x}10^{-10}$	3.7 (3.4, 4.2)
Hepatocellular carcinoma	$7\text{x}10^{-7}$	10 (10,11)	NA	NA
Lung adenocarcinoma	$< 2\text{x}10^{-16}$	13 (12, 14)	$< 2\text{x}10^{-16}$	5.0 (4.4, 5.6)
Pancreatic adenocarcinoma	$2\text{x}10^{-15}$	4.7 (4.3, 5.0)	0.009	2.6 (2.4, 2.8)

*μ_2 and μ_1 are the means of the total UMI counts from the tumor cell high-UMI cluster and tumor cell low-UMI cluster, respectively.

We also examined the cell cycle phase for each cell in the tumor clusters (**SI Table 4**) using the Seurat package. Cluster and patient numbers match **Fig. 1B**.

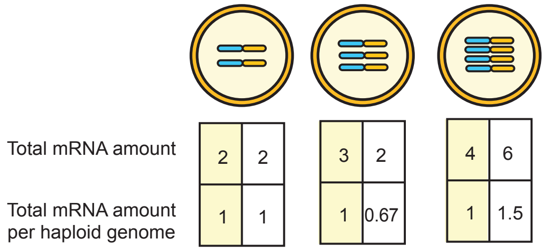
SI Table 4. Cell cycle states of the tumor cell clusters across four cancer types. The 95% confidence intervals for the odds ratios were calculated using Fisher's exact tests.

Cancer type	Patient id	Tumor cell cluster	Cell cycle phase		Odds Ratio for G1/S in high-UMI tumor cell cluster (95% CI)
			No. of cells in G1/S	No. of cells in G2/M	
Colorectal adenocarcinoma	Patient 1	Cluster 3	711	97	Cluster 3 vs. Cluster 1: 2.5 (2.0, 3.2) Cluster 2 vs. Cluster 1: 0.67 (0.49, 0.94) Cluster 3 vs. Cluster 2: 3.8 (2.6, 5.5)
		Cluster 2	124	64	
		Cluster 1	1800	626	
	Patient 2	Cluster 2	1342	363	0.48 (0.33, 0.68)
		Cluster 1	318	41	
Hepatocellular carcinoma	Patient 1	Cluster 2	25	1	10.9 (1.1, 549.6)
		Cluster 1	13	6	
Lung adenocarcinoma	Patient 1	Cluster 2	465	32	11.2 (7.6, 17.0)
		Cluster 1	489	378	
	Patient 2	Cluster 2	632	429	0.18 (0.15, 0.22)
		Cluster 1	1414	172	
Pancreatic adenocarcinoma	Patient 1	Cluster 2	430	32	2.1 (1.3, 3.6)
		Cluster 1	247	39	
	Patient 2	Cluster 2	1537	392	0.33 (0.27, 0.41)
		Cluster 1	1790	152	

2. TUMOR-SPECIFIC TOTAL MRNA EXPRESSION IN BULK SEQUENCING DATA

2.1. A mathematical model for tumor-specific total mRNA expression

2.1.1. Model



SI Figure 3. **Illustration of ploidy-adjusted total mRNA amount per cell.** Example of three types of cells, with ploidy = 2, 3, and 4. Under the scenario of linear dosage effects, as shown in the boxes with a yellow background, suppose their corresponding total mRNA amounts are 2, 3, and 4, then the ploidy-adjusted, or per haploid genome, total mRNA amount would be 1, 1, and 1. Under the scenario of dosage compensation, the second cell has a total mRNA amount of 2 and a per haploid value of 0.67. Under the scenario of dosage transgression, the third cell has a total mRNA amount of 6 and a per haploid genome value of 1.5.

For any group of cells, we use a ploidy-adjusted *GTL* (*pGTL*) to denote the average global mRNA transcript level per cell per haploid genome, which follows $pGTL = \sum_{c=1}^{C} (\sum_{g=1}^{G} u_{gc}/p_c)/C$ (**SI Figure 3**). Here p_c is the ploidy, i.e., the number of copies of the haploid genome in cell *c*. However, the cell level ploidy p_c is usually not measurable. Hence, in practice, we use average ploidy ψ of the corresponding cell group to approximate it: $pGTL \approx \sum_{c=1}^{C} \sum_{g=1}^{G} u_{gc}/(C\psi)$. For non-tumor cells, which are commonly diploid, this assumption is assured.

In the analysis of bulk RNAseq data from mixed tumor samples, we are interested in comparing tumor with non-tumor cell groups. We denote tumor cells by group *T*, and non-tumor cells by group *N*. Therefore, we define a tumor-specific total mRNA expression score (TmS) to reflect the ratio of total mRNA transcript level per haploid genome of tumor cells to that of the surrounding non-tumor cells, i.e., $TmS_{tumor} = pGTL_T / pGTL_N$, simplified as TmS from here forward. It is necessary to calculate this ratio in order to cancel out technical effects presented in sequencing data that confound with both $pGTL_T$ and $pGTL_N$. Let $T_g = \sum_{c=1}^{C_T} u_{gc}$ and $N_g = \sum_{c=1}^{C_N} u_{gc}$ denote the total number of transcripts of gene g across all cells from tumor and non-tumor cells, let C_T and C_N denote the number of tumor and non-tumor cells, and let ψ_T and ψ_N represent the average ploidy of tumor and non-tumor cells, respectively. Under the assumption that the tumor cells have a similar ploidy, we can derive TmS without using single-cell-specific parameters as

$$TmS = \frac{\sum_{g=1}^{G} T_g/(C_T\psi_T)}{\sum_{g=1}^{G} N_g/(C_N\psi_N)}. \qquad \text{Eq.S1}$$

Here we further introduce a tumor-specific total mRNA expression proportion $\pi = (\sum_{g=1}^{G} T_g) / (\sum_{g=1}^{G} T_g + \sum_{g=1}^{G} N_g)$ and a tumor cell proportion of tumor cells (termed "tumor purity") $\rho = C_T / (C_T + C_N)$. Note that deconvolution of just the gene expression data will not provide information on the total number of tumor cells and non-tumor cells, but only the sum of total mRNA expression across all cells of each cell type.

Using these deconvolution parameters, we rewrite Eq.S1 as

$$TmS = \frac{\psi_N \pi (1-\rho)}{\psi_T \rho (1-\pi)}. \qquad \text{Eq.S2}$$

Additionally, we can define a ploidy-unadjusted *TmS* by removing the ploidy terms.

2.1.2. Estimation

It is a common practice to assume the ploidy of non-tumor cells ψ_N equals to 2[13,14]. Hence, we have

$$\widehat{TmS} = \frac{2\hat{\pi}(1-\hat{\rho})}{\hat{\psi}_T \hat{\rho}(1-\hat{\pi})}. \qquad \text{Eq.S3}$$

In what follows, we use *TmS* to represent $\widehat{TmS}$, for the sake of simplicity.

Estimation of tumor-specific total mRNA transcript level π using high-throughput RNA sequencing has not been possible due to several technical and analytical factors including: 1) the need to account for technical artifacts introduced by varied library size, which currently involves normalization procedures across samples; 2) total mRNA transcripts per cell are confounded with technical artefacts so that normalization procedures adjust for both effects at once, consequently losing the ability to evaluate the downstream global transcriptome feature[15]; and 3) a limited focus on estimating cell proportions by popular methods[16–18].

Using deconvolution to partition tumor and non-tumor cells within the same sample under the same experimental conditions provides a mathematical means to cancel out the effect of technical artefacts while maintaining the effect of cell-type-specific total mRNA counts. We use the DeMixT model[19] to estimate tumor-specific total mRNA expression proportions. For sample *i* and across any gene *g*, we have

$$Y_{ig} = \pi_i T'_{ig} + (1-\pi_i) N'_{ig} \qquad \text{Eq.S4}$$

where Y_{ig} represents the scale normalized expression matrix from mixed tumor samples, T'_{ig} and N'_{ig} represent the normalized relative expression of gene *g* within tumor and surrounding non-tumor cells, respectively. The estimated tumor-specific total mRNA expression proportions $\hat{\pi}$ is the desirable quantity for Eq.S3.

Computational deconvolution methods, e.g., ASCAT[14] and ABSOLUTE[13], have been developed to perform allele-specific copy number analysis and to estimate tumor purity ρ and ploidy ψ_T from tumor DNA sequencing data. Such statistical methods jointly model the distribution of *logR* and *B* allele (or variant allele) frequency (BAF) across germline SNPs, with tumor purity and allele-specific copy number as parameters of interest. Then the tumor purity and ploidy (the average tumor copy number) can be estimated through minimizing the loss function or maximizing the likelihood. Below, we provide a detailed description for these methods using the ASCAT model as an example.

Sequence read counts at known SNP loci were computed from tumor DNA sequencing data. The $logR_i$ can be computed from the total read counts in the tumor versus normal for the *i*th SNP, which provides information on the ratio of total copy number between the tumor and the normal. Specifically, $logR_i$ can be expressed as[14]

$$logR_i = \gamma log_2\left(\frac{2(1\text{-}\rho)+\rho\left(n_{A,i}+n_{B,i}\right)}{2(1\text{-}\rho)+\psi_T}\right), \qquad \text{Eq.S5}$$

where ρ is the tumor purity, ψ_T is the tumor ploidy, γ is a constant depends on which DNA sequencing technology is used. $n_{A,i}$ and $n_{B,i}$ stand for the allele-specific copy number of A allele and B allele for the i^{th} SNP in tumor cells, respectively.

On the other hand, allelic imbalance can be inferred from the BAF_i for i^{th} SNP. The BAF_i can be expressed as[14]

$$BAF_i = \frac{1\text{-}\rho+\rho n_{B,i}}{2(1\text{-}\rho)+\rho\left(n_{A,i}+n_{B,i}\right)}. \qquad \text{Eq.S6}$$

Based on Eq.S5 and Eq.S6, the allele-specific copy number can be expressed as a function of the tumor purity and ploidy. Specifically, we have

$$\hat{n}_{A,i} = \frac{\rho\text{-}1+2^{\frac{logR_i}{\gamma}}(1-BAF_i)(2(1\text{-}\rho)+\psi_T)}{\rho};$$

$$\hat{n}_{B,i} = \frac{\rho\text{-}1+2^{\frac{logR_i}{\gamma}}BAF_i(2(1\text{-}\rho)+\psi_T)}{\rho}.$$

Allele-specific piecewise constant fitting (ASPCF)[14] was then applied to both $logR_i$ and BAF_i simultaneously, which enforced the change points to occur at the same genomic locations. Consequently, a segmentation of the genome was obtained, each segment corresponding to a genomic region between two adjacent change points. Using the ASPCF smoothed $logR_i$ and BAF_i, the final values for $\hat{\rho}$ and $\hat{\psi}_T$ were obtained through the optimization, such that the allele-specific copy number estimates $\hat{n}_{A,i}$ and $\hat{n}_{B,i}$ were as close to nonnegative integers as possible for germline heterozygous SNPs.

2.2. Improved estimation using DeMixT

Many computational deconvolution methods have been developed to estimate the cell type proportions through transcriptome data; however, most of them focus on the cellular proportion and not the global gene expression level of each cell type, due to lack of appropriate normalization approaches. The DeMixT[19] model is unique in aiming to estimate the global tumor-specific gene expression level relative to the normal reference in the context of admixed tumor samples. ISOpure[20] is the other model that presents similar objectives as the DeMixT model. The following issues and our proposed solution are generally applicable to both models.

The identifiability analysis of model parameters is a major issue for high dimensional models. Due to technical limitations, given a certain amount and quality of experimental data, not all model parameters are guaranteed for unambiguous estimation. Frequently, only a subset of model parameters are identifiable based on the available data, with the rest of the parameters considered unidentifiable. Confidence intervals can be derived for identifiable parameters, which contain the true value of the parameter with a desired probability[21]. Fortunately, with the DeMixT model, there is hierarchy in model identifiability in which the cell-type specific global gene expression proportions π are the most identifiable parameters, requiring only a subset of genes with identifiable expression distributions. Therefore, our goal is to select an appropriate set of genes as input to DeMixT that optimizes the estimation of the tumor-specific mRNA expression proportions. In general, genes are expressed at different levels, which, due to different numerical ranges, can affect tumor-specific global gene expression proportion estimation. We found that including genes that are not differentially expressed between the tumor and non-tumor components within the bulk sample, or genes with large variance in expression within the non-tumor component, can introduce large biases into the estimated tumor-specific mRNA expression proportions. By applying a profile likelihood approach to detect the identifiability of model parameters[22], we systematically evaluated the identifiability for all available genes based on the data, and selected the most identifiable genes for the estimation of proportions. As a result, the accuracy of the estimated proportions has been improved. As a general method, the profile likelihood-based gene selection strategy can be extended to any method that uses maximum likelihood estimation. Furthermore, we employed an additional virtual spike-in strategy to improve model identifiability.

2.2.1. Likelihood model for DeMixT

In the DeMixT model[19] (Eq.S4), we assumed that the observed expression level Y_{ig} is a linear combination of two hidden components T_{ig} (tumor, in place of T'_{ig} from now on) and N_{ig} (non-tumor, in place of N'_{ig} from now on), where gene $g = 1,2,\ldots,G$, sample $i = 1,2,\ldots,S$, and π_i is the tumor-specific mRNA expression proportions. We assume each hidden component follows the $\log_2$-normal distribution, i.e., $T_{ig} \sim LN\left(\mu_{Tg}, \sigma^2_{Tg}\right)$ and $N_{ig} \sim LN\left(\mu_{Ng}, \sigma^2_{Ng}\right)$.

Fitting the deconvolution model in Eq.S1 can be formally defined as an optimization problem that seeks to identify optimal estimates for sample-level, tumor-specific mRNA expression proportions π_i, and gene-level parameters. Denote the full parameter set $(\boldsymbol{\pi},\boldsymbol{\mu_T},\boldsymbol{\sigma_T})$, where $\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_S)$, $\boldsymbol{\mu_T} = (\mu_{T1}, \mu_{T2},\ldots, \mu_{TG})$, $\boldsymbol{\sigma_T} = (\sigma_{T1}, \sigma_{T2},\ldots,\sigma_{TG})$. The full log-likelihood of the DeMixT model can be written as

$$l(\boldsymbol{\pi},\boldsymbol{\mu_T},\boldsymbol{\sigma_T})=\textstyle\sum_{i=1}^{S} \sum_{g=1}^{G} log(f(Y_{ig}|\pi_i,\mu_{Tg},\sigma_{Tg})),$$

where $f\left(Y_{ig}|\pi_i,\mu_{Tg},\sigma_{Tg}\right)=\frac{1}{2\pi\sigma_{Ng}\sigma_{Tg}}\int_0^{Y_{ig}}\frac{1}{t(Y_{ig}-t)}exp(-\frac{(log2(t)-\mu_{Ng}-log2(1-\pi_i))^2}{2\sigma^2_{Ng}}-\frac{(log2\left(Y_{ig}-t\right)-\mu_{Tg}-log2(\pi_i))^2}{2\sigma^2_{Tg}})dt.$

The DeMixT model applies an optimization method, iterated conditional modes (ICM)[23], to maximize the full log-likelihood function and estimate all distribution parameters $(\boldsymbol{\mu_T},\boldsymbol{\sigma_T})$ and proportions $\boldsymbol{\pi}$.

