Differing total mRNA expression shapes the molecular and clinical phenotype of cancer

Shaolong Cao[1+], Jennifer R. Wang[2+], Shuangxi Ji[1], Peng Yang[1,3], Jingxiao Chen[1], Matthew D. Montierth[1,4], John Paul Shen[5], Jaewon James Lee[6,7,8], Paola A Guerrero[6,7], Kaixian Yu[9], Julie Livingstone[10,11,12,13], Vinayak Bhandari[14], Shawna M Hubert[15], Najat C. Daw[16], P. Andrew Futreal[17], Eleni Efstathiou[18], Bora Lim[19], Andrea Viale[17], Jianjun Zhang[15], Anirban Maitra[6,7,20], Scott Kopetz[5], Peter Campbell[21], Terrence P. Speed[22,23], Paul C. Boutros[10,11,12,13,14], Alfonso Urbanucci[24], Hongtu Zhu[9], Jonas Demeulemeester[25,26], Peter Van Loo[25] and Wenyi Wang[1*]

[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 Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[10]Department of Human Genetics, University of California, Los Angeles, CA, USA.

[11]Department of Urology, University of California, Los Angeles, CA, USA.

[12]Institute for Precision Health, University of California, Los Angeles, CA, USA.

[13]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, USA.

[14]Department of Medical Biophysics, University of Toronto, Canada.

[15]Department of Thoracic Head Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[16]Department of Pediatrics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[17]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[18]Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[19]Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[20]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[21]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK.

[22]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia.

[23]Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia.

[24]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway.

[25]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK.

[26]Department of Human Genetics, KU Leuven, Leuven, Belgium.

[+] Authors contributed equally

[*] Correspondence: wwang7@mdanderson.org. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center. 1400 Pressler, Houston TX 77030.

Abstract

Cancers can vary greatly in their transcriptomes. In contrast to alterations in specific genes or pathways, the significance of differences in tumor cell total mRNA content is poorly understood. Studies using single-cell sequencing or model systems have suggested a role for total mRNA content in regulating cellular phenotypes. However, analytical challenges related to technical artifacts and cellular admixture have impeded examination of total mRNA expression at scale across cancers. To address this, we evaluated total mRNA expression using single cell sequencing, and developed a computational method for quantifying tumor-specific total mRNA expression (TmS) from bulk sequencing data. We systematically estimated TmS in 5,181 patients across 15 cancer types and observed close correlations with clinicopathologic characteristics and molecular features, where high TmS generally accompanies high-risk disease. At a pan-cancer level, high TmS is associated with increased risk of disease progression and death. Moreover, TmS captures tumor type-specific effects of somatic mutations, chromosomal instability, and hypoxia, as well as aspects of intratumor heterogeneity. Taken together, our results suggest that measuring total mRNA expression offers a broader perspective of tracking cancer transcriptomes, which has important clinical and biological implications.

Introduction

Cells of different origins and states can present large differences in quantities of multiple RNA species[1–8]. Variation in total mRNA amount, i.e., the sum of detectable mRNA transcripts across all genes per cell, has been linked indirectly to cancer progression as a result of *MYC* activation[9,10] and aneuploidy[11,12]. More broadly, reprogramming of the transcriptional landscape is a critical hallmark of cancer[13]. Over the last two decades, we have come to understand how changes in the expression of specific genes or pathways affect tumor progression[14–16] and prognosis[17,18]. Recently, single-cell RNA sequencing (scRNAseq) has revealed that the total number of expressed genes per cell, and total mRNA content, were more predictive of cellular phenotype, such as developmental stages in normal cells and differentiation states in cancer cells, than alterations in specific genes and pathways[19,20]. Total mRNA expression per cell may therefore represent an important feature of cancer transcriptomes that has been largely overlooked.

Single-cell sequencing allows quantification of total mRNA expression in individual tumor cells[21–23]. However, high cost and sample quality requirements have prohibited its application to large cohorts of human tumors. Bulk tumor RNA sequencing data on the other hand can be readily obtained in a clinical setting, but total mRNA expression information is masked during standard data analysis. Specifically, methods for differential expression analysis typically assume that total mRNA content is constant across samples. As such, variation in total mRNA transcript levels is removed by normalization, together with technical biases such as read depth and library preparation[4,24–26]. A further obstacle in cancer studies is the inability to directly measure tumor-specific mRNA, as the data often contains reads from both tumor and admixed normal cells.

The intrinsic mixing of distinct cell types during bulk RNA sequencing presents an opportunity for analyzing cellular populations using a common technical baseline within each sample, while maintaining the cell-type specific total mRNA levels. Building upon prior work in bulk transcriptome deconvolution[27–29] and in modelling tumor ploidy[30,31], we here 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 scrutinized total mRNA expression using single-cell sequencing data across four cancer types[32,33], as well as in matched bulk RNA and DNA data from 5,292 patient samples across 15 cancer types from the TCGA, ICGC[34] and TRACERx studies[35,36]. Our analyses revealed that variation in total mRNA expression is a phenotypic feature of tumor cells that captures tumor behavior in a cancer type-specific fashion and predicts prognosis.

Results

Diversity in total mRNA expression is a hallmark of tumor cells

Total mRNA expression can be estimated directly by single-cell RNA sequencing (scRNAseq) via examination of two features, total unique molecular identifier (UMI) counts, which quantifies the total number of observed mRNA transcripts per cell, and the number of expressed genes per cell (gene counts). We quantified total UMI and gene counts in scRNAseq data generated from human colorectal, liver, lung[32] and pancreatic tumors[33] (**Methods and Supplementary Information (SI)**). We observe larger variability in total UMI and gene counts in tumor cells compared to non-tumor cells (epithelial, stromal and immune cells), which generally present a smaller dynamic range (**Fig. 1A, Fig. S1A, Table S1**). Consistent with previous reports[32,37,38], we find multiple clusters of tumor and non-tumor cells with distinct total UMI and gene counts, indicative of diversity in total mRNA content (**Fig. 1B**, **Methods and SI**). Across all clusters, UMI counts are highly correlated with gene counts, as expected (median Spearman $r = 0.96$, **Fig. S1B**). While this may be technical at least in part, it also suggests overlap in the underlying biology of the two features. Every patient sample contains one tumor cell cluster with total UMI and gene counts similar to those of stromal or immune cells. Tumor samples from seven patients across four cancer types present additional tumor cell clusters with higher total UMI and gene counts (i.e., high-UMI clusters). In four patients with shorter time to disease progression (colon, liver and pancreas cancers), or with advanced stage disease (lung cancer), a high-UMI tumor cell cluster is where UMI counts are significantly higher than any other cell clusters. As a result, an increased average total UMI count is observed across tumor cells in these samples (**Fig. 1B**, **SI**). The observed fold change in total UMI counts between tumor cell clusters range from 4.7 to 25, which are higher than what would be expected from a whole-genome duplication event[39,40] (adjusted P value < 0.001, **Methods and SI**). These findings suggest that high diversity in total mRNA expression is a distinctive feature of tumor cells and may relate to clinical characteristics. Given that the observed differences in total mRNA expression between tumor cell clusters are large and consistent, we hypothesize that variation in an average total mRNA expression among tumor cells may be detected across tumor samples using bulk sequencing. This hypothesis is corroborated by the scRNAseq generated pseudo-bulk data, where ratios of total mRNA expression for tumor versus non-tumor component are great than one in eight out of nine patient samples (**Fig. S1C, Methods and SI**), demonstrating detectable differences between tumor and non-tumor cells. We therefore set out to quantify tumor-specific total mRNA expression as a metric to track tumor phenotype.

Estimating tumor-specific total mRNA expression from bulk sequencing data

Cell type-specific total mRNA expression is not directly measurable using bulk sequencing data because tumor samples typically contain mixtures of tumor and non-tumor cells[4]. However, assuming technical

effects are similar for tumor and non-tumor cells within a sample, we can estimate the ratio of total mRNA expression between these two cellular populations. We formulate a normalized total mRNA expression score (TmS), estimated from the ratio of total tumor-specific mRNA levels over total non-tumor-specific mRNA levels, and incorporating tumour purity and ploidy (**Fig. 2A**, **Methods and SI**). Deconvolution methods (e.g., DeMixT[29]) are needed to derive tumor-specific total mRNA proportions from bulk RNA sequencing data, while purity and ploidy can be estimated from DNA sequencing data (e.g. ASCAT[30] and ABSOLUTE[31]) (**Fig. 2A and Fig. S2A-F, SI**). We developed a profile likelihood-based approach to select top ranking genes that maximized the DeMixT model identifiability of tumor-specific total mRNA proportions (**Methods and SI**). The selected genes form global transcription signature gene sets. Included genes are cancer specific and distributed across the genome (**Fig. S3A**). Across cancer types, 54-68% (mean = 62%) of signature genes are housekeeping or essential genes[41,42] (**Fig. S3B**). Furthermore, signature genes are enriched for genes that can play a role in transcriptional regulation[43] (**Fig. S3C,D, Methods and SI**). Compared to non-selected genes, signature genes are enriched for ATAC-seq peaks within their promoter regions in 279 (90%) of 310 samples across 13 cancer types, in keeping with the known contribution of chromatin accessibility to transcriptional dynamics[44] (**Fig. S3E, Methods and SI**). The cancer-type matched scRNAseq data provides additional evidence in the utility of signature genes to estimate total mRNA expression (Spearman *r* of total signature gene expression *vs.* total UMI counts is between 0.92 and 0.98 across nine patients, **Fig. S3F-H, Methods and SI**).

TmS is associated with prognostic clinicopathologic characteristics

Across all 15 TCGA cancer types where input data for DeMixT were available, we obtained TmS values and found considerable variation (**Fig. 2B, methods**, **SI**), with most cancer types demonstrating a wide TmS range (**Table S2**). TmS values are above 1 in 2,628 out of 5,031 (52%) patient samples (**Fig. 2B**). TmS is dependent upon the background normal reference tissue and cannot be used to make quantitative comparisons across cancer types. Nevertheless, for tumors derived from the same tissue, comparisons can be made between known histopathologic and/or molecular subtypes. Consistent trends are observed between subtypes of head and neck squamous cell carcinoma, breast carcinoma, renal papillary carcinoma, and prostate adenocarcinoma, where prognostically favorable subtypes are enriched in tumors with lower TmS and vice versa (**Fig. 3A-D**). In head and neck squamous cell carcinoma, the prognostically favorable human papillomavirus (HPV)-positive subtype has lower median TmS than the HPV-negative subtype (*P* value = 0.006, **Fig. 3A**). Similarly, triple negative receptor status is associated with higher TmS in breast carcinoma (adjusted *P* value = 4×10^{-36}, **Fig. 3B**), in keeping with this subtype's known propensity for aggressive behavior. Additional associations with molecular features and subtypes in breast cancer are shown in **Fig. 3B**. Subtypes of renal papillary carcinoma also show significant differences in TmS, where the more aggressive Type II tumors have higher TmS compared to Type I (*P*

value = 1 x 10^{-5}, **Fig. 3C**)[45]. In prostate adenocarcinoma, TmS is associated with tumor grade as assessed by the Gleason score, where high TmS tumors are enriched for Gleason scores of 8 and above (*P* value = 0.002, **Fig. 3D**). TmS also correlates with Tumor-Node-Metastasis (TNM) stage in some cancer types, although this relationship is not consistently observed in head and neck, thyroid, breast, colorectal, lung, and liver cancers (**Fig. 4A**). Weak to moderate correlations (up to 0.4) are observed between TmS and proliferation markers, *MKI67* and *PCNA*, across cancers (**Fig. 3E**). There are no correlations between TmS and other clinical characteristics, including age and sex (**Fig. S4B**).

TmS refines prognostication across cancer types

We examined the association of TmS with overall survival (OS) and progression-free interval (PFI) across TCGA (**Methods and SI**). In this pan-cancer analysis, high TmS was associated with reduced OS and PFI compared to low TmS (**Fig. 4A**). We next examined each cancer type in the context of overall TNM stage classification, which is used across cancers for predicting prognosis and treatment decision-making. Analysis stratified by early (I/II) *vs.* advanced (III/IV) stages revealed three different patterns for the differing effects of TmS by stage (**Fig. 4A-E, Fig. S4C-E**). The first group of tumors show consistent effects across stages (**Fig. 4A, Fig. S4c**). This includes thyroid, lung adeno, colorectal, hepatocellular, stomach adeno, and renal clear cell carcinomas, where high TmS is associated with higher risk of death and/or disease progression within both early and late stage tumors. In head and neck squamous cell, lung squamous cell, and bladder urothelial carcinomas, high TmS is associated with reduced survival in early stage tumors only, while for late stage tumors, high TmS is associated with improved survival (**Fig. 4B**, **Fig. S4D**). An opposite pattern is observed in breast and renal papillary carcinomas, where high TmS is associated with poor prognosis in late stage tumors, but improved survival in early stage tumors (**Fig. 4C**, **Fig. S4E)**. TmS remains significantly associated with survival outcomes in Cox regression models, after adjusting for known prognostic characteristics including subtype, stage and age, except in hepatocellular carcinomas, where only a trend was observed (**Fig. 4D**, **Table S3, SI**).

TmS as a prognostic feature in prostate adenocarcinoma

For prostate adenocarcinoma, the Gleason score is a commonly used prognostic marker. Gleason 6 tumors are typically indolent, while more variable outcomes are observed for intermediate (Gleason 7) and high-grade (Gleason 8+) tumors. Survival analyses showed that TmS can further stratify patients within subgroups defined by the Gleason score. At 5 years, 4.3% of Gleason 7 and 34% of Gleason 8+ patients with low TmS progressed, while 13% and 57% of high TmS tumors progressed for Gleason 7 and 8+ patients, respectively (**Fig. 4E**, **Table S4A**, **Methods and SI**).

To validate our findings, we examined an independent cohort of 79 patients with early-onset prostate adenocarcinoma (ICGC-EOPC)[34] (**Methods and SI**). As this cohort contains predominantly Gleason 7

tumors, we are not powered to detect a relationship between TmS and Gleason score. However, similar to TCGA, high TmS is associated with reduced progression-free survival in Gleason 7 and Gleason 8+ tumors (**Fig. 4F, Table S4B**). Although the rate of disease progression is generally higher within this cohort for Gleason 7 tumors compared to those in TCGA (**Fig. 4E-F**, **SI**), low TmS tumors demonstrate a reduction of over 50% in progression rate compared to high TmS tumors (13% versus 36%, **Table S4A**). TmS shows similar independent effect sizes in multivariable Cox regression analyses across Gleason categories (**Table S4B**). Using TCGA as a training set, we built a risk prediction model for disease progression in prostate cancer. Applied to the ICGC-EOPC data, the model demonstrates high discrimination (integrated AUC 0.81; 95%CI: [0.67, 0.91]) and calibration (5-year Integrated Brier Score: 0.19) (**Table S4B, Fig. S4F**). Overall, our findings demonstrate that TmS provides additional prognostic value beyond Gleason score and may be used to refine risk stratification in patients with prostate cancer.

TmS captures cancer-specific genomic dysregulation and hypoxia

We hypothesize that tumor-specific total mRNA expression may be regulated through a plethora of genomic, epigenetic and transcriptomic alterations (**Fig. 5A**). We therefore examined driver mutations (**Fig. 5B, Methods and SI**), which are expected to alter tumor cell phenotypes[46]. Driver mutations (nonsense, missense and splice-site SNVs and indels) in *TP53* are significantly associated with higher TmS in breast, lung, prostate and stomach cancers. Cancer-specific negative correlations with TmS are also identified, including *MAP3K1* and *PIK3CA* driver mutations in breast carcinoma and *RAS* driver mutations in papillary thyroid carcinoma. Expanding the somatic mutation analysis to include all non-synonymous mutations (SNVs and indels) across all genes (33,909 cancer-gene pairs) and using logistic regression models to adjust for covariates such as tumor mutation burden (**Fig. 5C, Methods and SI**), we re-captured the same significant cancer-gene pairs plus one additional pair, negative correlation with *FGFR3* in bladder urothelial carcinoma.

