Title

Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression

Journal

Nature Biotechnology, 40(11)

ISSN

1087-0156

Authors

Cao, Shaolong
Wang, Jennifer R
Ji, Shuangxi
et al.

Publication Date

2022-11-01

DOI

10.1038/s41587-022-01342-x

Peer reviewed

ARTICLES
https://doi.org/10.1038/s41587-022-01342-x

nature biotechnology

Check for updates

OPEN

Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression

Shaolong Cao[1,31], Jennifer R. Wang[2,31], Shuangxi Ji[1,31], Peng Yang[1,3], Yaoyi Dai[1,4], Shuai Guo[1], Matthew D. Montierth[1,4], John Paul Shen[5], Xiao Zhao[2], Jingxiao Chen[1], Jaewon James Lee[6,7,8], Paola A. Guerrero[6,7], Nicholas Spetsieris[9], Nikolai Engedal[10], Sinja Taavitsainen[11], Kaixian Yu[12], Julie Livingstone[13,14,15,16], Vinayak Bhandari[17], Shawna M. Hubert[18], Najat C. Daw[19], P. Andrew Futreal[20], Eleni Efstathiou[9], Bora Lim[21], Andrea Viale[20], Jianjun Zhang[18], Matti Nykter[11], Bogdan A. Czerniak[22], Powel H. Brown[23], Charles Swanton[24], Pavlos Msaouel[7,9], Anirban Maitra[6,7,22], Scott Kopetz[5], Peter Campbell[25], Terence P. Speed[26,27], Paul C. Boutros[13,14,15,16,17], Hongtu Zhu[12], Alfonso Urbanucci[10], Jonas Demeulemeester[28,29], Peter Van Loo[28,30] and Wenyi Wang[1,12] ✉

Single-cell RNA sequencing studies have suggested that total mRNA content correlates with tumor phenotypes. Technical and analytical challenges, however, have so far impeded at-scale pan-cancer examination of total mRNA content. Here we present a method to quantify tumor-specific total mRNA expression (TmS) from bulk sequencing data, taking into account tumor transcript proportion, purity and ploidy, which are estimated through transcriptomic/genomic deconvolution. We estimate and validate TmS in 6,590 patient tumors across 15 cancer types, identifying significant inter-tumor variability. Across cancers, high TmS is associated with increased risk of disease progression and death. TmS is influenced by cancer-specific patterns of gene alteration and intra-tumor genetic heterogeneity as well as by pan-cancer trends in metabolic dysregulation. Taken together, our results indicate that measuring cell-type-specific total mRNA expression in tumor cells predicts tumor phenotypes and clinical outcomes.

Reprogramming of the transcriptional landscape is a critical hallmark of cancer, which accompanies cancer progression, metastasis and resistance to treatment[1,2]. Recent single-cell studies revealed that expansion of cell state heterogeneity in cancer cells arises largely independently of genetic variation[3–9], bringing new conceptual insights into longstanding topics of cancer cell plasticity[10] and cancer stem cells[11,12]. Assessing these clinically relevant topics[13,14] in large patient cohorts, however, has been difficult due to the high cost and sample quality requirements associated with single-cell technologies. As bulk tumor RNA and DNA sequencing data are already available from large patient series with clinical outcomes, in silico approaches to analyze human tissues may expedite our understanding of tumor heterogeneity.

Some features of transcriptional diversity are more easily quantified in bulk tissues than others. For example, previous approaches to build cellular differentiation hierarchies are not suitable for

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Department of Head and Neck Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Department of Statistics, Rice University, Houston, TX, USA. [4]Baylor College of Medicine, Houston, TX, USA. [5]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [6]Sheikh Ahmed Center for Pancreatic Cancer Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [7]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [8]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [9]Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [10]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway. [11]Faculty of Medicine and Health Technology, Tampere University and Tays Cancer Center, Tampere, Finland. [12]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [13]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [14]Department of Urology, University of California, Los Angeles, Los Angeles, CA, USA. [15]Institute for Precision Health, University of California, Los Angeles, Los Angeles, CA, USA. [16]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, USA. [17]Department of Medical Biophysics, University of Toronto, Toronto ON, Canada. [18]Department of Thoracic Head Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [19]Department of Pediatrics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [20]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [21]Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [22]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [23]Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [24]The Francis Crick Institute, London, UK. [25]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. [26]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VC, Australia. [27]School of Mathematics and Statistics, The University of Melbourne, Melbourne, VC, Australia. [28]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. [29]Department of Human Genetics, KU Leuven, Leuven, Belgium. [30]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [31]These authors contributed equally: Shaolong Cao, Jennifer R. Wang, Shuangxi Ji. ✉e-mail: wwang7@mdanderson.org

large-scale human tissue studies where the individual cell identify is lost. These approaches also further require known cell-type-specific genetic markers[15]. Single-cell studies recently demonstrated that the total number of expressed genes per cell can be more predictive of cellular phenotype, such as developmental status, than alterations in any specific genes or pathways[16,17]. Total number of expressed genes in single cells enabled insights in tumorigenesis of breast[16], colon[18], pancreas[19] and blood[20]. In bulk tissues, variation in total mRNA amount—that is, the sum of detectable mRNA transcripts across all genes per cell—has been indirectly linked to cancer progression and de-differentiation as a result of *MYC* activation[21,22] or aneuploidy[23,24]. With current limitations in our knowledge of marker genes across cancers, total mRNA expression per tumor cell may represent a robust and measurable pan-cancer feature that warrants a systematic evaluation in patient cohorts.

Measuring such a feature in human tissues at-scale poses several analytical challenges, as total tumor cell mRNA expression information is masked during standard bulk data analysis, thus requiring deconvolution. Variation in total mRNA transcript levels is removed by routine normalization, together with technical biases, including read depth and library preparation[25–28]. DNA and RNA sequencing data generated from cancer studies contain reads from both tumor and admixed normal cells. Furthermore, copy number aberrations, such as gain or loss of chromosomal copies (that is, ploidy) in tumor cells, affect gene expression through dosage effects[24].

In this study, building upon prior work in bulk transcriptome deconvolution[29–31] and in modeling tumor ploidy[32,33], we created a measure of tumor-specific total mRNA expression (TmS), which captures the ratio of total mRNA expression per haploid genome in tumor cells versus surrounding non-tumor cells. We first scrutinized total mRNA expression using single-cell data from ten patients across four cancer types[34–36] and then calculated TmS in matching bulk RNA and DNA data from 6,580 patients across 15 cancer types from four large independent cohorts: The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC)[37], the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[38] and Tracking Non-Small-Cell Lung Cancer Evolution through Therapy (TRACERx)[39,40]. Our analyses revealed that variation in total mRNA expression is a robust and prognostic feature across cancers.

Results

Diversity in total mRNA expression across cancer cells. To motivate a model-based quantification of total mRNA expression in bulk tissue, we first analyzed single-cell RNA sequencing (scRNA-seq) data generated from 48,913 cells of ten patients with colorectal ($n=3$), liver ($n=3$)[34], lung ($n=2$)[35] or pancreatic ($n=2$)[36] cancers (Fig. 1a, Extended Data Fig. 1a, Methods and Supplementary Note 1.1). Total unique molecular identifier (UMI) counts of a cell can be modeled as total mRNA molecule counts multiplied by transcript capture efficiency[41]. Following recent studies[9,16], demonstrating gene counts as important markers of cellular differentiation, we further propose to use UMI counts to study tumor behavior in human cancers. We observed strong correlations between total UMI counts and gene counts (the number of detectably expressed genes per cell) across all cell types in the ten tumor samples (median Spearman $r=0.95$ and median absolute deviation (MAD) $=0.04$; Extended Data Fig. 1b), in agreement with a prior study in non-cancerous tissues[16]. This supports total UMI counts having a similar utility as gene counts in characterizing tumor cellular phenotype. By investigating the difference of total UMI count distributions in different cell types, we observed a larger variability in tumor cells compared to non-tumor cells (epithelial, stromal and immune cells) (*F*-test for variances, adjusted *P* values <0.02; Extended Data Fig. 1c,d). Consistent with previous reports[35,42], we found multiple clusters within tumor and non-tumor cells presenting distinct total UMI and gene counts (Fig. 1b,c, Extended Data Fig. 2a, Methods and Supplementary Note 1.2). High-UMI tumor cells generally demonstrate lower cell cycle activity—that is, non-cycling cells[43]—compared to low-UMI tumor cells (Extended Data Fig. 2c and Supplementary Note 1.2.3). Hence, UMI count is not a surrogate measure for proliferation. Trajectory inference using Monocle[44–46] shows distinct gene expression states among these clusters (Fig. 1d and Extended Data Fig. 2b). Tumor cells of high-UMI cluster show a less differentiated state[16] (adjusted *P* values <0.001; Fig. 1d, Extended Data Fig. 2b and Methods). For instance, in patients with a worse survival outcome (colon, liver and pancreas cancers) or advanced-stage disease (lung cancer), the high-UMI tumor cell clusters present a stem-like cell state as predicted by CytoTRACE[16] (Fig. 1c,d, Extended Data Fig. 2b and Supplementary Table 1) and demonstrate an enrichment for stemness and the epithelial–mesenchymal transition (EMT) genes (out of 18,617 gene sets[47,48] investigated; Supplementary Table 2 and Methods). The above observations support the significance of measuring total UMI counts and mRNA content across tumor cells[9,16].

To support the feasibility of quantifying tumor-specific total mRNA expression in bulk tissues, we pooled the scRNA-seq data to generate pseudo-bulks. As single-cell identity is lost in bulk tissues, we introduce the average total UMI counts per cell for each cell type. To allow for inter-patient comparisons and remove potential technical artifacts still contained in the UMI count measure, we further introduce the ratio of the average total UMI counts for tumor versus non-tumor cells for each sample. Using this bulk-level metric, we observed increased tumor mRNA content in the four patients with advanced disease and worse survival outcomes, as compared to other samples within each cancer type (Fig. 1e; adjusted *P* values <0.001). This led us to hypothesize that quantification of average tumor-specific total mRNA expression in bulk sequencing data may track tumor phenotype and clinical behavior.

Estimating tumor-specific total mRNA expression. To quantify the average tumor-specific total RNA expression across a large number of patient samples, we employ three steps in a sequential deconvolution of matched DNA/RNA sequencing data (Fig. 2a, Methods and Supplementary Note 2.1). (1) We estimate the ratio of total RNA expression between two cellular populations, tumor versus non-tumor cells, to cancel out technical effects. This ratio can be estimated as an odds of transcript proportions (π), based on a set of robust intrinsic tumor signature genes. (2) We divide the total RNA expression by their relative cell fractions to calculate a per-cell total RNA content for tumor and non-tumor cells separately. This step requires matched DNA data from which the tumor cell proportion—that is, purity (ρ)—as well as ploidy (Ψ_T) are estimated. (3) We divide the above metric by ploidy (for both components), thereby adjusting for the dosage effect of chromosomal copies on gene expression. We thus calculate our final quantitative metric: the per-cell, per-haploid genome total RNA expression for tumor—that is, TmS—as $[\pi\,(1-\rho)\,\Psi_N]/[\rho\,(1-\pi)\,\Psi_T]$. The parameters ρ and Ψ_T can be derived using DNA sequencing or single-nucleotide polymorphism (SNP) array data (for example, using ASCAT[32], ABSOLUTE[33] or Sequenza[49]; Extended Data Fig. 3a–h and Methods). The parameter π can be derived using RNA sequencing or microarray data (for example, using DeMixT[31]). A major challenge in estimating π is that the unobserved tumor-specific and non-tumor-specific expression levels of many genes present multimodal distributions across tumor subtypes, which would introduce large estimation biases (Extended Data Fig. 4a–d and Methods). To address this issue and obtain more robust π estimates, we introduce a profile likelihood of the DeMixT model to rank genes for each study cohort and identify top-ranked genes as an intrinsic tumor signature gene set, where genes follow a unimodal distribution with low variance across the hidden tumor component and are differentially expressed from the non-tumor

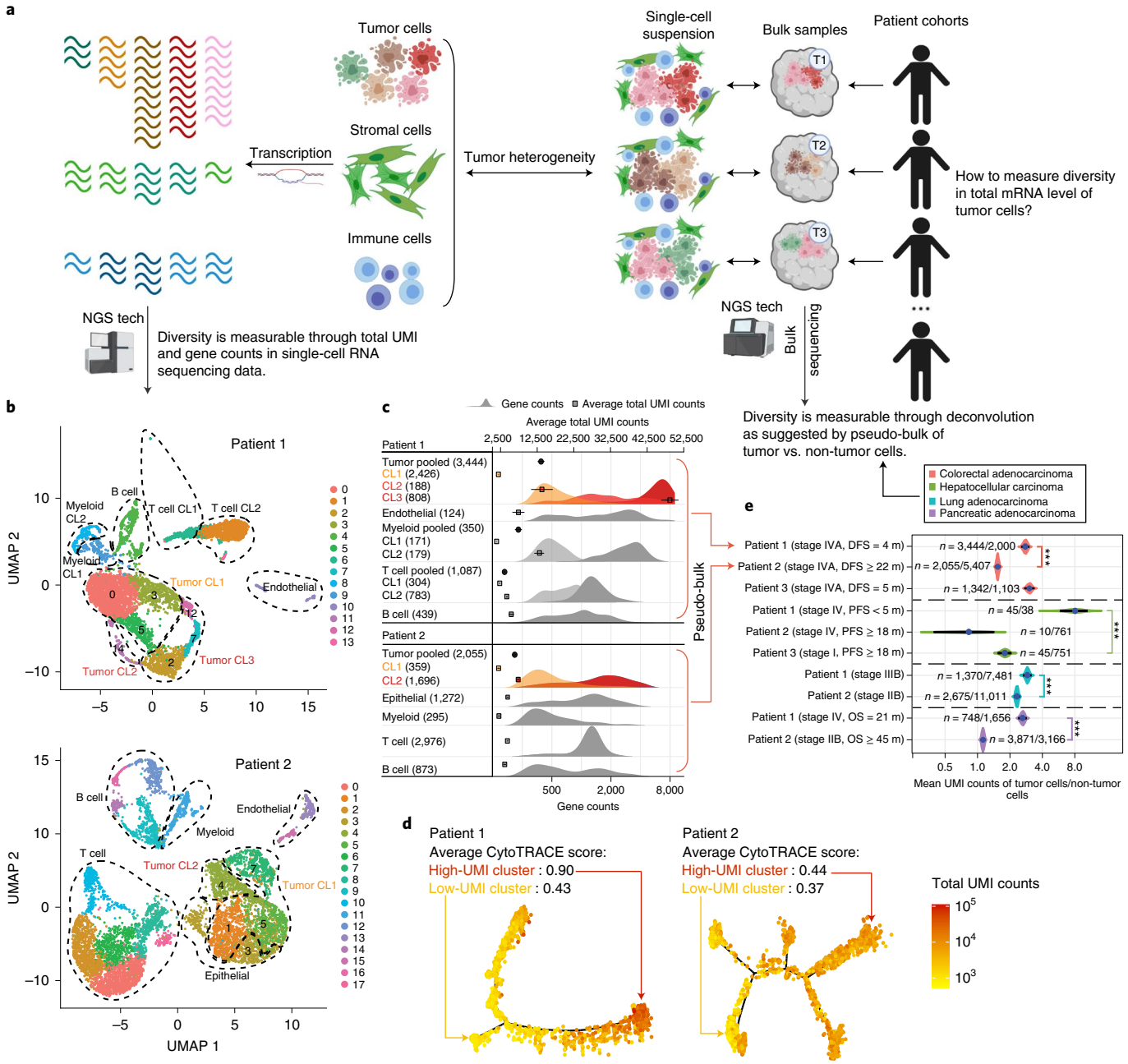


Fig. 1 | High diversity of total mRNA expression in cancer cells. a, Illustration of diversity in total mRNA levels in tumor cells versus other cell types. **b**, UMAP plots of scRNA-seq data from two patients with colorectal cancer. Tumor cell clusters are bolded in both samples. Dashed circles indicate groups of cells that are similar in total UMI and gene counts, which are merged for simplicity. **c**, Distributions of gene counts and total UMI counts by cell type in scRNA-seq data from the two patients shown in **b**. The top *x* axis annotates total UMI counts (with mean and 95% CI). The bottom *x* axis annotates gene count distribution (density). Density curves are colored for tumor cells and shown in grayscale for non-tumor cells. Clusters with higher gene counts are shown in darker shades. Numbers of cells analyzed are indicated in parentheses. Tumor cell clusters are highlighted by the same colors as those in **b**. **d**, Monocle-inferred trajectories for tumor cells from the two patients. Cells on the trees are colored by total UMI counts. Average differentiation scores by CytoTRACE for high-UMI and low-UMI clusters are provided. **e**, Ratios of mean total UMI counts of tumor cells to non-tumor cells (n = number of tumor cells / number of non-tumor cells) and 95% CIs in pooled scRNA-seq data (pseudo-bulk) from ten patients with colorectal ($n = 3$, including patients 1 and 2 shown in **b**–**d**), hepatocellular ($n = 3$), lung ($n = 2$) and pancreatic ($n = 2$) cancers. DFS, disease free survival; PFS, progression-free survival; OS, overall survival. The Benjamini–Hochberg-adjusted P values for two-sided Wilcoxon rank-sum tests comparing the ratios between patient samples are indicated by asterisks (*$P < 0.05$, **$P < 0.01$ and ***$P < 0.001$). UMAP, uniform manifold approximation and projection.

component (Extended Data Fig. 4c,d, Methods and Supplementary Note 2.2). Simulation studies confirmed more robust π estimation when only the intrinsic tumor signature genes are used to perform transcriptome deconvolution (Supplementary Note 2.2).