2.2.2. Optimized model identifiability

Based on the most stringent definition, for a parametric model $l(\boldsymbol{Y} \mid \theta)$, θ is identifiable if, $l(\boldsymbol{Y} \mid \theta_1) = l(\boldsymbol{Y} \mid \theta_2) => \theta_1 = \theta_2$. However, this rigorous identifiability is difficult to validate for a general high-dimensional and non-convex model, which is the case of the DeMixT model. Thus, for a parameter θ, we use the confidence interval $[\theta^-, \theta^+]$ to measure its identifiability[22].

In the DeMixT model, if we select genes with small confidence intervals of μ_{Tg} based on profile likelihood, which indicate high identifiability, the corresponding gene g will be more stable and reliable, so will the inferred tumor-specific mRNA expression proportions ($\boldsymbol{\pi}$). As a result, the length of confidence interval of μ_{Tg} serves as an estimable quantity with which we can evaluate the gene g's identifiability and prioritize genes to increase the estimation quality of π_i, μ_{Tg}, σ_{Tg}.

The profile likelihood is preferred to compute confidence intervals of parameters that often have better small-sample properties than those based on asymptotic standard errors calculated from the full likelihood[24]. Assume the kth gene's mean parameter μ_{Tk} is the parameter of interest. The definition of the profile likelihood function of μ_{Tk} is:

$$l_{\mu_{Tk}}(\mu_{Tk}=x|\boldsymbol{\pi},\boldsymbol{\mu_T},\boldsymbol{\sigma_T}) = max\{\sum_{i=1}^{S}[\sum_{g\neq k}^{G} log\ \left(f\left(\pi_i,\mu_{Tg},\sigma_{Tg}\right)\right) + log\ \left(f(\pi_i,\mu_{Tk}=x,\sigma_{Tk})\right)\ ]\}$$

The confidence interval of a profile likelihood function can be constructed through inverting a likelihood-ratio test[25]. Assume the null hypothesis as H$_0$: $\mu_{Tk} = x$, and the maximum likelihood estimator of (π_i, μ_{Tg}, σ_{Tg}) are ($\hat{\pi}_i$, $\hat{\mu}_{Tg}$, $\hat{\sigma}_{Tg}$). The null hypothesis will not be rejected at the α level of significance if and only if $2\left[l(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{T}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{T}}) - l_{\mu_{Tk}}(\mu_{Tk} = x \mid \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{T}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})\right] \leq \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ stands for $1-\alpha$ percentile of χ^2 distribution with a degree of freedom equal to 1. Since maximized likelihood $l(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{T}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})$ and model parameters $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{T}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{T}}$ can be estimated by running the DeMixT model on all available gene sets, for any given x, we are able to investigate the profile log-likelihood function $l_{\mu_{Tk}}(\mu_{Tk} = x \mid \hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})$. Consequently, we can estimate the lower and upper bound of confidence interval $[\mu^-_{Tk}, \mu^+_{Tk}]$ as

$$\mu^-_{Tk} = \min_x \{x|2\left[l(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}}) - l_{\mu_{Tk}}(\mu_{Tk} = x \mid \hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})\right] \leq \chi^2_{1-\alpha}(1)\}$$
$$\mu^+_{Tk} = \max_x \{x|2\left[l(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}}) - l_{\mu_{Tk}}(\mu_{Tk} = x \mid \hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})\right] \leq \chi^2_{1-\alpha}(1)\}$$

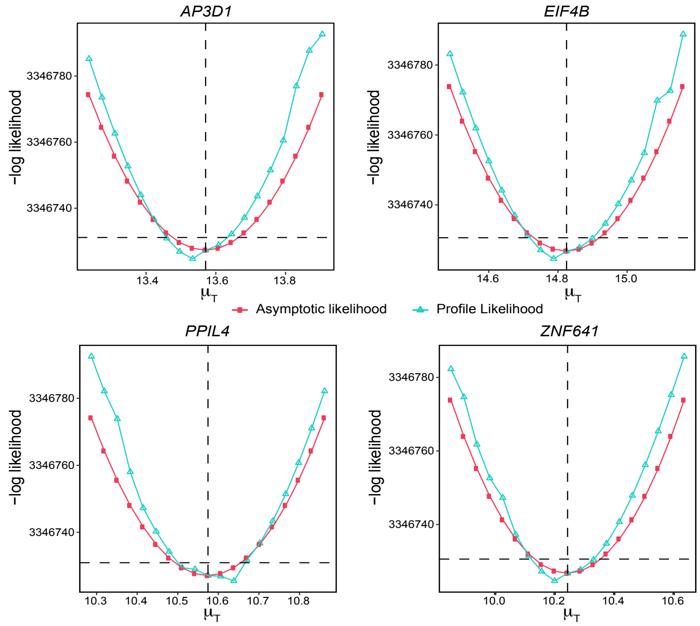
Following the same procedure, we can derive the confidence interval of μ_{Tk} for all available genes.

In real data analysis, the actual profile likelihood function of μ_{Tk} is intractable and prone to noise, since the algorithm can be easily trapped by local minimal solutions when calculating the profile likelihood. In addition, calculating the actual profile likelihood function of all μ_{Tk} across 20,000 genes is generally infeasible due to computational limits. An asymptotic approximation is necessary in order to quickly evaluate the profile likelihood function. If the measurement noise is small and the sample size is large enough, asymptotic confidence intervals are good approximations of the actual confidence intervals[22]. The asymptotic profile likelihood function can be derived from the observed Fisher information of the log likelihood, denoted as $H(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})$. Then the asymptotic α level confidence interval of μ_{Tk} can be written as follows[22]

$$\mu_{Tk}^{\pm} = \widehat{\mu_{Tk}} \pm \sqrt{2\chi^2_{1-\alpha}(1)\ H(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})^{-1}_{k,k}}\,. \qquad \text{Eq.S7}$$

We compared the actual profile likelihood function with the asymptotic profile likelihood function for a random set of 20 genes in real data (the TCGA prostate adenocarcinoma dataset) and observed good performance of the approximation profile likelihoods (**SI Figure 4**). With 35 randomly selected genes, we calculated the root mean squared error (RMSE) between the confidence intervals from the true and asymptotic profile likelihoods as 0.05.

Under the assumption that the maximum likelihood estimator is the global minimum of the full likelihood, the asymptotic profile likelihood-based confidence interval of μ_{Tk}, as shown in Eq.S7, can be used to measure the identifiability of the corresponding gene k.



SI Figure 4. Asymptotic profile likelihoods for 4 genes using 259 samples from the TCGA prostate cancer dataset. Comparison of asymptotic and actual profile likelihoods of μ_T for 4 randomly selected genes in the TCGA prostate adenocarcinoma data. The red curve shows the true profile likelihood of the corresponding parameter. The blue curve shows an asymptotic approximation of the profile likelihood of the corresponding parameter.

We hence introduce a metric, the gene selection score, as the width of an asymptotic profile likelihood-based 95% confidence interval of μ_{Tk} for gene k

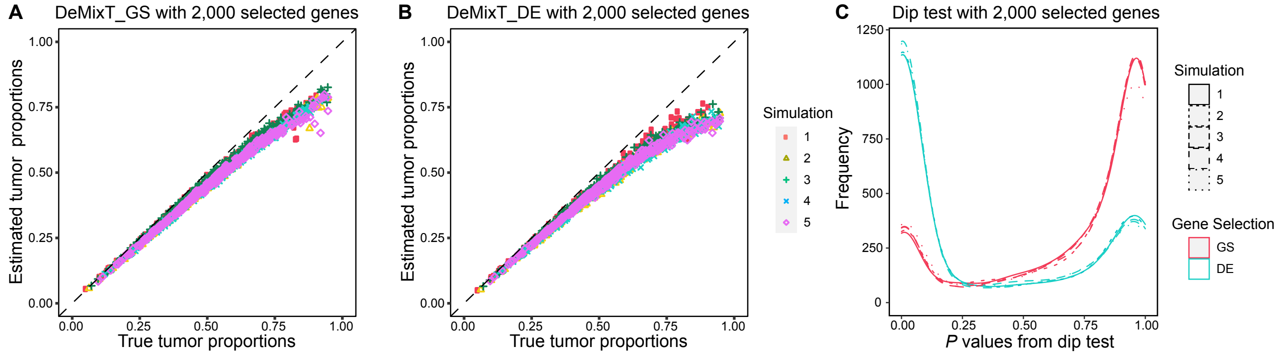
$$gene\ selection\ score_k = 2\sqrt{2\chi^2_{1-\alpha}(1)\ H\left(\hat{\pi},\hat{\mu}_T,\hat{\sigma}_T\right)^{-1}_{k,k}}.$$

Genes with a lower score have a smaller confidence interval, hence higher identifiability in their corresponding parameters. Genes will be ranked based on the gene selection score from the smallest to the largest. A subset of genes that are ranked on top will be used for parameter estimation. In the DeMixT R package (freely available from Bioconductor), our proposed profile likelihood-based gene selection approach is included as function DeMixT_GS.

We validated the accuracy of the proposed gene selection method through simulations. The DeMixT model assumes every gene g has a shared mean (μ_{Tg}) and variance (σ_{Tg}) parameters across all tumor samples. However, in real data, this assumption might be violated occasionally, due to the fact that some genes are significantly differentially expressed in different subtypes of the cancer. For example, the PAM50 genes are known to be differentially expressed in different molecular subtypes in breast cancer, e.g., Basal, Her2, LumA, and LumB subtypes. Therefore, our simulation aimed to assess the performance of the proposed gene selection method by simulating a subset of genes whose distribution is bimodal or trimodal, to mimic the subtype-specific differentially expressed genes. We denote these genes as subtype genes. The detailed simulation design is described below.

We simulated a dataset with expression levels from 15,000 genes and 300 mixed tumor samples, plus 100 normal reference samples. For the mixed tumor samples, the true distribution of tumor-specific mRNA expression proportions was simulated from a normal distribution (mean = 0.55, Standard Deviation (SD) = 0.2) and truncated at endpoints of 0.05 and 0.95. We generated the expressions of the 15,000 genes for the pure tumor T_{ig} and normal references N_{ig} with distributions $log_2\left(T_{ig}\right) \sim N\left(\mu_{Tg}, \sigma^2_{Tg}\right)$ and $log_2\left(N_{ig}\right) \sim N\left(\mu_{Ng}, \sigma^2_{Ng}\right)$, where i denotes sample, $i=1, \cdots, S, g=1, \cdots, G$, where $\mu_{Ng}, \mu_{Tg} \sim N\left(7,1.5^2\right)$ and $\sigma_{Ng}, \sigma_{Tg} \sim U(0.1,0.8)$. The 100 normal reference samples were simulated by $log_2\left(N_{ig}\right) \sim N\left(\mu_{Ng}, \sigma^2_{Ng}\right)$, $i=1, \cdots, 100, g=1, \cdots, G$. To simulate the subtype genes that are expressed differentially in certain tumor subtypes, we randomly drew a subset of genes G^* (2,000 in total) and split samples into subgroups S_1, S_2, and S_3 with corresponding $\mu_{T_1g}, \mu_{T_2g}, \mu_{T_3g}$, where $\mu_{T_sg} \sim N\left(7,1.5^2\right)$ and $S= S_1+S_2+ S_3$. If $g \in G^*$, $log_2\left(T_{ig}\right) \sim N\left(\mu_{T_kg}, \sigma^2_{Tg}\right)$, $k=1, 2,3$, $i \in S_k$, $g \in G^*$. Then we mixed the T_{ig} and N_{ig} component expression linearly at the generated tumor-specific mRNA expression proportions according to the DeMixT model: $Y_{ig}= \pi_i T_{ig}+(1-\pi_i)N_{ig}$, where $G=15,000$, $S=300$. The simulation procedure was repeated five times. The estimated tumor-specific mRNA expression proportions were shown in **SI Figure 5A**.

Under this simulated scenario, our proposed gene selection method successfully ranked the subtype genes lower than others, whereas the other gene selection method, DeMixT_DE, failed to do so. DeMixT_DE (also provided as an option for gene selection in the DeMixT R package) ranks genes based on the two-sided t-test statistic between mixed tumor and normal samples, where genes with larger t-statistics are ranked on the top. In essence, this method is used by most deconvolution methods to pre-select genes. Across simulations where we selected 2000, 3000, 4000, and 5000 genes, DeMixT_GS always outperformed DeMixT_DE in estimating proportions. Furthermore, we observed DeMixT_DE underestimated the proportions when the true tumor-specific mRNA expression proportions are high (**IS Figure 5B**). Such bias persisted, but was reduced with DeMixT_GS (**SI Figure 5A**). Across five simulations, when we selected the top 2,000 genes, DeMixT_DE selected an average of 1,150 subtype genes (>50%), and DeMixT_GS only selected an average of 277 subtype genes (**SI Figure 5C**). The dip test[26] was used to measure the unimodality of the distribution of gene expression. The test statistic was designed to test multimodality of a random variable based on the maximum difference between the empirical distribution and the unimodal distribution of all observed data points. Hence using profile likelihood to select genes proved to be a better approach.