Next, we examined broad-scale genomic alterations, including tumor mutation burden (TMB), chromosomal instability (CIN), whole genome duplication (WGD) and the degree of hypoxia as defined by a previously described gene expression signature[47]. TMB and CIN showed low to moderate correlations with TmS (Pearson r = -0.01 to 0.46, **SI**), suggesting that they may contribute to tumor-specific total mRNA expression in certain cancer types but are not universal determinants. In contrast, a dichotomized hypoxia score is significantly associated with TmS across all 13 cancer types with available data (**Fig. S5D**). Specifically, high TmS is correlated with low hypoxia in head and neck cancers, and with high hypoxia in the remaining 12 cancer types.

To study further contributions to TmS-mediated patient prognosis (**Fig. 5A**), We compared the distributions of each feature, i.e., TMB, CIN, hypoxia score, as well as *TP53* mutation rate (nonsynonymous SNVs and indels) across four patient groups, where prognostic differences were

identified: early stage and low TmS, early stage and high TmS, advanced stage and low TmS, advanced stage and high TmS (**Fig. 5D**, **Methods and SI**). In three cancer types (bladder, ER positive breast cancer, lung adenocarcinoma), all four features differ significantly across patient groups. In contrast, no differences are observed in any feature within triple negative breast cancer. In the remaining cancer types, at least one feature differs across patient groups. Overall, these findings show that the prognostic effect of TmS cannot be explained solely by shared genomic alterations across cancer types, supporting the notion that total tumor mRNA expression levels track a cellular phenotype resulting from a combination of genomic and microenvironmental factors.

Intra- and inter-tumor heterogeneity measured by total mRNA expression

TmS may contribute to intratumor heterogeneity, which serves as a reservoir for tumor evolution, treatment resistance and progression. In the scRNAseq data we identified tumor cell subclusters with different total mRNA expression levels in seven out of nine patients (**Fig. 1B**). TRACERx, a multi-region study of early-stage lung cancer[35], provides an opportunity to further evaluate this phenomenon across spatial regions (**Fig. 6A**). Across 94 regions from 30 patients (2-6 regions per sample), a wide range of TmS values is seen (**Fig. 6B**). Regions with higher percentage of subclonal copy number alterations (top 50%) present higher TmS (adjusted P = 0.004, **Fig. 6C**). Overall TmS shows a significantly higher correlation with ongoing chromosomal instability[48] (Spearman r = 0.44) than static chromosomal instability (difference in r = 0.20, 95% CI: 0.04, 0.37, **Fig. 6D**, **Methods and SI**). Summarized across regions from the same tumor sample, the percentage of subclonal copy number alterations is highly correlated with maximum TmS (TmS_{max}, Spearman r = 0.69), and moderately correlated with the range of TmS, Spearman r = 0.49, **Fig. S6A**). A smaller range of TmS is predictive of linear evolutionary relationship between regions sampled (AUC = 0.83, **Methods and SI**). Variable selection of all measures (**Fig. 6A**) shows that subclonal copy number alterations, range of TmS and the number of regions sampled together can predict values of TmS_{max} (**Fig. 6E, Methods and SI**). Moreover, TmS_{max}, but not TmS_{med}, is associated with the risk of recurrence of death (**Fig. 6F-G**). A high percentage of subclonal copy-number alterations is known to associate with a higher risk of recurrence or death in the TRACERx study (**Fig. S6B**), while adding TmS_{max} allows further discrimination of outcomes (**Fig. 6H, Fig. S6C**). High TmS_{max} remains associated with higher risk of recurrence or death when 22 additional patients with a single sample per tumor are included (**Fig. S6D-E**). In summary, the spatial and evolutionary diversity within tumor samples, as well as patient prognostication, can be captured by intra- and inter-patient TmS heterogeneity, respectively.

Discussion

Here we describe a key RNA feature, the total mRNA expression of tumor cells per haploid genome copy, that depicts clinically relevant phenotypes in cancer. Using a DNAseq and RNAseq joint deconvolution

metric TmS, we demonstrated that distinctive signals in total mRNA expression from tumor cell populations are detectable in bulk sequencing data from single- and multi-region tumor samples. Association of TmS with genomic features and hypoxia in TCGA suggests a complex relationship that is cancer specific and highly dependent upon additional contexts such as tumor subtype and stage. This is in keeping with our hypothesis that total mRNA expression level is a key characteristic of the dynamically changing tumor cellular phenotype, influenced by various molecular events that drive tumor development.

Regulation of total mRNA expression in tumor cells is currently not well understood. *MYC* dysregulation and aneuploidy can perturb cancer transcriptomes at scale though likely other mechanisms play a role as well[9,49]. We found that TmS correlates with different genetic alterations, including driver mutation, mutation burden and chromosomal instability, as a function of tumor type, suggesting that total mRNA expression is not governed by a single mechanism. This is in keeping with the unique complement of genetic alterations required for oncogenic gene expression dysregulation in distinct tissues[50]. Across cancers, high TmS was most frequently associated with hypoxia. Interestingly, Choudhry et al. showed that hypoxia leads to upregulated mRNA levels in breast cancer cells by releasing promoter-paused RNApol2[51]. Our observations illustrate the potential joint influence of genetic and microenvironmental factors in shaping tumor cell total mRNA content. Through TmS, tumor-specific total mRNA expression can now be quantified at scale, allowing for the discovery of additional mechanisms upstream as well as downstream.

While high TmS is generally associated with aggressive disease, clinical context remains important to evaluate its prognostic implications, as the direction of the prognostic effect inverted by stage in five out of thirteen cancer types. Given that early and advanced stage tumors are often treated using distinct modalities, this effect may in part be underpinned by a differential response of tumors with low *vs.* high total mRNA expression to treatment conditions. Indeed, transcriptional amplification in the context of *MYC* dysregulation has been linked to increased sensitivity to chemotherapeutic agents[52–55]. Additional studies incorporating data from clinical trials will be needed to elucidate how stage-specific and treatment-related factors interact with TmS to determine patient outcome.

Total mRNA expression variability has implications for routine differential expression analysis, where both total and relative abundance[4] of gene expression should be assessed[56]. Conceptually, analogous to DNA ploidy measuring the total DNA content per chromosomal copy, the total mRNA content per chromosomal copy can be considered the “ploidy of the transcriptome”, which is a key parameter hitherto hidden in most RNA-based assays. While our current work focuses on mRNA, the concepts developed here can readily be applied to the quantification of other RNA species (i.e. rRNA, miRNA, piRNA etc), further illuminating the cancer transcriptome. Enhanced attention to “transcriptome ploidy” will likely enable

better phenotypic characterization and a deeper biological understanding of transcriptional dysregulation in cancer and other diseases.

Methods

A detailed description of the methods used in this paper and additional results are described in **Supplementary Information**. Here, we provide a summary of the methods and analysis.

Total mRNA expression in single-cell RNA sequencing data

The single-cell RNA sequencing datasets generated from nine patients (**Table S1**) were preprocessed in a uniform fashion, including quality control, cell clustering, cell type annotation, and tumor cell identification. Cell type was annotated using known marker genes[32,33,57–59]. Tumor cells were identified based on the inferred presence of somatic copy number alterations by inferCNV[60]. Within each cell type, we further merged Seurat[61] identified clusters that are not significantly different in gene counts (Wilcoxon rank-sum test, α=0.001).

To enable comparison between different scRNAseq datasets in each cancer type, we performed scale normalization of the raw UMI counts, ensuring total UMI counts per cell across all cells are the same for different samples from the same study. Specifically, let $UMI_i = \{UMI_{igc}\}_{GxCi}$ be a matrix of raw UMI counts for the scRNAseq data for sample i being investigated, with genes g on the rows and cells c on the columns. G denotes the total number of genes; C_i is the number of cells in sample i. Then, the normalized UMI matrix UMI_i, denoted as UMI_i^{norm}, is calculated as $UMI_i^{norm} = UMI_i / r_i$, where, $r_i = \frac{UMI_i^{sum}/C_i}{baseline}$, $baseline = median\{UMI_1^{sum}/C_1, UMI_2^{sum}/C_2, \ldots, UMI_n^{sum}/C_n\}$, $UMI_i^{sum} = \sum_{c=1}^{C_i} \sum_{g=1}^{G} UMI_{igc}$.

Given a cell cluster, we let u_{gc} denote the amount of mRNA of gene g in cell c. The average total mRNA amount per cell is $\sum_{c=1}^{C} (\sum_{g=1}^{G} u_{gc})/C$. For scRNAseq data, we assume the UMI_{gc} from gene g, cell c is proportional to the total mRNA u_{gc} of gene g in that cell, with a constant k_g that represents technical effects: $UMI_{gc} = k_g * u_{gc}$. The constant k_g is introduced because every single-cell sequencing platform presents a <100% capture efficiency for mRNA, and such efficiency varies across different platforms[62]. Under the assumption that the technical effect k_g remains constant across cells and is often evaluated as an average effect across genes within the same platform, we can evaluate total mRNA expression in the scRNAseq data using the average total UMI counts, which is $\sum_{c=1}^{C} (\sum_{g=1}^{G} UMI_{gc})/C$. Average total UMI counts serve as a reasonable surrogate to compare total mRNA expression across cells that are generated from the same experiment.

We compared the distributions of total UMI counts between tumor cell clusters from the same patient sample using Wilcoxon rank-sum tests and corrected for multiple testing using Benjamini-Hochberg[63] (BH) method. We further examined cell cycle state of each tumor cell cluster using Seurat.

A full description is provided in **Section 1** in **Supplementary Information**.

A mathematical model for tumor-specific total mRNA expression estimation

Model. We have developed a tumor-specific total mRNA expression score (TmS) to estimate the ratio of total tumor mRNA expression per haploid genome to that of the surrounding non-tumor cells from bulk sequencing data (**Fig. 2A**). Let $T_g=\sum_{c=1}^{C_T} u_{gc}$ and $N_g=\sum_{c=1}^{C_N} u_{gc}$ denote the total number of transcripts of gene g across all tumor cells and non-tumor cells, C_T and C_N denote the number of tumor and non-tumor cells, and ψ_T and ψ_N denote the average ploidy of tumor and non-tumor cells, respectively. Under the assumption that the tumor cells have a similar ploidy, we can derive TmS without using single-cell-specific parameters as

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

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

$$TmS = \frac{\psi_N\pi(1-\rho)}{\psi_T\rho(1-\pi)} \qquad \text{Eq. 2}$$

The tumor-specific mRNA expression proportion π derived from the tumor can be estimated using DeMixT[29] as $\hat{\pi}$; the tumor purity ρ and ploidy ψ_T can be estimated using ASCAT[30], ABSOLUTE[31] or Sequenza[64] based on the matched DNA sequencing data as $\hat{\rho}$ and $\hat{\psi}_T$, respectively; the ploidy of non-tumor cells ψ_N was assumed to be 2[30,31]. Hence, we have

$$\widehat{TmS} = \frac{2\hat{\pi}(1-\hat{\rho})}{\hat{\psi}_T\hat{\rho}(1-\hat{\pi})} \qquad \text{Eq. 3}$$

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

Improved estimation using DeMixT. Many computational deconvolution methods have been developed to estimate the cell type proportions through transcriptome data; however, most of them focus on the cellular proportion and not the global gene expression level of each cell type, due to lack of appropriate normalization approaches. The DeMixT[29] model is unique in aiming to estimate the global tumor-specific

gene expression level relative to the normal reference in the context of admixed tumor samples. ISOpure[28] is the other model that presents similar objectives as the DeMixT model.

The identifiability analysis 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 global gene expression proportions π are the most identifiable parameters, requiring only a subset of genes with identifiable expression distributions. Therefore, our goal is to select an appropriate set of genes as input to DeMixT that optimizes the estimation of the tumor-specific mRNA expression proportions. In general, genes are expressed at different levels, which, due to different numerical ranges, can affect tumor-specific global gene expression proportion estimation. We found that including genes that are not differentially expressed between the tumor and non-tumor components within the bulk sample, or genes with large variance in expression within the non-tumor component, can introduce large biases into the estimated tumor-specific mRNA expression proportions. By applying a profile likelihood approach to detect the identifiability of model parameters[65], we systematically evaluated the identifiability for all available genes based on the data, and selected the most identifiable genes for the estimation of proportions. As a result, the accuracy of the estimated proportions has been improved. As a general method, the profile likelihood-based gene selection strategy can be extended to any method that uses maximum likelihood estimation.

Briefly, in the DeMixT model, for sample i and across any gene g, we have

$$Y_{ig}=\pi_i T'_{ig}+(1-\pi_i)N'_{ig} \qquad \text{Eq. 4}$$

where Y_{ig} represents the scale normalized expression matrix from mixed tumor samples, T'_{ig} and N'_{ig} represent the normalized relative expression of gene g within tumor and surrounding non-tumor cells, respectively. The estimated tumor-specific total mRNA expression proportions $\hat{\pi}$ is the desirable quantity for Eq.3. We assume each hidden component follows the $\log_2$-normal distribution, i.e., $T'_{ig} \sim LN\left(\mu_{Tg}, \sigma^2_{Tg}\right)$ and $N'_{ig} \sim LN\left(\mu_{Ng}, \sigma^2_{Ng}\right)$. We will use notation T and N and drop ' from now on. The identifiability of a gene k in the DeMixT model is measured by the confidence interval $[\mu^-_{Tk}, \mu^+_{Tk}]$ of its mean expression μ_{Tk}. The definition of the profile likelihood function of μ_{Tk} is

$$l_{\mu_{Tk}}(\mu_{Tk}=x|\boldsymbol{\pi},\boldsymbol{\mu_T},\boldsymbol{\sigma_T}) = max\{\sum_{i=1}^{S}[\sum_{g\neq k}^{G} log\left(f\left(\pi_i,\mu_{Tg},\sigma_{Tg}\right)\right) + log\left(f\left(\pi_i,\mu_{Tk}=x,\sigma_{Tk}\right)\right)]\} \qquad \text{Eq. 5}$$

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

The confidence interval of a profile likelihood function can be constructed through inverting a likelihood-ratio test[66]. However, in real data analysis, the actual profile likelihood function of μ_{Tk} is intractable and prone to noise; calculating the actual profile likelihood function of all genes (~20,000) is generally infeasible due to computational limits. An asymptotic approximation was adopted to quickly evaluate the profile likelihood function. If the measurement noise is small and the sample size is large enough, asymptotic confidence intervals are good approximations of the actual confidence intervals[65]. The asymptotic profile likelihood function can be derived from the observed Fisher information of the log likelihood, denoted as $H(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}})$. Then the asymptotic α level confidence interval of μ_{Tk} can be written as[65]

$$\mu_{Tk}^{\pm} = \widehat{\mu_{Tk}} \pm \sqrt{2\chi^2_{1-\alpha}(1)\, H\left(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}_{\boldsymbol{T}},\hat{\boldsymbol{\sigma}}_{\boldsymbol{T}}\right)^{-1}_{k,k}} \qquad \text{Eq. 6}$$

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

$$\textit{gene selection score}_k = 2\sqrt{2\chi^2_{1-\alpha}(1)\, H\left(\hat{\pi},\hat{\mu}_T,\hat{\sigma}_T\right)^{-1}_{k,k}} \qquad \text{Eq. 7}$$

Genes with a lower score have a smaller confidence interval, 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 and a subset of genes top-ranked are used to improve the estimation of tumor-specific mRNA expression proportions.

We further use virtual spike-ins to improve the estimation of tumor-specific mRNA expression proportions. We observed an imbalance of true proportions, e.g., median proportion > 0.5, can result in biases in proportion estimation that cannot remedied by gene selection. The virtual spike-ins are generated based on expression profiles from normal reference samples and added to the tumor samples such that there are roughly the same number of samples with tumor proportions below and above 50%.

A full description is provided in **Section 2.2** in **Supplementary Information**.

Datasets of bulk sequencing data from patient samples

TCGA. Raw read counts of high throughput RNA 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, uterine corpus endometrial carcinoma) were downloaded from the Genomic Data Commons Data Portal

(https://portal.gdc.cancer.gov/). ATAC-seq data[44], tumor purity and ploidy data[67,68], and annotations of driver mutation and indels[46] were downloaded for these samples. A CONSORT diagram is provided for the dataset (**Fig. S7A,B**).