We benchmarked the performance of TmS estimation using total RNA sequencing data generated from mixed cell populations with known proportions[31], resulting in accurate separation of the H1092 lung cancer cell transcriptome from that of cancer-associated

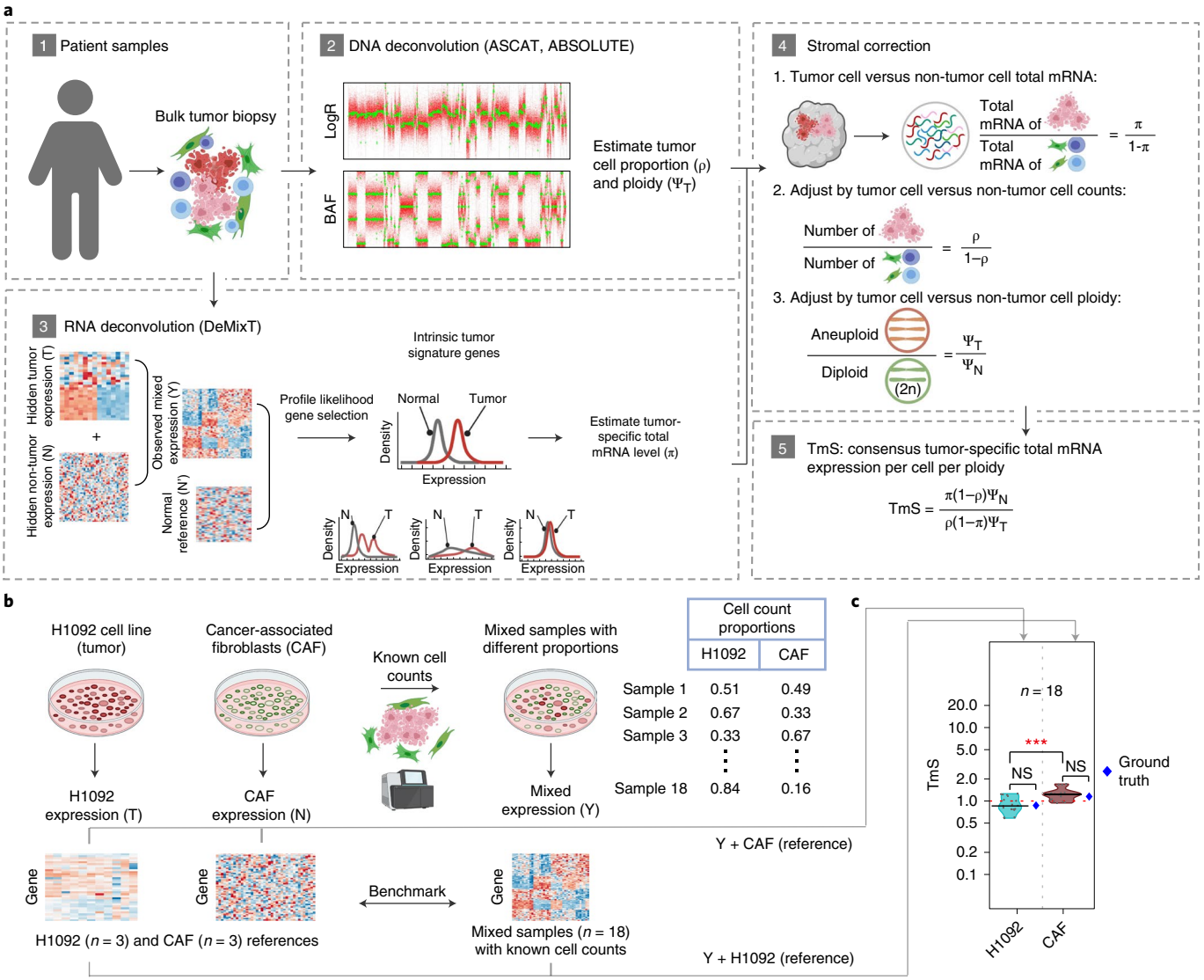


Fig. 2 | Analysis workflow to measure tumor-specific total mRNA expression and benchmarking. a, Calculation of TmS begins with deconvolution using matched DNA sequencing and RNA sequencing data. ASCAT and/or ABSOLUTE are used to estimate tumor purity and ploidy from the DNA sequencing data, whereas DeMixT estimates tumor-specific mRNA proportion from the RNA sequencing. **b**, Benchmarking using bulk RNA sequencing data from in vitro cell lines containing tumor and non-tumor cells, H1092 (human lung cancer) and cancer-associated fibroblasts (CAFs). **c**, Distribution of TmS in 18 mixed cell line samples estimated under two scenarios using DeMixT: (1) three pure CAF samples as known reference and (2) three pure H1092 samples as known reference. The true TmS values for H1092 and CAF are provided in blue dots. They are measured as the ratio of total RNA amount (in ng μl^{-1}) in 1 million cells: 0.87 for H1092 and 1.2 for CAF. The median estimates of TmS are (0.86, 1.2), with MADs of (0.24, 0.18) for H1092 and CAF cells, using the other cell line as the baseline. The *P* values for two-sided Wilcoxon rank-sum tests comparing TmS between groups are indicated (not significant (NS): $P > 0.05$; $^{***}P < 0.001$).

fibroblasts (CAFs) (Fig. 2b,c, Supplementary Table 3, Extended Data Fig. 5a and Methods).

TmS as a measure of tumor-specific total mRNA expression. We calculated TmS across 15 TCGA cancer types, the early-onset prostate cancer (EOPC) cohort from the ICGC, the METABRIC study and the TRACERx study (Fig. 3a,b, Methods and Supplementary Note 2.3). The intrinsic tumor signature genes selected for TmS estimation largely overlap across cancers (Extended Data Fig. 5b) and are enriched in housekeeping, essential[50,51], cancer hallmark[47] and transcriptional regulation pathway genes (RNA splicing and degradation and protein degradation; Extended Data Fig. 5c). As expected, selected genes also demonstrated increased chromatin accessibility[52] versus non-selected genes (Extended Data Fig. 5d). These pan-cancer consistencies support the biological underpinning of TmS as well as our profile-likelihood-based approach for selecting stably and differentially expressed genes in tumor cells. Moreover, all cancer types studied demonstrated a much wider TmS range in patient samples compared to the variance of TmS derived using a homogeneous tumor cell population in the benchmarking study (Fig. 2c versus Fig. 3b; *F*-test for variances, adjusted *P* values < 0.001 for all cancer types). These findings suggest that considerable variation in tumor-specific total mRNA expression exists among patient samples (Fig. 3b, Supplementary Table 4, Methods and Supplementary Note 2.3).

To serve as a meaningful measure, we expect TmS to capture alterations in tumor-specific total mRNA expression attributable to a variety of interacting biological processes (Extended Data Fig. 6a).

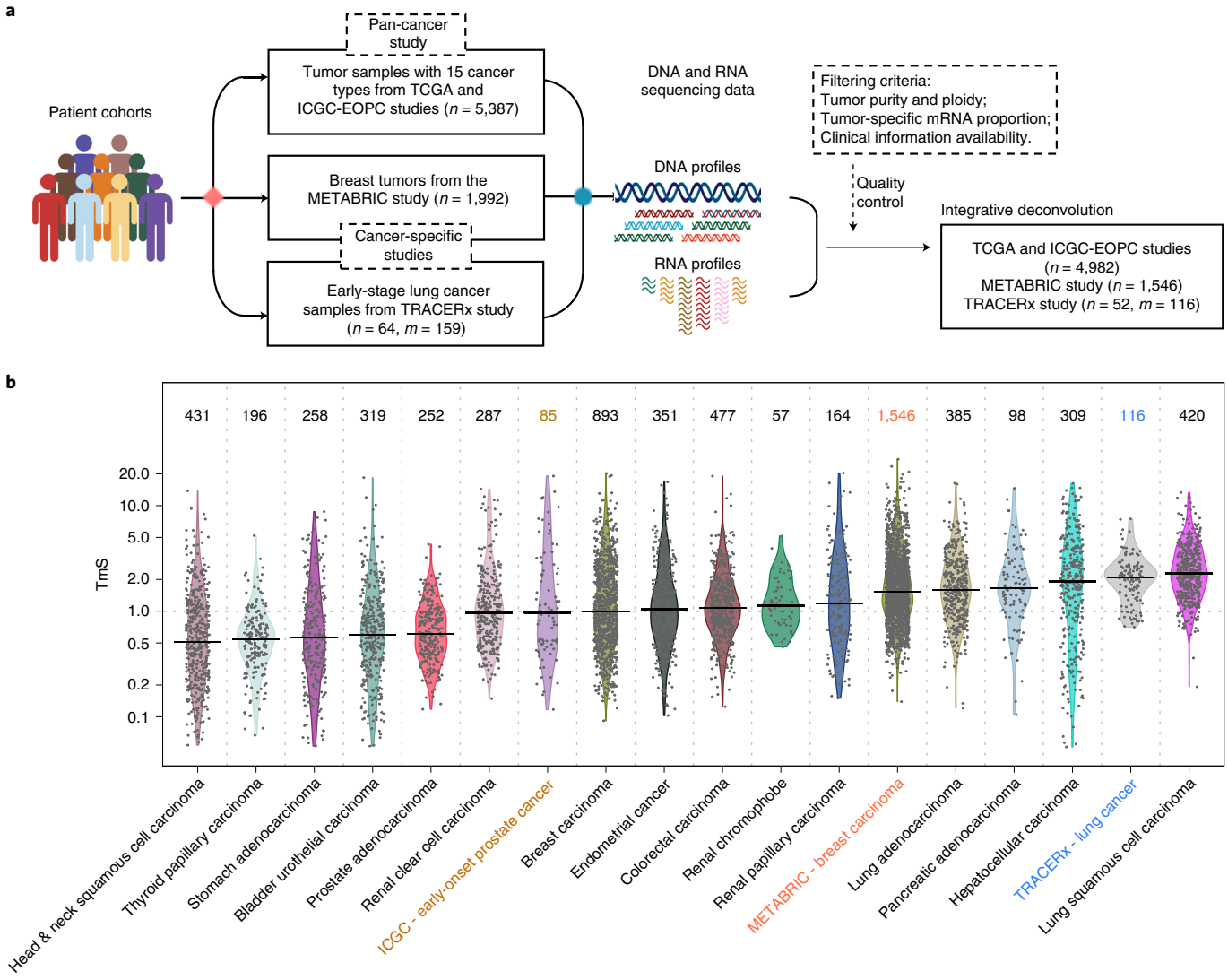


Fig. 3 | Estimation of tumor-specific total mRNA expression in bulk sequencing data. a, Diagram for the TmS calculation in TCGA, ICGC-EOPC, METABRIC and TRACERx datasets. The number of patients is denoted by *n*. When there are more than one tumor sample for each patient, the number of tumor samples is denoted by *m*. **b**, Distribution of TmS in 6,644 tumor samples from 6,580 patients across 15 cancer types in TCGA, ICGC-EOPC, METABRIC and TRACERx. The number of tumor samples for each cancer type is indicated above each violin plot.

We evaluated biological correlates of tumor-specific total mRNA expression across 4,982 patients from 15 cancer types in TCGA. Because *MYC* dysregulation is a known mechanism of global transcriptional amplification across cancers, we first evaluated the relationship between TmS and *MYC* expression and found a positive correlation in several cancer types[53], including breast carcinoma and renal papillary carcinoma (Spearman $r = 0.17$ and 0.21, respectively; Supplementary Note 2.3.2). We further examined genetic alterations, which may affect transcriptional activity, including driver mutations, tumor mutation burden (TMB), chromosomal instability (CIN) and whole-genome duplication (WGD) status (Methods and Supplementary Note 2.3.2). Significant associations were identified in some cancer types, suggesting that these genetic features may contribute to tumor-specific total mRNA expression in certain cancers but are not pan-cancer determinants (Extended Data Fig. 6b–e and Supplementary Note 2.3.2). Although we did not identify other pan-cancer genetic determinants of TmS, we found a pervasive upregulation of metabolic pathways in high-TmS samples across cancers. Specifically, the pentose phosphate pathway is the most frequently upregulated (significant in 12 of 15 cancers), followed by the glucose metabolism pathway (significant in seven of 15 cancers) (Extended Data Fig. 6f,g), in line with their roles in nucleotide synthesis and tumor metabolic reprogramming[54,55], respectively. These findings further validate the TmS metric in measuring tumor-specific total mRNA expression and support that the large inter-patient variation observed in TmS may be an important feature of tumor cells.

Tumor cell total mRNA expression refines prognostication. To understand the significance of TmS variation across patient samples, we first examined TmS in the context of histopathologic and molecular subtypes across cancers. Although many tumor subtypes have been described across cancers, we specifically examined five cancers where these subtypes have been most unequivocally shown to harbor differential biology and clinical significance. We observed consistent trends across subtypes of head and neck squamous cell carcinoma, renal papillary carcinoma[56], bladder urothelial carcinoma[57–59] and prostate adenocarcinoma, where prognostically favorable subtypes are enriched in tumors with lower TmS and vice versa (Fig. 4a–d and Methods). Similarly, in breast carcinoma,

triple-negative receptor status is associated with higher TmS, in keeping with this subtype's known propensity for aggressive behavior (TCGA: adjusted $P = 5 \times 10^{-36}$, Fig. 4e; METABRIC: adjusted $P = 9 \times 10^{-28}$, Fig. 4f). However, we found that TmS is not a surrogate for histopathologic or molecular subtype, tumor cellular proliferation or pluripotency genes[60] (Supplementary Note 2.3.2.5), suggesting that variation in TmS captures unique aspects of tumor biology that affects aggressiveness.

To further evaluate the potential utility of TmS to enable clinically relevant patient stratification, we examined the association of TmS with survival outcomes in TCGA and ICGC-EOPC (Methods and Supplementary Notes 3.1 and 3.2). In pan-cancer analyses, high TmS is associated with reduced overall survival (OS) and progression-free interval (PFI) (Fig. 4g, Extended Data Fig. 7a and Supplementary Table 5), which is robust to sample size differences across cancer types (Supplementary Note 3.2). TmS is independent of other clinical characteristics, including age and sex (Supplementary Note 2.3.2.5). Although TmS correlates with tumor-node-metastasis (TNM) stage in some cancer types, this relationship is not consistently observed across cancers (Supplementary Note 2.3.2.5). After feature selection and adjusting for known prognostic characteristics, including tumor subtype, stage and age (Methods), TmS was independently significantly associated with survival outcomes in all evaluable cancer types, except for estrogen receptor (ER)-positive breast carcinoma (Fig. 4h, Extended Data Fig. 7b–o, Supplementary Table 5 and Supplementary Notes 3.1 and 3.2). This association is retained, but weaker, when genome ploidy adjustment of TmS is omitted (Extended Data Fig. 8).

When patients are stratified by TNM stage classification, the prognostic effect of TmS differs between early (I/II) and advanced (III/IV) stage. Because early-stage versus advanced-stage tumors are generally treated using different therapeutic modalities, we hypothesized that the prognostic effect of TmS is modified by treatment. Given that the TCGA and ICGC studies did not consistently include chemotherapy and radiotherapy information[61], we identified a cohort of patients where chemotherapy and/or radiotherapy are generally not indicated (https://www.nccn.org/guidelines/category_1; Supplementary Table 6). Among these patients treated without systemic therapy, high TmS remains associated with worse PFI (Extended Data Fig. 9a,b).

In METABRIC, where treatment information is well-annotated, high TmS is associated with improved disease-free survival (DFS) in patients with early-stage triple-negative breast carcinoma (TNBC) treated with chemotherapy ($n = 118$, hazard ratio (HR) = 0.5, 95% confidence interval (CI): 0.28, 0.89, log-rank $P = 0.02$; Fig. 4i,j, Extended Data Fig. 9c and Supplementary Table 7). This is consistent with prior observations that high-risk breast tumors may respond better to chemotherapy[62,63]. This inversed relationship between high TmS and improved survival can be appreciated across all patients with TNBC in METABRIC with marginal significance ($n = 214$, HR = 0.7, 95% CI: 0.44, 1.12, log-rank $P = 0.1$; Fig. 4i and Supplementary Table 7), likely reflecting that most of these patients received systemic therapy. The same inversed relationship is observed in TNBC in TCGA (Fig. 4h and Supplementary Table 5).

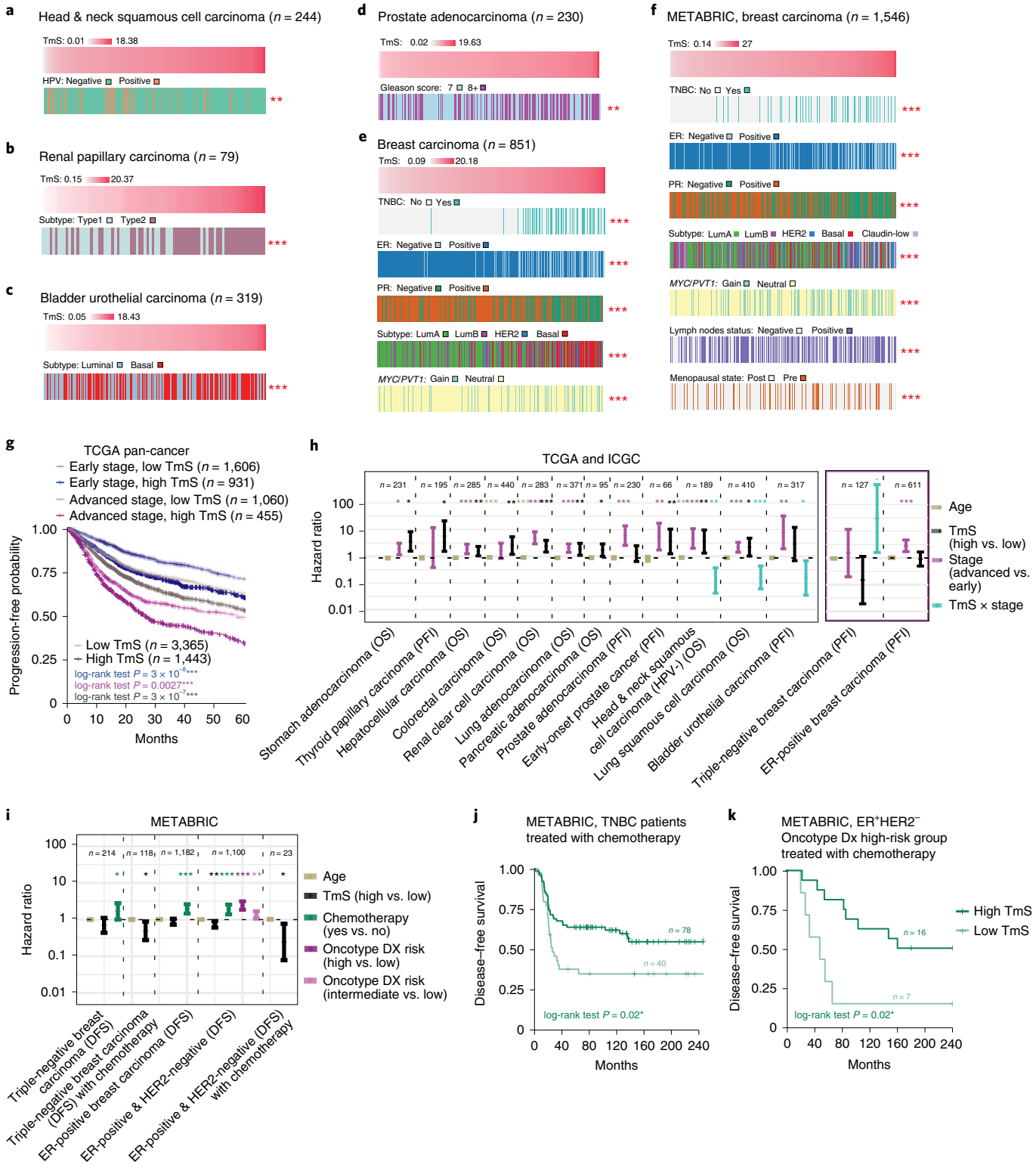
Furthermore, in METABRIC, we found that high TmS is associated with improved DFS for patients with ER^+HER2^- breast cancer, after adjusting for chemotherapy and Oncotype Dx risk status ($n = 1{,}100$, HR = 0.74, 95% CI: 0.60, 0.91, log-rank $P = 0.004$; Fig. 4i and Supplementary Table 7). Oncotype Dx risk score is routinely used clinically as a biomarker to estimate the risk of ER^+HER2^- tumors[64]. Within patients who were classified as high risk by Oncotype Dx and treated with chemotherapy, high TmS remains associated with better survival ($n = 23$, HR = 0.25, 95% CI: 0.08, 0.77, log-rank $P = 0.02$; Fig. 4i,k and Extended Data Fig. 9d). Patients with low TmS appeared to not have benefited from chemotherapy, suggesting the potential need for alternative therapy for this subgroup of patients. In summary, our findings suggest a unique utility of TmS in identifying and stratifying high-risk patients for treatment selection in breast cancer, which may be expandable to other cancer types.