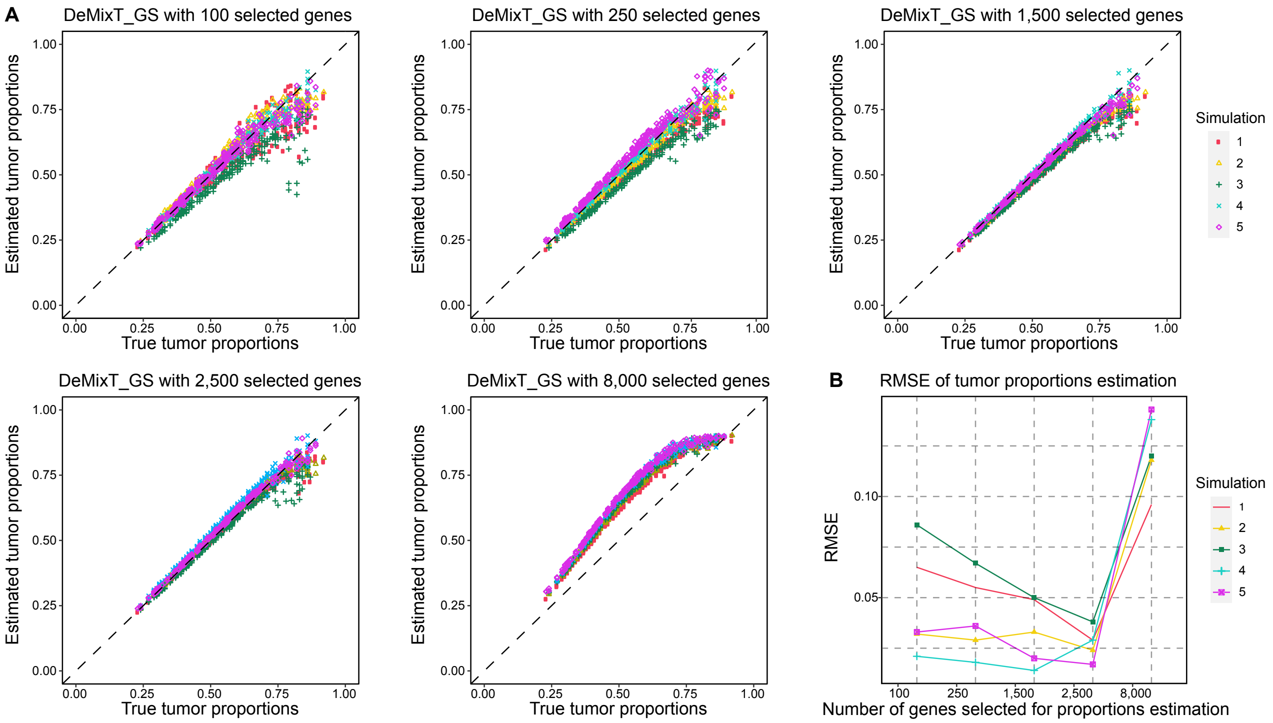


SI Figure 5. Profile likelihood-based gene selection (DeMixT_GS) improves tumor-specific mRNA expression proportions estimation. (**A**) Scatter plot of true versus estimated tumor-specific mRNA expression proportions using the DeMixT_GS method from 2,000 top genes with the smallest gene selection score. (**B**) Scatter plot of true versus estimated tumor-specific mRNA expression proportions using the DeMixT_DE method from 2,000 top genes with the smallest *P* values of differential expressed genes between mixed tumor and normal samples. (**C**) Density of *P* values based on a dip test for selected genes by DeMixT_GS and DeMixT_DE methods. The dip test was applied to indicate the distribution of gene expression for selected genes based on the DeMixT_GS and DeMixT_DE methods, respectively. A small *P* value of the dip test suggests the corresponding gene is not unimodally distributed, which violates the model assumption of $\log_2$-normal distribution across samples.

We observed that the number of genes selected by DeMixT_GS influences the performance of DeMixT. To estimate the optimal number of selected genes, we simulated gene expressions of 269 mixed samples and 100 normal references with 10,000 genes, mimicking the real data scenario presented in the TCGA prostate adenocarcinoma dataset. The true tumor-specific mRNA expression proportions were set as the tumor cell proportions derived from ASCAT. We generated the expressions of 8,000 of the 10,000 genes for the mixed samples and references with distributions $log_2\left(T_{ig}\right) \sim N\left(\mu_{Tg}, \sigma^2_{Tg}\right)$ and $log_2\left(N_{ig}\right) \sim N\left(\mu_{Ng}, \sigma^2_{Ng}\right)$, respectively, where $\mu_{Ng}, \mu_{Tg} \sim N\left(7,1.5^2\right)$ and σ_{Ng}, σ_{Tg} were sampled with replacement from the observed standard deviations from the normal samples of TCGA prostate adenocarcinoma. For the expressions of the remaining 2,000 genes, we randomly drew a subset of genes G^* (2,000 in total) and split samples into subgroups S_1, S_2, and S_3 with corresponding $\mu_{T_1g}, \mu_{T_2g}, \mu_{T_3g}$, where $\mu_{T_sg} \sim N\left(7,1.5^2\right)$ and $S= S_1+S_2+ S_3$. If $g \in G^*$, $log_2\left(T_{ig}\right) \sim N\left(\mu_{T_kg}, \sigma^2_{Tg}\right)$, $k=1, 2,3, i \in S_k, g \in G^*$. Then we mixed the T_{ig} and N_{ig} component expression linearly at the generated tumor-specific mRNA expression proportions according to the DeMixT model: $Y_{ig}= \pi_i T_{ig}+(1-\pi_i)N_{ig}$, where $G=10,000$, $S=269$.

The accuracies of tumor-specific mRNA expression proportion estimation based on 100, 250, 1,500, 2,500 and 8,000 genes selected by the proposed DeMixT_GS were compared. **(SI Figure 6)**. Accurate tumor-specific mRNA expression proportion estimation, as measured by the RMSE, was achieved with 1,500 or 2,500 genes. In real data, we used either the top 1,500 or top 2,500 genes to estimate the tumor-specific mRNA expression proportion.

In summary, our proposed profile likelihood-based gene selection approach was shown to substantially improve the tumor-specific mRNA expression proportions estimation by ranking the identifiability of genes in the DeMixT model.

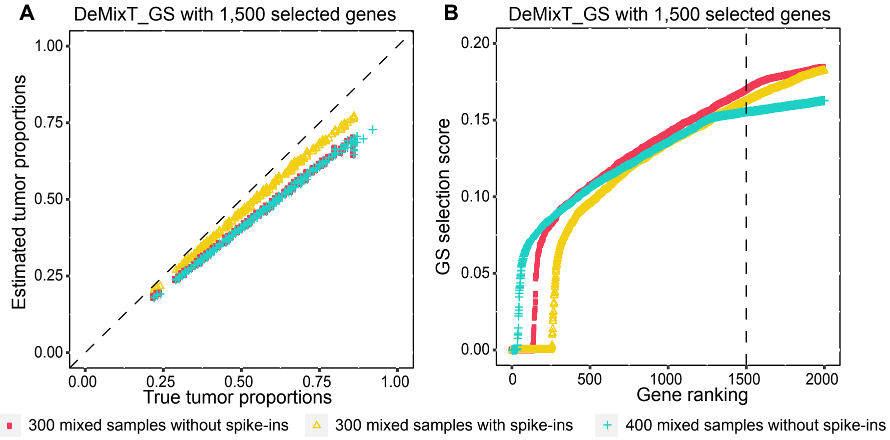


Supplementary Figure 6. Optimization of the number of genes selected for DeMixT_GS to improve mRNA expression proportions estimation. (**A**) Scatter plot of true versus estimated tumor-specific transcript proportions using DeMixT_GS method with different numbers of top-ranking genes with the smallest gene selection score. (**B**) Root Mean Square Error (RMSE) was calculated for each simulation scenario.

Virtual spike-ins to improve identifiability

In the studies of simulation, we further observed that the tumor-specific mRNA expression proportions estimation is unbiased only when the true proportions are centered around 0.5. When the true proportions are skewed towards the high end (i.e., median above 0.5), which is expected to occur frequently in real data (tumor samples with a low percentage of tumor cells are already discarded), the DeMixT estimation procedure, after careful gene selection, still underestimates the high proportions (**SI Figure 5A**).

To resolve this issue, we introduce a step to enforce the center of true proportions to be around 0.5. We will simulate additional “tumor” samples, i.e. spike-ins, with close to 0% of tumor-specific mRNA expression proportions, so that there are roughly the same number of samples with tumor proportions below and above 50%, i.e., $S_P + \left|\{i \mid \rho_i<0.5\}\right| \cong |\{i \mid \rho_i \geq 0.5\}|$, where S_P represents the number of spike-ins, ρ_i represent tumor purity of sample i, and the $|\cdot|$ represent cardinality of a set. For the cancer type whose median tumor purity is below 0.5, we set S_P at 5. The spike-ins are generated based on gene expression profiles observed from the input data of normal reference samples. We conducted a simulation study to assess the utility of spike-ins. We simulated 100 mixed samples and 100 normal reference samples with 8,000 genes based on the aforementioned simulation parameters with five replicates. Tumor-specific mRNA expression proportions were simulated from a normal distribution (mean = 0.55, SD = 0.2) and truncated at endpoints of 0.05 and 0.95. $\mu_{Ng}, \mu_{Tg} \sim N\left(7,1.5^2\right)$ and $\sigma_{Ng},\sigma_{Tg} \sim U(0.1,\ 0.8)$. The expression level of spike-ins is denoted as P_{jg}. We simulate $P_{jg} \sim LN\left(\hat{\mu}_{Ng},\ \hat{\sigma}_{Ng}^2\right)$, for gene $g = 1,2,\ldots,G$ and sample $j = 1,2,\ldots,S_P$. The spike-ins were then combined with mixed tumor samples. We ran DeMixT on the combined samples while fixing the mRNA expression proportions for the spike-ins at 0.01. We found adding spike-ins can reduce biases in the estimation of tumor-specific mRNA expression proportions (**SI Figure 7A**). Such utility from adding spike-ins was made by lowering gene selection scores in the top-ranking genes (**SI Figure 7B**).



SI Figure 7. Adding spike-ins reduces systematic bias of estimated tumor-specific mRNA expression proportions. (**A**) Scatter plot of true versus estimated tumor-specific mRNA proportions using DeMixT_GS method under different strategies of adding spike-ins. (**B**) Distribution of gene selection score to top 2,000 genes under different strategies of adding spike-ins, where the x-axis represents the sorted genes by the gene selection score from low to high and the y-axis represents the estimated gene selection score for the corresponding genes.

2.3. Tumor-specific total mRNA expression in patient samples

2.3.1. Datasets

The Cancer Genome Atlas (TCGA) data

Publicly available transcriptome profiling HT-seq raw read counts from 7,054 tumor samples from 15 cancer types in TCGA (breast adenocarcinoma, bladder urothelial carcinoma, colorectal cancer (colon adenocarcinoma + rectum adenocarcinoma), head and neck squamous cell carcinoma, kidney chromophobe, kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, stomach adenocarcinoma, thyroid carcinoma, uterine corpus endometrial carcinoma) were downloaded from the GDC data portal (v14.0)[27] (https://portal.gdc.cancer.gov/). They were generated through the standard RNAseq analysis pipeline (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/) by aligning reads to the GRCh38 reference genome and then by quantifying the mapped reads. We downloaded the clinical annotation data including overall survival (OS), progression free interval (PFI), pathologic stage, age, and sex of patients across 15 cancer types from the GDC data portal (https://gdc.cancer.gov/about-data/publications/pancanatlas). Somatic mutation data of the 15 cancer types were downloaded from the re-annotated mutation annotation file (MAF) format at the GDC (https://gdc.cancer.gov/about-data/publications/mc3-2017). ABSOLUTE tumor purity and ploidy data were downloaded from Aran et al[28]. ASCAT tumor purity and ploidy data were downloaded from Alexandrov et al[29]. Driver mutation and indels annotation were downloaded from the TCGA pan-cancer driver mutation database: http://intogen.org/download version 2016.5[30]. NarrowPeak format ATAC-seq data for TCGA samples was obtained from Corces et al[31]. NarrowPeak files were annotated using the R package chipseeker[32]. Peaks outside of promoter regions (-2kb to 1kb of transcription start sites) were excluded. For breast adenocarcinoma, molecular subtype, triple negative status, status of hormone receptor, were obtained from Koboldt et[33]. The copy number alternation status of *MYC* and *PVT1* were called by GISTIC[34] using the SNP6 DNA microarray data from breast carcinoma in TCGA, were obtained from cBioPortal (https://www.cbioportal.org/)[35]. For prostate adenocarcinoma, the Gleason score was obtained from Abeshouse et al[36]. For head and neck squamous cell carcinoma, the HPV status was obtained from Lawrence et al[37]. In renal papillary carcinoma, the molecular subtypes was obtained from Linehan et al[38]. A CONSORT diagram is provided for the dataset (**Fig. S7A,B**).

International Cancer Genome Consortium – Early Onset Prostate Cancer (ICGC-EOPC) data[39]

Matched RNAseq and whole genome sequencing (WGS) data from 121 tumor samples and 9 adjacent normal samples from 96 patients, the corresponding clinical data including biochemical recurrence (BCR), and Gleason scores were downloaded from an early-onset (treatment age < 55) prostate cancer patient cohort[39]. Among these 96 patients, there were 13 with a Gleason score = 3+3, 58 with a Gleason score = 3+4, 11 with a Gleason score = 4+3, 1 with a Gleason score = 4+4, 6 with a Gleason Score = 4+5, 6 with a Gleason score = 5+4, and 1 with a Gleason Score = 5+5.

For the early-onset prostate adenocarcinoma dataset, the gene expression read counts from the ultra-deep total RNAseq and relevant clinical data of 121 tumors samples from 96 patients were obtained from Gerhauser et al [39]. RNA reads were aligned to the human GRCh37 reference genome using BWA and SAMtools. Uniquely mapped reads were annotated using Ensembl v62. DNA library preparation and WGS was performed on Illumina sequencers[40] with a median insert size of 310 bp (sd 57 bp) and a median WGS coverage of 61-fold for tumor and 38-fold for germline control samples. WGS data was aligned to the GRCh37 reference genome using BWA-MEM[41] according to Pan Cancer Analysis of Whole Genomes (PCAWG) protocol (https://doi.org/10.1101/161638). The clinical data contain the biochemical relapse (BCR) interval, Gleason score, pathologic stage, and mutation clonal status. DNAseq-based purity and ploidy estimates for 113 samples from 89 patients were determined by Sequenza[42]. A CONSORT diagram is provided for the dataset (**Fig. S7C**).