ICGC-EOPC. In this cohort, matched RNA sequencing data and whole genome sequencing data, as well as clinical data including biochemical recurrence, Gleason score and pathologic stage, from 121 tumor regions and 9 adjacent normal samples from 96 patients (age at treatment < 55), were downloaded from Gerhauser et al.[34] A CONSORT diagram is provided for the dataset (**Fig. S7C**).

TRACERx. A total of 159 tumor samples from 64 patients with matched RNA sequencing data and whole exome sequencing data were obtained from the TRACERx cohort[35,36]. Tumor purity and ploidy estimates for these samples were determined by Sequenza[64]. Clinical information of progression free survival and per region segmented copy number data were downloaded from Jamal-Hanjani et al[35]. A CONSORT diagram is provided for this dataset to demonstrate the filtering steps (**Fig. S7D**).

GTEx. RNA sequencing data from 42 normal prostate samples, 67 normal thyroid samples and 20 normal lung samples without significant pathology in the corresponding tissue types were downloaded from the GTEx Data Portal (https://www.gtexportal.org/home)[69].

A detailed description of above datasets is available in **SI, Section 2.3.1**.

Tumor-specific total mRNA expression in TCGA

Estimation of tumor-specific mRNA expression proportions. For each cancer type, we filtered out poor quality tumor and normal samples using a hierarchical clustering model based on the top 1,000 differentially expressed genes selected from the two types of samples. We then selected available adjacent normal samples as reference for the tumor deconvolution using DeMixT (**SI, Section 2.3.2**). Based on a simulation study and observed distributions of gene selection scores in real data (**SI, Section 2.2.2**), we chose the top 1,500 or 2,500 genes (varies across cancer types) to estimate tumor-specific mRNA expression proportions.

Consensus of TmS estimation. It is possible for DNA based deconvolution methods ASCAT and ABSOLUTE to provide different tumor purity ρ and ploidy ψ pairs for the same sample. These typically differ by a whole-genome duplication event. To calculate one final set of TmS values for a maximum number of samples, we took a consensus strategy. We first calculated TmS values for TCGA samples with tumor purity and ploidy estimates derived from both ABSOLUTE and ASCAT, and then fitted a linear regression model on the $\log_2$-transformed TmS values calculated with ASCAT by using the $\log_2$-transformed TmS values calculated with ABSOLUTE as a predictor variable. We removed 264 samples with a Cook's distance $\geq 4/n$ (n is the number of total samples) and calculated the final

$TmS=2^{(log_2(TmS_{ASCAT})+log_2(TmS_{ABSOLUTE}))/2}$ (**Figs. S2F and 7A**). See a full description in **Section 2.3.2.1** in **Supplementary Information**.

Global transcription signature genes. For each of the 15 cancer types from TCGA, we conducted gene set enrichment analyses on Hallmark and KEGG pathways[43] for all the available genes ordered by their gene selection scores calculated by DeMixT using GSEA[43] and g:Profiler[70]. We combined the outputs of GSEA and g:Profiler and only the pathways with adjusted *P* value < 0.05 from both GSEA and g:Profiler were considered as significantly enriched. We also used GeneMANIA[71] to identify functional pathways enriched in the overlap between individual signature gene sets across cancer types.

The 75th percentile of normalized peak scores across all peaks within the promoter region was selected for each gene as its peak score, and genes with normalized peak scores less than 1 were excluded. For each sample, we calculated the mean of the peak scores of all signature genes. A null distribution of mean peak scores was generated by calculating means from 1,000 random subsets of genes with the matching number of the signature genes from all genes. *P* values assessing the significance of the deviation of the observed mean score for signature genes from the null distribution were calculated as the percentile of the permuted means being greater than or equal to the observed mean. Within cancer types, *P* values were adjusted using the BH method.

A full description is provided in **Section 2.3.2.2** in **Supplementary Information**.

Validation using scRNAseq data. Using scRNAseq data from four matching cancer types, we compared the expression levels of signature genes in tumor versus non-tumor cells within each patient and those in tumor cells across patient samples. We also made pseudo-bulk data to compare with gene expression in bulk data (**Section 2.3.2.3** in **Supplementary Information**).

Statistical analysis. Kruskal-Wallis tests were used to compare the distribution of TmS between subgroups defined by each clinical variable. The *P* values from Kruskal-Wallis tests were adjusted using BH correction across all available clinical variables within the corresponding cancer type.

We evaluated the association between TmS and survival outcome (overall survival and progression free interval) across 15 cancer types in TCGA. To ensure sufficient sample size in each group, we summarized pathologic stages into two categories: early (I/II) and advanced (III/IV). For prostate cancer, we used Gleason score (Gleason Score = 7 versus 8+) instead of early and advanced stages. We used a recursive partitioning survival tree model, rpart[72], to find the optimal TmS cutoff 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 pathological stages were performed. We then fitted multivariate

Cox Proportional Hazard models with age, TmS, stage, and an interaction term of TmS and Stage (TmS x stage) as predictors of overall survival and progression free interval for each TCGA cancer type.

We compared associations with survival outcomes of the three metrics that follow: TmS x ploidy = ploidy-unadjusted TmS, in four cancer types: head and neck squamous cell carcinoma, lung squamous cell carcinoma, renal clear cell carcinoma, and ER-positive breast carcinoma.

For each cancer type within TCGA, we considered genes which had driver mutations (including nonsense, missense and splice-site SNVs and indels) in at least 10 samples. For each of these genes, samples were labelled as "driver mutant" if they carried at least one driver mutation in that gene or "wild type" otherwise[46]. We also implemented an agnostic search over all genes for the 15 cancer types to identify among non-silent mutations (including SNVs and indels), those that were significantly associated with TmS. We applied two statistical tests to evaluate the difference between the "mutant" and "wild type" samples. We fitted a linear regression model using $\log_2$-transformed TmS as the dependent variable and mutation status as a predictor: $log_2(TmS)=b_0+b_1\, log_2(TMB)+b_2 MUT$, where TMB represents tumor mutation burden. $MUT=1$ if the sample has at least one mutation in the candidate gene, and $MUT=0$ otherwise. The *P* values were calculated by a t-test of the regression coefficient b_2. The *P* values of each gene based on Wilcoxon rank-sum test and t-test were adjusted by BH correction based on the number of candidate genes within the corresponding cancer type.

A full description is provided in **Section 2.3.2.4** in **Supplementary Information**.

Tumor-specific total mRNA expression in ICGC-EOPC

We applied rpart[72] to iteratively partition samples by TmS and the Gleason score (Gleason Score = 7 versus 8+). We fitted multivariate Cox Proportional Hazard models with age, TmS, stage, and the interaction term of TmS and Gleason score (TmS x Gleason score) as predictors for progression free interval analysis of TCGA prostate adenocarcinoma samples**.**

As an external validation for risk prediction with TmS, we evaluated risk prediction models trained on TCGA prostate adenocarcinoma samples and predicted disease progression risk for patients from the EOPC study. We compared the prediction performance between a baseline model, containing only age and Gleason score as covariates, and the "age and TmS x Gleason score model" (termed "TmS" model), consisting of age, TmS, Gleason score and TmS x Gleason score as covariates. We evaluated the discriminatory ability of the TmS model using Uno's estimator of cumulative AUC (iAUC)[73] and constructed 95% confidence intervals by bootstrap resampling with 1,000 times. To measure the calibration ability of the TmS model, we also calculated the 5-year IBS (Integrated Brier Score)[74].

A full description is provided in **Section 2.3.3** in **Supplementary Information**.

Tumor-specific total mRNA expression in TRACERx

Association of regional TmS with measures of chromosomal instability. We first calculated the TmS of each region for patients in TRACERx. We then calculated the percentage of copy number alteration (CNA) burden (percentage of genome affected by CNAs) per region, the percentage of subclonal CNA per region, and the percentage of subclonal CNA per patient as measures of chromosomal instability. Here, a subclonal CNA is defined as a CNA only existing in a subset of regions. A full description is provided in **Section 2.3.4.** in **Supplementary Information**.

We defined the evolutionary relationship in two regions from the same patient as either linear or branching, and for each evolutionary relationship per patient, we defined the *range of TmS* to be $log_2(TmS_{max})$ - $log_2(TmS_{min})$ across regions. We fitted linear regression models by taking $log_2(TmS_{max})$ as the response variable. The predictor variables including percentage of subclonal CNA, number of regions, range of TmS, evolutionary relationship and their interactions as predictors. The best model was selected by stepwise adding or dropping one predictor that achieves the best AIC (Akaike's Information Criteria) (**Fig. 6E**).

We applied rpart[72] to partition 30 patients with multiple-region samples into two groups using TmS_{max} as the variable (**Fig. 6F**). As a negative control, we ranked the patients by TmS_{med} (median TmS across regions for each patient), and assigned the same number of patients as TmS_{max} into two groups (**Fig. 6G**). As the percentage of subclonal CNA was shown to be highly associated with survival outcomes in TRACERx[35], we used both TmS_{max} and percentage of subclonal CNA as variables in rpart, and separate the 30 patients into groups (**Fig. 6H**). As a negative control, we separated the patients into groups by sorting TmS_{med} and percentage of subclonal CNA successively into groups with the same number of patients as TmS_{max} (**Fig. S6C**). Log-rank tests comparing survival outcomes between groups were performed.

A full description is provided in **Section 2.3.4** in **Supplementary Information**.

Data availability

Count matrices of the hepatocellular carcinoma single cell RNA sequencing data were downloaded from the Gene Expression Omnibus (GEO) with the accession code GSE125449. The raw read counts and cell type annotations of the lung adenocarcinoma single cell RNA sequencing data were downloaded from the ArrayExpress under accessions E-MTAB-6149. Raw read counts of high throughput 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.

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

Raw read counts of RNAseq data in GTEx were downloaded from https://www.gtexportal.org/home.

All other relevant data are available from the corresponding author upon reasonable request.

Code availability

All code used for analyses was written in R version 3.6.1 and will be made available. The core computational pipelines developed for estimating tumor-specific mRNA expression proportion are available in R package DeMixT1.4.0, which can be downloaded from https://www.bioconductor.org/packages/release/bioc/html/DeMixT.html.

Funding Acknowledgments:

S.C. is supported by the Norman Jaffe Professorship in Pediatrics Endowment Fund, MD Anderson Colorectal Cancer Moon Shot Program, and NIH R01CA183793. J.R.W. is supported by American Thyroid Association/ThyCa grant, Mark Foundation for Cancer Research ASPIRE award. S.J. is supported by Human Cell Atlas Seed Network - Breast, Chan Zuckerberg Institute, MD Anderson Colorectal Cancer Moon Shot Program, NIH R01CA183793. Peng Yang is supported by NIH R01. J.C. is supported by NIH R01CA158113. P.V.L. and J.D. are supported in part by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). They are also supported in part by the Medical Research Council (grant number MR/L016311/1). J.D. is supported in part by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 703594-DECODE) and the Research Foundation – Flanders (FWO, Grant No. 12J6916N). J.P.S. is supported by the Cancer Prevention and Research Institute of Texas as a CPRIT Scholar in Cancer Research and by National Institutes of Health (K22CA234406). J.J.L. is supported by the NIH (T32CA009599). 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), Kadoorie Charitable Foundation, NIH/NCI (U01 CA224044-01A1, 1R01CA231465-01A1). B.L. is supported by the Single cell transcriptome of IBC cells and surrounding microenvironment, SWOG HOPE foundation, Human Cell Atlas Seed Network - Breast, Chan Zuckerberg Institute. P.C.B. is supported by the NIH/NCI under awards number P30CA016042, 1U01CA214194-01 and 1U24CA248265-01. A.U. is supported by the Norwegian Cancer Society (grant number 198016-2018). J.Z. is supported by the MD Anderson Physician Scientist Program, the MD Anderson Lung Cancer Moon Shot Program and the Cancer Prevention and Research Institute of Texas 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 the National Institutes of Health (NIH U01CA196403, U01CA200468, U24CA224020, P50CA221707). W.W. is supported by Human Cell Atlas Seed Network - Retina, Chan Zuckerberg Institute, NIH R01CA183793, NIH R01CA239342, NIH R01CA158113, P30CA016672.
This study makes use of data generated by The TRAcking Non-small Cell Lung Cancer Evolution Through Therapy (Rx) (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 Cancer Research UK and coordinated through the Cancer Research UK and UCL Cancer Trials Centre.

Competing interests:

A.M. receives royalties for a pancreatic cancer biomarker test from Cosmos Wisdom Biotechnology, and this financial relationship is managed and monitored by the UTMDACC Conflict of Interest Committee.

A.M. is also listed as an inventor on a patent that has been licensed by Johns Hopkins University to Thrive Earlier Detection. J.Z. reports research funding from Merck, Johnson and Johnson, and consultant fees from BMS, Johnson and Johnson, AstraZeneca, Geneplus, OrigMed, Innovent outside the submitted work.

References

1. Coate, J. E. & Doyle, J. J. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: An example from a plant allopolyploid. *Genome Biol. Evol.* **2**, 534–546 (2010).
2. Aanes, H. et al. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res.* **21**, 1328–1338 (2011).
3. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
4. Lovén, J. et al. Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).
5. Miettinen, T. P. et al. Identification of transcriptional and metabolic programs related to mammalian cell size. *Curr. Biol.* **24**, 598–608 (2014).
6. Padovan-Merhar, O. et al. Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Mol. Cell* **58**, 339–352 (2015).
7. Marinov, G. K. et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).
8. Coate, J. E. & Doyle, J. J. Variation in transcriptome size: are we getting the message? *Chromosoma* **124**, 27–43 (2015).
9. Lin, C. Y. et al. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
10. Nie, Z. et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**, 68–79 (2012).
11. Macaulay, I. C. et al. G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
12. 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).
13. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
14. Strzyz, P. Cancer biology: TGFβ and EMT as double agents. *Nat. Rev. Mol. Cell Biol.* **17**, 202–203 (2016).

15. 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).
16. Hoxhaj, G. & Manning, B. D. The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism. *Nat. Rev. Cancer* **20**, 74–88 (2020).
17. Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 15149–15154 (2001).
18. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
19. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
20. Athanasiadis, E. I. et al. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat. Commun.* **8**, 1–11 (2017).
21. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
22. Torre, E. et al. Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. *Cell Syst.* **6**, 171–179 (2018).
23. Wang, J. et al. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6437–E6446 (2018).
24. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 31–36 (2001).
25. 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).
26. Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
27. Ahn, J. et al. DeMix: Deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* **29**, 1865–1871 (2013).
28. Quon, G. et al. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.* **5**, 29 (2013).
29. Wang, Z. et al. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, 451–460 (2018).
30. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
31. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
32. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat.*

Med. **24**, 1277–1289 (2018).

33. Ma, L. *et al.* Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. *Cancer Cell* **36**, 418–430 (2019).
34. Gerhauser, C. et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell* **34**, 996–1011 (2018).
35. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
36. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
37. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
38. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
39. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
40. Kim, C. et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e13 (2018).
41. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
42. Dempster, J. M. et al. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nat. Commun.* **10**, 1–14 (2019).
43. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
44. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
45. Zhang, J. et al. International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).
46. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
47. Bhandari, V. et al. Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).
48. Watkins, T. B. K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* (2020). doi:10.1038/s41586-020-2698-6.
49. Sheltzer, J. M., Torres, E. M., Dunham, M. J. & Amon, A. Transcriptional consequences of

aneuploidy. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 12644–12649 (2012).