Intra-tumor and inter-tumor heterogeneity in total mRNA expression. Intra-tumor heterogeneity serves as a reservoir for tumor evolution, treatment resistance and progression. Although intra-tumor heterogeneity can be identified using scRNA-seq (Fig. 1b,c and Extended Data Fig. 1a), the evolutionary relationships of tumor cell subpopulations cannot be readily inferred from scRNA-seq data alone. We, therefore, used TRACERx, a multi-region study of early-stage lung cancer evolution[39], to evaluate the potential utility of TmS for quantifying transcriptomic intra-tumor heterogeneity (Fig. 5a).

We calculated TmS using matched whole-exome sequencing (WES) and RNA sequencing data generated from 116 evolutionarily and spatially distinct regions across 52 patients, 30 of whom have two or more regions sampled (94 regions total) (Figs. 3b and 5b and Extended Data Fig. 10a). Subclonal copy number alterations (CNAs) and phylogenetic relationships of cancer subclones have been determined for these regions[39]. We first investigated the relationship between TmS and subclonal CNA, as determined by TRACERx. Across all 94 regions, TmS correlates better with the fraction of CNAs that are subclonal—that is, CNAs identified in only some regions of the tumor—than the fraction of the genome affected by CNA events (difference in Spearman $r = 0.20$, 95% CI: 0.04, 0.37; Fig. 5c,d and Methods). This suggests that TmS tracks ongoing chromosomal instability[65], reflecting intra-tumor heterogeneity, rather than the total CNA burden. To summarize across regions, we calculated the median and maximum of TmS, TmS_{med} and TmS_{max}, as well as the range of TmS (maximum – minimum

Fig. 4 | TmS is associated with known prognostic characteristics and refines prognostication in addition to stage. a–f, Clinicopathologic annotations for TCGA head and neck (**a**); TCGA renal papillary (**b**); TCGA bladder urothelial (**c**); TCGA prostate (**d**); TCGA breast (**e**); and METABRIC breast (**f**) cancers. Receptor status is indicated as follows: ER, estrogen; PR, progesterone; TNBC, triple-negative. Tumor samples are ordered by TmS from low to high. Benjamini–Hochberg-adjusted *P* values for Kruskal–Wallis tests comparing TmS across clinicopathologic subgroups are indicated by asterisks. For *MYC/PVT1* copy number status, 'Gain' indicates either *MYC* or *PVT1* amplification, and 'Neutral' indicates that no copy number alterations were detected. **g**, Kaplan–Meier curves of PFI for TCGA samples. Gray lines denote summary Kaplan–Meier curves of patients with high versus low TmS across all cancer types. Kaplan–Meier curves are further grouped into four groups by TmS and pathologic stage. *P* values of log-rank tests between high- versus low-TmS groups are indicated by asterisks. **h**, Forest plot of HRs (center points) and 95% CIs (error bars) of multivariate Cox proportional hazard models for OS or PFI in TCGA. Models are adjusted for age, TmS (high versus low), stage (advanced versus early) as well as an interaction term of TmS × stage, where applicable (see details in Supplementary Table 5). **i**, Forest plot of HRs (center points) and 95% CIs (error bars) of multivariate Cox proportional hazard models with age, TmS (high versus low), chemotherapy (yes versus no), Oncotype Dx risk classification (high versus intermediate versus low) as predictors for DFS in METABRIC (see details in Supplementary Table 7). For **h** and **i**, *P* values of two-sided Wald tests for the covariates are indicated by asterisks. Kaplan–Meier curves of DFS grouped by TmS (high versus low) for METABRIC TNBC (**j**) and ER^+HER2^- (**k**) patients treated with chemotherapy. *P* values of log-rank tests between high- versus low-TmS groups are indicated by asterisks. For all *P* values, significance levels are denoted as follows: $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$. HPV, human papillomavirus.

TmS across regions) per patient (Extended Data Fig. 10a). As expected, TmS_{med} is highly correlated with TmS_{max} across patients (Spearman $r=0.61$). However, TmS_{max} shows a higher correlation with the total fraction of subclonal CNAs than TmS_{med} or the range of TmS (Spearman $r=0.69$ versus 0.44 and 0.49; Extended Data Fig. 10b). Furthermore, TmS_{max} can be best explained, in a multiple linear regression, by the total fraction of subclonal CNAs (coefficient = 2.9, $P<0.001$, regression goodness-of-fit $R^2=0.7$; Fig. 5e, Methods and Supplementary Note 3.3). Additionally, in a logistic regression model, a smaller range of TmS per patient is predictive of linear evolutionary relationship between the regions sampled (area under the curve (AUC) = 0.83; Supplementary Note 3.3). These findings support the utility of measuring TmS per tumor region to quantify transcriptomic intra-tumor heterogeneity and, more specifically, its variation over evolutionary relationships.



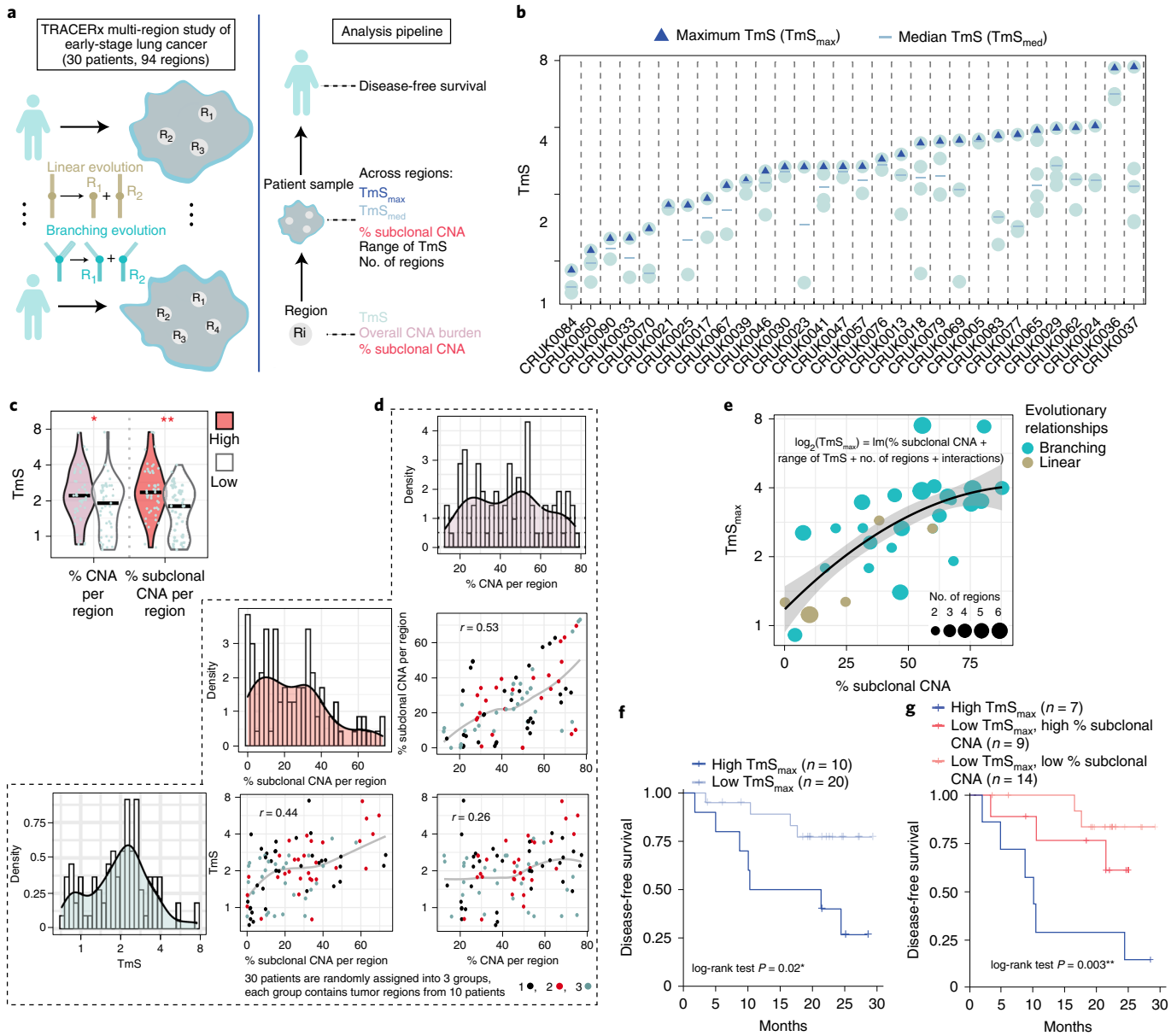


Fig. 5 | Regional estimation of TmS identifies spatial heterogeneity and refines prognostication in early-stage lung cancer. a, Illustration of the TRACERx multi-region study and a multi-level analysis pipeline. **b**, Distribution of TmS for 94 tumor regions from 30 TRACERx patients with at least two regions sampled. Blue triangles denote the maximum TmS for each patient. Blue '-' denotes the median TmS for each patient. **c**, Distributions of TmS per region with high or low % CNA burden (left) and % subclonal CNA per region (right). The number of regions is 47 for each group. Benjamini–Hochberg-adjusted *P* values of two-sided Wilcoxon rank-sum tests are indicated by asterisks. **d**, Pairwise scatter plots and histograms of % CNA, % subclonal CNA and TmS per region across 94 regions. Different colors annotate three randomly assigned patient groups, demonstrating that the correlation between TmS and % subclonal CNA per region is not driven by a subset of patients. Spearman correlation coefficient *r* values are shown, and the gray lines represent a LOESS fit. **e**, Scatter plot showing TmS_{max} versus total % subclonal CNA in each patient ($n=30$). The regression line and its 95% confidence band are colored in black and gray, respectively. Patients are colored by the evolutionary relationship. **f**, Kaplan–Meier survival curves of DFS stratified by TmS_{max}. **g**, Kaplan–Meier survival curves of DFS stratified by both TmS_{max} and % subclonal CNA. *P* values are obtained by log-rank tests between high- versus low-TmS groups. For all *P* values, significance levels are denoted as follows: $^*P<0.05$, $^{**}P<0.01$ and $^{***}P<0.001$.

Following the multi-cohort single-sample analyses, we hypothesized that the tumor region harboring subclones with highest TmS is most predictive of prognosis in early-stage lung cancer. Confirming this hypothesis, we observed that high TmS_{max} is associated with worse DFS (log-rank $P=0.02$; Fig. 5f), which is also consistent with our findings from TCGA in lung cancer. Patient stratification using both TmS_{max} and fraction subclonal CNA allows further discrimination of clinical outcomes (log-rank $P=0.003$; Fig. 5g), with a Cox regression concordance index of 0.75 (TmS_{max} and fraction subclonal CNA) versus 0.66 (fraction subclonal CNA only; Extended Data Fig. 10c). When 22 additional patients with a single region per tumor are included, high TmS_{max} remains associated with higher risk of recurrence or death (log-rank $P=0.005$; Extended Data Fig. 10d). High TmS_{med} shows a similar trend, although not statistically significant (log-rank $P=0.3$; Extended Data Fig. 10e).

In summary, variation in tumor total mRNA expression appears to be synergistic with recently acquired DNA alterations during evolution. A multi-region design, by measuring average tumor-specific

total mRNA expression for each region, can improve the resolution of the TmS quantification, thus enabling assessment of transcriptomic intra-tumor heterogeneity and further prognostication of early-stage lung cancer.

Discussion

Our study identifies TmS, a robust and measurable feature of tumor phenotype, from bulk tumor tissues. TmS is clinically and molecularly relevant across cancer types. Although single-cell technology can depict tumor cell populations with distinct gene expression states (a microscopic view), questions remain on how these populations coexist and interact to affect patient outcomes[10]. Average signals across all tumor cells summarize the magnitudes and fractions of each tumor cell population. It is known, mathematically, that in distributions such as Poisson and Exponential, the mean and the variance are highly correlated. In such scenarios, the average measures provide essential information for the entire distribution. Here we demonstrate that, indeed, the average value of tumor-specific total mRNA expression is informative when used to investigate both inter-tumor and intra-tumor heterogeneity and is also predictive of clinical outcomes in patients with cancer (a macroscopic view).

Using the lens of diversity in total mRNA expression, our study sheds light on cancer cell plasticity, previously evaluated in only a few tumors or in model systems[14]. To achieve a pan-cancer analysis that complements single-cell-based studies[16,18–20], we developed and calculated TmS, as an integrative RNA and DNA deconvolution metric for bulk tissues, in 6,580 patient samples from 15 cancer types. Association of TmS with transcriptional regulators, genetic features, metabolism as well as evolutionary relationships supports a consistent and biologically meaningful measurement of a bulk-level feature of tumor phenotype. We further report the ability of TmS to refine prognostication within each of the 12 cancer types with staging information and sufficient sample size.

Although high tumor cell total mRNA expression is generally associated with high-risk disease, clinical context remains important to evaluate its prognostic implications, as the direction of the prognostic effect was inverted by stage in four of 12 cancer types examined. Given that different tumor types and stages are often treated using distinct modalities, the inverted effect may, in part, be underpinned by a differential response of tumors with low versus high total mRNA expression to treatment. We validated the inverted effect in breast cancer subtypes in TCGA using the METABRIC cohort study in which treatment information was well-documented. Our findings are consistent with prior reports describing subsets of patients with aggressive cancer subtypes that respond favorably to systemic therapy[63,66,67]. Identifying which patients may benefit from specific systemic therapies remains a challenge, and TmS may serve to identify these patients as well as others requiring alternative treatments. Additional studies incorporating data from clinical trials will be needed to elucidate how stage-specific and treatment-related factors interact with tumor cell total mRNA expression to determine patient outcome and to help select the most effective treatments for low- and high-TmS tumors.

Conceptually, analogous to DNA ploidy measuring the average number of haploid genomes in tumors, the average total mRNA content per haploid genome can be considered the 'ploidy of the transcriptome'. Total mRNA content is a key parameter of tumor heterogeneity and phenotype plasticity, previously hidden in most RNA-based assays. Although our current work focuses on interpretation of mRNA, the methodology developed here can readily be applied to the quantification of other RNA species (for example, rRNA, miRNA and piRNA), further illuminating the cancer transcriptome. Enhanced attention to 'transcriptome ploidy' will enable better phenotypic characterization and a deeper biological understanding of transcriptional dysregulation in cancer and other diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01342-x.

Received: 7 June 2021; Accepted: 29 April 2022;
Published online: 13 June 2022

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Quintanal-Villalonga, Á. et al. Lineage plasticity in cancer: a shared pathway of therapeutic resistance. *Nat. Rev. Clin. Oncol.* **17**, 360–371 (2020).
3. Marjanovic, N. D. et al. Emergence of a high-plasticity cell state during lung cancer evolution. *Cancer Cell* **38**, 229–246 (2020).
4. LaFave, L. M. et al. Epigenomic state transitions characterize tumor progression in mouse lung adenocarcinoma. *Cancer Cell* **38**, 212–228 (2020).
5. Stewart, C. A. et al. Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat. Cancer* **1**, 423–436 (2020).
6. Halbritter, F. et al. Epigenomics and single-cell sequencing define a developmental hierarchy in Langerhans cell histiocytosis. *Cancer Discov.* **9**, 1406–1421 (2019).
7. Guo, W. et al. Single-cell transcriptomics identifies a distinct luminal progenitor cell type in distal prostate invagination tips. *Nat. Genet.* **52**, 908–918 (2020).
8. Domingues, A. F. et al. Loss of Kat2a enhances transcriptional noise and depletes acute myeloid leukemia stem-like cells. *eLife* **9**, e51754 (2020).
9. Teschendorff, A. E. & Feinberg, A. P. Statistical mechanics meets single-cell biology. *Nat. Rev. Genet.* **22**, 459–476 (2021).
10. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).
11. Batlle, E. & Clevers, H. Cancer stem cells revisited. *Nat. Med.* **23**, 1124–1134 (2017).
12. Morral, C. et al. Zonation of ribosomal DNA transcription defines a stem cell hierarchy in colorectal cancer. *Cell Stem Cell* **26**, 845–861 (2020).
13. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* **20**, 1349–1360 (2018).
14. Gupta, P. B., Pastushenko, I., Skibinski, A., Blanpain, C. & Kuperwasser, C. Phenotypic plasticity: driver of cancer initiation, progression, and therapy resistance. *Cell Stem Cell* **24**, 65–78 (2019).
15. Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148**, 33–45 (2012).
16. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
17. Athanasiadis, E. I. et al. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat. Commun.* **8**, 2045 (2017).
18. Chen, B. et al. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell* **184**, 6262–6280 (2021).
19. Grünwald, B. T. et al. Spatially confined sub-tumor microenvironments in pancreatic cancer. *Cell* **184**, 5577–5592 (2021).
20. Frede, J. et al. Dynamic transcriptional reprogramming leads to immunotherapeutic vulnerabilities in myeloma. *Nat. Cell Biol.* **23**, 1199–1211 (2021).
21. Lin, C. Y. et al. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
22. Nie, Z. et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**, 68–79 (2012).
23. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
24. Upender, M. B. et al. Chromosome transfer induced aneuploidy results in complex dysregulation of the cellular transcriptome in immortalized and cancer cells. *Cancer Res.* **64**, 6941–6949 (2004).
25. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* **98**, 31–36 (2001).
26. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
27. Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
28. Lovén, J. et al. Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).
29. Ahn, J. et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* **29**, 1865–1871 (2013).