TRAcking Cancer Evolution through therapy (Rx) (TRACERx) data[43,44]

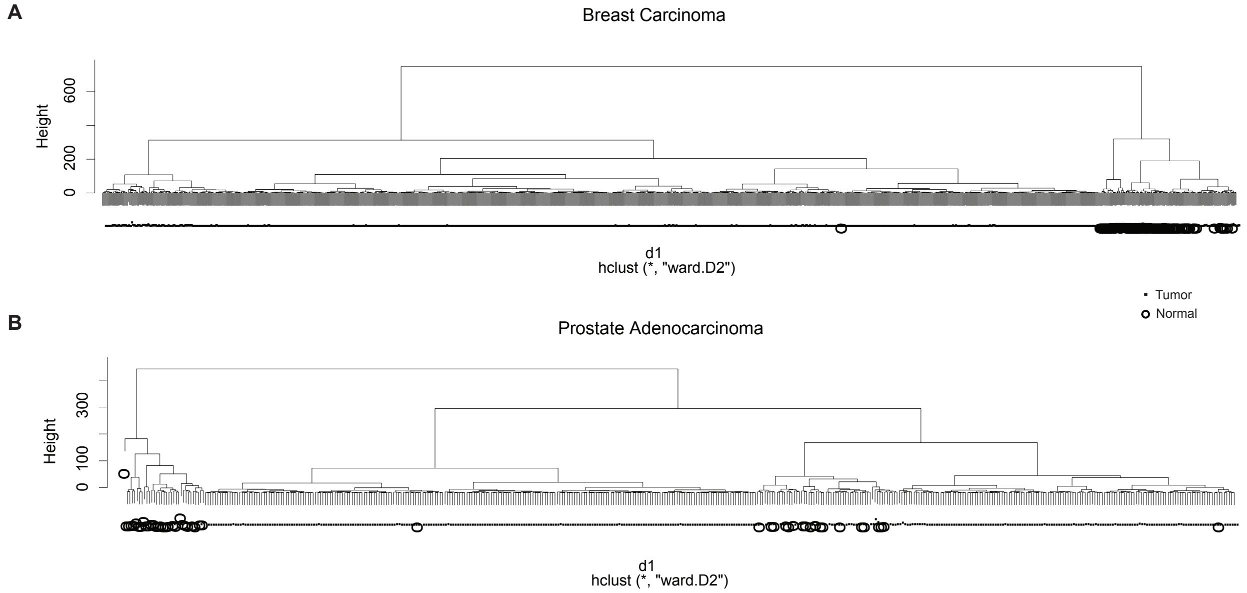
For this cohort of 100 patients, multi-region RNAseq and WES data were sequenced from the same tissue[45]. The WES data were aligned to the human hg19 reference genome by the Ion Torrent Torrent Suite software. SAMtools mpileup (0.1.19)[46] was used to locate non-reference positions in tumor and germline samples. A combination of picard tools (1.107), GATK (2.8.1) and FastQC (0.10.1) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were used to perform quality control. RNA data were generated using a modification of the AllPrep kit (Qiagen) and assessed by TapeStation (Agilent Technologies). The STAR package[47] (version 2.5.2b) was used to perform alignment and map reads to the human hg19 reference genome.

We downloaded clinical information including information of whether recurrence occurred, progression free survival (time to recurrence in the paper[43]) and per region segmented copy number data of the 100 patients from Jamal-Hanjani et al.[43]. DNAseq-based purity and ploidy estimates for 327 tumor samples from the 100 patients were determined by Sequenza[42]. RNAseq-based tumor-specific mRNA proportions for 159 tumor samples from 64 patients were estimated by DeMixT[19]. The other 168 samples with only DNAseq data and no matching RNAseq data were removed. In the end, we focus on 30 patients with multi-region samples (m = 94) and 52 patents with both single and multi-region samples (m = 116) for the downstream analysis. A CONSORT diagram is provided for this dataset to demonstrate the filtering steps (**Fig. S7D**).

Genotype-Tissue Expression (GTEx) data[48]

Publicly available transcriptome profiling raw read counts data which is based on the Illumina TruSeq RNA sequencing platform from normal prostate samples, normal thyroid samples, and normal lung samples were downloaded from the GTEx data portal (https://www.gtexportal.org/home). RNAseq data were aligned to the human GRCh37 reference genome using Tophat (v1.4.1)[49]. Gencode (v12)[50] was used as a transcriptome model for the alignment as well as all gene and isoform quantifications. Of all the samples provided by GTEx, we selected 42 normal prostate samples, 67 normal thyroid samples, and 20 normal lung samples without significant pathology in the corresponding tissue types.

2.3.2. TCGA



SI Figure 8. Hierarchical clustering of read count data from adjacent normal and mixed tumor samples for breast and prostate adenocarcinoma in TCGA as examples. Dendrograms of the hierarchical clustering results based on the top 1,000 differentially expressed feature genes for breast carcinoma (**A**) and prostate adenocarcinoma (**B**), respectively. Circles represent adjacent normal samples and dots represent mixed tumor samples.

Data pre-processing

To estimate the tumor-specific mRNA expression proportions (π) for each sample, we used the two-component mode of DeMixT for 15 TCGA cancer types where sufficient normal reference samples were available (the minimum number of normal samples is seven). For each cancer type, the following quality control was performed on both the tumor and normal samples to remove any suspicious samples. For each gene, we first used the Wilcoxon rank-sum test to test for differential expression between normal and tumor samples. The top 1,000 genes with the smallest *P* values were selected as the feature genes. The first two principal component scores of the feature genes were extracted for hierarchical clustering using Euclidean distance and the Ward method. We separated samples into two groups using the "cutree" function. In general, one cluster contained tumor samples and the other contained normal samples. Any samples that were clustered outside of its general group label, e.g., tumor samples clustered within the normal sample cluster or normal samples in the tumor cluster, were considered as suspicious samples and filtered out (**SI Figure 8**).

SI Table 5 summarizes the selected sample numbers for the 15 cancer types before and after quality control. Scale normalization at the seventy-fifth percentile based on the DSS package[51] was then applied to the post quality-control tumor and normal samples. Next, we applied two criteria to filter out spurious genes. First, we filtered out genes with a zero count in either the mixed tumor or normal samples. Second, we filtered out genes with a large variance ($\hat{\sigma}^2_{Ng}$ > 0.6) in the normal reference samples. Here, the standard deviation of a gene is calculated as $\hat{\sigma}^2_{Ng} = sd(log_2(R_{.g}))$, where $R_{.g}$ is the normalized expression of gene *g* for normal reference samples.

SI Table 5. Summary of sample sizes for 15 TCGA cancer types.

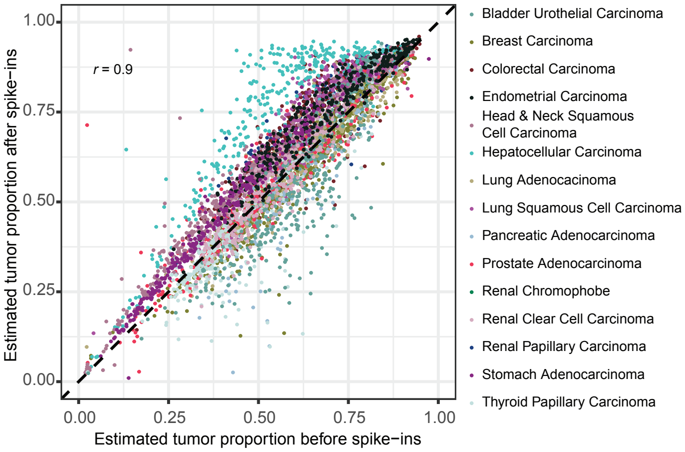
Cancer type	Original number of normal samples	Original number of tumor samples	Number of normal samples after quality control	Number of tumor samples after quality control
Bladder urothelial carcinoma	19	414	17	385
Breast carcinoma	113	1101	98	1032
Colorectal carcinoma	51	633	43	598
Head & neck squamous cell carcinoma	44	500	31	494
Renal chromophobe	24	65	23	64
Renal clear cell carcinoma	72	538	66	495
Renal papillary carcinoma	32	288	26	276
Hepatocellular carcinoma	50	371	50	362
Lung adenocarcinoma	59	532	57	446
Lung squamous cell carcinoma	49	502	48	486
Pancreatic adenocarcinoma	4	177	7*	142
Prostate adenocarcinoma	52	498	47	295
Stomach adenocarcinoma	32	375	32	299
Thyroid papillary carcinoma	57	502	55	418
Endometrial carcinoma	35	549	26	524

*Pancreatic adenocarcinoma is the only cancer type with increased normal samples, for pseudo-normal samples are added, which are tumor samples of stromal tissue with scant tumor presence.

For each cancer type, we applied the DeMixT algorithm (DeMixT_DE) to the quality-controlled expression data together with simulated spike-ins as input data, to generate initial tumor-specific mRNA expression proportions π_0. We used ASCAT estimated tumor purities as an informed prior to calculate a reasonable number for S_P. With other datasets in general, we set $S_P = max(50, 0.3*Sample\ size)$, as the default option of the DeMixT_GS function. Results from the TCGA datasets across 15 cancer types were largely consistent with small to moderate changes from with or without spike-ins (**SI Figure 9).**

We then used these $\pi_0's$ as initial values in the profile likelihood calculation on all genes to calculate gene selection scores. We ranked all genes based on their gene selection scores from the smallest to the largest. Based on a simulation study (**SI Figure 7**) and observed distributions of gene selection scores in real data, we chose the top 1,500 or 2,500 genes to ensure accuracy in proportion estimation (**SI Figure 6B**). Within each cancer type, we used the spike-ins as benchmarking samples and evaluated the RMSE of the estimated proportions of the spike-ins with either the top 1,500 or top 2,500 genes ($\hat{\pi}_{1500}(Sp)$ and

$\hat{\pi}_{2500}(Sp)$). If RMSE($\hat{\pi}_{1500}(Sp)-0$)<RMSE($\hat{\pi}_{2500}(Sp)-0$), we used the results of the top 1,500 genes, i.e., the tumor proportions $\pi=\pi_{1500}$; otherwise, $\pi=\pi_{2500}$. The finalized tumor proportions are shown in **Fig. S2A**. In general, the RMSEs were small (median = 0.02 across 15 cancer types), and the two sets of tumor proportions, $\hat{\pi}_{1500}$ and $\hat{\pi}_{2500}$, were consistent within each cancer type.



SI Figure 9. Comparison of tumor-specific mRNA expression proportions with and without spike-ins across 15 TCGA cancer types. A scatter plot of estimated tumor-specific mRNA expression proportions using the DeMixT_GS method for 5,031 TCGA samples across 15 cancer types with and without spike-ins. The x axis represents the estimated tumor-specific transcript proportions without spike-ins and the y-axis represents the estimated tumor-specific transcript proportions with spike-ins.

2.3.2.1. Consensus TmS estimation

For DNA-based deconvolution methods such as ASCAT and ABSOLUTE, there could be multiple tumor purity ρ and ploidy ψ pairs that have similar likelihoods. Consequently, the estimated tumor purity ρ and ploidy ψ for some samples would be ambiguous without a unique solution. To be specific, both of ASCAT and ABSOLUTE can accurately estimate the product of purity and ploidy $\rho\psi$; however, they sometimes lack power to identify ρ and ψ separately, as they use BAF (B allele frequency) of germline SNPs, which only measures the total proportions of alternative allele. TmS is derived from estimates of tumor purity and ploidy via their product, hence automatically deals with any ambiguity in tumor purity and ploidy estimation, ensuring the robustness of the TmS calculation. We validated this point by showing that among 20% of all TCGA samples, the agreement between TmS values calculated from ASCAT and ABSOLUTE was substantially improved, as compared to those for the ploidy or purity individually (**Fig. S2D-F**). To calculate one final set of TmS values for a maximum number of samples, we took a consensus strategy. We first calculated TmS values for 5,295 TCGA samples with matched tumor-specific mRNA expression proportions and ABSOLUTE or ASCAT derived tumor purity and ploidy estimates. We then fitted a linear regression model on $\log_2$-transformed TmS calculated by ASCAT using $\log_2$-transformed TmS calculated by ABSOLUTE as a predictor variable. We removed samples with a Cook's distance ≥ $4/n$ (n is the number of total samples) (**Fig. S2F**), and for the remaining samples, which were the majority, we calculated the final TmS as: $TmS=2^{(log_2(TmS_{ASCAT})+log_2(TmS_{ABSOLUTE}))/2}$. These TmS estimates were used throughout the paper (**SI Table 6**). A CONSROT diagram demonstrates the sample exclusion for TmS in TCGA (**Fig. S7A).**

SI Table 6. Summary of sample sizes for 15 TCGA cancer types before and after consensus TmS estimation.

Cancer type	Number of samples before consensus analysis	Number of samples after consensus analysis	Number of samples removed
Bladder urothelial carcinoma	350	328	22
Breast carcinoma	932	916	16
Colorectal carcinoma	499	490	9
Head & neck squamous cell carcinoma	449	443	6
Renal chromophobe	59	59	0
Renal clear cell carcinoma	299	295	4
Renal papillary carcinoma	192	169	23
Hepatocellular carcinoma	333	317	16
Lung adenocarcinoma	399	395	4
Lung squamous cell carcinoma	440	431	9
Pancreatic adenocarcinoma	105	101	4
Prostate adenocarcinoma	266	259	7
Stomach adenocarcinoma	272	265	7
Thyroid papillary carcinoma	297	202	95
Endometrial carcinoma	403	361	42

2.3.2.2. Global transcription signature genes

We found the genomic locations of the selected genes covered 22 autosomes and the X chromosome (**Fig. S3A**) across 15 cancer types, which is expected for an unbiased gene set to measure global gene expression.