50. Haigis, K. M., Cichowski, K. & Elledge, S. J. Tissue-specificity in cancer: The rule, not the exception. *Science.* **363**, 1150–1151 (2019).
51. Choudhry, H. et al. Extensive regulation of the non-coding transcriptome by hypoxia: Role of HIF in releasing paused RNApol2. *EMBO Rep.* **15**, 70–76 (2014).
52. Liu, T. et al. MYC predetermines the sensitivity of gastrointestinal cancer to antifolate drugs through regulating TYMS transcription. *EBioMedicine* **48**, 289–300 (2019).
53. Chen, R. et al. Although c-MYC contributes to tamoxifen resistance, it improves cisplatin sensitivity in ER-positive breast cancer. *Int. J. Oncol.* **56**, 932–944 (2020).
54. Arango, D., Corner, G. A., Wadler, S., Catalano, P. J. & Augenlicht, L. H. C-myc/p53 interaction determines sensitivity of human colon carcinoma cells to 5-fluorouracil in vitro and in vivo. *Cancer Res.* **61**, 4910–4915 (2001).
55. Pereira, C. B. L. et al. MYC Amplification as a Predictive Factor of Complete Pathologic Response to Docetaxel-based Neoadjuvant Chemotherapy for Breast Cancer. *Clin. Breast Cancer* **17**, 188–194 (2017).
56. Cai, H. et al. Identification and characterization of genes with absolute mRNA abundances changes in tumor cells with varied transcriptome sizes. *BMC Genomics* **20**, 134 (2019).
57. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
58. 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).
59. Hashimoto, K. et al. Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24242–24251 (2019).
60. 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).
61. 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).
62. Wang, Y. J. et al. Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues. *bioRxiv* 541433 (2019). doi:10.1101/541433
63. Benjamini, Y. & Hochberg, Y. Controlling for the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
64. Favero, F. et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
65. Raue, A. et al. Structural and practical identifiability analysis of partially observed dynamical

models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).

66. Venzon, D. J. & Moolgavkar, S. H. A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Appl. Stat.* **37**, 87–94 (1988).
67. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
68. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
69. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
70. Reimand, J. et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
71. Franz, M. et al. GeneMANIA update 2018. *Nucleic Acids Res.* **46**, W60–W64 (2018).
72. Therneau, T. M. & Atkinson, E. J. An Introduction to Recursive Partitioning Using the RPART Routines. *Mayo Found. Tech. Rep.* **61**, 452 (1997).
73. 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).
74. Mogensen, U. B., Ishwaran, H. & Gerds, T. A. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *J. Stat. Softw.* **50**, 1–23 (2012).

Figure legends:

Fig. 1. High diversity of total mRNA expression in tumor cells. (**A**) Total gene expression in tumor cells compared to non-tumor cells (epithelial, stromal, immune cells). 100 cells of each cell type were randomly selected from a patient with Stage IIIB lung adenocarcinoma. In each heatmap, expressed genes (UMI count > 0) are shown in black, and non-expressed genes (UMI count = 0) in gray. Cells (rows) and genes (columns) are ordered from high to low total number of expressed genes and number of expressing cells, respectively. Bar plots show the total number of expressed genes, i.e., gene counts, and total UMI counts in the corresponding cells. (**B**) Distributions of gene counts and total UMI counts by cell type in scRNAseq data from nine patients with colorectal, hepatocellular, lung and pancreatic cancers. Patients are ordered by pathological stage or survival outcome. The top x-axis annotates total UMI counts (means and 95% Confidence Intervals, CIs). The bottom x-axis annotates gene count distribution (density). Density curves are colored for tumor cells and shown in gray-scale for non-tumor cells. Clusters with higher gene counts are shown in darker shades. The numbers in the parentheses indicate the number of cells analyzed.

Fig. 2. Estimation of tumor-specific total mRNA expression in bulk sequencing data. (**A**) The definition of TmS and its analysis pipeline using matched DNAseq and RNAseq data. (**B**) Distributions of TmS in 5,031 tumor samples across 15 cancer types in TCGA. The number of patient samples for each cancer type is indicated on the top of each violin plot.

Fig. 3. TmS is associated with known prognostic characteristics. (**A-D**) Clinicopathologic annotations are shown for (A) head & neck squamous cell carcinoma, (B) breast carcinoma, (C) renal papillary carcinoma, and (D) prostate adenocarcinoma samples. Tumor samples are ordered by TmS from low to high. The Benjamini–Hochberg adjusted *P* values for Kruskal-Wallis tests comparing TmS between clinicopathologic subgroups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). For breast carcinoma, triple-negative breast cancers (TNBC) are shown on the second row. *MYC*/*PVT1* copy number alterations are shown on the sixth row, where “Gain” indicates that either *MYC* or *PVT1* was amplified and “Neutral” indicates that no copy number alternations were detected. (**E**) Spearman correlation coefficients between TmS and the expression levels of *MKI67* and *PCNA* across 15 cancer types.

Fig. 4. TmS refines prognostication on pathological stages. (**A**) Kaplan-Meier (KM) curves of overall survival (OS) and progression-free interval (PFI) for TCGA samples. Gray lines denote summary KM curves of patients with high versus low TmS across all cancer types. KM curves are further grouped by

TmS and pathological stage into four groups. *P* values of log-rank tests between high versus low TmS groups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). Three patterns of associations were identified (summarized in D). Pattern I is the most prevalent and shown in A. (**B-C**) KM curves for representative cancers (top and bottom panels) in Pattern II (B) and Pattern III (C). (**D**) Forest plot of hazard ratios and 95% of CIs of multivariate Cox proportional hazard models with Age, TmS (High vs. Low), Stage (Advanced vs. Early) and TmS x Stage as predictors, and with OS or PFI as response variable (See detail in **table S3**). (**E**) KM survival curves of PFI for the TCGA prostate adenocarcinoma study cohort. (**F**) KM survival curves of PFI for the ICGC-EOPC study cohort. In E & F, patients are grouped by Gleason score and TmS.

Fig. 5. TmS captures cancer-specific dysregulations of genomic features and hypoxia. (**A**) Illustration of a potential role for total mRNA expression as a mediator in the pathway between genomic alterations, epigenetic events and prognosis. Pathologic stage may be combined with TmS to define a clinically informative phenotype of tumor samples. The solid circles indicate that the corresponding data are evaluated. (**B**) Distributions of TmS for TCGA samples with or without driver mutations across the most frequently mutated genes across cancer types. (**C**) Volcano plot showing $\log_2$-fold change in TmS for samples with non-silent mutations in a given gene vs those without. Cancer-gene pairs with adjusted *P* values < 0.05 are highlighted in red points. (**D**) Heatmaps of median 5-year survival, tumor mutation burden, chromosomal instability, hypoxia, and *TP53* non-silent somatic mutation rate across patient groups and cancer types. Patient groups are defined as in Fig. 4D-E, with cancer types following three distinct patterns in survival outcome. Adjusted *P* values are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001).

Fig. 6. Regional 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 2 regions sampled. Blue triangles denote the maximum TmS for a patient. Blue "-" denote the median TmS for a patient. (**C**) Pairwise scatter plots and histograms of % CNA, % subclonal CNA, and TmS per region. Spearman correlation coefficient *r*'s are shown. Different colors annotate three randomly assigned patient groups. The gray lines represent a loess fit. (**D**) Distributions of TmS per each region with high or low % CNA burden per region (left), % subclonal CNA per region (right), with regions grouped at the median values. Adjusted *P* values of Wilcoxon rank-sum tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). (**E**) Scatter plot showing TmS_{max} versus the percent subclonal CNA. A regression fit and 95%CI is shown. (**F-G**) KM survival curves of disease-free probability stratified by TmS_{max} or TmS_{med}. (**H**) KM survival curves of disease-free probability stratified by both TmS_{max} and % subclonal CNA.

Fig. S1. Using gene counts and total UMI counts to measure the global gene expression heterogeneity in tumor cells. (**A**) Illustration of expressed genes in tumor cells compared to non-tumor cells (epithelial, stromal, and immune cells), randomly selected from each patient sample. 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 expressions of each gene, respectively. The barplots provide the distributions of total number of expressed genes, i.e., gene counts, and total UMI counts in the corresponding cells. (**B**) Smoothed scatter plots show the correlations between gene counts and total UMI counts in cell clusters from each patient sample. Patient samples are arranged in the same order of cell clusters as (A). In each smoothed scatter plot, the Spearman correlation coefficient is labeled on the top (r). (**C**) Ratios of mean total UMI counts of tumor cells to non-tumor cells and 95% confidence intervals in pseudo-bulk data, which are made by pooling scRNAseq data.

Fig. S2. Consensus estimation of TmS from matched RNA- and DNAseq data in TCGA. (**A**) Distributions of tumor-specific mRNA expression proportions estimated by DeMixT across cancer types. (**B-C**)**,** Distributions of tumor cell proportions estimated by (B) ASCAT or (C) ABSOLUTE across cancer types. (**D**) Smoothed scatter plot of tumor ploidy estimates from ABSOLUTE versus ASCAT across all samples. Gray points correspond to 997 samples that presented inconsistent tumor ploidy (and purity) estimates between the two methods. (**E**) TmS estimates using either ABSOLUTE or ASCAT-derived purity and ploidy estimates with or without ploidy adjustment for the 977 discordant samples from (D). Blue and gray points correspond to, respectively, TmS prior to and after ploidy adjustment. Ploidy adjustment improved consistency between the ABSOLUTE and ASCAT results. (**F**) 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 ≥ $4/n$, 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 deemed consistent and robust across two DNAseq deconvolution methods.

Fig. S3. Global transcription signature genes across cancer types. (**A**) Karyotype plots showing the genomic locations of signature genes for each cancer type. Signature genes are presented as dots colored by cancer type. An overall gene density track is shown in gray shades underneath the dots. The density of signature genes is consistent with the overall gene density. (**B**) Proportions of global transcription signature genes in five gene categories across 15 cancer types. (**C**) Heatmap of normalized

enrichment scores of enriched Hallmark pathways across 15 cancer types. (**D**) Heatmap of normalized enrichment scores of enriched KEGG pathways across 15 cancer types. For C & D, significantly enriched pathways are those with an adjusted *P* values <0.05 from both GSEA and g:Profiler. Pathways are ordered by the average normalized enrichment score across 15 cancer types from top to bottom. (**E**) M-A plot comparing ATAC-seq peak scores of signature genes (signature) versus other genes (non-signature) from matched tumor samples in each cancer type. Samples with adjusted *P* values < 0.05 from permutation tests are shown as dots. Samples above the horizontal dashed line have significantly higher ATAC-seq peaks score in signature genes compared to those in non-signature genes. For A-E, all 15 TCGA cancer types are listed in the same order and annotated using colored squares as shown in the legend. (**F**) Distributions of mean signature gene UMI count per cell for tumor and non-tumor cells in scRNAseq data across four cancer types. For each cancer type, patient samples were ordered by disease stage from advanced to early or by progression outcome from poor to good. (**G**) Distributions of mean signature and non-sginature gene UMI count per tumor cell from scRNAseq data across four cancer types. For F & G, adjusted *P* values from Wilcoxon rank-sum tests are indicated by asterisks. (**H**) Distributions of the ratio of mean UMI counts for signature genes per cell for tumor cells versus non-tumor cells from scRNAseq data across four cancer types. The adjusted *P* values from Kruskal-Wallis tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). For F-H, patients are in the same order.

Fig. S4. TmS refines prognostication on pathological stages. (**A**) Distributions of TmS for TCGA samples within early (stage I and II) and advanced (stage III and IV) pathological stages across 15 cancer types. Adjusted *P* values of Wilcoxon rank-sum tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). (**B**) Distributions of TmS for female and male patient samples in TCGA across 15 cancer types. None of the adjusted *P* values of Wilcoxon rank-sum tests comparing TmS between the two groups reached significance at a confidence level of 0.05. Brown circles (read out on the right y-axis) represent Spearman correlation coefficients between TmS and age within the same sex and cancer type. The red dotted horizontal line represents TmS equal to 1 (left y axis) and correlation equal to 0 (right y axis). None of the adjusted *P* values for correlation tests reached significance at a confidence level of 0.05. (**C**) KM survival curves for individual cancer types with pattern I. (**D**) KM survival curves for bladder urothelial carcinoma with pattern II. (**E**) KM survival curves for renal papillary carcinoma with pattern III. (**F**) Predicted integrated AUC (iAUC) curves with 95% confidence intervals for patients with Gleason score ≥ 7 in the early-onset prostate adenocarcinoma validation cohort. The “Age and Gleason score model” was trained on the TCGA prostate adenocarcinoma data with Gleason score and age as predictors. The “Age and TmS x Gleason score model” was trained on the TCGA prostate adenocarcinoma data with

age and subgroups defined by TmS and Gleason score as predictors. The d-iAUC represents the difference in iAUC in the validation dataset using the two models.

Fig. S5. Association of TmS with cancer-specific genomic dysregulations and hypoxia in TCGA. A-D, Distributions of TmS for patient samples with (A) high or low tumor mutation burden (TMB); (B) high or low chromosomal instability score; (C) with or without a whole genome duplication event; (D) high or low hypoxia score. Cutoffs in (a,b,d) are set at the median. Adjusted *P* values of Wilcoxon rank-sum tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001).

Fig. S6. Regional TmS identifies spatial heterogeneity and refines prognostication in patients with early-stage lung cancer. (**A**) Pairwise scatter plots and histograms of number of regions per patient, range of TmS, % subclonal CNA per patient, TmSmax, and TmSmed. Spearman correlation coefficient *r*'s are shown. The gray lines represent a loess fit. (**B**) KM survival curves of disease-free probability for the 30 patients stratified by % subclonal CNA: high versus low. (**C**) KM survival curves of disease-free probability for 30 TRACERx patients with multi-region sequencing stratified by both TmS_{med} and percent subclonal CNA: (1) high TmS_{med}, (2) low TmS_{med} and high percent subclonal CNA and (3) low TmS_{med} and low percent subclonal CNA. (**D**) 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. (**E**) KM survival curves of disease-free probability for 52 patients stratified into two groups by TmS_{max}: high versus low TmS_{max}.

Fig. S7. CONSORT diagrams for data exclusions in TmS calculation and downstream analysis. (**A**) CONSORT diagram for TmS calculation in TCGA datasets. (**B**) CONSORT diagram for survival analysis in TCGA datasets. (**C**) CONSORT diagram for TmS calculation in ICGC-EOPC dataset. (**D**) CONSORT diagram for TmS calculation in TRACERx dataset.

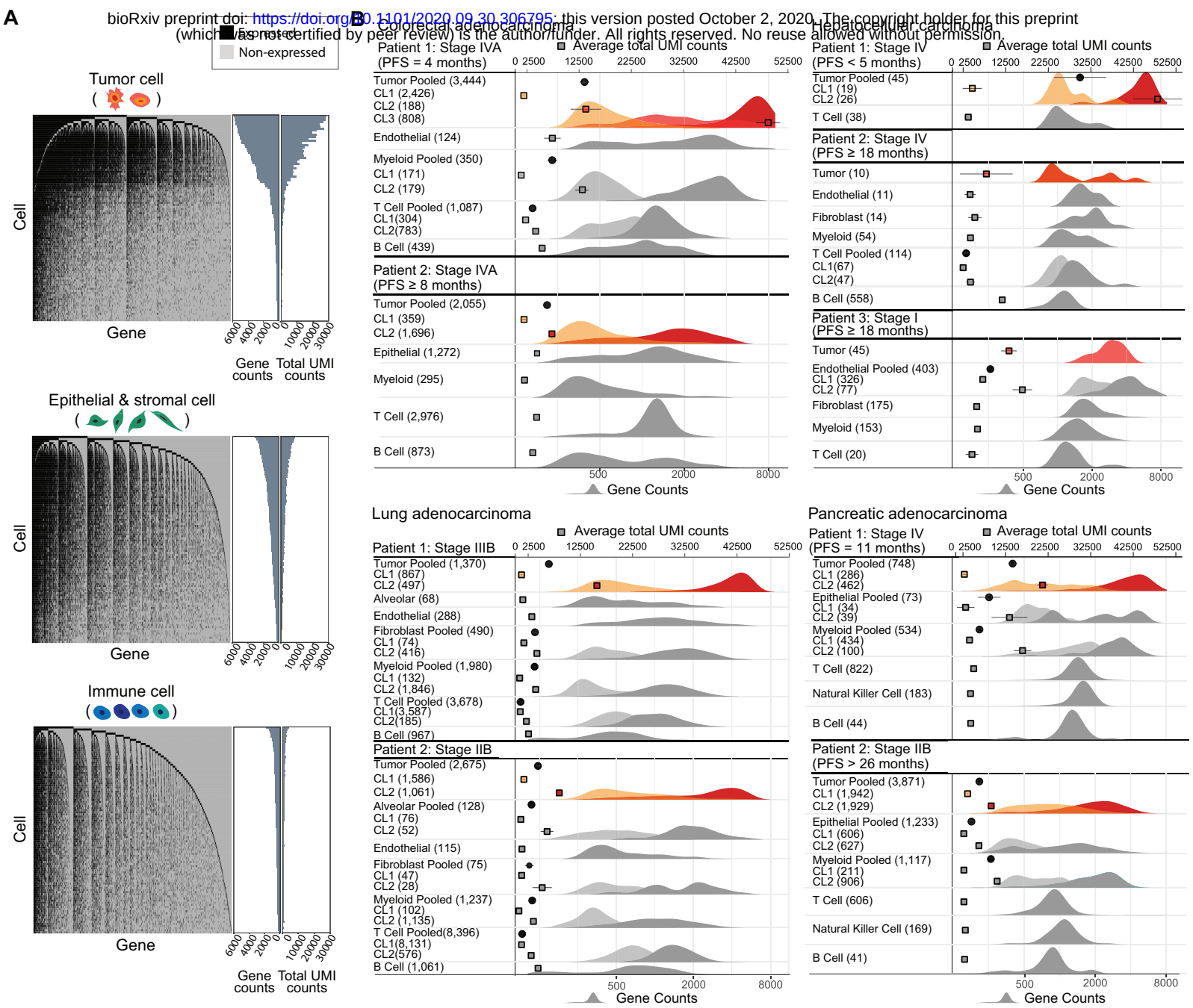

Fig. 1. High diversity of total mRNA expression in tumor cells. (**A**) Total gene expression in tumor cells compared to non-tumor cells (epithelial, stromal, immune cells). 100 cells of each cell type were randomly selected from a patient with Stage IIIB lung adenocarcinoma. In each heatmap, expressed genes (UMI count > 0) are shown in black, and non-expressed genes (UMI count = 0) in gray. Cells (rows) and genes (columns) are ordered from high to low total number of expressed genes and number of expressing cells, respectively. Bar plots show the total number of expressed genes, i.e., gene counts, and total UMI counts in the corresponding cells. (**B**) Distributions of gene counts and total UMI counts by cell type in scRNAseq data from nine patients with colorectal, hepatocellular, lung and pancreatic cancers. Patients are ordered by pathological stage or survival outcome. The top x-axis annotates total UMI counts (means and 95% Confidence Intervals, CIs). The bottom x-axis annotates gene count distribution (density). Density curves are colored for tumor cells and shown in gray-scale for non-tumor cells. Clusters with higher gene counts are shown in darker shades. The numbers in the parentheses indicate the number of cells analyzed.