30. Quon, G. et al. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.* **5**, 29 (2013).
31. Wang, Z. et al. Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *iScience* **9**, 451–460 (2018).
32. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
33. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
34. Ma, L. et al. Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer Cell* **36**, 418–430 (2019).
35. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
36. 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. *Clin. Cancer Res.* **27**, 5912–5921 (2021).
37. Gerhauser, C. et al. Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* **34**, 996–1011 (2018).
38. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
39. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
40. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
41. Wang, J. et al. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl Acad. Sci. USA* **115**, E6437–E6446 (2018).
42. Hosein, A. N. et al. Cellular heterogeneity during mouse pancreatic ductal adenocarcinoma progression at single-cell resolution. *JCI Insight* **4**, e129212 (2019).
43. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
44. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
45. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
46. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
47. Liberzon, A. et al. The Molecular Signatures Database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
48. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, 721–728 (2019).
49. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
50. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
51. Dempster, J. M. et al. Agreement between two large pan-cancer CRISPR–Cas9 gene dependency data sets. *Nat. Commun.* **10**, 5817 (2019).
52. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
53. Lorenzin, F. et al. Different promoter affinities account for specificity in MYC-dependent gene regulation. *eLife* **5**, e15161 (2016).
54. Pavlova, N. N. & Thompson, C. B. The emerging hallmarks of cancer metabolism. *Cell Metab.* **23**, 27–47 (2016).
55. Vander Heiden, M. G. & DeBerardinis, R. J. Understanding the intersections between metabolism and cancer biology. *Cell* **168**, 657–669 (2017).
56. Linehan, W. M. et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).
57. Miettinen, T. P. et al. Identification of transcriptional and metabolic programs related to mammalian cell size. *Curr. Biol.* **24**, 598–608 (2014).
58. Dadhania, V. et al. Meta-analysis of the luminal and basal subtypes of bladder cancer and the identification of signature immunohistochemical markers for clinical use. *EBioMedicine* **12**, 105–117 (2016).
59. Guo, C. C. et al. Assessment of luminal and basal phenotypes in bladder cancer. *Sci Rep.* **10**, 9743 (2020).
60. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
61. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).
62. Carey, L. A. et al. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clin. Cancer Res.* **13**, 2329–2334 (2007).
63. Gianni, L. et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. *J. Clin. Oncol.* **23**, 7265–7277 (2005).
64. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
65. Watkins, T. B. K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
66. Msaouel, P. et al. Updated recommendations on the diagnosis, management, and clinical trial eligibility criteria for patients with renal medullary carcinoma. *Clin. Genitourin. Cancer* **17**, 1–6 (2019).
67. Barlin, J. N. et al. Validated gene targets associated with curatively treated advanced serous ovarian carcinoma. *Gynecol. Oncol.* **128**, 512–517 (2013).

Methods

Additional details and results are described in the Supplementary Notes. Here, we summarize the key aspects of the analysis.

Total mRNA expression in scRNA-seq data. *Dataset.* We collected scRNA-seq data from ten patients, comprising three with colorectal adenocarcinoma, three with hepatocellular carcinoma, two with lung adenocarcinoma and two with pancreatic adenocarcinoma (Supplementary Table 1). A full description is provided in Supplementary Note 1.1. The three colorectal adenocarcinoma patient samples were obtained with informed consent and were approved by the Human Subjects Protection Office, the Clinical Research Committee as well as five separate institutional review boards at MD Anderson Cancer Center, in accordance with the Declaration of Helsinki.

Quality control, clustering, cell type annotation and normalized UMI. For each sample, we first filtered out cells based on number of genes expressed, total UMI counts and proportion of total UMI counts derived from mitochondrial genes. We also removed cells that were detected as doublets. After the quality control, 48,913 cells remained from the ten human tumor samples. Within each patient sample, highly variable genes were detected and used for principal component analysis (PCA). Cells were then clustered with the Seurat package[68]. Cell type was annotated using known marker genes[34,35,69–71]. Tumor cells were identified based on the inferred presence of somatic CNAs by inferCNV[72]. We further merged Seurat[68]-identified clusters that were not significantly different in gene counts, which is the total number of expressed genes (Wilcoxon rank-sum test, $\alpha = 0.001$; Fig. 1b). A full description is provided in Supplementary Note 1.2.1.

To enable comparison among different scRNA-seq samples within the same study, we performed scale normalization to ensure that the total UMI count per cell was comparable across different samples from the same study. A full description is provided in Supplementary Note 1.2.2.

Trajectory and gene set enrichment analyses. We applied Monocle 2 (version 2.14.0)[44–46] to construct single-cell trajectories and used the CytoTRACE (version 0.3.3) score to measure the differentiation state of tumor cells[16]. To compare CytoTRACE scores among the tumor cell clusters from patient samples within the same cancer type, we integrated tumor cells from patients 1, 2 and 3 from colorectal cancer and patients 1 and 2 from each of the lung and pancreatic cancers using ComBat (version 3.20.0)[73] embedded in CytoTRACE, which corrects for batch effects. We quantified gene set enrichment for the high-UMI versus low-UMI tumor cell clusters using the GeneOverlap R package (version 1.24.0)[74]. A comprehensive set of signatures with 18,617 human gene sets (containing at least four genes) was compiled from the Molecular Signatures Database (version 6.2)[47] and CellMarker[48]. A full description is provided in Supplementary Note 1.2.4.

Pseudo-bulk analysis. We pooled normalized scRNA-seq data to form pseudo-bulk samples and estimated the ratio of the mean total UMI counts of tumor cells to that of the non-tumor cells for each sample. The 95% CIs were constructed by bootstrapping the same numbers of tumor and non-tumor cells with 1,000 repetitions.

Tumor-specific total mRNA expression in bulk sequencing data. *A mathematical model for tumor-specific total mRNA expression estimation.* For any group of cells, we use S to denote the average global mRNA transcript level per cell per haploid genome, which follows $S = \sum_{c=1}^{C}\left(\sum_{g=1}^{G} u_{gc}/p_c\right)/C$. Here, u_{gc} denotes the number of mRNA transcripts of gene g in cell c; G is the total number of genes; C is the number of cells; and p_c is the ploidy—that is, 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: $S \approx \sum_{c=1}^{C}\sum_{g=1}^{G} u_{gc}/(C\Psi)$. For non-tumor cells, which are commonly diploid, this assumption is assured.

In the analysis of bulk RNA sequencing data from mixed tumor samples, we are interested in comparing tumor to non-tumor cell groups. We let T denote tumor cells and N denote non-tumor cells. Therefore, we define a TmS to reflect the ratio of total mRNA transcript level per haploid genome of tumor cells to that of the surrounding non-tumor cells—that is, $\mathrm{TmS}_{\mathrm{tumor}} = S_T / S_N$, simplified as TmS from here forward. It is necessary to calculate this ratio to cancel out technical effects presented in sequencing data that confound with both S_T and S_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 mRNA transcripts of gene g across all cells from tumor and non-tumor cells; let $T_+ = \sum_{g=1}^{G} T_g$, $N_+ = \sum_{g=1}^{G} N_g$, C_T and C_N denote the total 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

$$\mathrm{TmS} = [T_+/(C_T\Psi_T)]/[N_+/(C_N\Psi_N)] = [T_+/N_+]/[(C_T\Psi_T)/(C_N\Psi_N)] \quad (1)$$

We further introduce the proportion of total bulk mRNA expression derived from tumor cells (hereafter 'tumor-specific mRNA proportion') $\pi = \left(\sum_{g=1}^{G} T_g\right)/\left(\sum_{g=1}^{G} T_g + \sum_{g=1}^{G} N_g\right)$ and the tumor cell proportion (hereafter 'tumor purity') $\rho = C_T/(C_T + C_N)$. We, thus, have

$$\begin{aligned}\mathrm{TmS} &= [\pi/(1-\pi)]/[(\rho/(1-\rho))(\Psi_T/\Psi_N)] \\ &= [\pi(1-\rho)\Psi_N]/[\rho(1-\pi)\Psi_T]\end{aligned} \quad (2)$$

The tumor-specific mRNA proportion π derived from the tumor can be estimated using DeMixT[31] as $\hat{\pi}$; the tumor purity ρ and ploidy Ψ_T can be estimated using ASCAT[32], ABSOLUTE[33] or Sequenza[49] based on the matched DNA sequencing data as $\hat{\rho}$ and $\hat{\Psi}_T$, respectively; and the ploidy of non-tumor cells Ψ_N was assumed to be 2 (refs. [32,33]). Hence, we have

$$\widehat{\mathrm{TmS}} = \frac{\hat{\pi}(1-\hat{\rho})\Psi_N}{\hat{\rho}(1-\hat{\pi})\hat{\Psi}_T} \quad (3)$$

In what follows, we use TmS to represent $\widehat{\mathrm{TmS}}$ for simplicity. A full description is provided in Supplementary Note 2.1.

Consensus of tumor purity and ploidy estimation. For DNA-based deconvolution methods such as ASCAT and ABSOLUTE, there could be multiple tumor purity ρ and ploidy Ψ_T pairs that have similar likelihoods. Both ASCAT and ABSOLUTE can accurately estimate the product of purity and ploidy $\rho\Psi_T$; however, they sometimes lack power to identify ρ and Ψ_T separately. TmS is derived from the product of tumor ploidy and the odds of tumor purity. Hence, it is potentially more robust to ambiguity in the tumor purity and ploidy estimation, ensuring the robustness of the TmS calculation. We illustrate this robustness by showing that the agreement between TmS values calculated from ASCAT and ABSOLUTE are substantially improved, as compared to the agreement between the ploidy values calculated from the two methods that was low among 20% of TCGA samples (Extended Data Fig. 3f,g). To calculate one final set of TmS values for a maximum number of samples, we take a consensus strategy. We first calculate TmS values with tumor purity and ploidy estimates derived from both ABSOLUTE and ASCAT and then fit a linear regression model on the $\log_2$-transformed $\mathrm{TmS}_{\mathrm{ASCAT}}$ by using the $\log_2$-transformed $\mathrm{TmS}_{\mathrm{ABSOLUTE}}$ as a predictor variable. We remove samples with Cook's distance $\geq 4/n$ ($n = 5{,}295$; Extended Data Fig. 3h) and calculate the final $\mathrm{TmS} = \sqrt{\mathrm{TmS}_{\mathrm{ASCAT}} \times \mathrm{TmS}_{\mathrm{ABSOLUTE}}}$.

Improved estimation of tumor-specific mRNA proportion. The identifiability of model parameters is a major issue for high-dimensional models. With the DeMixT model, there is hierarchy in model identifiability in which the cell-type-specific mRNA proportions 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 proportions (π). In general, genes expressed at different numerical ranges can affect estimation of π. We found that including genes that are not differentially expressed between the tumor and non-tumor components, differentially expressed across tumor subtypes in different samples or with large variance in expression within the non-tumor component can introduce large biases to the estimated π. On the other hand, the tumor component is hidden in the mixed tumor samples, hence preventing a differential expression analysis between mixed and normal samples from finding the best genes. By applying a profile-likelihood-based approach to detect the identifiability of model parameters[75], we systematically selected the top-ranking identifiable genes for the estimation of π. As a general method, the profile-likelihood-based gene selection strategy can be extended to any method that uses maximum likelihood estimation. We also employed a virtual 'normal' spike-in strategy to balance proportion distributions, which further improved the deconvolution performance. A full description is provided in Supplementary Note 2.2.

Profile-likelihood-based gene selection. In brief, in the DeMixT model, for sample $i \in (1, 2, \ldots, M)$ and gene $g \in (1, 2, \ldots, G)$, we have

$$Y_{ig} = \pi_i T'_{ig} + (1-\pi_i) N'_{ig} \quad (4)$$

where Y_{ig} represents the scale-normalized expression count matrix observed from mixed tumor samples, and 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 mRNA proportion $\hat{\pi}$ is the desirable quantity for Eq. 3. We assume each hidden component follows the $\log_2$-normal distribution—that is, $T'_{ig} \sim LN\left(\mu_{Tg}, \sigma^2_{Tg}\right)$ and $N'_{ig} \sim LN\left(\mu_{Ng}, \sigma^2_{Ng}\right)$. We will use notation T and N and drop the $'$ sign from now on. The identifiability of a gene k in the DeMixT model is measured by the CI $[\mu^-_{Tk}, \mu^+_{Tk}]$ around the mean expression μ_{Tk}. The definition of the profile likelihood function of μ_{Tk} is

$$\begin{aligned}&l_{\mu_{Tk}}(\mu_{Tk} = x|\pi, \mu_T, \sigma_T) \\ &= \max_{\pi_i, \mu_{Tg}, \sigma_{Tg}, \sigma_{Tk}} \left\{\sum_{i=1}^{M}\left[\sum_{g\neq k}^{G} \log\left(f\left(\pi_i, \mu_{Tg}, \sigma_{Tg}\right)\right) + \log\left(f(\pi_i, \mu_{Tk} = x, \sigma_{Tk})\right)\right]\right\}\end{aligned} \quad (5)$$

where

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

is the likelihood function of the DeMixT model.

The CI of a profile likelihood function can be constructed through inverting a likelihood-ratio test[76]. However, calculating the actual profile likelihood function of all genes (~20,000) is generally infeasible due to computational limits. We adopted an asymptotic approximation to quickly evaluate the profile likelihood function[75], using the observed Fisher information of the log-likelihood, denoted as $H(\hat{\pi}, \hat{\mu}_T, \hat{\sigma}_T)$. Then, the asymptotic α-level CI of μ_{Tk} can be written as[75]

$$\mu_{Tk}^{\pm} = \widehat{\mu_{Tk}} \pm \sqrt{2\chi_{1-\alpha}^2(1)H(\hat{\pi}, \hat{\mu}_T, \hat{\sigma}_T)_{k,k}^{-1}} \quad (6)$$

We hereby introduce a gene selection score to represent the length of an asymptotic profile-likelihood-based 95% CI of μ_{Tk} for gene k,

$$\text{gene selection score}_k = 2\sqrt{2\chi_{0.05}^2(1)H(\hat{\pi}, \hat{\mu}_T, \hat{\sigma}_T)_{k,k}^{-1}} \quad (7)$$

Genes with a lower score have a smaller CI, hence higher identifiability for their corresponding parameters in the DeMixT. Genes are ranked based on the gene selection scores 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, our proposed profile-likelihood-based gene selection approach is included as function 'DeMixT_GS'. A full description is provided in Supplementary Note 2.2.2. We performed a simulation study, mimicking the TCGA prostate adenocarcinoma dataset, to validate the proposed gene selection method. A full description is provided in Supplementary Note 2.2.3. The implementation of virtual 'normal' spike-ins and a simulation study is provided in Supplementary Note 2.2.4.

TmS validation using bulk RNA sequencing data from mixed cell lines. We validated TmS estimates using an experimental dataset from a previous mixed cell line study (GSE121127)[31] and selected a subset of 18 mixed samples with negligible RNA content from the immune component. Lung adenocarcinoma in humans (H1092) and CAF cells were mixed at different cell count proportions (Supplementary Table 3) to generate each bulk sample, plus three additional samples of 100% H1092 or 100% CAF. The raw reads were generated from paired-end total RNA Illumina sequencing and mapped to the human reference genome build 37.2 from the National Center of Biotechnology Information through TopHat[77]. SAMtools[78] was applied to remove improperly mapped and duplicated reads. Picard tools were used to sort the cleaned SAM files according to their reference sequence names and create an index for the reads. The gene-level expression was quantified using the R packages GenomicFeatures and GenomicRanges.

For each cell line, we measured total RNA amount (in ngμl^{-1}) for 1 million cells in three repeats using the Qubit RNA Broad Range Assay Kit (Life Technologies). The true TmS values of H1092 or CAF were then derived as a ratio of the total RNA amount per cell between the two cell types—specifically, $\text{TmS}_{\text{H1092}} = \frac{\text{total RNA amount per cell of H1092}}{\text{total RNA amount per cell of CAF}} = 0.87$ and $\text{TmS}_{\text{CAF}} = \frac{\text{total RNA amount per cell of CAF}}{\text{total RNA amount per cell of H1092}} = 1.2$. We estimated the RNA proportion of H1092 and CAFs using DeMixT (DeMixT_GS function with 4,000 genes selected) under two scenarios: (1) three pure CAFs samples were used as reference; and (2) three pure H1092 samples were used as reference. To estimate TmS values, we used the known cell counts to calculate ρ values.

TmS estimation in patient cohorts. A full description of all datasets is provided in Supplementary Note 2.3.1.

TCGA datasets. Raw read counts of high-throughput mRNA sequencing data, clinical data and somatic mutations from 7,054 tumor samples across 15 TCGA cancer types (breast carcinoma, 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 and uterine corpus endometrial carcinoma) were downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). ATAC-seq data[52], tumor purity and ploidy data[79,80] and annotations of driver mutation and indels[81] were downloaded for these samples.

Estimation of tumor-specific mRNA proportions from RNA sequencing data. For each cancer type, we filtered out poor-quality tumor and normal samples that were likely misclassified. We then selected available adjacent normal samples as reference for the tumor deconvolution using DeMixT. Based on simulation studies (Supplementary Note 2.2.3) and observed distributions of gene selection scores in real data, we chose the top 1,500 or 2,500 genes (varies across cancer types) to estimate tumor-specific mRNA proportions (π). For each cancer type, the selected 1,500 or 2,500 genes are defined as intrinsic tumor signature genes. We added varying numbers of virtual spike-in samples depending on cancer types. We additionally removed samples with extreme estimates of π, >85% or ranked at the top 2.5 percentile of all samples within each cancer type to mitigate the remaining underestimation when π is close to 1. A full description is provided in Supplementary Note 2.3.2.1.

Consensus TmS estimation. We calculated a consensus TmS as $\text{TmS} = \sqrt{\text{TmS}_{\text{ASCAT}} \times \text{TmS}_{\text{ABSOLUTE}}}$ and removed 264 of 5,295 TCGA samples that deviated from our consensus model, as described previously. A full description on sample exclusions is provided in Supplementary Note 2.3.2.2.

Intrinsic tumor signature genes. For each cancer type, the selected genes used for estimating π are called intrinsic tumor signature genes. We conducted gene set enrichment analyses (GSEAs) on hallmark pathways and KEGG pathways[47] for these genes ranked with their gene selection scores from small to large using GSEA[82] and g:Profiler[83]. We further evaluated the chromatin accessibility of intrinsic tumor signature genes using ATAC-seq data from TCGA samples[52]. For each sample, we calculated the mean of the peak scores of selected genes and compared it with the corresponding permuted null distribution for each cancer type. A full description is provided in Supplementary Note 2.3.2.3.