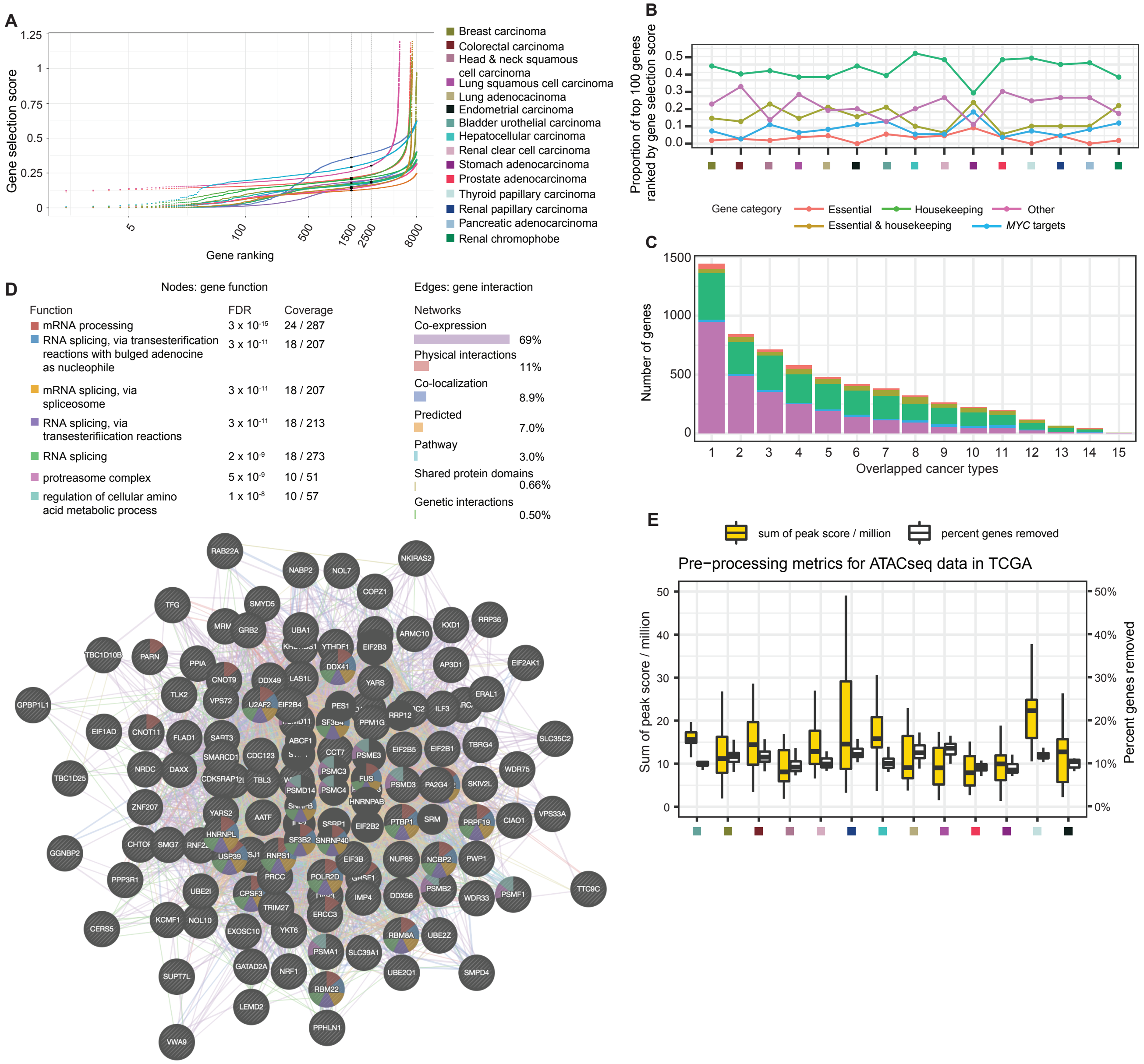
For each cancer type, as well as consistently across 15 cancer types, we found that 54-68% (mean = 62%) of global transcription signature genes are housekeeping genes[52] or essential genes[53], and 3.5-7.2% (mean = 5.4%) are MYC targets genes (**Fig. S3B**). The common pan-cancer essential genes are derived from a total of 147 cancer cell lines and 16,733 genes that were screened independently by both the Sanger and Broad institutes[53]. We observed similar proportions of each gene group among the top 100 genes in terms of gene selection scores for each cancer type (**SI Figure 10B**).

We conducted gene set enrichment analyses on Hallmark pathways[54] and KEGG pathways[54] for all the available genes with their gene selection scores calculated by DeMixT using GSEA[54] and g:Profiler[55]. For each cancer type, the genes were ranked according to their gene selection scores from the smallest to the largest and then fed into GSEA in the "pre-ranked" mode and g:Profiler. For GSEA, we adopted permutation tests (1,000 times) to generate a normalized enrichment score (NES) and an adjusted *P* value for each candidate pathway. Specifically, for each pathway, a running-sum statistic was calculated by walking down the ranked list of the genes from top to bottom. The enrichment score (ES) is the maximum deviation from zero for the running-sum statistic. Then the normalized enrichment score is calculated as NES = ES/mean(ES's of all permutations). A gene set permutation which randomly permute gene labels was then implemented to estimate the null distribution of the NES[54]. The nominal *P* values and *P* values adjusted for multiple testing for pathways are estimated by comparing observed NES against the null distribution. g:Profiler detects statistically significantly enriched pathway for the given gene list by implementing hypergeometric tests. This technique starts from the top-ranked genes to the bottom-ranked genes in the list and identifies the optimal subset where the hypergeometric statistic is the largest. For each candidate pathway, a nominal *P* values is calculated by the hypergeometric test, and adjusted for multiple testing by the Benjamini-Hochberg method. We combined results from GSEA and g:Profiler: only the pathways with adjusted *P* value < 0.05 from both GSEA and g:Profiler were considered as significantly enriched and shown in **Fig. S3C,D**. As a result, the minimum NES of significantly enriched Hallmark pathways and KEGG pathways are above 1.74 and 1.70, respectively.

We also checked the percent overlap between individual signature gene sets across cancer types. A total of 114 genes were repeatedly selected in at least 13 out of 15 cancer types (**SI Figure 10C**). We used GeneMANIA[56] to evaluate the functional relationship of these genes and found they encode for proteins related to mRNA processing, namely binding and splicing mRNA, which is consistent with findings from KEGG pathways (**SI Figure 10D**). Seven genes (*CPSF3*, *EIF2B1*, *LAS1L*, *ELAVL1*, *SMARCD1*, *UBE2Z* and *WDR46*) were consistently selected within the transcriptional signature genes across all 15 TCGA cancer types. Endonuclease activity of *CPSF3* is associated with pre-mRNA processing. *EIF2B1* is a translation factor. *LAS1L* is associated with ribosome biogenesis. *ELAVL1* is associated with RNA binding. *SMARCD1* belongs to the SWI/SNF complex. *UBE2Z* is associated with ubiquitin ligase. *WDR46* is associated with ribosomal RNA processing. In summary, the seven genes encode for proteins related to fundamental processes previously linked to cancer such as RNA splicing, ribosomal RNA processing, chromatin remodeling, and protein translation.

We further evaluated the chromatin accessibility of signature genes using ATAC-seq data TCGA samples[31]. For each sample, peak scores (-log10(p-value)) were scaled by dividing each individual peak score by the sum of all of the peak scores in the given sample divided by 1 million. These scaling values ranged from 1.4 to 67.4 across cancer types (**SI Figure 10E**). The 75th percentile of normalized peak scores across all peaks within the promoter region was selected for each gene as a representative peak score, and genes with normalized peak scores less than 1 were excluded, with 7.1% to 20.4% of genes excluded across cancer types (**SI Figure 10E**). For each sample, we calculated the mean of the peak scores of all signature genes. A null distribution of mean peak scores was generated by calculating means from 1,000 random subsets of genes with the matching number of the signature genes from all genes. *P* values assessing the significance of the deviation of the observed mean score for signature genes from the null distribution were calculated as the percentile of the permuted means being greater than or equal to the observed mean. Within cancer types, *P* values were adjusted for multiple testing using the Benjamini-Hochberg procedure. A total of 259 (84%) out of 310 samples across 13 cancer types presented a significant difference in peak score >1 (10 fold difference in *P* values between signature and non-signature genes), and 121 samples (40%) showing a significance difference in peak score >3 (1,000 fold change in *P* values between signature and non-signature genes) (**Fig. S3E**), indicating higher chromatin accessibility in the global transcription signature genes.

In summary, we reasoned that the global transcription signature genes provided reasonable genome-wide coverage and were associated with transcriptional regulation and mRNA processing across cancer types. As such, the global transcription signature provided representative gene sets to track global gene expression across cancer types.

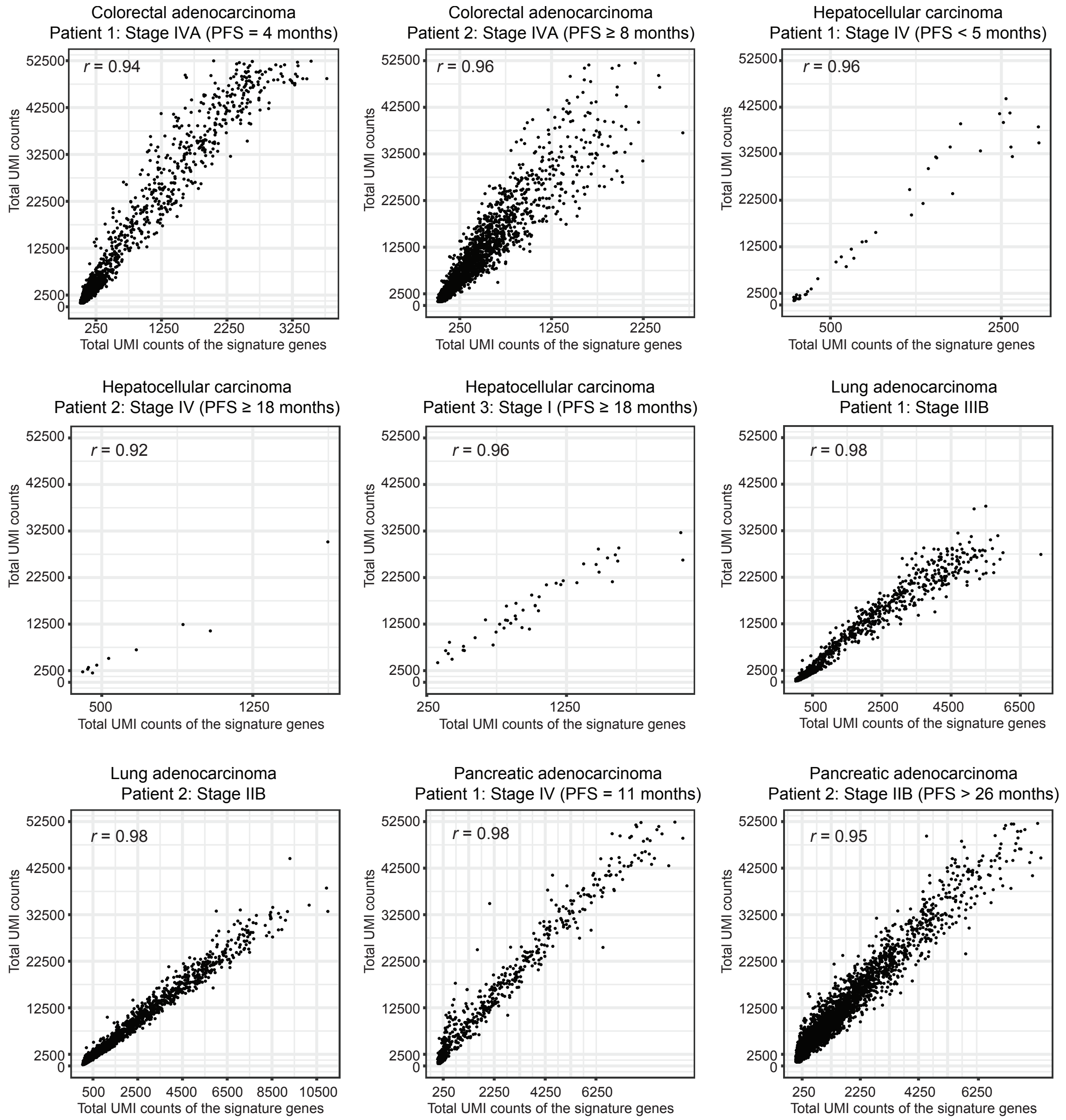


Supplementary Figure 10. Validation of gene selection to represent a global transcription signature across cancer types. (**A**) Ordered gene selection scores of all genes from low to high across 15 cancer types. The black solid dots represent the cutoffs for gene selection in the corresponding cancer type. (**B**) Proportions of the top 100 genes in five gene categories across 15 cancer types. (**C**) Histogram of the numbers of genes selected in five gene categories as they overlap across 15 cancer types. The y axis represents the total number of genes and the x axis represents how many times the same genes were selected across cancer types. (**D**) The association network and enrichment statistics of the top 114 repeatedly selected genes using GeneMANIA. Top seven gene functional enriched pathways are shown on the nodes with different colors. Seven types of gene interactions, i.e., co-expression, physical interaction, co-localization, predicted, pathway, shared protein domains, and genetic interactions, are shown on the edges with different colors. (**E**) Boxplots showing sums of peak scores across samples / 1 million (yellow) and percentages of genes removed per sample (white). The sums of peak scores were used as scaling factors to normalize the ATAC-seq data. The genes with normalized scores <1 were removed from downstream analysis. For A,B,E, all 15 TCGA cancer types are annotated using colored squares as shown in the legend of A.

2.3.2.3. Validation using scRNAseq data.

Using the aforementioned scRNAseq data from four cancer types (**Section 1.2**), we found the glonal transcription signature genes were consistently expressed at higher levels in tumor cells than in non-tumor cells in all patient samples (**Fig. S3F,G**). Wilcoxon rank-sum tests were used to compare the distribution of mean UMI counts of the global transcription signature genes per cell in tumor cells to that in non-tumor cells within each patient (**Fig. S3F**). The same tests were used to compare the distribution of mean UMI counts of the global transcription signature genes per cell to that of the non-global transcription signature genes per tumor cell within each patient (**Fig. S3G**). The *P* values from Wilcoxon rank-sum tests were adjusted using the Benjamini-Hochberg correction across the nine patients. Furthermore, within each cancer type, the signature genes were significantly highly expressed in patient samples with worse prognoses (**Fig. S3H**). Kruskal-Wallis tests were used to compare the ratio of mean UMI counts of the global transcription signature genes per cell of tumor cells to that of non-tumor cells between patients within each cancer type. The *P* values from the Kruskal-Wallis tests were adjusted for multiple testing using the Benjamini-Hochberg correction across the four cancer types. We also calculated the correlations between the total UMI counts of the signature genes and the total UMI counts of all genes for each tumor cell. The Spearman correlation coefficients (*r*) are between 0.92-0.98 across the nine patients (**SI Figure 11**). Therefore, we posit the global transcription signature genes represented important gene pathways in global gene expression regulation and tumor progression.