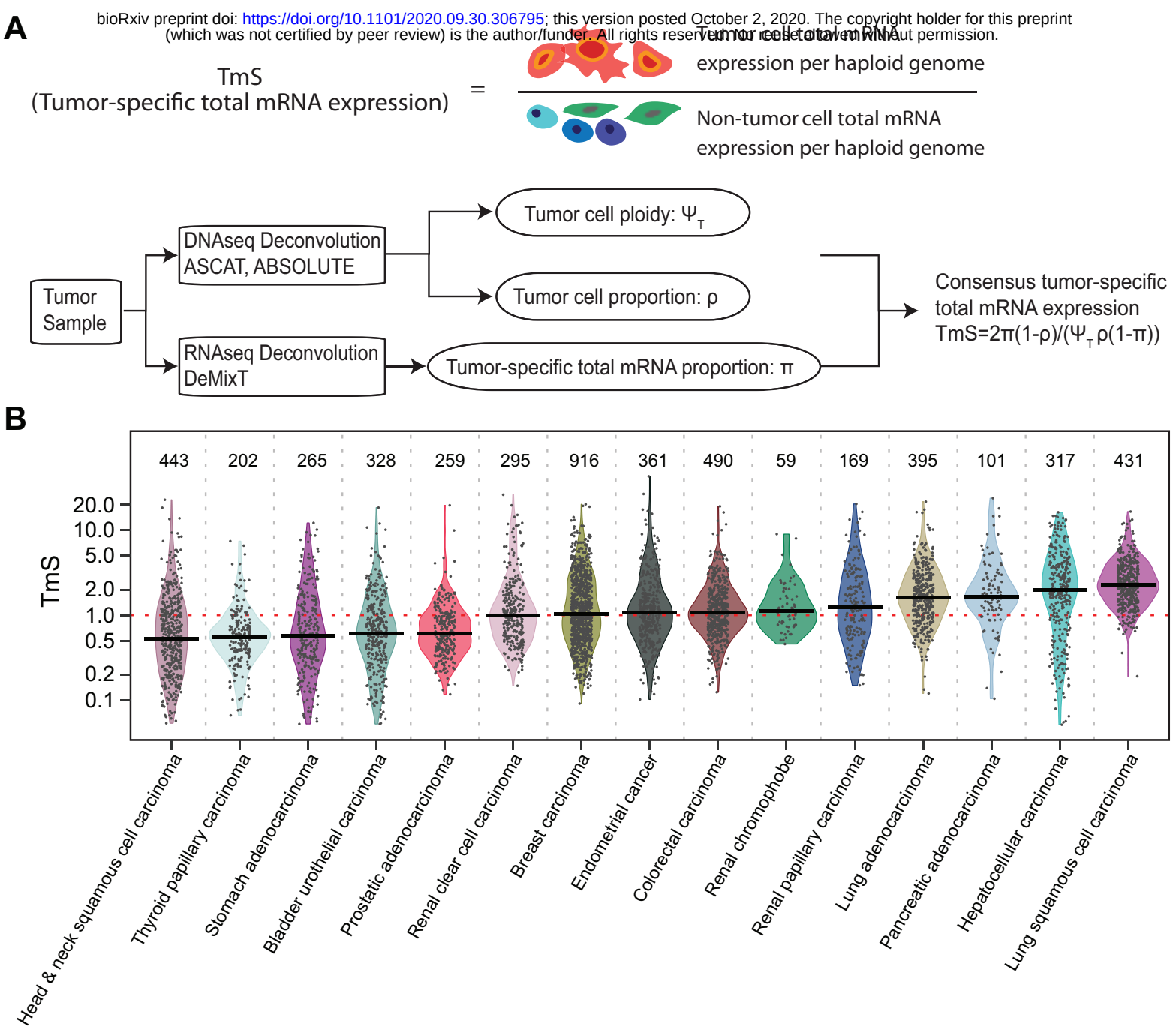

Fig. 2. Estimation of tumor-specific total mRNA expression in bulk sequencing data. (**A**) The definition of TmS and its analysis pipeline using matched DNAseq and RNAseq data. (**B**) Distributions of TmS in 5,031 tumor samples across 15 cancer types in TCGA. The number of patient samples for each cancer type is indicated on the top of each violin plot.

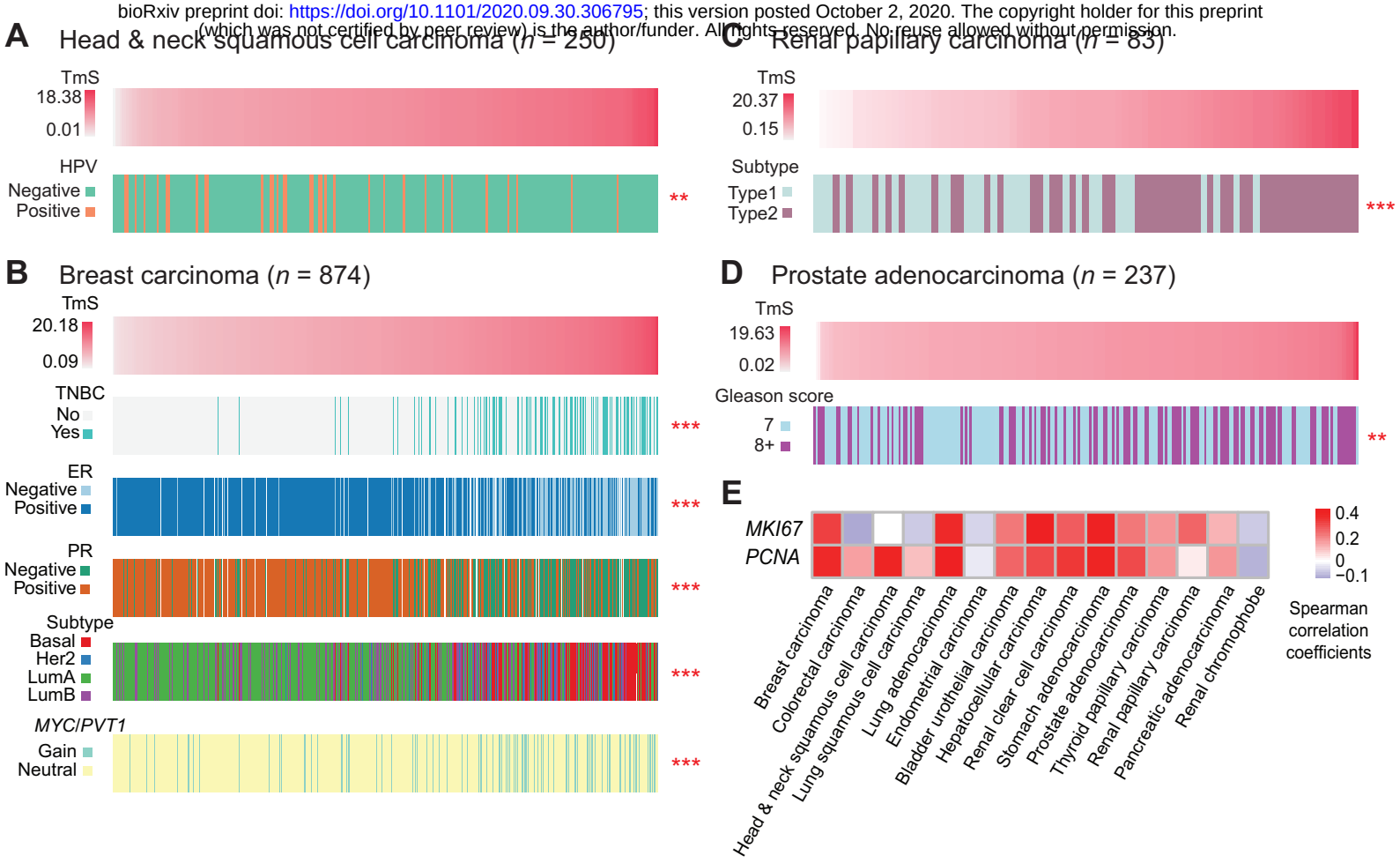

Fig. 3. TmS is associated with known prognostic characteristics. (**A-D**) Clinicopathologic annotations are shown for (A) head & neck squamous cell carcinoma, (B) breast carcinoma, (C) renal papillary carcinoma, and (D) prostate adenocarcinoma samples. Tumor samples are ordered by TmS from low to high. The Benjamini–Hochberg adjusted *P* values for Kruskal-Wallis tests comparing TmS between clinicopathologic subgroups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). For breast carcinoma, triple-negative breast cancers (TNBC) are shown on the second row. *MYC*/*PVT1* copy number alterations are shown on the sixth row, where “Gain” indicates that either *MYC* or *PVT1* was amplified and “Neutral” indicates that no copy number alternations were detected. (**E**) Spearman correlation coefficients between TmS and the expression levels of *MKI67* and *PCNA* across 15 cancer types.

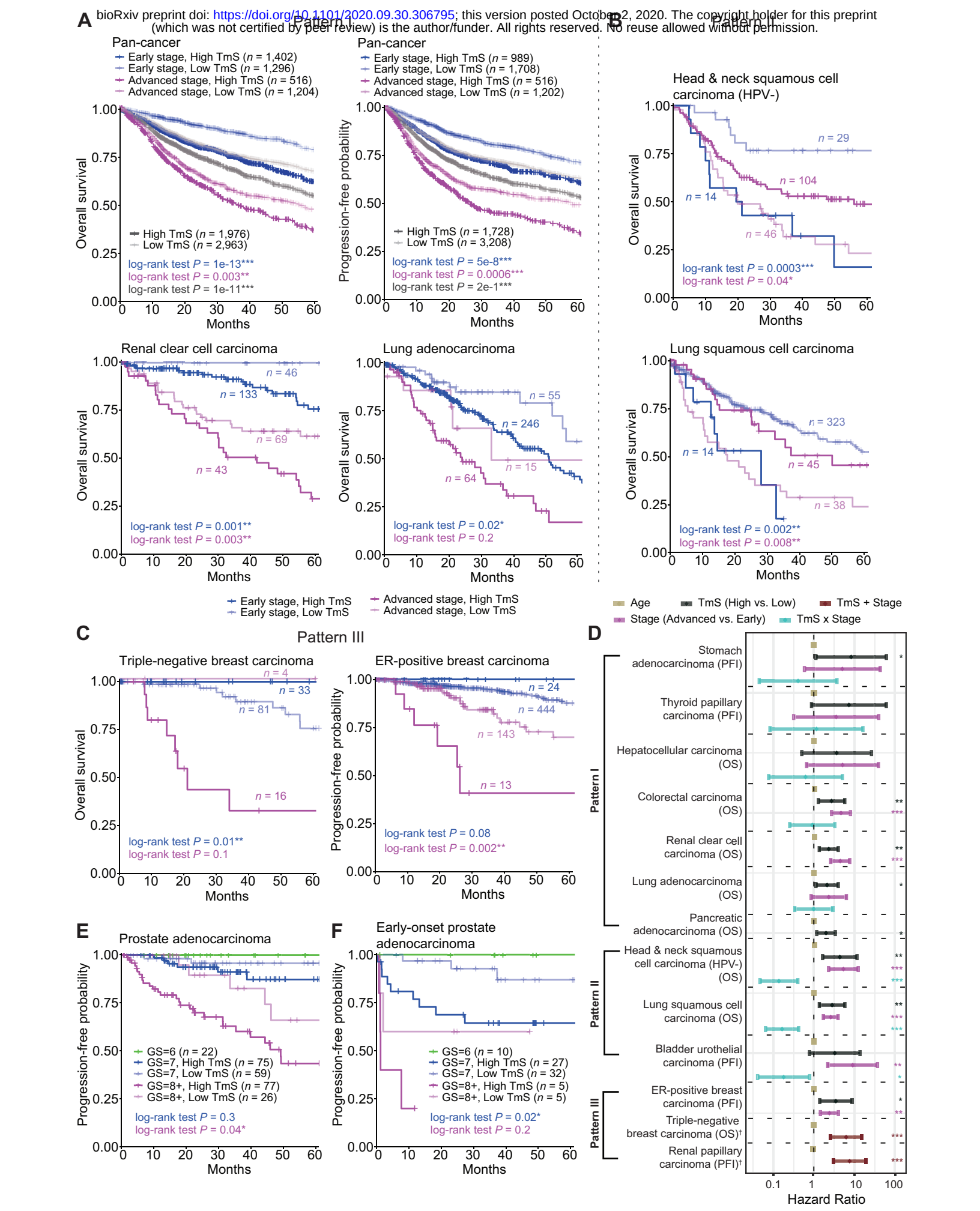

Fig. 4. TmS refines prognostication on pathological stages.

Fig. 4. (cont'd): (**A**) Kaplan-Meier (KM) curves of overall survival (OS) and progression-free interval (PFI) for TCGA samples. Gray lines denote summary KM curves of patients with high versus low TmS across all cancer types. KM curves are further grouped by TmS and pathological stage into four groups. *P* values of log-rank tests between high versus low TmS groups are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). Three patterns of associations were identified (summarized in D). Pattern I is the most prevalent and shown in A. (**B-C**) KM curves for representative cancers (top and bottom panels) in Pattern II (B) and Pattern III (C). (**D**) Forest plot of hazard ratios and 95% of CIs of multivariate Cox proportional hazard models with Age, TmS (High vs. Low), Stage (Advanced vs. Early) and TmS x Stage as predictors, and with OS or PFI as response variable (See detail in table S3). (**E**) KM survival curves of PFI for the TCGA prostate adenocarcinoma study cohort. (**F**) KM survival curves of PFI for the ICGC-EOPC study cohort. In E & F, patients are grouped by Gleason score and TmS.