Association of TmS with genetic alterations and metabolism. We searched among driver mutations (including nonsense, missense and splice-site single-nucleotide variants (SNVs) and indels)[81] as well as all non-synonymous mutations (including SNVs and indels) over all genes for the 15 cancer types to identify those that were significantly associated with TmS. We investigated 24 cancer–gene pairs for the driver mutation analysis and 32,894 cancer–gene pairs for the non-synonymous mutation analysis. We applied a Wilcoxon rank-sum test to each candidate gene to compare the distributions of TmS of the samples with mutations versus without mutations. We also fitted a linear regression model on TmS to adjust for TMB. The P values of each gene were adjusted for multiple testing using Benjamini–Hochberg correction across all candidate genes within the corresponding cancer type. See Supplementary Note 2.3.2.4 for further details.

TMB was calculated by counting the total number of somatic mutations based on the consensus mutation calls (MC3)[84]. 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[32]. For samples present in both TCGA and Pan-Cancer Analysis of Whole Genomes (PCAWG), the consensus copy number was derived from published results[85]. Tumor samples that had undergone whole-genome duplication (WGD) were identified based on homologous copy number information[33].

For each cancer type from TCGA, we conducted GSEAs[82] on the metabolism of carbohydrate pathways (the Reactome database[86]). The genes were ranked by the Spearman correlation coefficient between their expression levels and TmS across samples; they were then put through GSEA in the 'pre-ranked' mode. For GSEA, we adopted permutation tests (1,000 times) to generate a normalized enrichment score (NES) for each candidate pathway. A hierarchical clustering on the expression levels of the Reactome pentose phosphate pathway (15 genes total, of which two genes were removed due to high-frequency zero counts across samples) for the tumor samples was performed using Euclidean distance and Ward linkage. The samples were then separated into two groups using the 'cutree' function. For each cancer type, a Wilcoxon rank-sum test was used to compare the distributions of TmS estimates between the two tumor sample groups. P values were adjusted for multiple testing using Benjamini–Hochberg correction across all cancer types.

ICGC-EOPC dataset. In this cohort, matched mRNA sequencing data and whole-genome sequencing data, as well as clinical data including biochemical recurrence, Gleason score and pathologic stage, from 121 tumor samples and nine adjacent normal samples from 96 patients (age at treatment <55 years) were downloaded from Gerhauser et al.[37] We used the nine available adjacent normal samples as the normal reference. The mRNA sequencing data came from three batches: batch 1 (17 patients and 25 samples), batch 2 (42 patients and 52 samples) and batch 3 (37 patients and 44 samples). We observed consistency and robustness of DeMixT results with or without batch effect correction. See Supplementary Notes 2.3.1 and 2.3.3 for further details.

METABRIC dataset. This dataset included 1,992 pairs of expression arrays and Affymetrix SNP 6.0 arrays profiled for tumor samples from 1,992 patients, which was divided into a discovery set (997 patients) and a validation set (995 patients)[38]. A total of 144 expression arrays for adjacent normal tissues were provided.

We applied the DeMixT deconvolution pipeline to the expression arrays of the combined discovery and validation sets, after batch effect correction, to estimate tumor-specific proportions using the adjacent normal samples as the reference. Affymetrix CEL files were processed by PennCNV[87] to obtain the LogR and B allele frequency (BAF) data, followed by both ASCAT[32] and Sequenza[49] to estimate tumor

purity and ploidy for each sample. The consensus TmS strategy was applied to obtain robust TmS estimations. In total, 1,664 patient samples with TmS remained after the above steps. We additionally removed 118 patient samples due to missing follow-up information of biochemical recurrence intervals or the PAM50 subtypes. A final cohort of 1,546 patient samples from both the discovery and validation sets was kept for downstream analyses. See Supplementary Notes 2.3.1 and 2.3.4 for further details.

TRACERx dataset. A total of 159 tumor samples from 64 patients with matched RNA sequencing data and WES data were downloaded[39,40,88] (see Supplementary Note 2.3.1 for further details). Tumor purity and ploidy were estimated from WES data by Sequenza[49]. We used RNA sequencing data from normal lung samples without significant pathology in the corresponding tissue types in the GTEx study as the reference for the deconvolution of tumor samples in this dataset (see Supplementary Note 2.3.5 for further details). Focusing on tumor samples with tumor purity > 0.15, we calculated TmS for 116 regions from 52 patient samples, among which 30 patients have at least two regions. We further performed association analysis of regional and sample-specific TmS with measures of chromosomal instability. We defined the subclonal CNA as a CNA presented only in a subset of regions. We further define the evolutionary relationship in two regions from the same patient as either linear or branched. For each evolutionary relationship per patient, we defined the 'range of TmS' as $log_2(TmS_{max}) - log_2(TmS_{min})$ across regions. We fitted linear regression models by taking $log_2(TmS_{max})$ as the response variable and the percentage of subclonal CNA, number of regions, range of TmS, evolutionary relationship and their interactions as predictors. The best model was selected by stepwise selection based on the Bayesian information criterion (BIC)[89]. See Supplementary Note 3.3 for further details.

Statistical analysis. *Batch effect correction.* For RNA sequencing data from multiple batches, we applied batch effect correction using ComBat[73] and limma[90] to combine RNA sequencing data in one pool before estimating tumor-specific mRNA proportions. See Supplementary Note 3.1 for further details on the robustness of TmS estimation.

Association with clinical variables. Kruskal–Wallis tests were used to compare the distribution of TmS between subgroups defined by each clinical variable. The *P* values from the Kruskal–Wallis tests were adjusted using Benjamini–Hochberg correction across all available clinical variables within the corresponding cancer type.

Association with survival outcomes. Associations with TmS were assessed in terms of OS, PFI and DFS depending on cancer type and study cohort. For TCGA, we used outcome measures that are recommended by Liu et al.[61]. If both OS and PFI were recommended, we used the more clinically relevant outcomes for an individual cancer type. We dichotomized pathologic stages into two categories: early (I/II) and advanced (III/IV). For prostate cancers, we used the Gleason score (Gleason score = 7 versus 8+) instead of early and advanced stages. Furthermore, we followed clinical guidelines and physician recommendations to identify tumor samples that were treated without systemic therapy (surgery only) in TCGA and used the corresponding meaningful outcome measures for the selected populations. For all association analyses with clinical outcomes across datasets, we used a recursive partitioning survival tree model, rpart[91], to find the optimal TmS cutoff (high versus low) separating different survival outcomes within each of the two stages defined above in each cancer type. Splits were assessed using the Gini index, and the maximum tree depth was set to 2. Log-rank tests between high- and low-TmS groups within early or advanced pathologic stages were performed. We performed sensitivity analysis on the TmS cutoff to confirm that a similar trend can be observed with other values. See Supplementary Note 3.2 for further details on the survival analysis and the identification of patients without systemic therapy.

Cox regression with model selection. We fitted multivariate Cox proportional hazard models with age, stage, TmS (high versus low) and other variables as predictors of OS, PFI or DFS for each dataset and calculated HRs and 95% CIs. We use the stepwise model selection method with BIC[89], where the baseline model includes age, stage and TmS predictors, and additional variables to select include the interaction term of TmS × stage.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The UMI counts of the hepatocellular carcinoma scRNA-seq data were downloaded from the Gene Expression Omnibus under accession number GSE125449. The UMI counts and cell type annotations of the lung adenocarcinoma scRNA-seq data were downloaded from the ArrayExpress under accession number E-MTAB-6149. The UMI counts of the colorectal adenocarcinoma scRNA-seq data are available at http://crcmoonshot.org/?page_id=189. FASTQ files of scRNA-seq data from pancreatic cancer is publicly available on the Gene Expression Omnibus under accession number GSE156405.

Raw read counts from the mixed cell line study were downloaded from the Gene Expression Omnibus under accession number GSE121127.

Raw read counts of RNA sequencing data, clinical data and somatic mutations from 7,054 tumor samples across 15 TCGA cancer types are available for download from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). ATAC-seq data for TCGA samples were downloaded from https://science.sciencemag.org/content/362/6413/eaav1898/tab-figures-data.

Clinical information of ICGC-EOPC was downloaded from https://www.sciencedirect.com/science/article/pii/S1535610818304823?via%3Dihub#gs1.

All primary METABRIC data, including Affymetrix SNP 6.0 CEL files and Illumina HT-12 gene expression arrays, are available at the European Genome-phenome Archive (EGAS00000000083) and may be downloaded from https://ega-archive.org/studies/EGAS00000000083. Clinical information of METABRIC was downloaded from https://www.cbioportal.org/study/clinicalData?id=brca_metabric.

Clinical information of TRACERx was downloaded from https://www.nejm.org/doi/full/10.1056/NEJMoa1616288#article_supplementary_material.

WES data of TRACERx were downloaded from https://ega-archive.org/studies/EGAS00001002247.

RNA sequencing data of TRACERx were downloaded from https://ega-archive.org/studies/EGAS00001003458.

TmS values of all samples and the identified intrinsic tumor signature genes for this study are available for download at https://github.com/wwylab/TmS.

All other relevant data are available from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

DeMixT used for estimating tumor-specific mRNA expression proportion is freely available as an R package and can be downloaded from https://github.com/wwylab/DeMixT. DeMixT version 1.2.2 was used to generate the results in this work. A tutorial for estimating TmS based on the DeMixT output is available at https://github.com/wwylab/TmS.

References

68. 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).
69. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
70. 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).
71. Hashimoto, K. et al. Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl Acad. Sci. USA* **116**, 24242–24251 (2019).
72. 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).
73. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
74. Shen, L. GeneOverlap: an R package to test and visualize gene overlaps. https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html (2022).
75. Raue, A. et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).
76. Venzon, D. J. & Moolgavkar, S. H. A method for computing profile-likelihood-based confidence intervals. *Appl. Stat.* **37**, 87–94 (1988).
77. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
78. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
79. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
80. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
81. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
82. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
83. Reimand, J. et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
84. 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).
85. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
86. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

87. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
88. Biswas, D. et al. A clonal expression biomarker associates with lung cancer mortality. *Nat. Med.* **25**, 1540–1548 (2019).
89. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
90. Ritchie, M. E. et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
91. Therneau, T. M. & Atkinson, E. J. *An Introduction to Recursive Partitioning Using the RPART Routine*. Technical report no. 61 (Mayo Clinic, section of statistics, Minnesota, 1997).

Acknowledgements

S.C. is supported by the Norman Jaffe Professorship in Pediatrics Endowment Fund, the MD Anderson Colorectal Cancer Moon Shot Program and National Institutes of Health (NIH) R01CA183793. J.R.W. is supported by an American Thyroid Association/ThyCa grant, a Mark Foundation for Cancer Research ASPIRE award and the Michael Petrick Anaplastic Thyroid Cancer Research Fund. S.J. is supported by Human Cell Atlas Seed Network-Retina by the Chan Zuckerberg Institute, the MD Anderson Colorectal Cancer Moon Shot Program, NIH R01CA183793 and Cancer Prevention and Research Institute of Texas (CPRIT) RP200383. P.Y. and M.D.M. are supported by NIH R01CA239342. S.G. is supported by Human Cell Atlas Seed Network-Retina by the Chan Zuckerberg Institute and the MD Anderson Prostate Cancer Moon Shot Program. J.C. is supported by NIH R01CA158113. J.P.S. was supported by National Cancer Institute (NCI) L30 CA171000, K22 CA234406 and P50CA221707, CPRIT RR180035 (to J.P.S.; J.P.S. is a CPRIT Scholar in Cancer Research) and the Col. Daniel Connelly Memorial Fund. J.J.L. is supported by NIH T32CA009599. S.T. and M.N. are supported by The Academy of Finland, the Cancer Society of Finland, the Sigrid Juselius Foundation and the Finnish Cultural Foundation. N.C.D. is supported by the Norman Jaffe Professorship in Pediatrics Endowment Fund. P.A.F. is supported, in part, by the Welch Foundation, MEI Pharma, Inc., Cancer Research United Kingdom (CRUK), the Kadoorie Charitable Foundation and NIH/NCI U01 CA224044 and R01CA231465. B.L is supported by the SWOG Hope Foundation, the Human Cell Atlas-Breast by the Chan Zuckerberg Institute, the US Department of Defense, the Breast Cancer Research Foundation and the NIH. P.M. is supported by a Career Development Award from the American Society of Clinical Oncology, a Research Award from KCCure, the MD Anderson Khalifa Scholar Award and the MD Anderson Physician-Scientist Award. P.C.B. is supported by NIH/NCI P30CA016042, 1U01CA214194-01 and 1U24CA248265-01. A.U. and N.E. are supported by the Norwegian Cancer Society (198016-2018). J.Z. is supported by the MD Anderson Physician-Scientist Award, the MD Anderson Lung Cancer Moon Shot Program, NIH/NCI R01CA234629-01 and U01-CA256780-01 and a CPRIT Multi-Investigator Research Award grant (RP160668). A.M. is supported by the MD Anderson Pancreatic Cancer Moon Shot Program, the Khalifa Bin Zayed Al-Nahyan Foundation and NIH U01CA196403, U01CA200468, U24CA224020 and P50CA221707. S.K. is supported by NIH P50CA221707. C.S. is the Royal Society Napier Research Professor (RSRP\R\210001). This work was supported by the Francis Crick Institute, which receives its core funding from CRUK (FC001169), the UK Medical Research Council (FC001169) and the Wellcome Trust (FC001169). This research was funded in part by the Wellcome Trust (FC001169). For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission. C.S. is funded by CRUK (TRACERx (C11496/A17786), PEACE (C416/A21999) and CRUK Cancer Immunotherapy Catalyst Network), CRUK Lung Cancer Centre of Excellence (C11496/A30025), the Rosetrees Trust, the Butterfield and Stoneygate Trusts, the Novo Nordisk Foundation (ID16584), the Royal Society Professorship Enhancement Award (RP/EA/180007), the National Institute for Health Research (NIHR) Biomedical Research Centre at University College London Hospitals, the CRUK-University College London Centre, the Experimental Cancer Medicine Centre and the Breast Cancer Research Foundation (BCRF 20-157). This work was supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-AACR-DT23-17 to S.M.D. and A.E.S.). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the scientific partner of SU2C. C.S. is in receipt of an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement no. 835297). P.V.L. and J.D. are supported by the Francis Crick Institute, which receives its core funding from CRUK (FC001202), the UK Medical Research Council (FC001202) and the Wellcome Trust (FC001202). For the purpose of open access, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. P.V.L. and J.D. are also supported by the Medical Research Council (MR/L016311/1). J.D. is supported by the European Union's Horizon 2020 Research and Innovation Programme (Marie Skłodowska-Curie grant agreement no. 703594-DECODE) and the Research Foundation–Flanders (FWO, grant no. 12J6916N). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support for the establishment of The Francis Crick Institute. P.V.L. is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). W.W. is supported by the Human Cell Atlas Seed Network-Retina by the Chan Zuckerberg Institute and NIH R01CA183793, R01CA239342, R01CA158113, P30CA016672 and P50CA221707. This study makes use of data generated by TRACERx Consortium and provided by the UCL Cancer Institute and The Francis Crick Institute.

The TRACERx study is sponsored by University College London, funded by CRUK and coordinated through CRUK and the UCL Cancer Trials Centre. This study makes use of data generated by METABRIC and provided by CRUK and the British Columbia Cancer Agency Branch. The METABRIC study is funded by CRUK, the British Columbia Cancer Foundation and the Canadian Breast Cancer Foundation BC/Yukon.

Author contributions

S.C., J.R.W. and S.J. developed computational methods, implemented pipeline for the statistical model, performed the data analysis and wrote the manuscript, in collaboration with all other authors. P.Y. conducted simulation experiments and analyzed the TRACERx dataset. S.G. assisted with scRNA-seq analysis. Y.D. analyzed the METABRIC dataset. M.D.M. performed ATAC-seq data analysis. J.P.S. and S.K. provided scRNA-seq data and advised on the data analysis for colorectal cancer. X.Z. advised on analysis of metabolic pathways using bulk RNA sequencing data. J.C. assisted with scRNA-seq data analysis. J.J.L., P.A.G. and A.M. provided scRNA-seq data and advised on data analysis for pancreatic cancer. K.Y., T.P.S. and H.Z. advised on statistical model development and implementation. N.S., N.E., S.T., J.L., V.B., E.E., M.N., P.C.B. and A.U. advised on data analysis for prostate cancer. S.M.H., P.A.F., J.Z. and C.S. advised on data analysis for the TRACERx dataset. B.L. and P.H.B. advised on data analysis for breast cancer. B.A.C advised on data analysis for bladder cancer. A.V. advised on the scRNA-seq data analysis. P.M. advised on data analysis for renal cancers. A.U. participated in GSEA for bulk RNA sequencing data and advised on ATAC-seq data analysis. P.C. advised on the initial concepts of the project. N.C.D. participated in manuscript writing. P.C.B., J.D. and P.V.L suggested improvements on data analysis, figure design and manuscript writing. J.D. performed data analysis on the DNA sequencing data. P.V.L. advised on the pan-cancer analysis and data interpretations. W.W. conceived the project, planned and supervised the work, developed computational methods, performed the data analysis and wrote the manuscript, in collaboration with all other authors. All authors contributed to the interpretation of results and commented on and approved the final manuscript.