We further pooled scRNAseq data to form pseudo-bulk samples. We calculated the ratio of the mean total UMI counts of tumor cells to that of the non-tumor cells for each of the nine scRNAseq patient samples. This ratio represents the ploidy-unadjusted TmS (defined in **Section 2.1.1**) in pseudo-bulk data. Within each patient sample, we also constructed the 95% confidence interval for the ratio using bootstrapping with 1,000 repetitions. For each bootstrap repetition, we sampled the same number of tumor cells as the original with replacement from the corresponding patient, and kept all non-tumor cells. The results are shown in **Fig. S1C.**



SI Figure 11. Correlations between the total UMI counts of the signature genes and the total UMI counts of all genes. Scatter plots of correlations measured for each tumor cell. The Spearman correlation coefficients (*r*) is shown on each panel.

2.3.2.4. Statistical analysis

Association with clinical variables

In breast carcinoma, the association of TmS with clinical variables including molecular subtype (Luminal A, Luminal B, Her2, and Basal), triple negative status, status of hormone receptor (ER and PR), and the copy number alternation status of *MYC* and *PVT1,* were tested by using all subjects who had both TmS and clinical variables. In prostate adenocarcinoma, the association of TmS and the Gleason score was tested by using all subjects who had both TmS and Gleason score. In head and neck squamous cell carcinoma, the association of TmS and HPV status was tested by using all subjects who had both TmS and HPV status. In renal papillary carcinoma, the association of TmS and molecular subtypes was tested by using all subjects who had both TmS and molecular subtypes. Kruskal-Wallis tests were used to compare the distribution of TmS between subgroups defined by each clinical variable. The *P* values from Kruskal-Wallis tests for the clinical variables were adjusted using Benjamini-Hochberg correction across all available clinical variables within the corresponding cancer type (**Fig. 3A-D**).

Association with survival outcomes

For the TCGA datasets, we used clinical data that passed at least one of the three quality control steps introduced from the TCGA pan-cancer clinical paper[57]. We used two survival outcomes, the overall survival (OS) and the progression-free interval (PFI). To ensure sufficient sample size in each category, we combined pathologic stages into two stage categories: early stage and advanced stage. The early stage includes Stage I, Stage IA, Stage IB, Stage IC, Stage II, Stage IIA, Stage IIB, and Stage IIC, while the advanced stage consists of Stage III, Stage IIIA, Stage IIIB, Stage IIIC, Stage IV, Stage IVA, Stage IVB, and Stage IVC. With prostate cancer, we used Gleason score (Gleason Score = 7 versus Gleason Score = 8+) instead of early and advanced stage. The CONSROT diagram that demonstrates the sample exclusion for survival analysis in TCGA is shown in **Fig. S7B.**

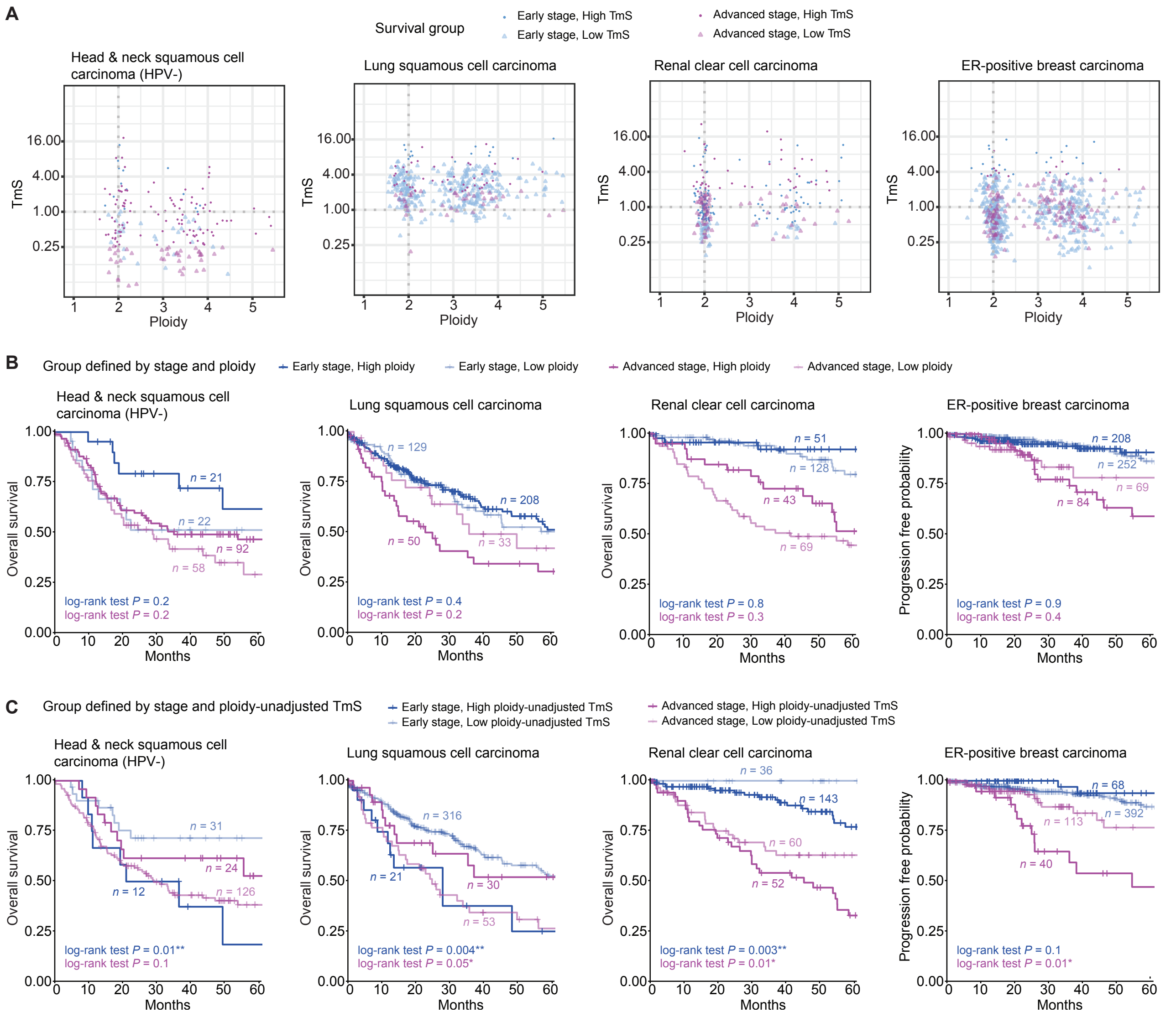
Due to the potential nonlinear relationship between TmS and survival outcomes, we used a recursive partitioning survival tree model, rpart[58], to find an optimized TmS cutoff that best separated differentiating survival outcomes within each of the two stages as defined above in each cancer type. The splitting criteria were Gini index, and the maximum tree depth was set to 2. The TmS cutoffs of early/advanced stage across cancers are shown in **SI Table 7**. For each cancer type, samples are divided by both TmS and pathological stage into four groups: (1) Early stage, High TmS, (2) Early stage, Low TmS, (3) Advanced stage, High TmS, and (4) Advanced stage, Low TmS. The Kaplan-Meier (KM) survival curves of overall survival and progression free interval for the corresponding groups defined by TmS and stage are shown in **Fig. 4** and **Fig. S4**. Log-rank tests between high and low TmS groups within early or advanced pathological stages were performed. We then fitted multivariate Cox Proportional Hazard models with age, TmS, stage, and the interaction term of TmS and Stage (TmS x Stage) as predictors for overall survival and progression free interval analysis for each cancer type (**Table S3**). We also divided samples for 14 cancer types combined (endometrial carcinoma was excluded due to missing pathological stage information) into four groups using the rpart (**Fig. 4A**). In addition, we applied the rpart to find an optimized TmS cutoff that best separated differentiating survival outcomes of 15 cancer types combined regardless of pathological stage (**Fig. 4A**). The splitting criteria were Gini index, and the maximum tree depth was set to 1.

SI Table 7. Summary of TmS cutoffs for early/advanced stage across cancers.

Cancer type	Overall survival		Progression free interval	
	Early stage	Advanced stage	Early stage (Gleason score=7 for prostate cancers)	Advanced stage (Gleason score>=8 for prostate cancers)
Pan-Cancer(14 cancer types)	1.10	1.72	1.65	1.72
Bladder urothelial carcinoma	0.15	NA	0.15	0.60
Triple-negative breast carcinoma	4.11	1.80	3.02	1.80
ER-positive breast carcinoma	2.14	NA	3.61	3.24
Colorectal carcinoma	1.94	4.52	NA	4.14
Head & neck squamous cell carcinoma (HPV-)	1.00	0.26	0.14	0.26
Renal chromophobe	2.21	3.95	1.79	0.74
Renal clear cell carcinoma	0.54	1.78	0.33	1.67
Renal papillary carcinoma	0.61	0.71	0.87	0.64
Hepatocellular carcinoma	0.16	1.81	NA	0.64
Lung adenocarcinoma	0.81	0.97	0.51	8.66
Lung squamous cell carcinoma	6.67	2.08	5.67	6.37
Pancreatic adenocarcinoma	1.83	NA	1.83	NA
Stomach adenocarcinoma	0.40	0.15	0.28	0.31
Thyroid papillary carcinoma	NA	NA	0.57	1.25
Prostate adenocarcinoma	NA	NA	0.50	0.48
Early-onset prostate adenocarcinoma (ICGC-EOPC)	NA	NA	1.25	0.84

TmS, ploidy, and ploidy-unadjusted TmS

We also compared effects of the three metrics that follow: TmS x ploidy = ploidy-unadjusted TmS values (**SI Figure 3**), in four cancer types: head and neck squamous cell carcinoma, lung squamous cell carcinoma, renal clear cell carcinoma, and ER-positive breast carcinoma. The distributions of TmS and ploidy estimates were independent across cancer types (**SI Figure 12A**). We then applied rpart to find optimized cutoffs for ploidy and ploidy-unadjusted TmS that best separate differentiating survival outcomes within each pathologic stage of each cancer type. Ploidy alone was not able to effectively distinguish survival outcomes (**SI Figure 12B**). Ploidy-unadjusted TmS scores roughly recapitulated TmS-defined survival groups, but to a lesser extent (**SI Figure 12C**). Consequently, the refined prognostication by TmS over stage was more pronounced for TmS as compared to TmS scores without ploidy adjustment across cancer types.



Supplementary Figure 12. Prognostication using ploidy or ploidy-unadjusted TmS on pathological stages in four cancer types. (**A**) Scatter plot of TmS (y axis) versus tumor ploidy (x axis) for samples from head and neck squamous cell carcinoma (HPV negative), lung squamous cell carcinoma, renal clear cell carcinoma, and ER-positive breast carcinoma cohorts. The samples were grouped into high versus low TmS within early or advanced pathological stages, with different groups shown in different colors. (**B**) KM survival curves of overall survival or progression free interval in four cancer types over patient groups defined by ploidy and stage. We grouped patients into high versus low ploidy based on a cutoff of 2.5 within early or advanced pathological stage. (**C**) KM survival curves of overall survival or progression free interval in four cancer types over patient groups defined by ploidy-unadjusted TmS and stage. *P* values of log-rank tests between high and low TmS groups within each of the three pairs of patient groups are shown with matching colors and are indicated by asterisk (* *P* < 0.05, ** < 0.01, *** < 0.001).

Association with genomic dysregulations and hypoxia

Tumor mutation burden (TMB) was calculated by counting the total number of all somatic mutations based on the consensus mutations calls (MC3)[59]. Chromosomal Instability (CIN) scores were calculated as the ploidy-adjusted percent of genome with an aberrant copy number state. ASCAT was used to calculate allele-specific copy numbers[14]. For samples present in both TCGA and PCAWG, the consensus copy number was derived from published results[60]. Tumor samples that had undergone WGD were identified based on homologous copy-number information[13]. Hypoxia scores were generated as described previously[61,62] using the Buffa Signature[63], which have been previously shown to be well-correlated with direct and transcriptional measures of hypoxia[64]. Signature scoring was done on Level 3 mRNA abundance data (2016-01-28 data release) as a single cohort of 7,791 patients to ensure comparability of scores across cancer types. For each gene in the signature, patients were median dichotomized. Patients with RNA abundance above the median were assigned a gene-score of +1, while those with RNA abundance below the median were assigned a gene-score of -1. The gene-scores for all signature genes were summed to generate a per-patient hypoxia-score. High values of this score suggest more hypoxia (lower levels of oxygen), while low values suggest less hypoxia (higher levels of oxygen).

We calculated the Spearman correlation coefficients between TmS, and hypoxia scores/TMB/CIN scores for each cancer type (**SI Table 8**). Within each cancer type, tumor samples were evenly split into high

SI Table 8. Spearman correlation coefficients between TmS and hypoxia scores/TMB/CIN scores across 15 cancer types.

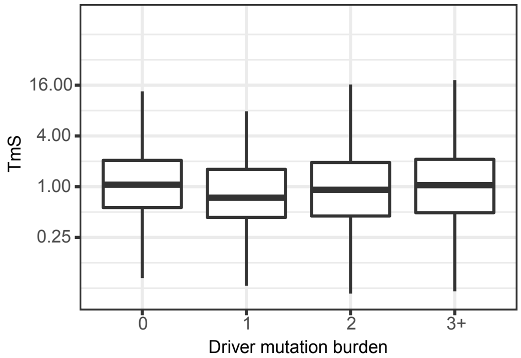
Cancer type	Hypoxia score	TMB	CIN
Breast carcinoma	0.65	0.35	0.46
Lung adenocarcinoma	0.61	0.3	0.23
Thyroid papillary carcinoma	0.44	0.066	0.1
Pancreatic adenocarcinoma	0.43	0.082	0.1
Renal clear cell carcinoma	0.41	0.17	0.35
Lung squamous cell carcinoma	0.38	0.04	0.051
Bladder urothelial carcinoma	0.35	0.21	0.24
Renal papillary carcinoma	0.35	-0.21	0.18
Colorectal carcinoma	0.3	-0.021	0.056
Prostate adenocarcinoma	0.26	0.072	0.3
Endometrial carcinoma	0.25	-0.062	0.16
Hepatocellular carcinoma	0.2	-0.066	-0.095
Head & neck squamous cell carcinoma (HPV+)	0.17	0.32	0.27
Head & neck squamous cell carcinoma (HPV-)	-0.08	-0.017	-0.081
Stomach adenocarcinoma	NA	-0.073	-0.15
Renal Chromophobe	NA	0.13	0.23

TMB/CIN score/WGD/Hypoxia score or low TMB/CIN score/WGD/hypoxia score, based on the median value of TMB/CIN score/WGD/Hypoxia score, respectively. The Wilcoxon rank-sum test was then used to compare the distributions of TmS estimates between tumor samples with high and low TMB/CIN score/Hypoxia score (**Fig. S5A,B,D**).