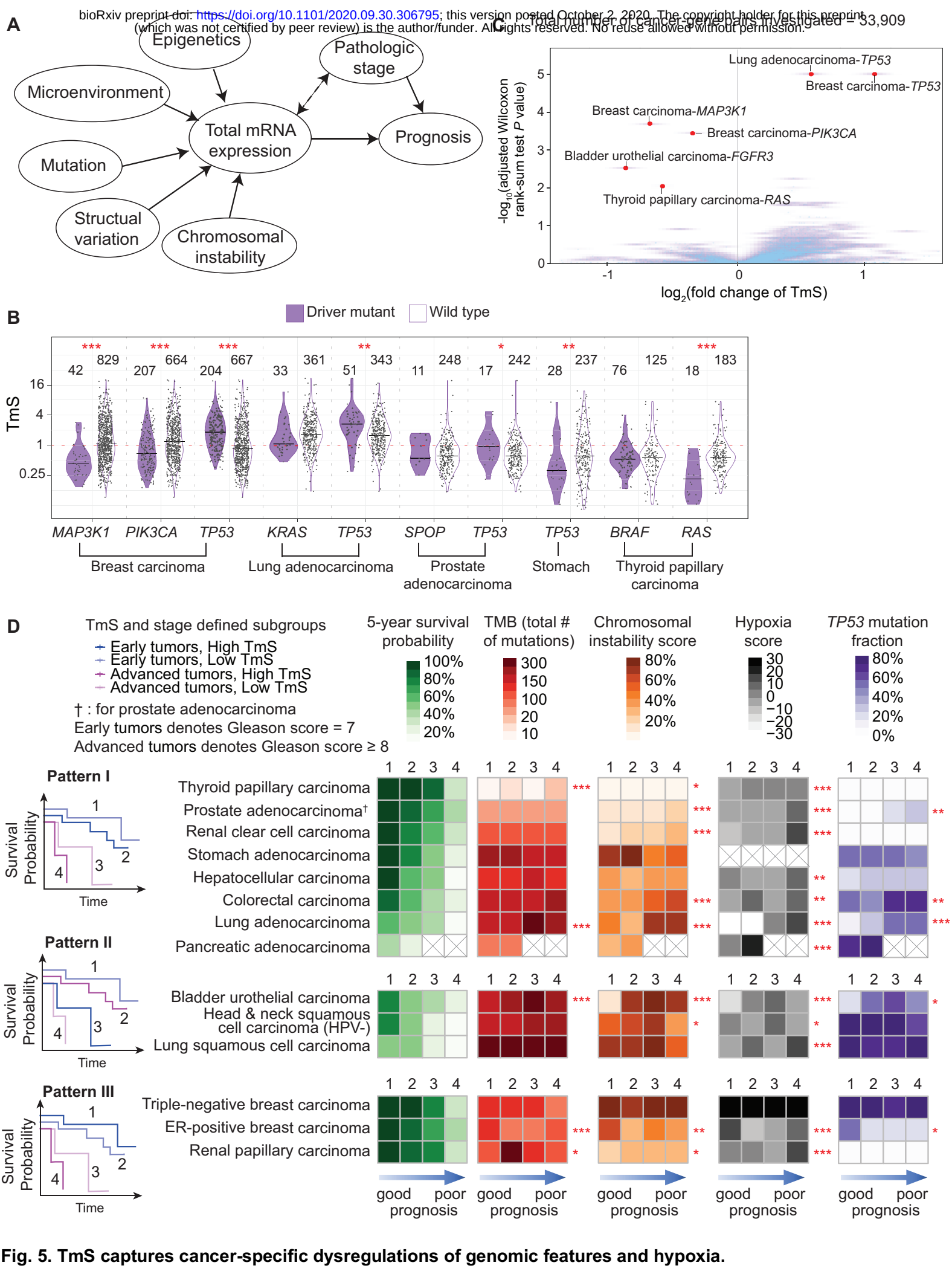

Fig. 5. TmS captures cancer-specific dysregulations of genomic features and hypoxia.

Fig. 5. (cont'd): (**A**) Illustration of a potential role for total mRNA expression as a mediator in the pathway between genomic alterations, epigenetic events and prognosis. Pathologic stage may be combined with TmS to define a clinically informative phenotype of tumor samples. The solid circles indicate that the corresponding data are evaluated. (**B**) Distributions of TmS for TCGA samples with or without driver mutations across the most frequently mutated genes across cancer types. (**C**) Volcano plot showing $\log_2$-fold change in TmS for samples with non-silent mutations in a given gene vs those without. Cancer-gene pairs with adjusted *P* values < 0.05 are highlighted in red points. (**D**) Heatmaps of median 5-year survival, tumor mutation burden, chromosomal instability, hypoxia, and *TP53* non-silent somatic mutation rate across patient groups and cancer types. Patient groups are defined as in Fig. 4D-E, with cancer types following three distinct patterns in survival outcome. Adjusted *P* values are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001).

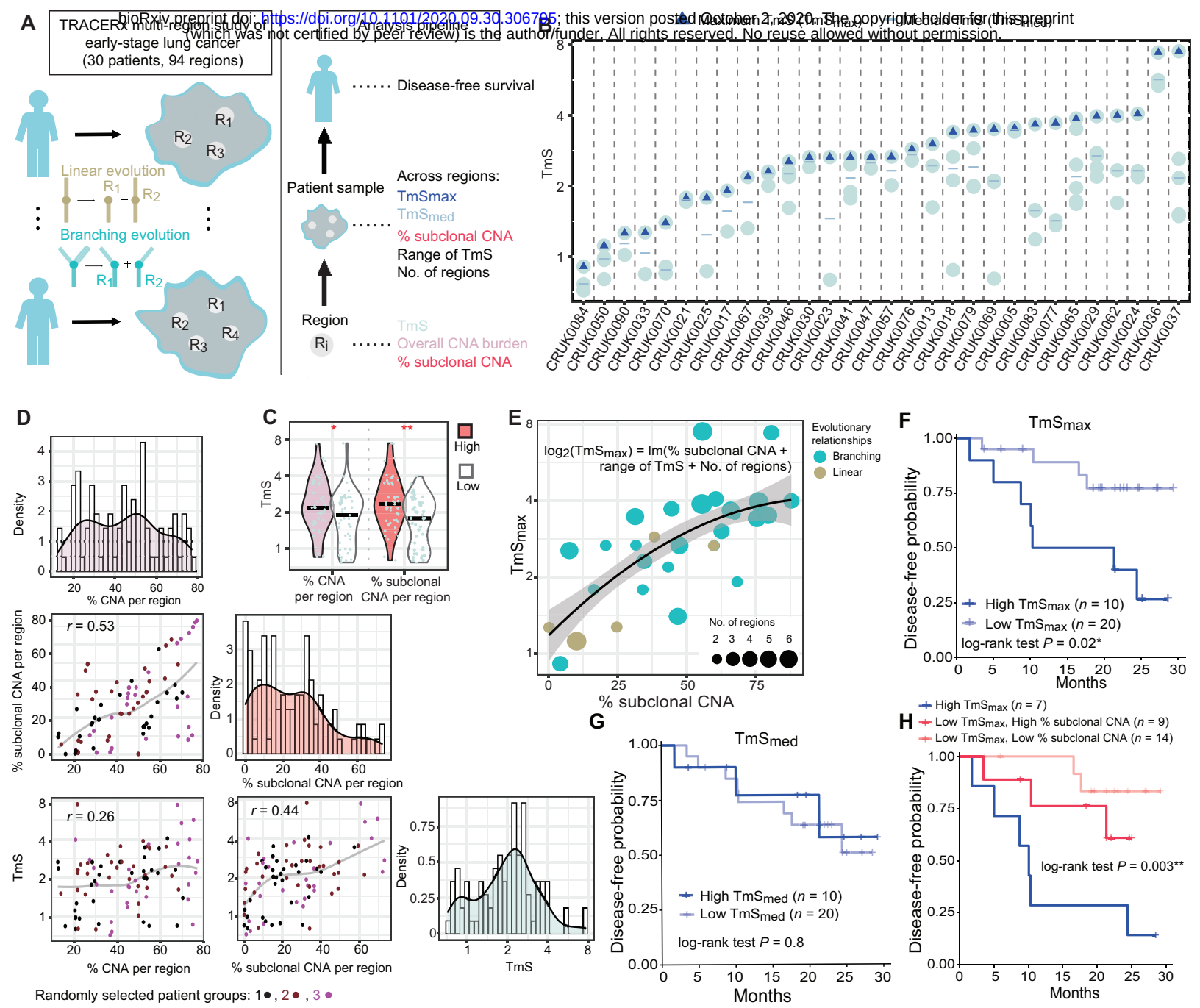

Fig. 6. Regional TmS identifies spatial heterogeneity and refines prognostication in early-stage lung cancer. (**A**) Illustration of the TRACERx multi-region study and a multi-level analysis pipeline. (**B**) Distribution of TmS for 94 tumor regions from 30 TRACERx patients with at least 2 regions sampled. Blue triangles denote the maximum TmS for a patient. Blue "-" denote the median TmS for a patient. (**C**) Pairwise scatter plots and histograms of % CNA, % subclonal CNA, and TmS per region. Spearman correlation coefficient *r*'s are shown. Different colors annotate three randomly assigned patient groups. The gray lines represent a loess fit. (**D**) Distributions of TmS per each region with high or low % CNA burden per region (left), % subclonal CNA per region (right), with regions grouped at the median values. Adjusted *P* values of Wilcoxon rank-sum tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). (**E**) Scatter plot showing TmS_{max} versus the percent subclonal CNA. A regression fit and 95%CI is shown. (**F-G**) KM survival curves of disease-free probability stratified by TmS_{max} or TmS_{med}. (**H**) KM survival curves of disease-free probability stratified by both TmS_{max} and % subclonal CNA.

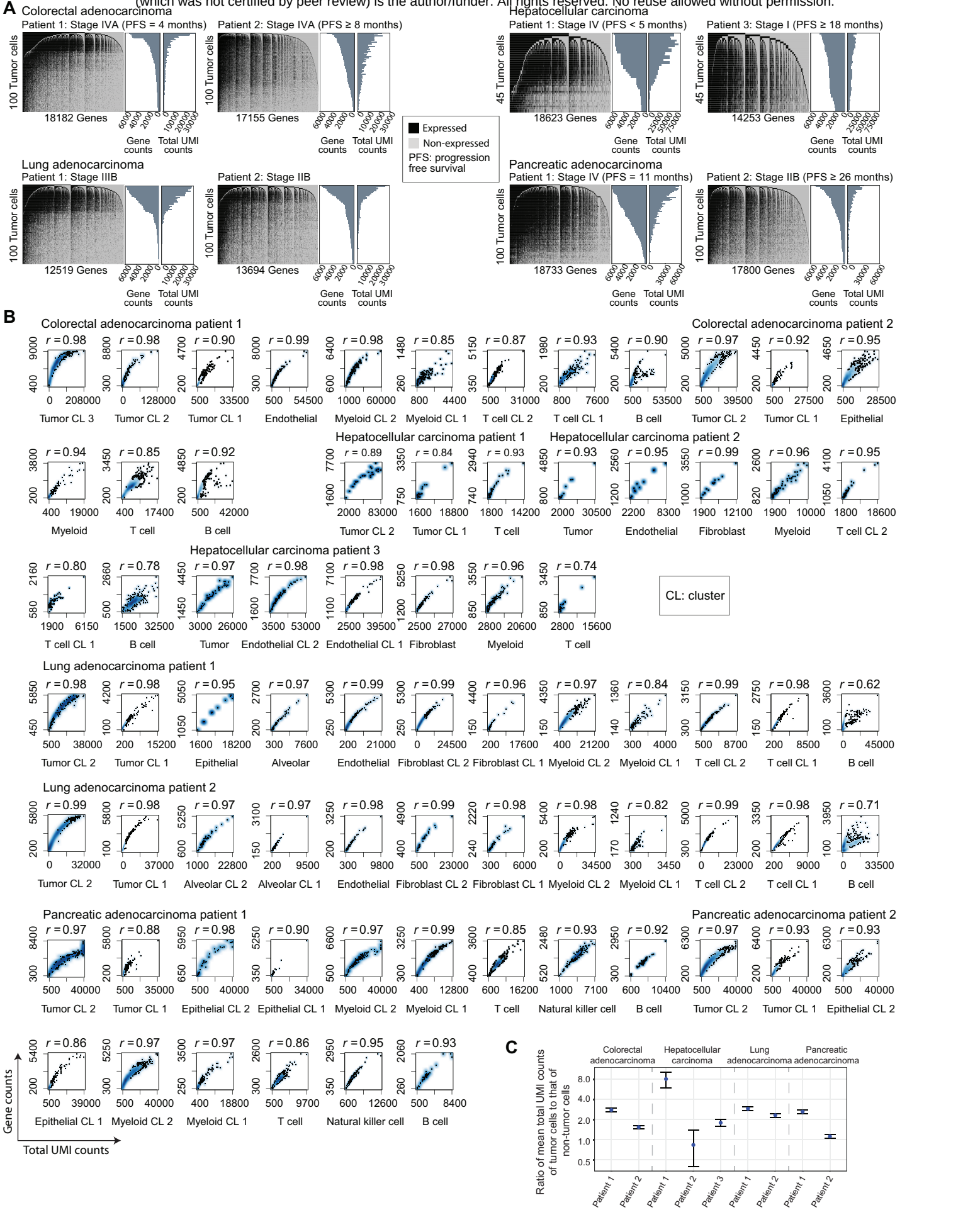

Fig. S1. Using gene counts and total UMI counts to measure the global gene expression heterogeneity in tumor cells.

Fig. S1. (cont'd): (**A**) Illustration of expressed genes in tumor cells compared to non-tumor cells (epithelial, stromal, and immune cells), randomly selected from each patient sample. 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 expressions of each gene, respectively. The barplots provide the distributions of total number of expressed genes, i.e., gene counts, and total UMI counts in the corresponding cells. (**B**) Smoothed scatter plots show the correlations between gene counts and total UMI counts in cell clusters from each patient sample. Patient samples are arranged in the same order of cell clusters as (A). In each smoothed scatter plot, the Spearman correlation coefficient is labeled on the top (*r*). (**C**) Ratios of mean total UMI counts of tumor cells to non-tumor cells and 95% confidence intervals in pseudo-bulk data, which are made by pooling scRNAseq data.