Competing interests

A.M. receives royalties for a pancreatic cancer biomarker test from Cosmos Wisdom Biotechnology. A.M. is also listed as an inventor on a patent that has been licensed by Johns Hopkins University to Thrive Earlier Detection. A.M. is a consultant for Freenome and Tezcat Biotechnology. J.Z. reports research funding from Merck and Johnson & Johnson and consultant fees from Bristol Myers Squibb (BMS), Johnson & Johnson, AstraZeneca, Geneplus, OrigMed and Innovent outside of the submitted work. P.M. has received honoraria for service on a Scientific Advisory Board for Mirati Therapeutics and BMS, non-branded educational programs supported by Exelixis and Pfizer and research funding for clinical trials from Takeda, BMS, Mirati Therapeutics and Gateway for Cancer Research. W.W. reports research funding from Curis, Inc. J.P.S. and W.W. report research funding from Celsius Therapeutics. J.P.S. is a paid consultant for Engine Biosciences. S.K. has ownership interest in MolecularMatch, Lutris and Iylon and is a consultant for Genentech, EMD Serono, Merck, Holy Stone, Novartis, Eli Lilly, Boehringer Ingelheim, Boston Biomedical, AstraZeneca/MedImmune, Bayer Health, Pierre Fabre, Redx Pharma, Ipsen, Daiichi Sankyo, Natera, HalioDx, Lutris, Jacobio, Pfizer, Repare Therapeutics, Inivata, GlaxoSmithKline, Jazz Pharmaceuticals, Iylon, Xilis, Abbvie, Amal Therapeutics, Gilead Sciences, Mirati Therapeutics, Flame Biosciences, Servier, Carina Biotechnology, Bicara Therapeutics, Endeavor BioMedicines, Numab Pharma and Johnson & Johnson/Janssen and receive research funding from Sanofi, Biocartis, Guardant Health, Array BioPharma, Genentech/Roche, EMD Serono, MedImmune, Novartis, Amgen, Eli Lilly and Daiichi Sankyo. P.A.F. reports research funding from MEI Pharma, Inc. P.H.B. owns stock in GeneTex. C.S. acknowledges grant support from AstraZeneca, Boehringer Ingelheim, BMS, Pfizer, Roche-Ventana, Invitae (previously Archer Dx—collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical. C.S. is an AstraZeneca Advisory Board member and Chief Investigator for the AZ MeRmaiD 1 and 2 clinical trials and is also chief investigator of the NHS Galleri trial. C.S. has consulted for Amgen, AstraZeneca, Pfizer, Novartis, GlaxoSmithKline, Merck, BMS, Illumina, Genentech, Roche-Ventana, GRAIL, Medicxi, Metabomed, Bicycle Therapeutics, Roche Innovation Centre Shanghai and the Sarah Cannon Research Institute. C.S. had stock options in Apogen Biotechnologies and GRAIL until June 2021; currently has stock options in Epic Bioscience and Bicycle Therapeutics; and has stock options in and is a co-founder of Achilles Therapeutics. C.S. holds various patents relating to assay technology for cancer; US patents relating to detecting tumor mutations and methods for lung cancer detection; and both a European and a US patent related to identifying insertion/deletion mutation targets. All is outside the submitted work. The remaining authors declare no competing interests.

Additional information

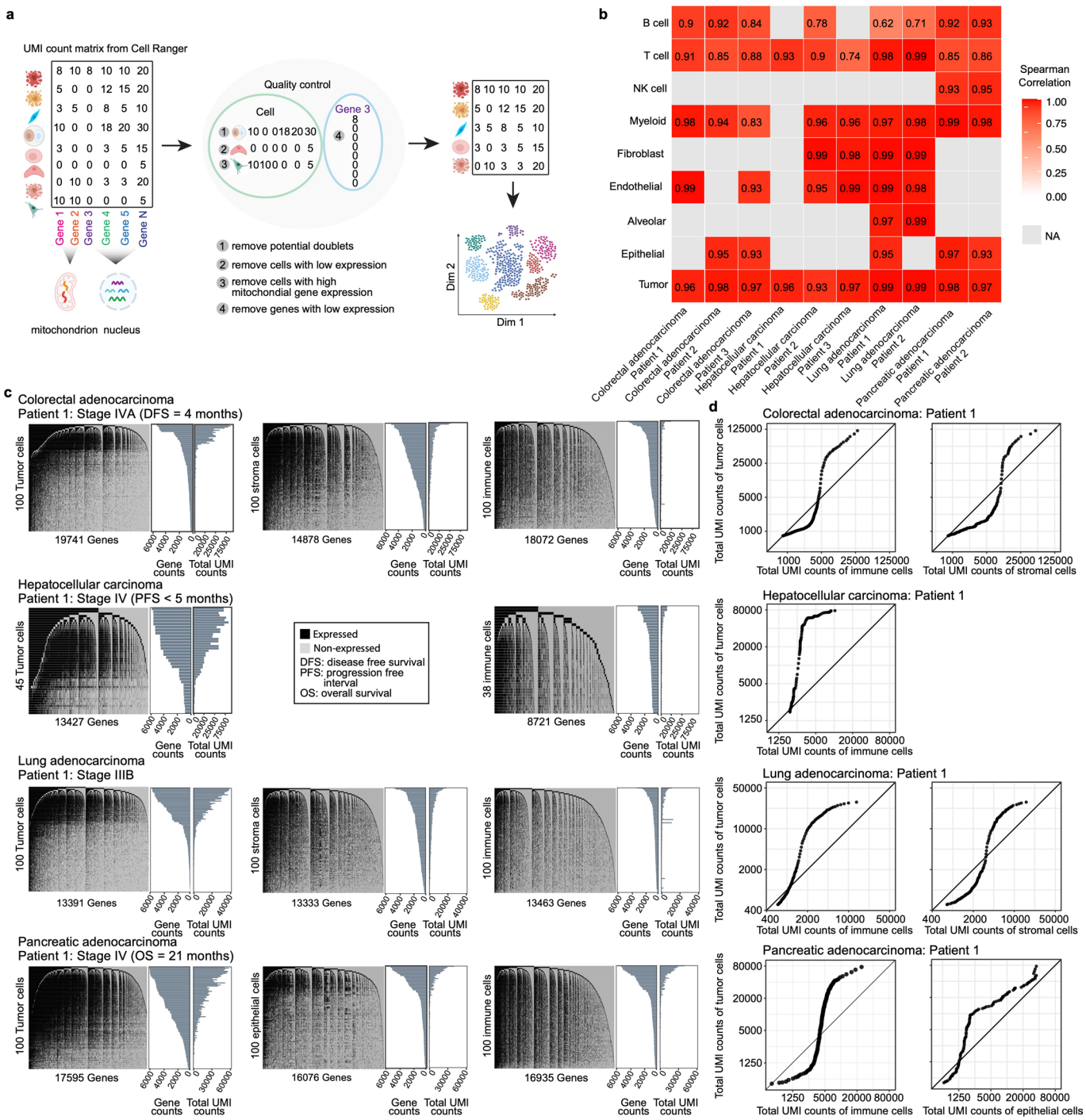
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01342-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01342-x.

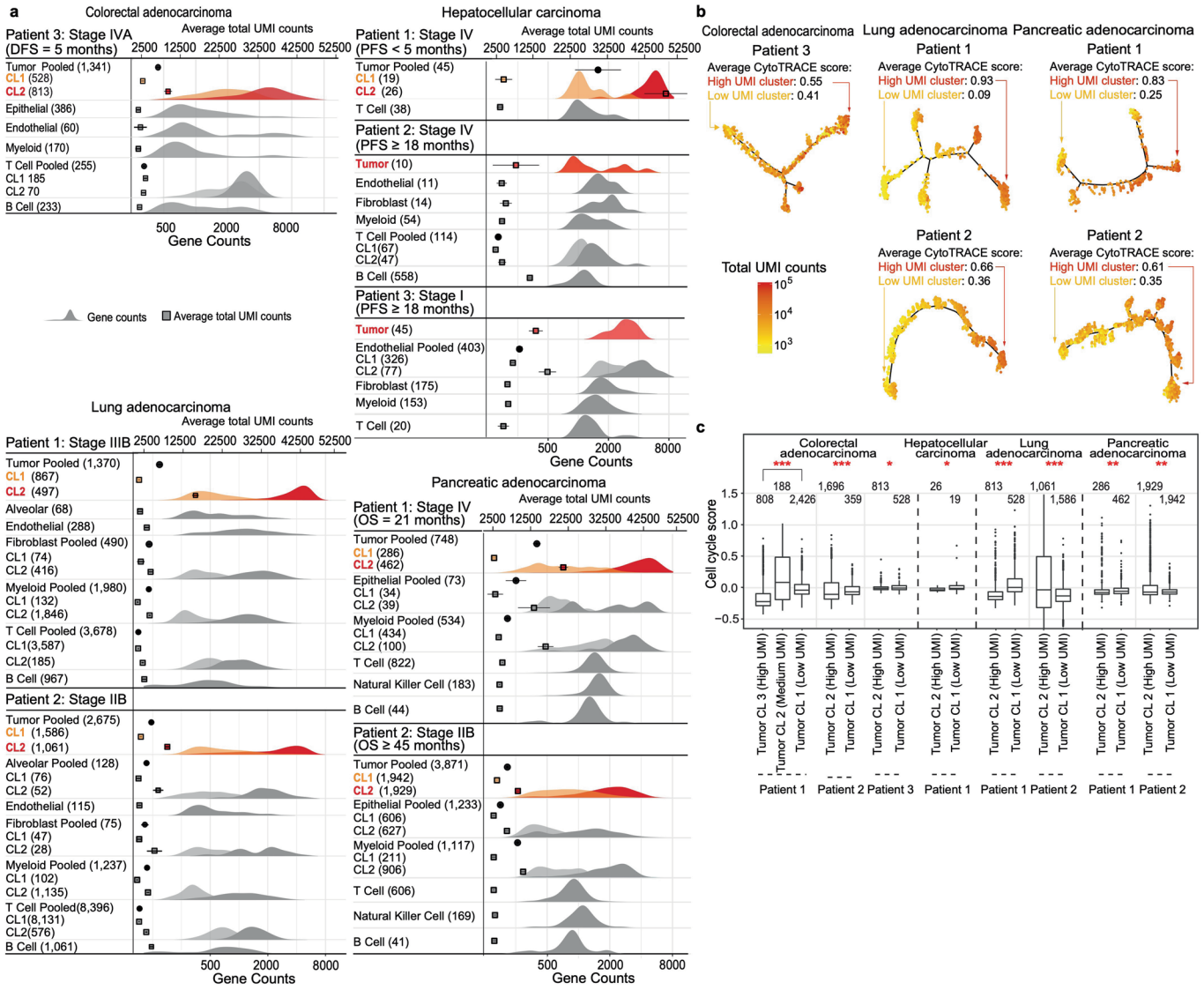
Correspondence and requests for materials should be addressed to Wenyi Wang.

Peer review information *Nature Biotechnology* thanks Wei Sun and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

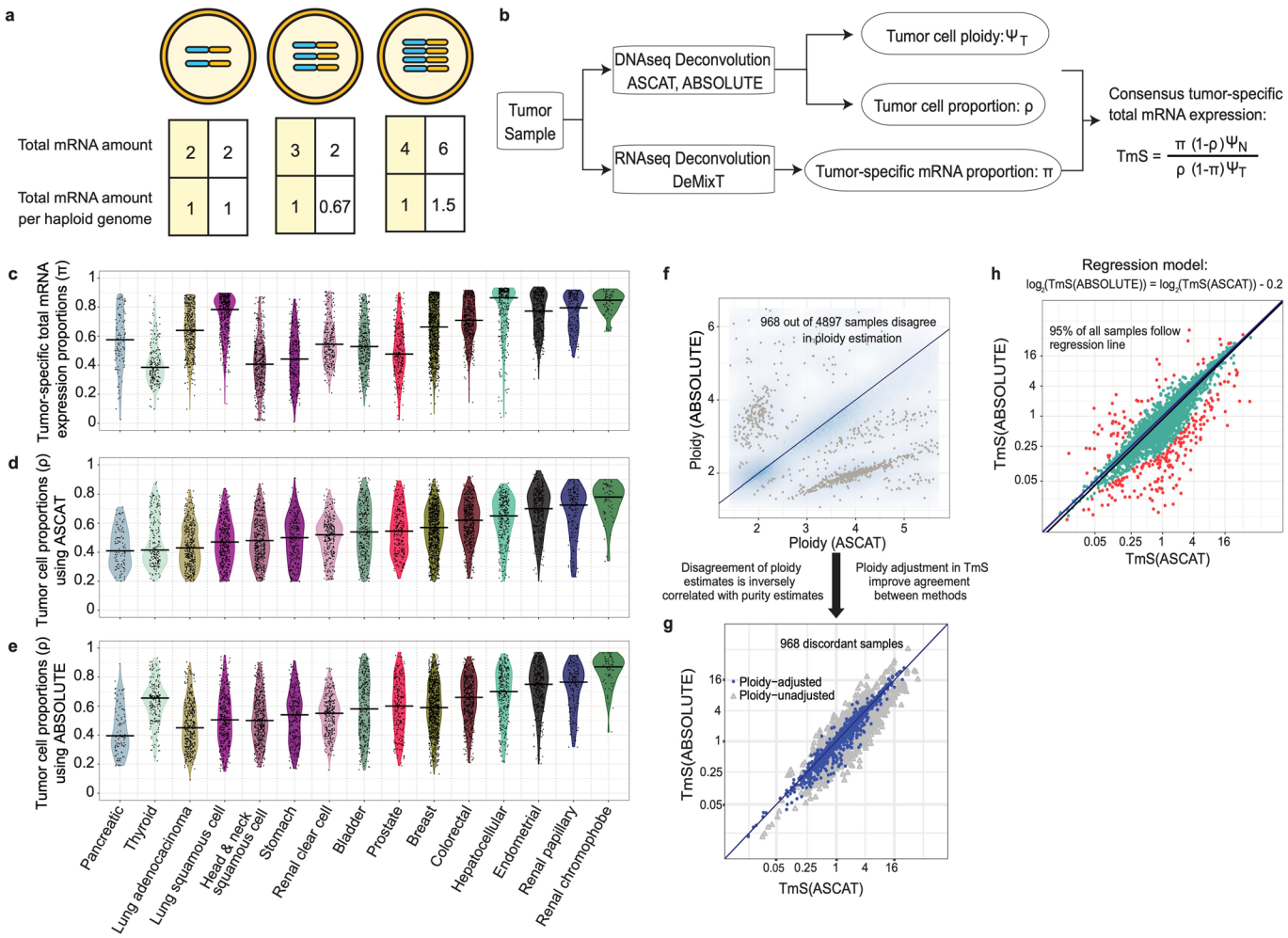
Reprints and permissions information is available at www.nature.com/reprints.



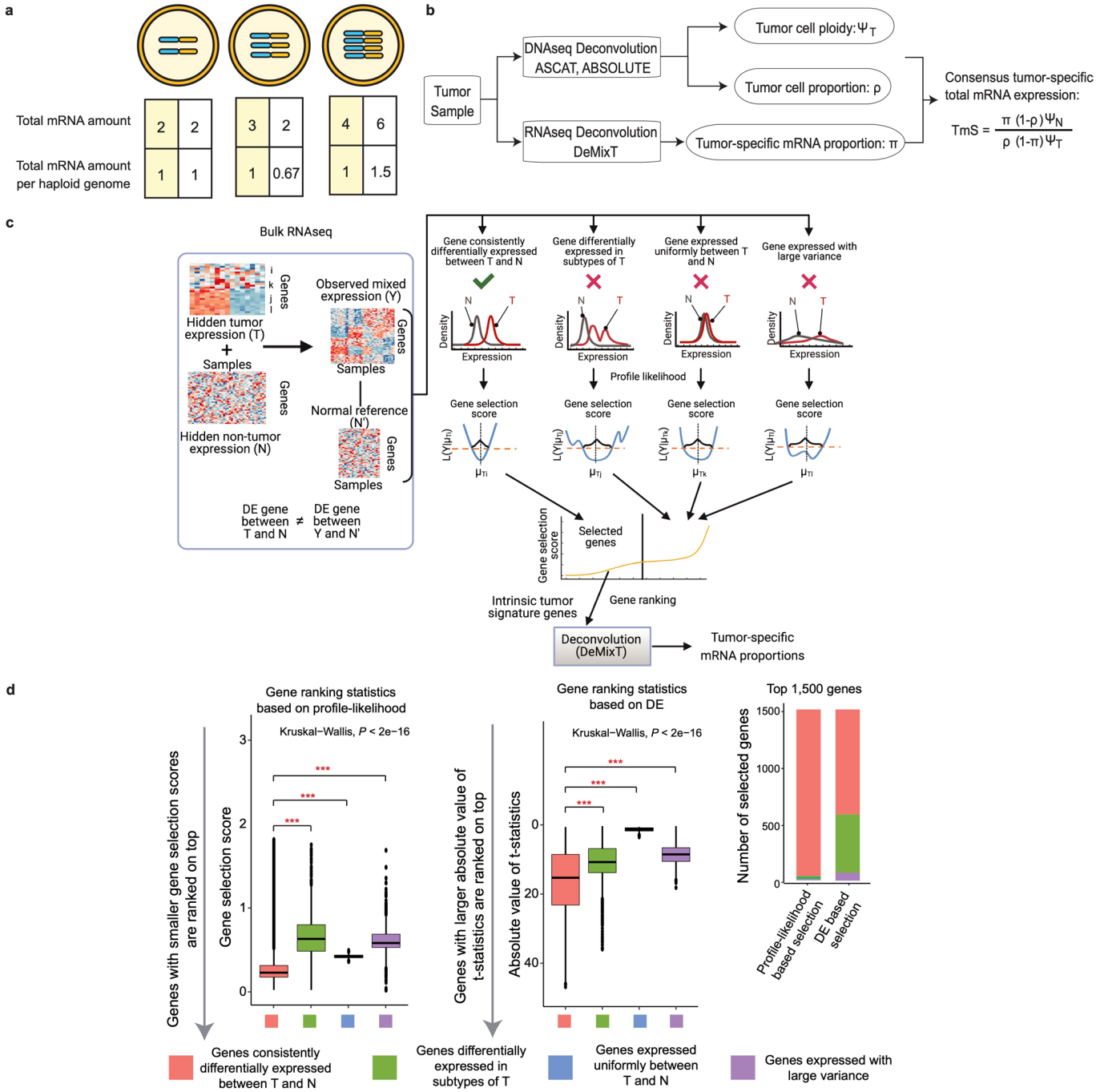
Extended Data Fig. 1 | High diversity of total mRNA expression in tumor cells. a, Flowchart of scRNA-seq data preprocessing. **b**, Heatmap showing the Spearman correlations between gene counts and total UMI counts across cell types in the ten patient samples. **c**, Illustration of expressed genes in tumor cells (left panels) compared to non-tumor cells: epithelial and stromal cells (middle panels) and immune cells (right panels). The data shown are based on cells randomly selected from each of the four 'patient 1' samples with colorectal, hepatocellular, lung and pancreatic cancers, who presented worse prognosis or advanced disease. In each heatmap, expressed genes (UMI count > 0) are shown in black, and non-expressed genes (UMI count = 0) are shown in gray. Cells in the rows and genes in the columns are ordered from high to low by the total numbers of expressed genes and the number of cells with detected expression of each gene, respectively. Barplots provide the corresponding distributions of gene counts and total UMI counts. **d**, Q-Q plots of total UMI counts in tumor cells compared to non-tumor cells for the same four 'patient 1' samples that were used as in **c**. For each patient, the $\log_2$ transformed total UMI counts of immune cells (left) or stromal/epithelial cells (right) are used as the theoretical quantiles, respectively.