We calculated median value of TMB/CIN score/Hypoxia score of samples from groups defined by TmS and stage in 14 cancer types (**Fig. 5D**). The Kruskal-Wallis test was used to test the distribution of the TMB/CIN score/Hypoxia score across these groups. The null hypothesis was that there is no difference of TMB/CIN score/Hypoxia score for each group defined by TmS and stage. The *P* values from the Kruskal-Wallis tests were adjusted for multiple testing using Benjamini-Hochberg correction across all cancer types.

Association with somatic SNVs and indels

Driver mutation annotation (including nonsense, missense and splice-site SNVs and indels) was obtained from a TCGA pan-cancer driver mutation database[30]. For each cancer type, we considered a gene as a candidate gene, if there were at least 10 samples containing driver mutations in that gene. For any given candidate gene, individuals were labelled as "Driver Mutant" if the tumor sample carried at least one driver mutations in that candidate gene or "WT" if no SNVs or indels were identified (**Fig. 5B**). The mutation types of *TP53* driver mutations are shown in the **SI Table 9**. We applied a Wilcoxon rank-sum test to each candidate gene to compare the distributions of TmS of the Driver Mutant and WT samples. The *P* values of each gene was adjusted for multiple testing using Benjamini-Hochberg correction across all candidate genes within the corresponding cancer type. We did not observe any significant association between TmS and driver mutation burden across 15 cancer types (**SI Figure 13**).



SI Figure 13. Distribution of TmS with respect to driver mutation burden across 15 TCGA cancer types. For each sample, the total number of driver mutation were calculated and categorized by four categories: no driver mutation ("0"), only one driver mutation ("1"), only two driver mutations ("2"), and equal or more than three driver mutations ("3+").

SI Table 9. Distribution of *TP53* driver somatic mutation type for 4 cancer types in Fig. 5B.

Cancer type	Driver mutation type	Frequency
Breast carcinoma	Frame_Shift_Del	16
Breast carcinoma	Frame_Shift_Ins	10
Breast carcinoma	Missense_Mutation	140
Breast carcinoma	Nonsense_Mutation	30
Breast carcinoma	Splice_Site	19
Lung adenocarcinoma	Frame_Shift_Del	1
Lung adenocarcinoma	Frame_Shift_Ins	1
Lung adenocarcinoma	Missense_Mutation	42
Lung adenocarcinoma	Nonsense_Mutation	13
Lung adenocarcinoma	Splice_Site	7
Prostate adenocarcinoma	Frame_Shift_Del	3
Prostate adenocarcinoma	Missense_Mutation	15
Prostate adenocarcinoma	Splice_Site	1
Stomach adenocarcinoma	Frame_Shift_Del	5
Stomach adenocarcinoma	Missense_Mutation	15
Stomach adenocarcinoma	Nonsense_Mutation	7
Stomach adenocarcinoma	Splice_Site	1

We also implemented an agnostic search over all genes for the 15 available cancer types to identify, among non-silent mutations (including SNVs and indels), those were significantly associated with TmS. We applied two statistical tests to evaluate the difference between the “mutant” and “wild type” samples. We first applied a Wilcoxon rank-sum test for each candidate gene to evaluate the difference between the distribution of TmS of the mutant and wild-type samples. We then fitted a linear regression model using $\log_2$-transformed TmS as the dependent variable and mutation status as a predictor: $log_2(TmS) = b_0 + b_1 log_2(TMB) + b_2 MUT$, where TMB represents tumor mutation burden. $MUT = 1$ if the sample has at least one mutation in the candidate gene, and $MUT = 0$ otherwise. The *P* values were calculated by a t-test of the regression coefficient b_2. The *P* values of each cancer-gene pair from both Wilcoxon rank-sum tests and t-tests were adjusted using Benjamini-Hochberg correction across all candidate genes within the corresponding cancer type. We also performed the same analysis for silent mutations and did not found any gene in any cancer type that is significantly different in terms of TmS based on the adjusted *P* values of Wilcoxon rank-sum test and t-test.

2.3.3. ICGC-EOPC

Data pre-processing

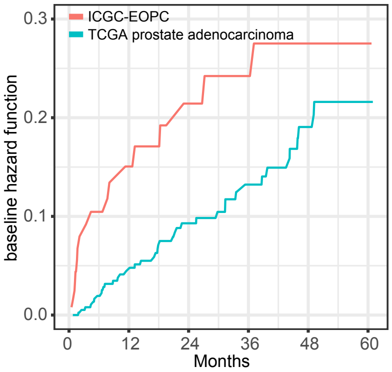
For the ICGC-EOPC dataset, all 121 tumor samples from 96 patients passed the quality control as described in **Section 2.3.2**. We then used the 9 available adjacent normal samples from this cohort as the normal reference for DeMixT to estimate the tumor-specific mRNA expression proportions for the tumor samples. The RNAseq data came from 3 batches - batch 1 (17 patients, 25 samples), batch 2 (42 patients, 52 samples) and batch 3 (37 patients, 44 samples). To evaluate and adjust for potential batch effects, we applied the DeMixT deconvolution pipeline in three scenarios: (1) ran all samples together; (2) ran each batch separately; (3) applied Combat[65] to adjust for batch effect then ran all samples together. The Spearman correlation coefficients between in tumor-specific mRNA expression proportions obtained in scenarios (1) and (2), scenarios (1) and (3), scenarios (2) and (3) were 0.9, 0.8, 0.8, respectively. Given the consistency across the three scenarios and the robustness of DeMixT to select signature genes that can capture global transcription level across samples from subtypes, we presented results from scenario (1).

To calculate TmS values, we additionally removed eight samples from seven patients without DNA based tumor purity and ploidy estimates based on Sequenza[42]. For six patients with multiple regions, we calculated the maximum TmS value across regions to represent patient-wise TmS value for downstream analysis, which was a summary strategy for multi-region samples consistent with our findings in the TRACERx dataset (see Section 2.3.4 for details). For survival analysis, an additional 10 patients were removed due to missing follow-up information of biochemical recurrence intervals. The CONSROT diagram that demonstrates the sample exclusion for TmS is shown below in **Fig. S7C.**

Survival prediction model using TmS as a prognostic feature

We applied the recursive partitioning survival tree model, rpart, to iteratively partition samples by TmS and the Gleason score (Gleason Score = 7 versus Gleason Score = 8+). The splitting criteria were Gini index, and the maximum tree depth was set to 2. The TmS cutoff of Gleason Score = 7/8+ groups are shown in **SI Table 7**. The summary statistics and percentage of disease progression risk of TmS and Gleason score defined groups comparing TCGA prostate adenocarcinoma and early-onset prostate adenocarcinoma are shown in **Table S4A**. We then fitted multivariate Cox Proportional Hazard models with age, TmS, stage, and the interaction term of TmS and Gleason score (TmS x Gleason score) as predictors for progression free interval analysis of TCGA prostate adenocarcinoma (**Table S4B**)**.**

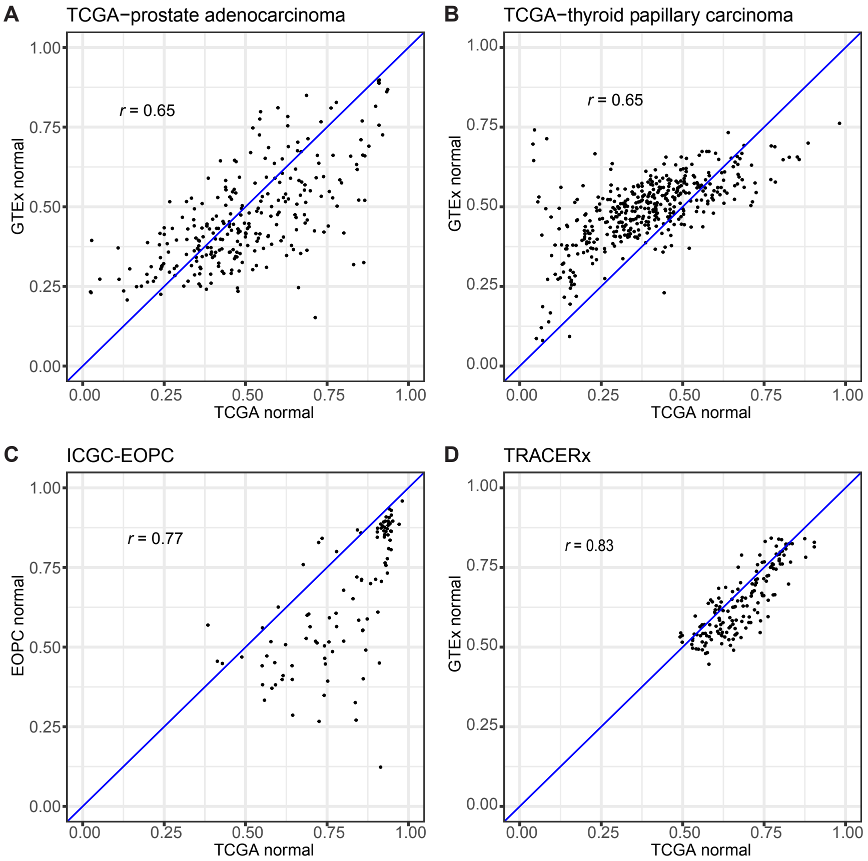
To perform an external validation for the risk prediction model with TmS, we first built a multivariant Cox regression model with age, TmS, Gleason score and TmS x Gleason score as predictors based on the TCGA prostate adenocarcinoma data. We then used the trained Cox model to predict disease progression risk for patients from the EOPC study, taking only the observed covariate values from these patients (**Table S4B**). We compared the prediction performances in these two studies between the “Age and Gleason score model”, which only contains covariates such as age and Gleason score, and the “Age and TmS x Gleason score model”, which contains covariates such as age, TmS, Gleason score and TmS x Gleason score. We evaluated the model’s discrimination ability using Uno’s estimator of cumulative AUC (iAUC)[59] for survival outcomes and constructed 95% confidence intervals of iAUC by bootstrap resampling with 1,000 repetitions (**Fig. S4G**). To measure the calibration ability of the TmS model, we also calculated the Integrated Brier Score (IBS)[66] for 5-year survival. The median iAUC, median IBS, and constructed 95% confidence intervals of d-iAUC and d-IBS are shown in **Table S4B**. For each bootstrap repetition, d-iAUC is defined as iAUC of “Age and TmS x Gleason score model” subtracted by the iAUC of the “Age and Gleason score model”. d-IBS is defined as IBS of “Age and TmS x Gleason score model” subtracted by the IBS of the “Age and Gleason score model”. The higher 5-year IBS for the external validation is mainly due to distinct baseline hazard functions between the testing and training datasets (**SI Figure 14**), as the hazard ratios for the TmS and Gleason score are similar for the two datasets (**Table S4B**).



SI Figure 14. Baseline hazard functions of Cox models for the TCGA prostate adenocarcinoma and the ICGC-EOPC datasets. The baseline is calculated for samples with a Gleason score of 7 and low TmS in each dataset.

2.3.4. TRACERx

This dataset does not contain RNAseq data from adjacent normal samples, which is required for running DeMixT. Instead, we used RNAseq data from normal lung samples which are available in the GTEx study. To mitigate the technical artefacts, such as batch effects, scale normalization was applied before deconvolution. The final tumor-specific mRNA expression proportions were estimated using the DeMixT_DE mode with the top 1,500 genes[19]. DNA-based tumor purity and ploidy were estimated by Sequenza[42].



SI Figure 15. DeMixT deconvolution using normal reference from different studies. (**A**) Scatter plot of DeMixT estimated tumor proportions for TCGA-prostate adenocarcinoma samples using GTEx normal (y axis) or TCGA normal (x axis) samples. (**B**) Scatter plot of DeMixT estimated tumor proportions of EOPC using EOPC normal (y axis) and TCGA normal (x axis) samples. (**C**) Scatter plot of DeMixT estimated tumor proportions of TCGA-thyroid papillary carcinoma samples using GTEx normal (y axis) and TCGA normal (x axis) samples. (**D**) Scatter plot of DeMixT estimated tumor proportion of TRACERx samples using GTEx normal (y axis) and TCGA normal (x axis) samples. Spearman correlation coefficients (*r*) between the two sets of tumor proportion estimates are shown on the top of each panel.

Deconvolution using normal reference samples from GTEx

We conducted a series of experiments across cancer types to evaluate the impact of technical artefacts such as batch effects to the proportion estimation when using a different cohort. We first applied GTEx expression data[48] from normal prostate samples as the normal reference to deconvolute the TCGA prostate cancer samples. Even though the overall performance of deconvolution was negatively impacted, the estimated proportions showed a reasonable correlation (Spearman correlation coefficient = 0.65) with those generated using TCGA normal prostate samples as the normal reference (**SI Figure 15A**). We repeated this experiment on the deconvolution of TCGA thyroid papillary carcinoma samples using RNAseq data from TCGA normal and GTEx normal thyroid samples as the reference, respectively. Again, the two sets of estimated tumor-specific mRNA expression proportions were highly correlated (Spearman correlation coefficient = 0.65) (**SI Figure 15B**). For the EOPC tumor samples where RNAseq data from 9 normal samples were available, we observed a higher correlation (Spearman correlation coefficient = 0.77) between the estimated tumor-specific mRNA expression proportions using EOPC normal and TCGA prostate normal samples on the deconvolution of EOPC tumor samples, respectively (**SI Figure 15C**). Furthermore, for the deconvolution of TRACERx tumor samples, we also observed a high correlation (Spearman correlation coefficient = 0.83) between the estimated tumor proportions using TCGA and GTEx normal lung samples as the reference, respectively. (**SI Figure 15D**).

We calculated TmS values for all regions (median number of regions per patient = 2, ranging from 1 to 6) in the TRACERx dataset. The CONSORT diagram demonstrating the sample exclusion for TmS is shown below in **Fig. S7D.**

Association of regional TmS with measures of chromosomal instability

We calculated the percentage of copy number alteration burden per region, the percentage of subclonal copy number alteration per region, and the percentage of subclonal copy number alteration per patient.

For each chromosomal segment *i* in tumor region *k*, we define the copy number alteration (gain and loss) event[43] as an indicator function I_{ik},

$$I_{ik}=\begin{cases} 1 & \text{if } \alpha>\log_2(2.5/2) \text{ or } \alpha<\log_2(1.5/2) \\ 0 & \text{otherwise} \end{cases},$$

where $\alpha=\frac{cnTotal_{ik}}{Ploidy_k}$ and $cnTotal_{ik}$ is the integer total copy number of this segment[43].