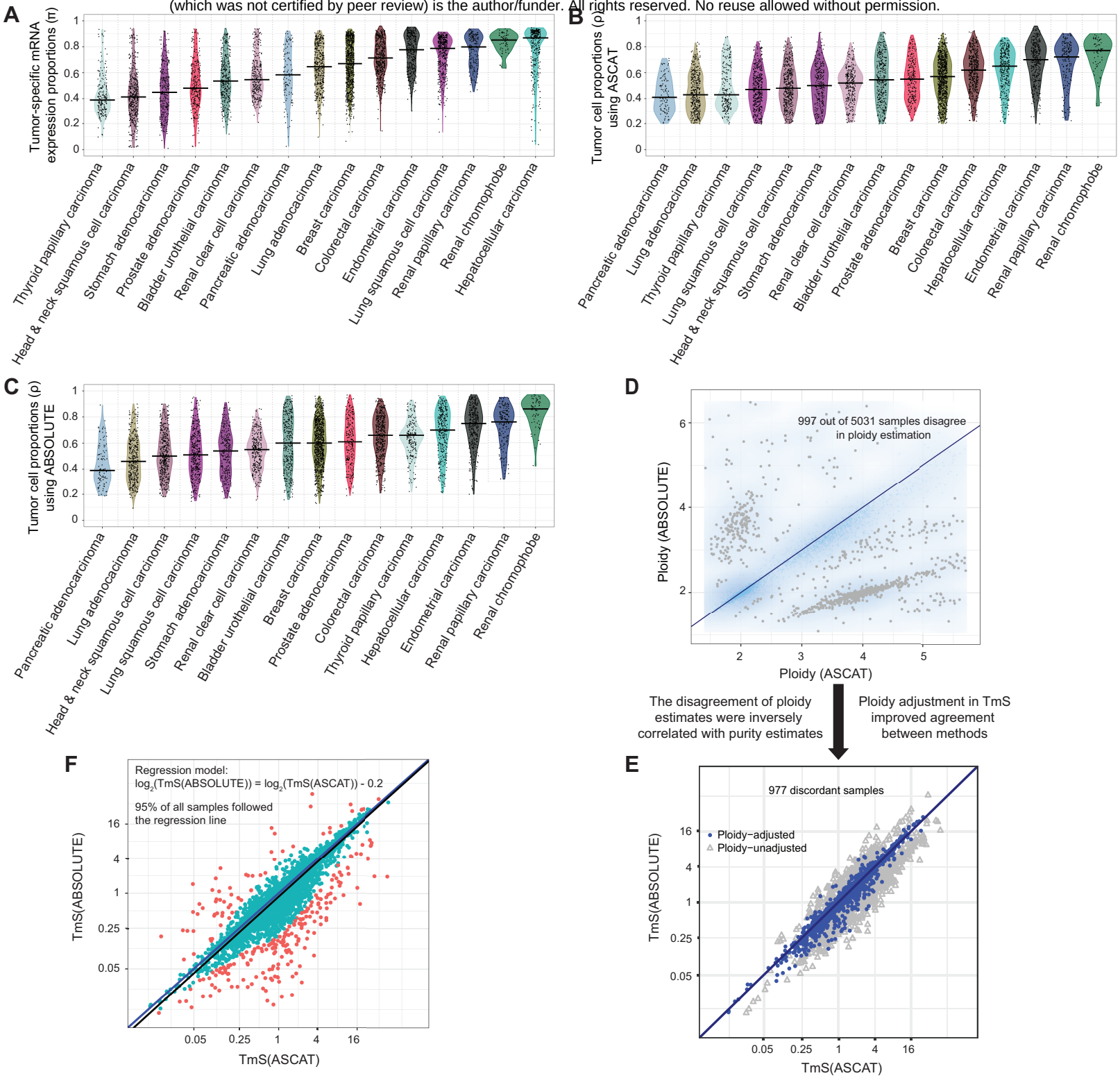

Fig. S2. Consensus estimation of TmS from matched RNA- and DNAseq data in TCGA. (**A**), Distributions of tumor-specific mRNA expression proportions estimated by DeMixT across cancer types. (**B-C**) Distributions of tumor cell proportions estimated by (**B**) ASCAT or (**C**) ABSOLUTE across cancer types. (**D**) Smoothed scatter plot of tumor ploidy estimates from ABSOLUTE versus ASCAT across all samples. Gray points correspond to 997 samples that presented inconsistent tumor ploidy (and purity) estimates between the two methods. (**E**) TmS estimates using either ABSOLUTE or ASCAT-derived purity and ploidy estimates with or without ploidy adjustment for the 977 discordant samples from (**D**). Blue and gray points correspond to, respectively, TmS prior to and after ploidy adjustment. Ploidy adjustment improved consistency between the ABSOLUTE and ASCAT results. (**F**) 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 ≥ 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 deemed consistent and robust across two DNAseq deconvolution methods.

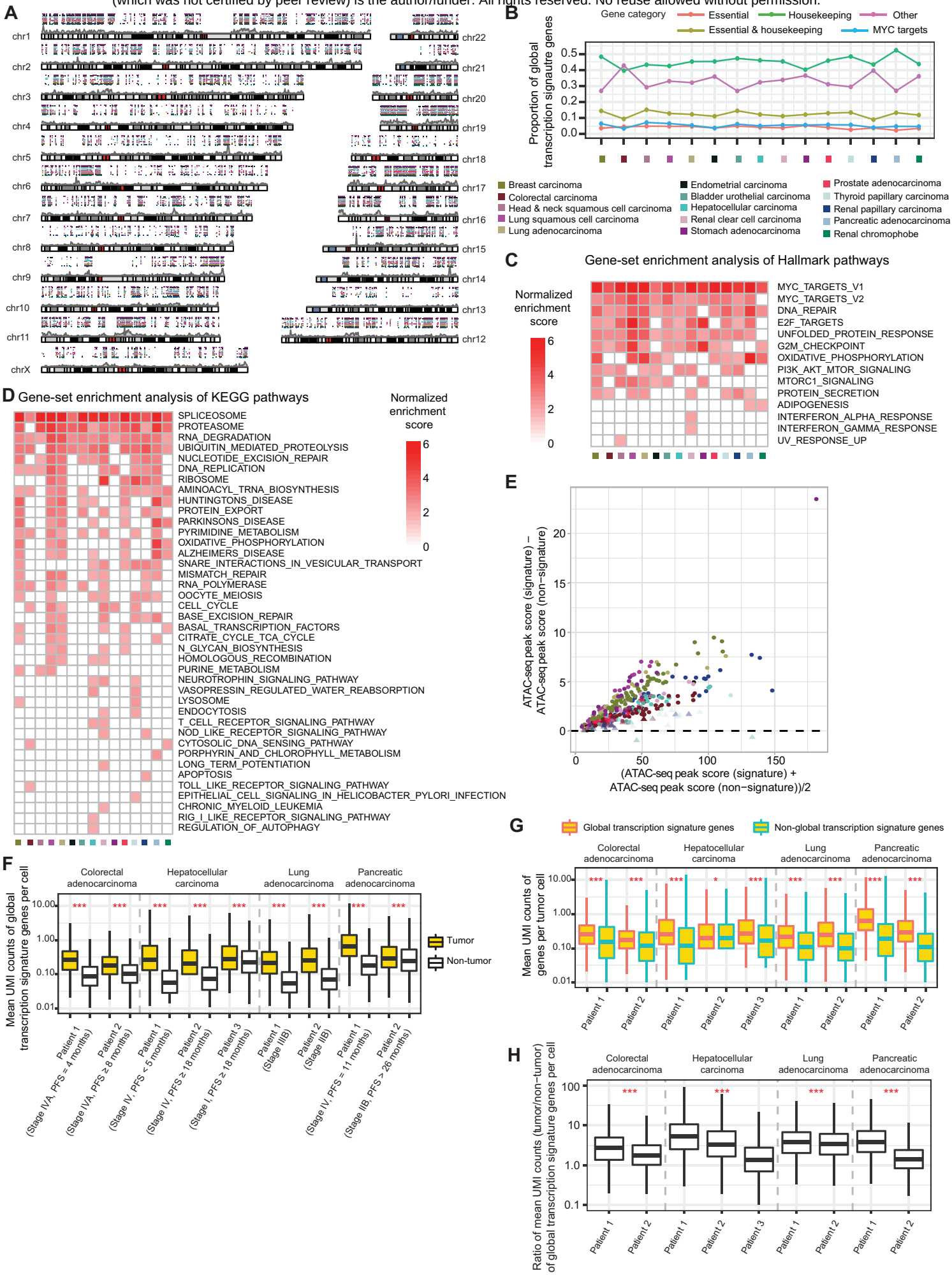

Fig. S3. Global transcription signature genes across cancer types.

Fig. S3. (cont'd): (**A**) Karyotype plots showing the genomic locations of signature genes for each cancer type. Signature genes are presented as dots colored by cancer type. An overall gene density track is shown in gray shades underneath the dots. The density of signature genes is consistent with the overall gene density. (**B**) Proportions of global transcription signature genes in five gene categories across 15 cancer types. (**C**) Heatmap of normalized enrichment scores of enriched Hallmark pathways across 15 cancer types. (**D**) Heatmap of normalized enrichment scores of enriched KEGG pathways across 15 cancer types. For (C & D), significantly enriched pathways are those with an adjusted *P* values <0.05 from both GSEA and g:Profiler. Pathways are ordered by the average normalized enrichment score across 15 cancer types from top to bottom. (**E**) M-A plot comparing ATAC-seq peak scores of signature genes (signature) versus other genes (non-signature) from matched tumor samples in each cancer type. Samples with adjusted *P* values < 0.05 from permutation tests are shown as dots. Samples above the horizontal dashed line have significantly higher ATAC-seq peaks score in signature genes compared to those in non-signature genes. For (A-E), all 15 TCGA cancer types are listed in the same order and annotated using colored squares as shown in the legend. (**F**) Distributions of mean signature gene UMI count per cell for tumor and non-tumor cells in scRNAseq data across four cancer types. For each cancer type, patient samples were ordered by disease stage from advanced to early or by progression outcome from poor to good. (**G**) Distributions of mean signature and non-sginature gene UMI count per tumor cell from scRNAseq data across four cancer types. For (F & G), adjusted *P* values from Wilcoxon rank-sum tests are indicated by asterisks. (**H**) Distributions of the ratio of mean UMI counts for signature genes per cell for tumor cells versus non-tumor cells from scRNAseq data across four cancer types. The adjusted *P* values from Kruskal-Wallis tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). For (F-H) patients are in the same order.

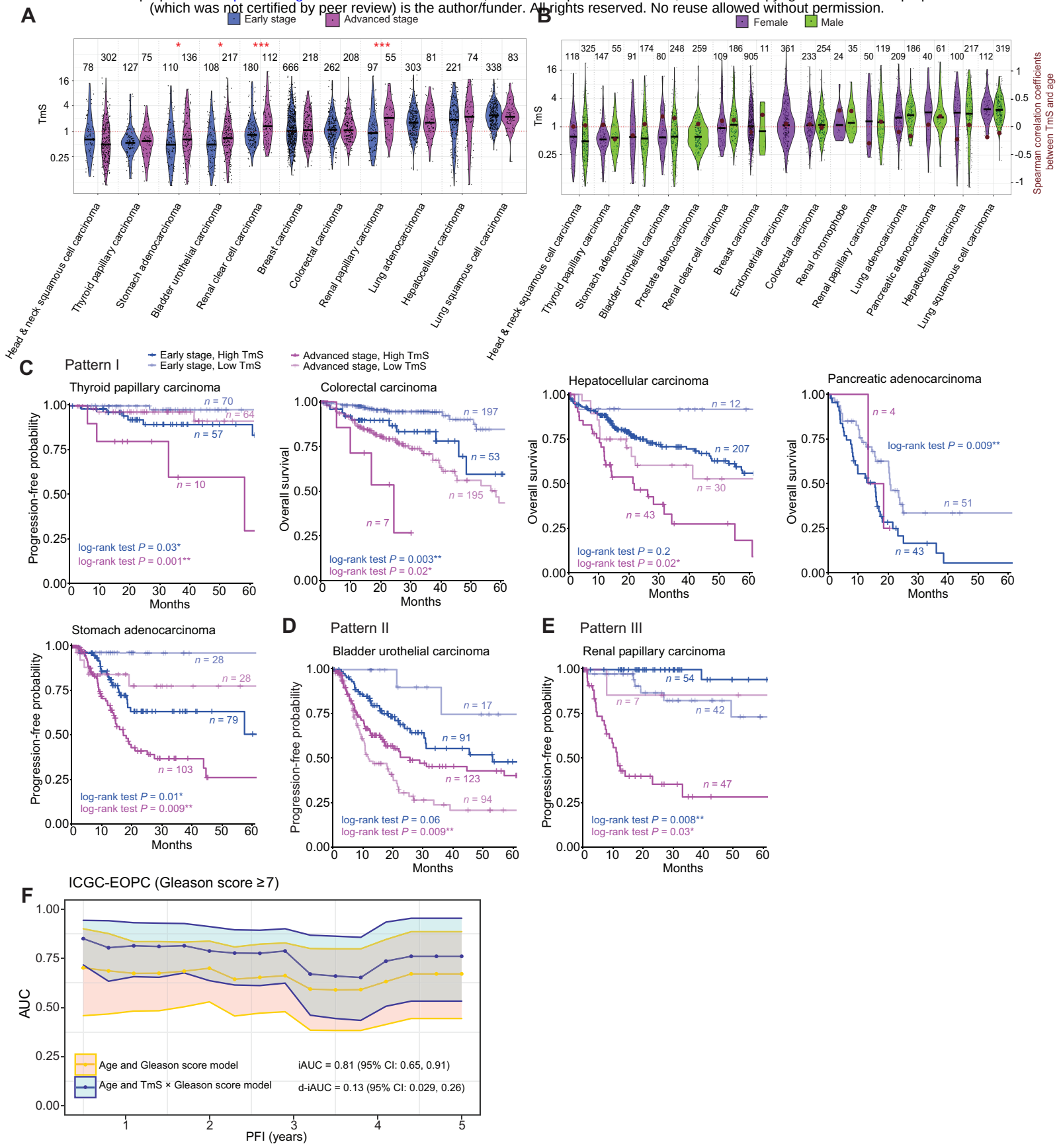

Fig. S4. TmS refines prognostication on pathological stages. (**A**) Distributions of TmS for TCGA samples within early (stage I and II) and advanced (stage III and IV) pathological stages across 15 cancer types. Adjusted *P* values of Wilcoxon rank-sum tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001). (**B**) Distributions of TmS for female and male patient samples in TCGA across 15 cancer types. None of the adjusted *P* values of Wilcoxon rank-sum tests comparing TmS between the two groups reached significance at a confidence level of 0.05. Brown circles (read out on the right y-axis) represent Spearman correlation coefficients between TmS and age within the same sex and cancer type. The red

dotted horizontal line represents TmS equal to 1 (left y axis) and correlation equal to 0 (right y axis). None of the adjusted *P* values for correlation tests reached significance at a confidence level of 0.05. (**C**) KM survival curves for individual cancer types with pattern I. (**D**) KM survival curves for bladder urothelial carcinoma with pattern II. (**E**) KM survival curves for renal papillary carcinoma with pattern III. (**F**) Predicted integrated AUC (iAUC) curves with 95% confidence intervals for patients with Gleason score ≥ 7 in the early-onset prostate adenocarcinoma validation cohort. The “Age and Gleason score model” was trained on the TCGA prostate adenocarcinoma data with Gleason score and age as predictors. The “Age and TmS x Gleason score model” was trained on the TCGA prostate adenocarcinoma data with age and subgroups defined by TmS and Gleason score as predictors. The d-iAUC represents the difference in iAUC in the validation dataset using the two models.

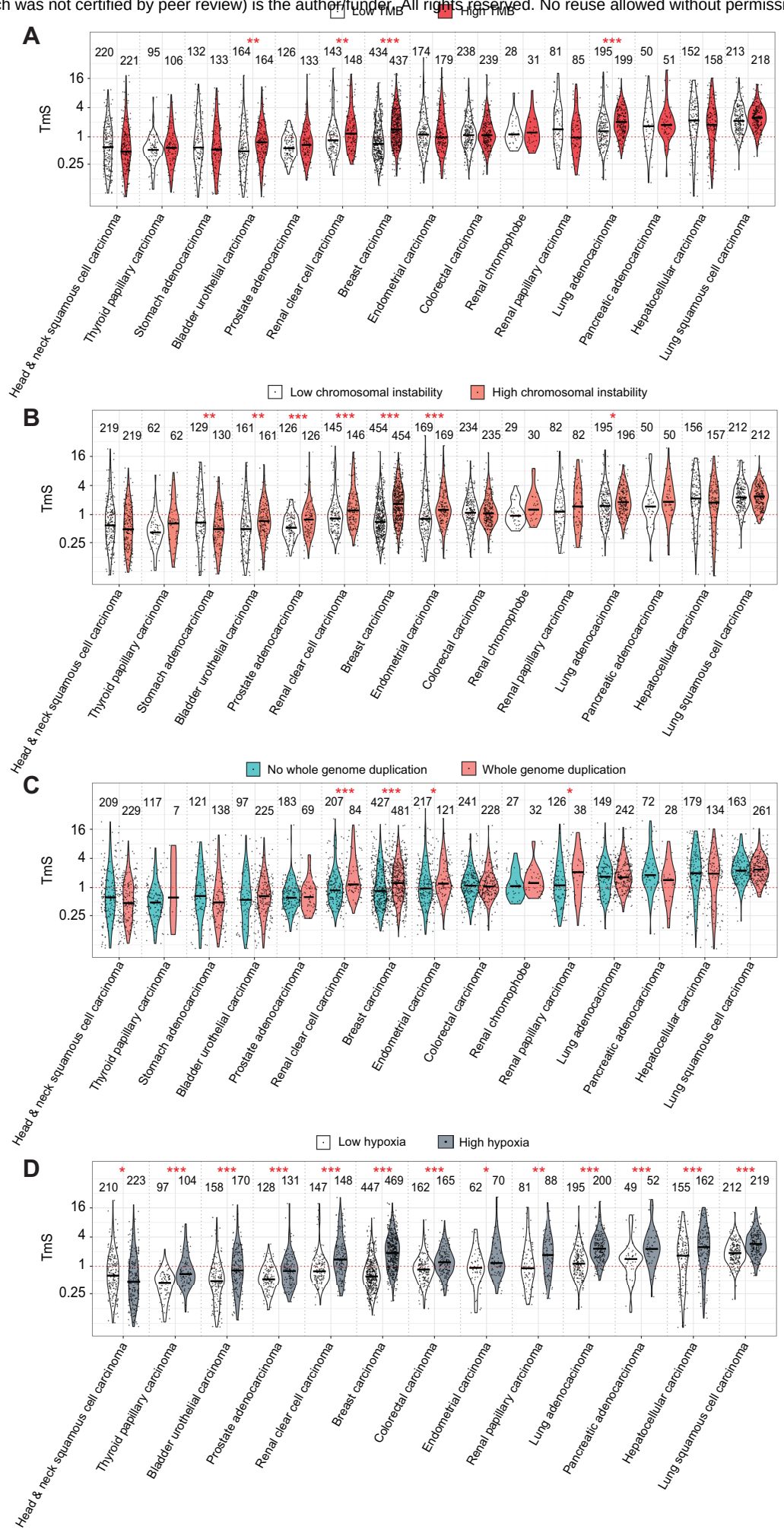

Fig. S5. Association of TmS with cancer-specific genomic dysregulations and hypoxia in TCGA.