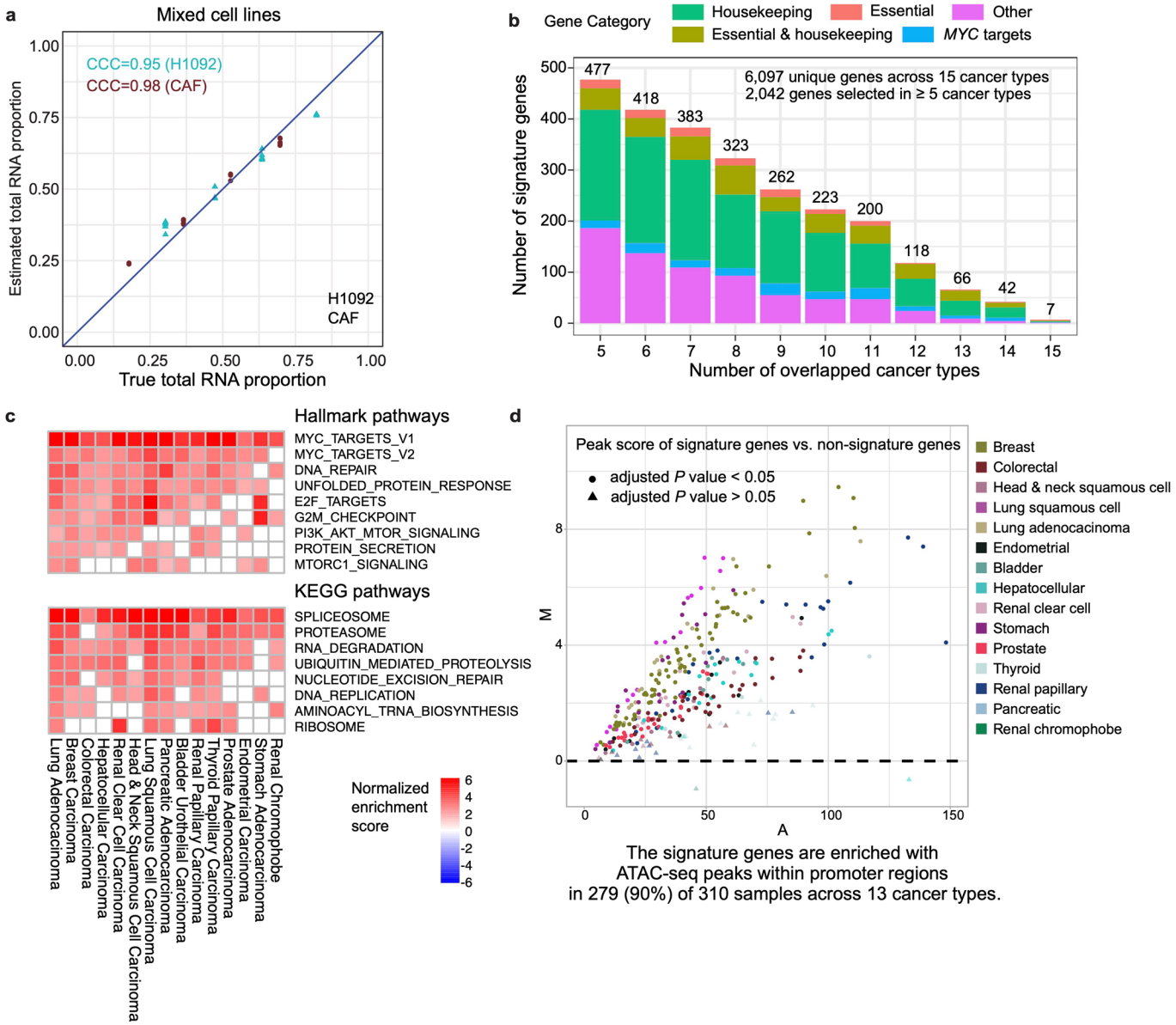
Extended Data Fig. 2 | Using total UMI counts and gene counts to measure global gene expression heterogeneity. a, Distributions of gene counts and total UMI counts by cell type in scRNA-seq data from eight remaining patients with colorectal, hepatocellular, lung or pancreatic cancers (in relation to Fig. 1). The top x-axis annotates total UMI counts (means and 95% CIs). The bottom x-axis annotates gene count distribution (density). Density curves are shown in color for tumor cells and in grayscale for non-tumor cells. Clusters with higher gene counts are shown in darker shades. Numbers in the parentheses indicate the number of cells analyzed. **b**, Monocle-inferred trajectories for tumor cells from five patients with colorectal, lung and pancreatic cancers. Cells on the trees are colored by total UMI counts. Average differentiation scores by CytoTRACE for high- and low-UMI count tumor cell clusters are labelled. **c**, Distribution of cell cycle scores in tumor cell clusters from eight scRNA-seq patient samples where multiple tumor cell clusters were presented. Cell cycle score is the sum of the S and G2/M scores as estimated by Seurat. *P* values of two-sided Wilcoxon rank-sum tests comparing the cell cycle scores across clusters are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). In the boxplots, whiskers represent the maximum and minimum values of cell cycle scores, the middle line in the box denotes median, and the bounds of the box stand for upper and lower quartiles.



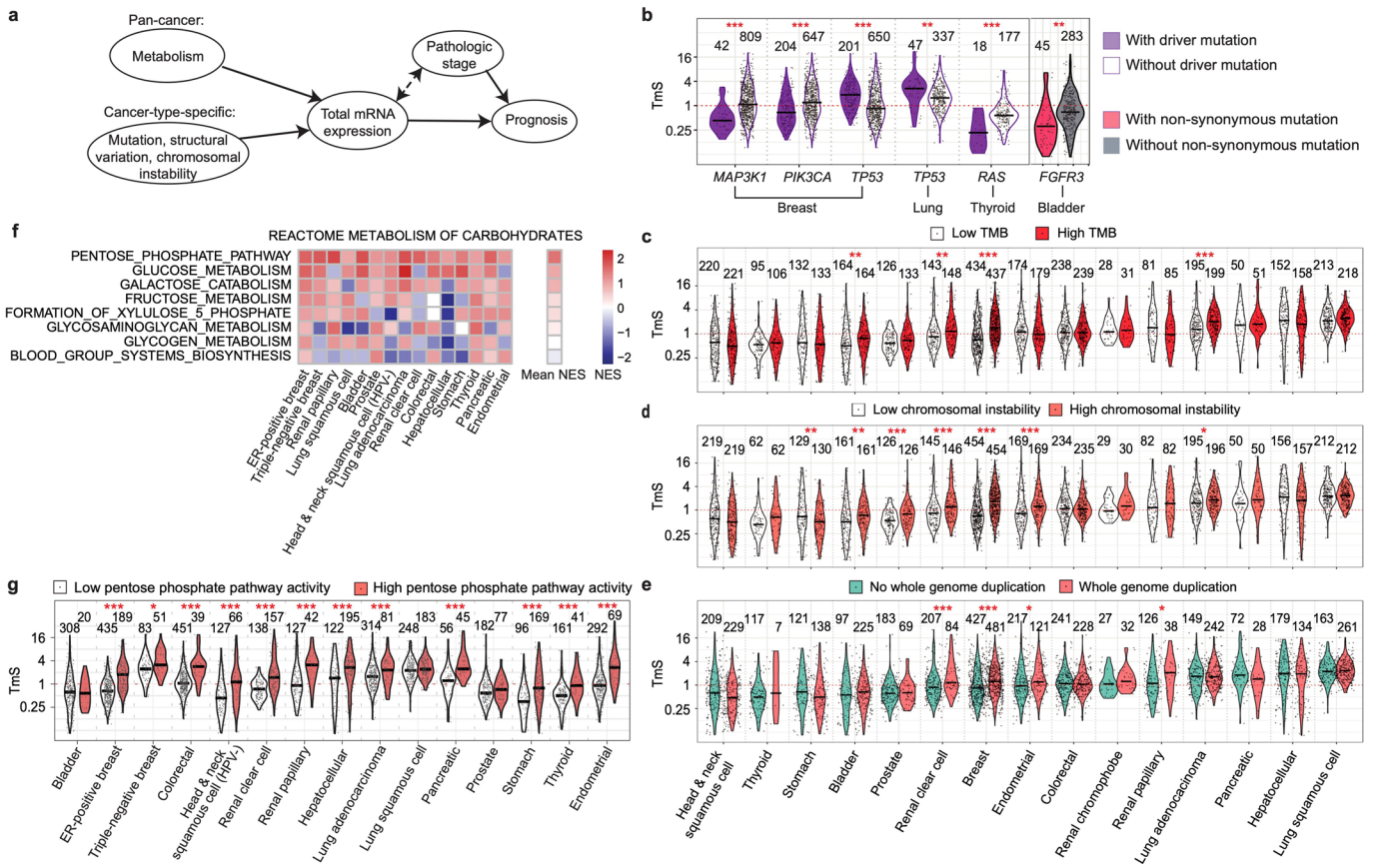
Extended Data Fig. 3 | Consensus estimation of TmS from matched RNAseq and DNAseq data in TCGA. a, Illustrative relationship between cells, ploidy and mRNA content. Three examples with ploidy of 2, 3, or 4 are given. Under the scenario of linear dosage effects, as shown in the boxes with a yellow background, if cellular 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, respectively. Under the scenario of dosage compensation, that is, more chromosomal copies but maintaining the same total dose, the second cell has a total mRNA amount of 2 and a per haploid genome value of 0.67. Under the scenario of dosage transgression, *that is*, more chromosomal copies with more dose per copy, the third cell has a total mRNA amount of 6 and a per haploid genome value of 1.5. **b**, Definition of TmS and its analytic pipeline. **c**, Distribution of tumor-specific mRNA proportions estimated by DeMixT across cancer types. **d-e**, Distributions of tumor cell proportions estimated by (**d**) ASCAT or (**e**) ABSOLUTE across cancer types. **f**, Smoothed scatter plot of tumor ploidy estimates from ABSOLUTE *vs.* ASCAT across all samples. Gray points correspond to 968 samples that presented inconsistent tumor ploidy (and purity) estimates between the two methods. **g**, TmS estimates using either ABSOLUTE or ASCAT-derived purity and ploidy estimates with or without ploidy adjustment for the 968 discordant samples from (**f**). Blue and gray points correspond to TmS prior to and after ploidy adjustment, respectively. Ploidy adjustment improved consistency between the ABSOLUTE and ASCAT results. **h**, Scatter plot of TmS calculated using the two methods. A linear regression model was fitted using $\log_2$(TmS estimated by ABSOLUTE) as the predicted variable and $\log_2$(TmS estimated by ASCAT) as the predictor variable. Red points are outliers with a Cook's distance $\geq 4/n$, where $n = 5{,}295$ for the total number of TCGA samples. Cyan points are the remaining samples (95%) that showed a good fit for the model and hence their TmS estimates are consistent and robust across two DNAseq deconvolution methods.



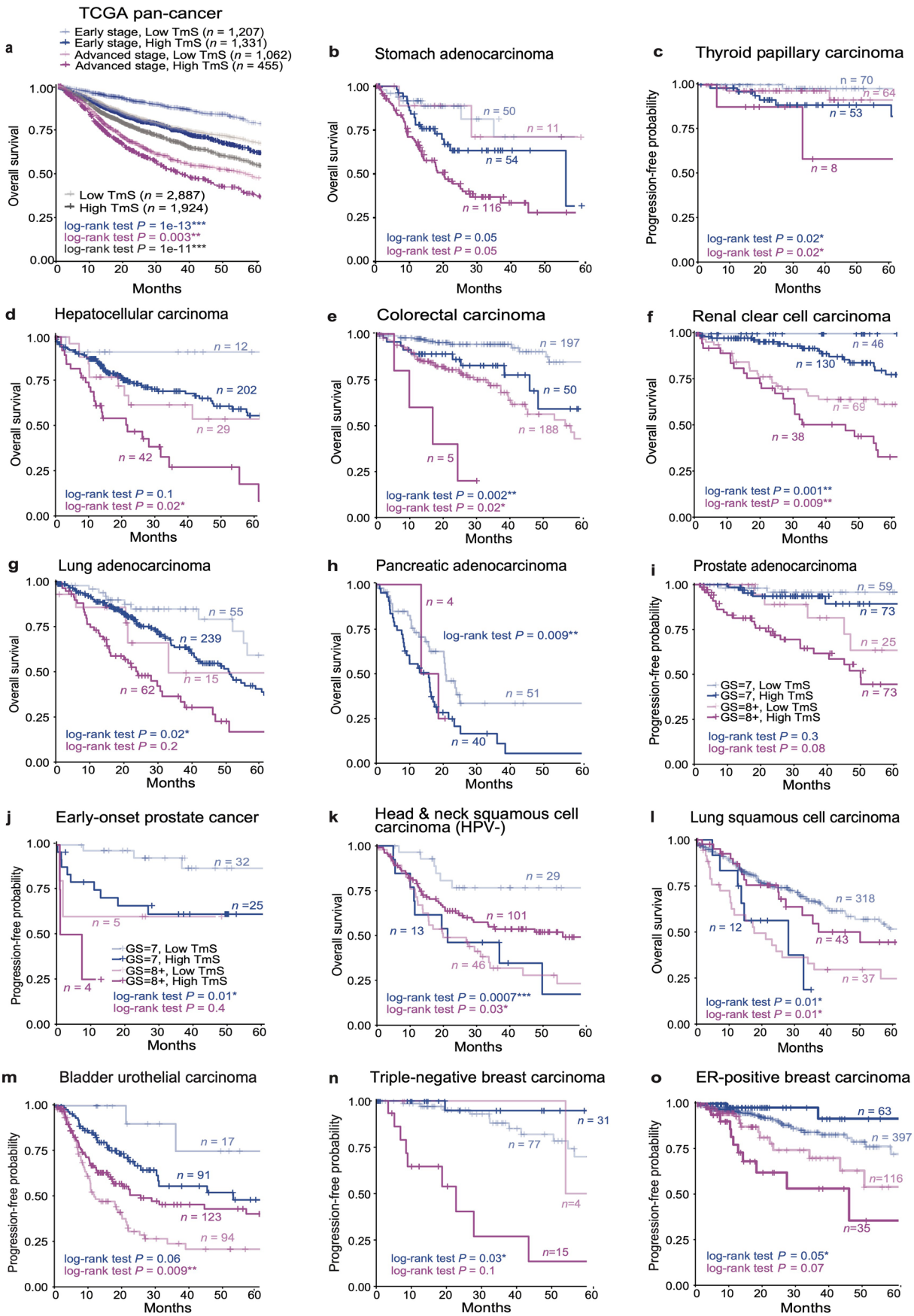
Extended Data Fig. 4 | Profile likelihood-based gene selection for RNAseq deconvolution. a-b, same as Extended Data Fig. 3a-b. **c**, Illustration of the RNAseq deconvolution workflow with intrinsic tumor signature genes selected using a profile-likelihood based gene selection approach. Three scenarios where genes with undesirable properties are included, leading to large estimation biases, are illustrated with red 'x' on top. Their corresponding gene selection scores are expected to be larger than genes with the desirable property for the DeMixT model-based deconvolution (illustrated with a green check on top). Therefore, when genes are ranked based on the gene selection score, as derived using profile likelihoods, selecting the top-ranked genes will reduce the biases in estimating tumor-specific mRNA proportions. **d**, Distributions of gene selection scores across four types of genes in a simulation study (**Supplementary Note 2.2**). For the profile-likelihood based gene selection, genes are ranked from the smallest to the largest score (left). For the DE based gene selection, genes are ranked from the largest to the smallest absolute t-statistics (middle). *P* values of Kruskal-Wallis (one-way ANOVA) test across all four gene groups are shown on top. *P* values of two-sided Wilcoxon rank-sum tests within pairs of gene groups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). The types of genes among the top 1,500 selected genes are shown (right) for the two rankings. Ideally only genes consistently differentially expressed between tumor (*T*) and normal (*N*), annotated in red, should be selected, corresponding demonstrating the lowest values in both panels as compared to genes annotated in other colors. This is achieved by the profile-likelihood method but not the DE method. In the boxplots, whiskers represent the maximum and minimum values of gene selection scores, the middle line in the box stands for median, and the bounds of the box stand for upper and lower quartiles.



Extended Data Fig. 5 | Validating the TmS measure through benchmarking and evaluating the biological relevance of intrinsic tumor signature genes in TCGA. a, Total mRNA proportion estimation for H1092 and CAF using DeMixT in the benchmarking study ($n=18$). The concordance correlation coefficient (CCC) for two variables x (true tumor-specific RNA proportions) and y (estimated tumor-specific RNA proportions) is expressed as $\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}$, where μ and σ^2 represent the mean and variance, and ρ is the Pearson correlation coefficient. **b**, Histogram of the number of overlapping genes across cancer types and their annotation categories. The y axis represents the total number of genes and the x axis represents the number of cancer types for which a gene was selected. **c**, Heatmap of normalized enrichment scores of top cancer hallmark pathways and KEGG pathways. Only pathways with a BH adjusted P value < 0.05 are colored. **d**, M-A plot comparing ATAC-seq peak scores of intrinsic tumor signature genes (signature) *vs.* other genes (non-signature) from matched tumor samples in each cancer type. Samples above the dashed line have higher ATAC-seq peak scores in intrinsic tumor signature genes compared to those in non-signature genes. Samples with BH adjusted P values < 0.05 from per-sample permutation tests are shown as circles.