We then define the percentage of CNA (copy number alteration) burden for each region as the percentage of genome affected by copy number alterations,

$$\textit{percentage of CNA burden}_k=\frac{\sum_{i=1}^{nS} D_i\times I_{ik}}{\sum_{i=1}^{nS} D_i}\times 100\%,$$

where *nS* and D_i denotes the number of shared segments and the length of shared segment *i* across regions, respectively.

Further, for each region, whether the segment *i* has a subclonal copy number alteration event is defined as

$$S_{ik}=\begin{cases} 1 & I_{ik}=1 \text{ and} \sum_{k=1}^{K} I_{ik} \neq K \\ 0 & \text{Otherwise} \end{cases},$$

where *K* is the total number of regions for a given tumor sample. We then introduce S_i as an indicator function representing the subclonal copy number alteration event on shared segment *i* across regions: $S_i = \bigcup_{k=1}^{K} S_{ik}$.

Besides, we define T_i as an indicator function which denotes whether there is a copy number alteration event (including clonal and subloncal) on shared segment *i*.

$$T_i=\begin{cases} 1 & 0 < \sum_{k=1}^{K} I_{ik} \leq K \\ 0 & \text{Otherwise} \end{cases}.$$

Therefore, the percentage of subclonal CNA for region *k* (percentage of subclonal CNA per region) is defined as

$$\text{percentage of subclonal } CNA_k = \frac{\sum_{i=1}^{nS} D_i \times S_{ik}}{\sum_{i=1}^{nS} D_i \times T_i} \times 100\%.$$

Correspondingly, the percentage of subclonal CNA for each patient is defined as

$$\text{percentage of subclonal } CNA = \frac{\sum_{i=1}^{nS} D_i \times S_i}{\sum_{i=1}^{nS} D_i \times T_i} \times 100\%.$$

Across regions, the Spearman correlation coefficient between $\log_2$(TmS) and percentage of subclonal CNA per region is 0.44; the Spearman correlation coefficient between $\log_2$(TmS) and copy number aberration burden per region is 0.26. The difference between these two correlation coefficients between is statistically significant (bootstrapping 1,000 times, mean difference = 0.2, 95% confidence interval: 0.04, 0.37).

Two subclonal structures in two regions can be linearly related to each other, or have a common ancestor, but develop a branching relationship, which is more common in this dataset (**Fig. 6A**). For example, a linear relationship can be described as a parent and child relationship, where two subclonal structures share overlapped segments and one structure evolves further than the other. For a branching relationship, two subclonal structures usually share a common node (ancestor), and two structures evolve in different directions. The subclonal structures of 5 out of 30 patients are defined as linear relationships. For each evolutionary relationship per patient sample, we defined the *range of TmS* = $\log_2(\text{maximum } TmS) - \log_2(\text{minimum } TmS)$ across regions (**SI Table 10**). We observed a strong correlation between $\log_2(\text{TmS}_{\text{max}})$ and percentage of subclonal CNA among 30 patients with multi-region sequencing data

(Spearman correlation coefficient *r* = 0.69, **Fig. S6A)**. To further explore the underlying relationship between $\log_2(\mathrm{TmS}_{max})$ and all variables (e.g., percentage of subclonal CNA, number of regions, range of TmS, evolutionary relationship and their interactions) across patients, we fit linear regression models by taking TmS_{max} as the response variable and others as predictors. The best model was selected by stepwise adding or dropping one predictor that achieves the best AIC (Akaike's Information Criteria, R function stepAIC) (**Fig. 6E, SI Table 11A**). We also adopted a logistic regression model by taking the evolutionary relationship as the response variable, and after the model selection (likelihood ratio test), percentage of subclonal CNA and range of TmS were chosen separately as predictor variables (**SI Table 11B-C**, **SI Figure 15**).

Association between TmS and survival outcomes

We used rpart[58] to find an optimized cutoff on TmS_{max} that best separates survival outcomes of the patients (cutoff = 3.5, log-rank test *P* value = 0.02, **Fig. 6F**). As a negative control, we ranked the patients using TmS_{med} across regions, then assigned the top 10 patients into one group and the rest of the patients go into the other group. The disease-free survival outcomes of the two groups are not significantly different (log-rank test *P* value = 0.8, **Fig. 6G**).

The percentage of subclonal CNA was shown to be highly associated with survival outcomes within this dataset[43], which was recapitulated in our analysis (**Fig. S6B**). Using both TmS_{max} and the percentage of subclonal CNA in the recursive partitioning survival tree model, we separated 30 patients into three groups with distinct survival outcomes (log-rank test *P* value = 0.003, **Fig. 6H**). In the negative control, we kept the number of groups and the number of patients in each group. We first used TmS_{med} to rank the patients, and assigned the top 7 into high TmS_{med} group. The rest of 23 were further sorted into two other groups with high (9 patients) and low (14 patients) percentage of subclonal CNA. The disease-free survival outcomes of the three groups are not significantly different (log-rank test P value = 0.7, **Fig. S6C).**

Finally, we added patients with only a single region sample to the patients with multiple-region samples, resulting in 52 patients (**Fig. S6D**), and separated them into two groups using TmS_{max} values and rpart. The KM survival curves of disease-free probability remained significantly different between the two groups (log-rank test *P* value = 0.005) (**Fig. S6E**).

SI Table 10. Evolutionary relationships for 30 TRACERx patients with multi-region samples.

Patient	Histology	Evolutionary Relationships	Region with Maximum TmS	Maximum TmS	Region with Minimum TmS	Minimum TmS	Range of TmS
CRUK0005	LUAD	Branching	R4	3.5	R3	3.4	0.050
CRUK0013	LUAD	Branching	R2	3.0	R3	1.6	0.90
CRUK0017	LUAD	Branching	R4	1.9	R1	1.3	0.58
CRUK0018	LUAD	Branching	R4	3.4	R2	0.88	2.0
CRUK0021	LUAD	Branching	R1	1.8	R2	1.7	0.050
CRUK0023	LUAD	Branching	R4	2.7	R1	0.80	1.7
CRUK0024	LUAD	Branching	R1	4.1	R4	2.2	0.89
CRUK0025	LUAD	Branching	R3	1.8	R1	0.87	1.0
CRUK0029	LUAD	Branching	R2	4.0	R6	2.2	0.85
CRUK0030	LUAD	Linear	R2	2.7	R3	2.4	0.14
CRUK0033	LUAD	Linear	R1	1.3	R2	0.85	0.58
CRUK0036	LUAD	Branching	R4	7.4	R2	5.4	0.47
CRUK0037	LUAD	Branching	R2	7.5	R3	1.5	2.3
CRUK0039	LUAD	Branching	R1	2.3	R2	2.0	0.19
CRUK0041	LUAD	Branching	R4	2.5	R1	1.8	0.48
CRUK0046	LUAD	Branching	R2	2.5	R1	1.6	0.65
CRUK0047	LUAD	Branching	R2	2.7	R1	2.4	0.16
CRUK0050	LUAD	Linear	R4	1.1	R3	0.98	0.19
CRUK0057	LUAD	Branching	R1	2.7	R2	2.0	0.40
CRUK0062	LUSC	Branching	R7	4.0	R2	1.7	1.2
CRUK0065	LUSC	Branching	R3	3.9	R1	1.7	1.2
CRUK0067	LUSC	Branching	R1	2.2	R3	1.3	0.73
CRUK0069	LUSC	Branching	R1	3.5	R3	0.81	2.1
CRUK0070	LUSC	Branching	R6	1.4	R1	0.85	0.72
CRUK0076	LUSC	Linear	R2	2.9	R4	2.6	0.16
CRUK0077	LUSC	Branching	R1	3.7	R2	1.4	1.5
CRUK0079	LUSC	Branching	R1	3.5	R3	2.0	0.81
CRUK0083	LUSC	Branching	R3	3.7	R1	1.2	1.6
CRUK0084	LUSC	Branching	R2	0.91	R3	0.72	0.34
CRUK0090	LUSC	Linear	R1	1.3	R2	1.0	0.30

SI Table 11. Summary of regression models

A. linear regression model with maximum TmS as response variable

Variable	Coefficient	T-statistics	Standard Error	*P* value
Intercept	0.3	0.7	0.5	0.5
% Subclonal CNA	2.8	4.5	0.6	0.0002***
Range of TmS	0.3	0.4	0.6	0.7
No. of Regions	-0.1	-1	0.1	0.3
% Subclonal CNA * Range of TmS	-1.8	-2.3	0.8	0.03*
Range of TmS * No. of Region	0.3	2.2	0.1	0.04*

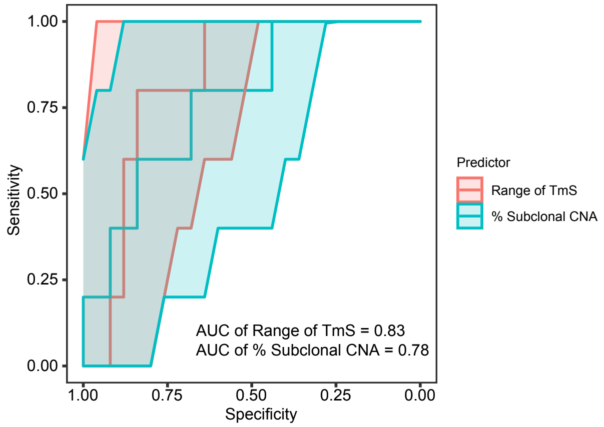
F-statistics	R-squared	Adjusted R-squared	RMSE	*P* value
9.4 on 5 and 24 DF	0.7	0.6	0.4	4x10^{-05} ***

B. Logistics regression model with Range of TmS as predictor and Evolutionary Relationships (Branching = 1, Linear = 0) as response variable

Variables	Coefficient	Z-statistics	Standard Error	*P* value
Range of TmS	3.3	2.7	1.3	0.008**

C. Logistics regression model with % Subclonal CNA as predictor and Evolutionary Relationships (Branching = 1, Linear = 0) as response variable

Variables	Coefficient	Z-statistics	Standard Error	*P* value
% Subclonal CNA	4.3	3.1	1.4	0.002**



SI Figure 15. ROC curves for predicting evolutionary relationships: branching versus linear. Two logistic models were used (**SI Table 11B-C**), with either the range of TmS or the percentage of subclonal CNA as the predictor. The 95% confidence intervals and area under the ROC curves (AUC) are provided.

REFERENCES

1. Ma, L. *et al.* Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. *Cancer Cell* **36**, 418–430 (2019).

2. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).

3. Lee, J. J. *et al.* Elucidation of tumor-stromal heterogeneity and the ligand-receptor interactome by single cell transcriptomics in real-world pancreatic cancer biopsies Running Title: scRNA-seq of pancreatic cancer real-world biopsies. *bioRxiv* 2020.07.28.225813 (2020). doi:10.1101/2020.07.28.225813

4. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

5. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).

6. 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).

7. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).

8. Hashimoto, K. *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24242–24251 (2019).

9. Wang, Y. J. *et al.* Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues. *bioRxiv* 541433 (2019). doi:10.1101/541433

10. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

11. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879–893 (2018).

12. Benjamini, Y. & Hochberg, Y. Controlling for the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

13. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413 (2012).

14. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.*

107, 16910–16915 (2010).

15. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).

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

17. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).

18. Li, B. *et al.* Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).

19. Wang, Z. *et al.* Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, 451–460 (2018).

20. Quon, G. *et al.* Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.* **5**, 29 (2013).

21. Brown, P. J. & Lehmann, E. L. *Theory of Point Estimation. Journal of the Royal Statistical Society. Series A (General)* (1984). doi:10.2307/2981857

22. Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).

23. Besag, J. On the Statistical Analysis of Dirty Pictures. *J. R. Stat. Soc. Ser. B* **48**, 259–279 (1986).

24. Cox, D. R. & Reid, N. A Note on the Calculation of Adjusted Profile Likelihood. *J. R. Stat. Soc. Ser. B* **55**, 467–471 (1993).

25. Venzon, D. J. & Moolgavkar, S. H. A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Appl. Stat.* **37**, 87–94 (1988).

26. Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *Ann. Stat.* **13**, 70–84 (1985).

27. Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, L. M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

28. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).

29. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science (80-.).* **354**, 618–622 (2016).

30. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance

of tumor alterations. *Genome Med.* **10**, 25 (2018).

31. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science (80-.).* **362**, (2018).

32. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

33. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61 (2012).

34. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

35. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **10**, 401–404 (2012).

36. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).

37. Lawrence, M. S. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).

38. Linehan, W. M. *et al.* Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).

39. Gerhauser, C. *et al.* Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell* **34**, 996–1011 (2018).

40. Weischenfeldt, J. *et al.* Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer. *Cancer Cell* **23**, 159–170 (2013).

41. Li, H. [Heng Li - Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* **1303**, (2013).

42. Favero, F. *et al.* Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).

43. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* (2017). doi:10.1056/NEJMoa1616288

44. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature* (2019). doi:10.1038/s41586-019-1032-7

45. Biswas, D. *et al.* A clonal expression biomarker associates with lung cancer mortality. *Nature*

Medicine **25**, 1540–1548 (2019).

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

47. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

48. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).

49. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

50. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* (2006). doi:10.1186/gb-2006-7-s1-s4

51. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243 (2013).

52. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

53. Dempster, J. M. *et al.* Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nat. Commun.* **10**, 1–14 (2019).

54. 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).

55. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).

56. Franz, M. *et al.* GeneMANIA update 2018. *Nucleic Acids Res.* **46**, W60–W64 (2018).

57. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416 (2018).

58. Therneau, T. M. & Atkinson, E. J. An Introduction to Recursive Partitioning Using the RPART Routines. *Mayo Found. Tech. Rep.* **61**, 452 (1997).

59. Uno, H., Cai, T., Tian, L. & Wei, L. J. Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* **102**, 527–537 (2007).

60. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

61. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**,

308–318 (2019).

62. Bhandari, V., Li, C. H., Bristow, R. G. & Boutros, P. C. Divergent mutational processes distinguish hypoxic and normoxic tumours. *Nat. Commun.* **11**, 1–10 (2020).

63. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).

64. Lalonde, E. *et al.* Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: A retrospective cohort study. *Lancet Oncol.* **15**, 1521–1532 (2014).

65. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

66. Mogensen, U. B., Ishwaran, H. & Gerds, T. A. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *J. Stat. Softw.* **50**, 1 (2012).