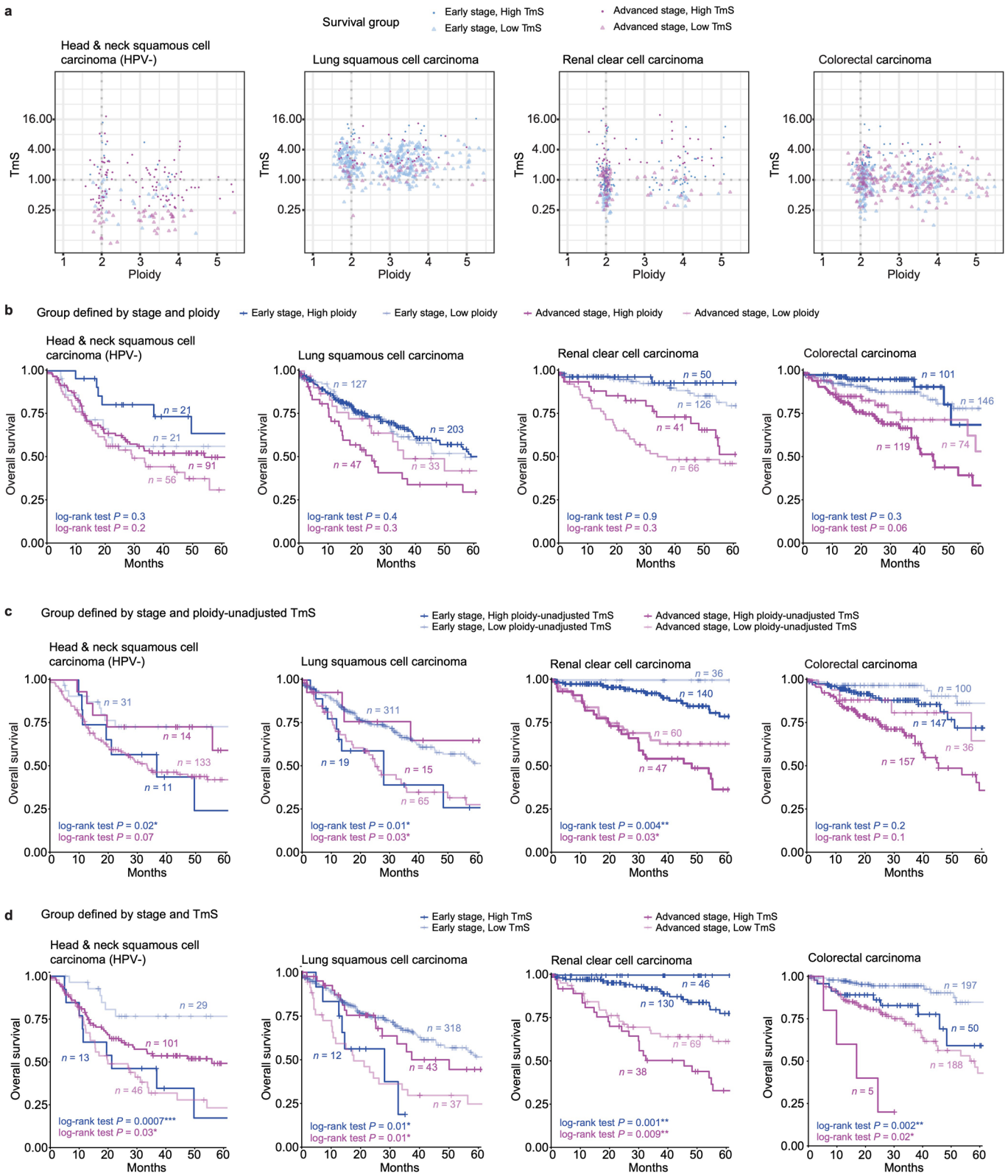


Extended Data Fig. 6 | TmS is associated with tumor genomic features and metabolic pathway activities across cancer types. a, Contributors to tumor-specific total mRNA expression. **b**, Distributions of TmS for TCGA samples with or without specific mutations in six cancer-gene pairs. The number of samples is indicated on the top. We performed an agnostic association analysis of TmS with all non-synonymous mutations (32,894 cancer-gene pairs, using logistic regression models), and concurrently a driver mutation-specific association analysis of TmS (24 cancer-gene pairs). We find 5 overlapping pairs out of 6 statistically significant pairs produced from each interrogation (BH adjusted *P* values < 0.01). The additional pair found through the agnostic search (*FGFR3* in bladder carcinoma in TCGA) was not identified in the driver mutation analysis due to a limited sample size. These associations in breast, lung, thyroid, and bladder cancers show that TmS can capture changes in tumor phenotypes induced by driver mutations in a cancer type-specific manner. Our observation also supports previous findings that the same driver mutations may not have the same prognostic effect across cancers, and their effects may be modified by additional tumor and/or treatment-related factors. **c-e**, Distribution of TmS for patient samples with (**c**) high or low tumor mutation burden (TMB); (**d**) high or low chromosomal instability score; (**e**) with or without a whole genome duplication event. Patient groups are categorized as high *vs.* low based on the median values of TMB and chromosomal instability scores in (**c**) and (**d**) respectively. **f**, Heatmap of normalized enrichment scores (NES) of Reactome metabolism of carbohydrates pathways across 15 cancer types in TCGA. Pathways are ordered by the mean NES across 15 cancer types, from high to low. **g**, Distribution of TmS for patient samples with high or low for pentose phosphate pathway activity, where patient groups are defined by hierarchical clustering of expression levels from 13 genes. For **b**-**d** and **g**, the BH adjusted *P* values for two-sided Wilcoxon rank-sum tests comparing TmS between corresponding groups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001).

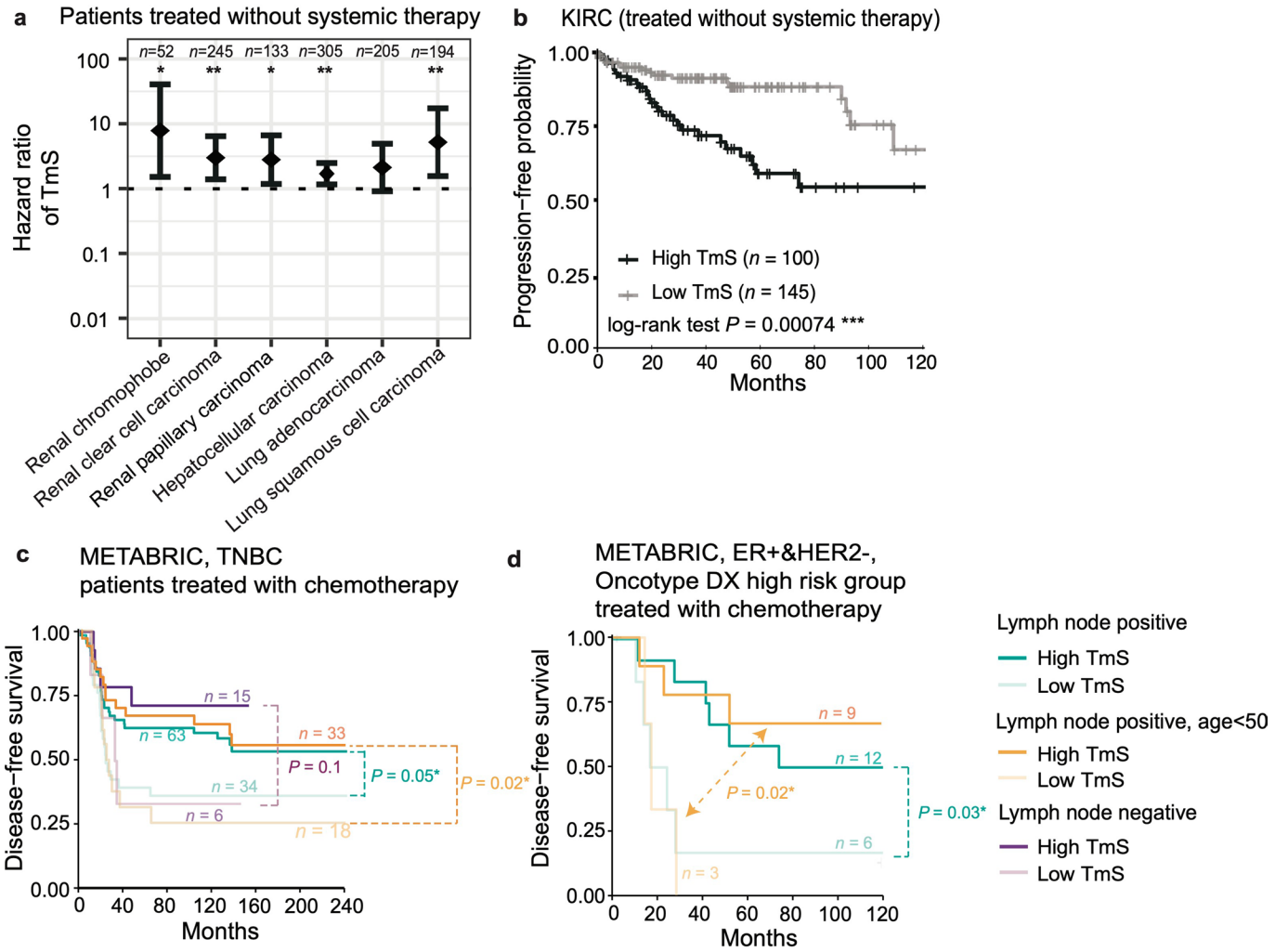


Extended Data Fig. 7 | See next page for caption.

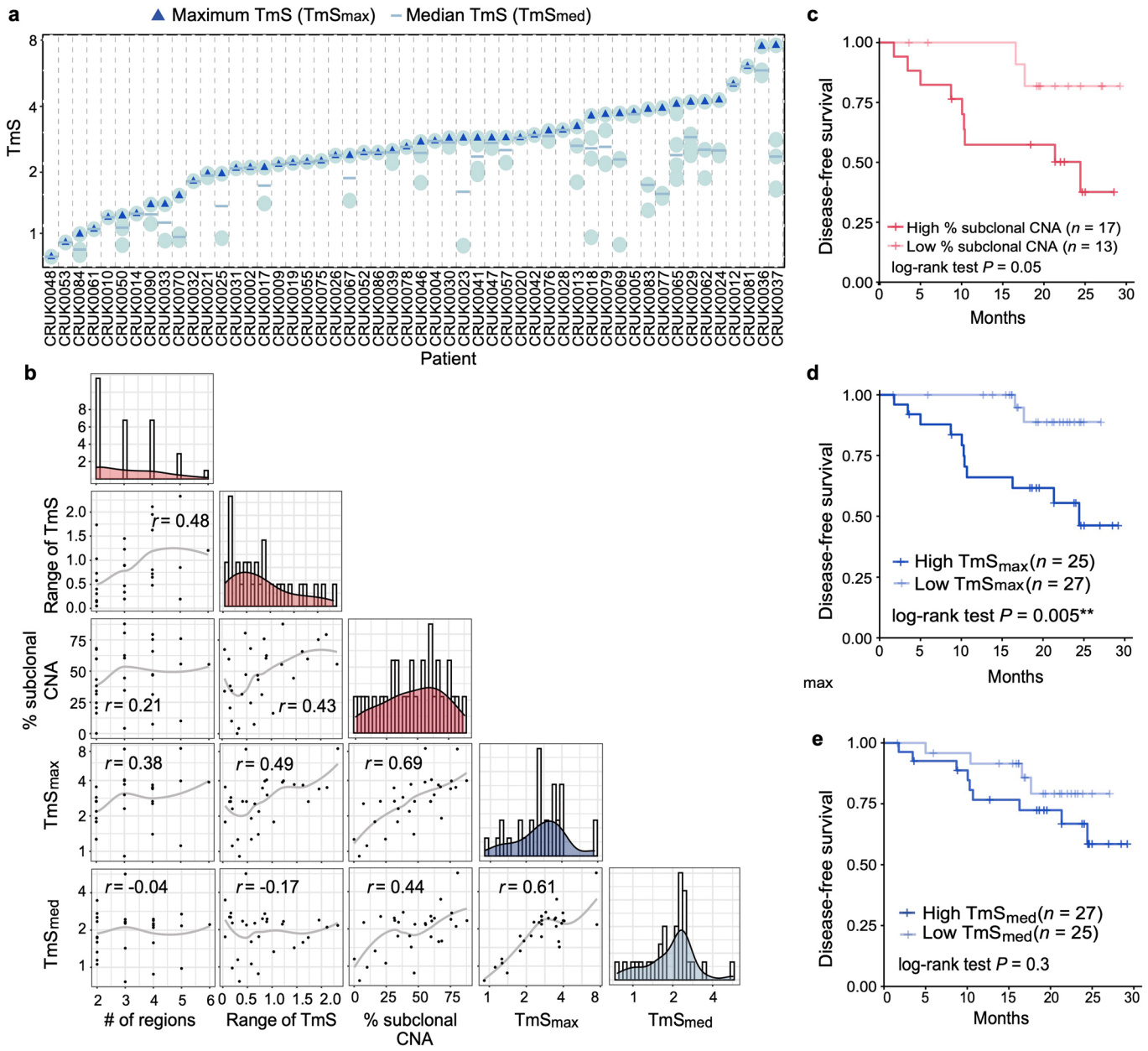
Extended Data Fig. 7 | TmS refines prognostication on pathological stages. a, KM curves of OS for TCGA pan-cancer. Gray lines denote summary KM curves of patients with high *vs.* low TmS across all cancer types. KM curves are further grouped by TmS and pathological stages into four groups. *P* values of log-rank tests between high *vs.* low TmS groups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). **b-o**, KM survival curves for individual cancer types.



Extended Data Fig. 8 | Prognostication using ploidy or ploidy-unadjusted TmS on pathological stages. a, Scatter plots of TmS (y axis) *vs.* tumor ploidy (x axis) for samples from TCGA patient cohorts with head-and-neck squamous cell carcinoma (HPV negative), lung squamous cell carcinoma, renal clear cell carcinoma, and colorectal carcinoma. The samples were grouped into high *vs.* low TmS within early or advanced pathological stages, with different groups shown in distinct colors. TmS shows no correlation with tumor ploidy, with Spearman correlation coefficients $r = -0.12$, 0.01, 0.08 and −0.02 for the four cancer types. **b**, KM survival curves of OS in four cancer types according to patient groups defined by ploidy and stage. We grouped patients into high *vs.* low ploidy based on a cutoff of 2.5 within early or advanced pathological stage. **c**, KM survival curves of overall survival in four cancer types over patient groups defined by ploidy-unadjusted TmS and stage. **d**, KM survival curves of OS in four cancer types for patient groups defined by TmS and stage. *P* values of log-rank tests between pairs of patient groups are shown with matching colors and are indicated by asterisk (* $P < 0.05$, ** < 0.01, *** < 0.001).



Extended Data Fig. 9 | TmS refines prognostication in cancer patients with and without systemic therapy. a, Forest plot of hazard ratios and 95% of CIs of TmS as predictor in patients treated without systemic therapy across 6 TCGA cancer types. *P* values of two-sided Wald tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). **b**, KM curves of PFI for renal clear cell carcinoma patients without systemic therapy. **c**, KM curves of DFS for METABRIC triple negative breast cancer patients who are treated with chemotherapy. KM curves are further grouped by TmS, Lymph node status and age into six groups. **d**, KM curves of DFS for METABRIC estrogen receptor (ER) positive and human epidermal growth receptor-2 (HER2) negative breast cancer patients who are classified as high risk by Oncotype Dx risk score and treated with chemotherapy. KM curves are further grouped by TmS and age under 50. For **b**, **c** and **d**, *P* values of log-rank tests between pairs of patient groups are shown with matching colors and are indicated by asterisk (* $P < 0.05$, ** < 0.01, *** < 0.001).



Extended Data Fig. 10 | Regional TmS identifies spatial heterogeneity and refines prognostication in patients with early-stage lung cancer. **a**, Distribution of TmS values for 116 tumor regions from 52 patients of the TRACERx study. Blue triangles denote the maximum TmS for a patient. Blue '-' denote the median TmS for a patient. **b**, Pairwise scatter plots and histograms of number of regions, range of TmS, % subclonal CNA, maximum of TmS across regions (TmS_{max}), and median of TmS across regions (TmS_{med}) per patient. The number of evaluated patients with at least 2 regions is 30. Spearman correlation coefficient *r*'s are shown, and the gray lines represent a loess fit. **c**, KM survival curves of DFS for the 30 patients stratified by % subclonal CNA: high versus low. **d-e**, KM survival curves of DFS for all 52 patients stratified into two groups by TmS_{max} (**d**) and (**e**) TmS_{med}, respectively. *P* values obtained by log-rank tests between high *vs.* low TmS groups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001).

nature research

Corresponding author(s): Wenyi Wang

Last updated by author(s): Apr 26, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No commercial, open source, or custom software or code was used for data collection.
Data analysis	Individual software components are as follows: R(v 4.1.2), DeMixT (v 1.2.2), ASCAT (v2.5.2), sequenza (v 3.0.0), Seurat (v 3.0), Monocle 2 (v 2.14.0), CytoTRACE (v 0.3.3) Downstream analyses and statistical analyses were performed using package as follows: rpart (v 4.1.15), survival (v 3.2.0), survminer (v 0.4.9), gprofiler2 (v 0.2.0), fgsea (v 1.14.0), DSS (v 2.42.0), ChIPseeker (v 3.11), ComBat (v 3.20.0), limma (v 3.50.1), GeneOverlap (v 1.24.0), ComplexHeatmap (v 2.4.3), RColorBrewer (v1.1.2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The UMI counts of the hepatocellular carcinoma single cell RNA sequencing data were downloaded from the Gene Expression Omnibus (GEO) with the accession code GSE125449. The UMI counts and cell type annotations of the lung adenocarcinoma single cell RNA sequencing data were downloaded from the ArrayExpress under accessions E-MTAB-6149. The UMI counts of the colorectal adenocarcinoma single cell RNA sequencing data are available at http://crcmoonshot.org/?page_id=189. FASTQ files of single-cell RNA sequencing data from pancreatic cancer will be publicly available on the GEO with the

accession code GSE156405.
Raw read counts from the mixed cell-line study were downloaded from GEO with accession code GSE121127.
Raw read counts of RNA sequencing data, clinical data, and somatic mutations from 7,054 tumor samples across 15 TCGA cancer types are available for download from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). ATACseq data for TCGA samples were downloaded from https://science.sciencemag.org/content/362/6413/eaav1898/tab-figures-data.
Clinical information of ICGC-EOPC was downloaded from https://www.sciencedirect.com/science/article/pii/S1535610818304823?via%3Dihub#gs1.
All primary METABRIC data including Affymetrix SNP 6.0 CEL files and Illumina HT 12 gene expression arrays, are available at the EGA (EGAS00000000083), and may be downloaded from https://ega-archive.org/studies/EGAS00000000083. Clinical information of METABRIC was downloaded from https://www.cbioportal.org/study/clinicalData?id=brca_metabric.
Clinical information of TRACERx was downloaded from https://www.nejm.org/doi/full/10.1056/NEJMoa1616288#article_supplementary_material.
WES data of TRACERx was downloaded from https://ega-archive.org/studies/EGAS00001002247.
RNAseq data of TRACERx was downloaded from https://ega-archive.org/studies/EGAS00001003458.
TmS values of all samples and the identified intrinsic tumor signature genes for this study are available for download at https://github.com/wwylab/TmS.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample size was determined by the availability of the data.
Data exclusions	Samples with missing RNAseq data or DNA sequencing data were excluded from the analysis.
Replication	All attempts at replication were successful.
Randomization	There are no experimental groups in this study.
Blinding	There are no experimental groups in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Demographically, the single-cell RNA sequencing datasets from three colorectal patients includes 1 male at age 45, 1 male at 37 and 1 female at age 63.
Recruitment	The three colorectal adenocarcinoma patients with single-cell RNA sequencing data were identified prior to surgery or biopsy and were asked to prospectively sign consent for participation on the IRB approved protocol LAB10-0982 after discussion of risks and benefits. Prospective consenting of patients was required due to the intent to utilize fresh tumor tissue. No known selection bias other than the fact that these patients went through surgery at MD Anderson.

Ethics oversight	The three colorectal adenocarcinoma patient samples were obtained with informed consent and were approved by the Human Subjects Protection Office, Clinical Research Committee as well as five separate Institutional Review Boards (IRB) at the MD Anderson Cancer Center, in accordance with the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.