Fig. S5. (cont'd): (**A-D**) Distributions of TmS for patient samples with (A) high or low tumor mutation burden (TMB); (B) high or low chromosomal instability score; (C) with or without a whole genome duplication event; (D) high or low hypoxia score. Cutoffs in (A, B, D) are set at the median. Adjusted *P* values of Wilcoxon rank-sum tests are indicated by asterisks (* $P < 0.05$, ** < 0.01, *** < 0.001).

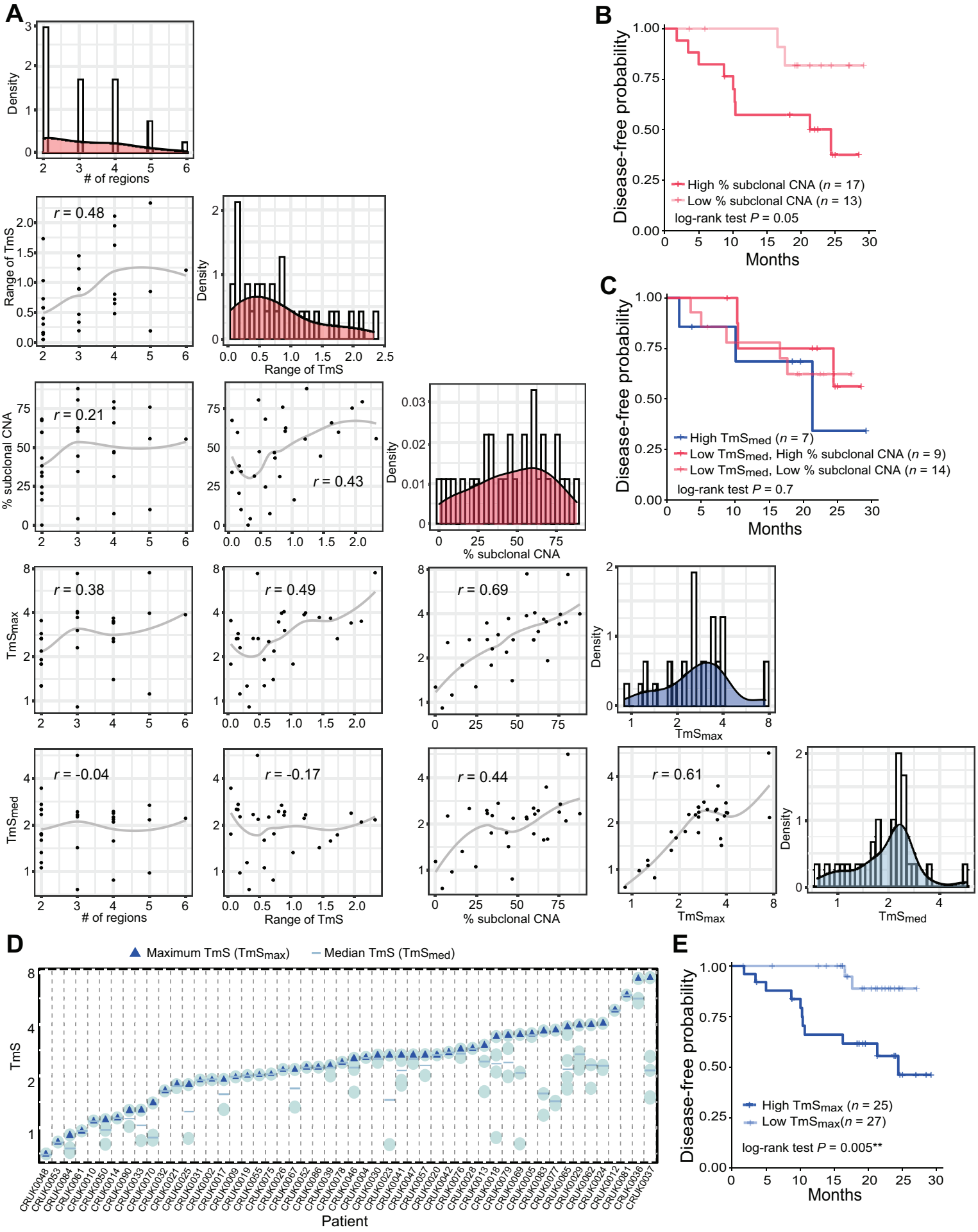

Fig. S6. Regional TmS identifies spatial heterogeneity and refines prognostication in patients with early-stage lung cancer.

Fig. S6. (cont'd): (**A**) Pairwise scatter plots and histograms of number of regions per patient, range of TmS, % subclonal CNA per patient, TmSmax, and TmSmed. Spearman correlation coefficient *r*'s are shown. The gray lines represent a loess fit. (**B**) KM survival curves of disease-free probability for the 30 patients stratified by % subclonal CNA: high versus low. (**C**) KM survival curves of disease-free probability for 30 TRACERx patients with multi-region sequencing stratified by both TmS_{med} and percent subclonal CNA: (1) high TmS_{med}, (2) low TmS_{med} and high percent subclonal CNA and (3) low TmS_{med} and low percent subclonal CNA. (**D**) 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. (**E**) KM survival curves of disease-free probability for 52 patients stratified into two groups by TmS_{max}: high versus low TmS_{max}.

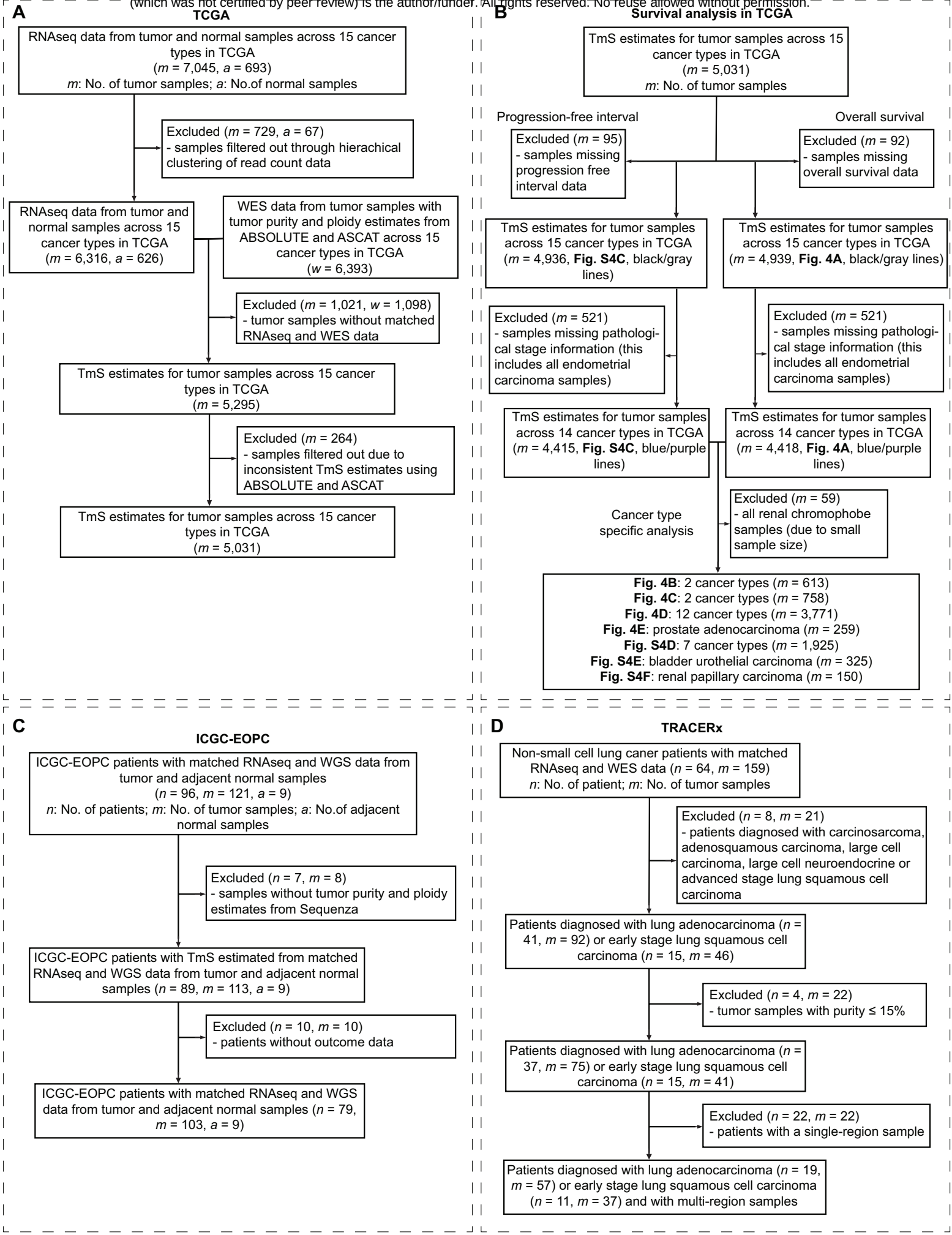

Fig. S7. CONSORT diagrams for data exclusions in TmS calculation and downstream analysis.

Fig. S7. (cont'd): (**A**) CONSORT diagram for TmS calculation in TCGA datasets. (**B**) CONSORT diagram for survival analysis in TCGA datasets. (**C**) CONSORT diagram for TmS calculation in ICGC-EOPC dataset. (**D**) CONSORT diagram for TmS calculation in TRACERx dataset.

Table S1 | Clinical information of the nine patients in the single cell RNA sequencing data analysis.

Cancer type	Patient id	Age	Sex	Treatment	Stage	Progression free survival (months)	Outcome	Tumor tissue collection	Details
Colorectal adenocarcinoma	Patient 1	45	M	FOLFOX/ panitumumab	IVA	4	Relapsed with multiple tumors in the liver	Surgical resection	Received 10 cycles of chemotherapy prior to surgery; regression grade 3 and 70% viable tumor; tumor is moderately differentiated; KRAS wild type
	Patient 2	63	F	FOLFOX/bev	IVA	≥ 8	No tumor on last scan	Surgical resection	Received 4 cycles of chemotherapy prior to surgery; regression grade 3 and 90% viable tumor; tumor is moderately differentiated; KRAS mutation
Hepatocellular carcinoma	Patient 1	65	M	Durvalumab/ tremelimumab	IV	< 5	Alive but with decreased progression	Surgical resection	Etiology is Hepatitis C virus
	Patient 2	63	M	Durvalumab/ tremelimumab	IV	≥ 18	Progression free	Surgical resection	Etiology is Hepatitis C virus
	Patient 3	74	M	Treatment-naïve	I	≥ 18	Progression free	Core needle biopsy	NA
Lung adenocarcinoma	Patient 1	68	M	Treatment-naïve	IIIB	NA	NA	Surgical resection	TNM stage: pT4N2M0
	Patient 2	64	F	Treatment-naïve	IIB	NA	NA	Surgical resection	TNM stage: pT2aN1M0
Pancreatic adenocarcinoma	Patient 1	62	F	Treatment-naïve	IV	16	Developed liver metastses	Fine needle aspiration	Alive; date of diagnosis is 12/18/2018; date of last follow up time is 5/6/2020; date of biopsy is 1/10/2019; developed liver metastases on 12/5/2019
	Patient 2	70	M	Treatment-naïve	IIB	≥ 26	Progression free	Fine needle aspiration	Alive; date of diagnosis is 1/7/2018; date of last follow up time is 3/23/2020; date of biopsy is 1/9/2019; date of surgery is 5/31/2019

Table S2 | Summary of the distributions of TmS across 15 cancer types in TCGA.

Cancer	Median TmS	MAD TmS	Sample size
Head & neck squamous cell carcinoma	0.53	0.54	443
Thyroid papillary carcinoma	0.55	0.33	202
Stomach adenocarcinoma	0.58	0.55	265
Bladder urothelial carcinoma	0.61	0.54	328
Prostate adenocarcinoma	0.62	0.40	259
Renal clear cell carcinoma	1.0	0.72	295
Breast carcinoma	1.0	0.92	916
Endometrial carcinoma	1.1	0.84	361
Colorectal carcinoma	1.1	0.67	490
Renal chromophobe	1.1	0.60	59
Renal papillary carcinoma	1.2	1.2	169
Lung adenocarcinoma	1.6	1.2	395
Pancreatic adenocarcinoma	1.7	1.2	101
Hepatocellular carcinoma	2.0	2.2	317
Lung squamous cell carcinoma	2.3	1.4	431
Overall	1.1	0.96	5,031

MAD: Median absolute deviation

Table S3 | Multivariate Cox proportional hazard models with Age, TmS, Stage and TmS x Stage as predictors for overall survival and progression-free interval analysis across cancer types in TCGA.

Cancer	Sample size	Clinical Outcome Endpoint	Variable	Hazard ratio	95% CI	P value (Wald test)	TmS prognostication type
Colorectal carcinoma	*n* = 452	OS	Age	1.04	(1.02, 1.06)	$5x10^{-5}$	Pattern I
			TmS (High vs. Low)	2.7	(1.3, 5.8)	0.01	
			Stage (Advanced vs. Early)	4.7	(2.7, 8.0)	$2x10^{-8}$	
			TmS x Stage	0.92	(0.25, 3.3)	0.9	
		PFI	Age	1.0	(0.98, 1.02)	1	
			TmS (High vs. Low in Advanced Stage)	2.4	(0.97, 6.0)	0.06	
			Stage/(TmS x Stage)	-	-	-	
Hepatocellular carcinoma	*n* = 292	OS	Age	1.0	(1.00, 1.03)	0.08	
			TmS (High vs. Low)	3.7	(0.51, 26)	0.2	
			Stage (Advanced vs. Early)	5.1	(0.67, 39)	0.1	
			TmS x Stage	0.62	(0.078, 5.1)	0.7	
		PFI	Age	1.0	(0.98, 1.0)	0.7	
			TmS (High vs. Low in Advanced Stage)	1.8	(0.83, 3.7)	0.1	
			Stage/(TmS x Stage)	-	-	-	
Lung adenocarcinoma	*n* = 380	OS	Age	1.0	(0.99, 1.03)	0.2	
			TmS (High vs. Low)	2.1	(1.2, 4.0)	0.02	
			Stage (Advanced vs. Early)	2.3	(0.86, 6.3)	0.1	
			TmS x Stage	0.99	(0.34, 2.9)	1.0	
		PFI	Age	1.0	(0.98, 1.01)	0.8	
			TmS (High vs. Low)	2.9	(0.93, 9.2)	0.1	
			Stage (Advanced vs. Early)	3.6	(1.1, 11.7)	0.04	
			TmS x Stage	1.2	(0.23, 6.4)	0.8	
Pancreatic adenocarcinoma	*n* = 98	OS	Age	1.0	(0.99, 1.0)	0.3	
			TmS (High vs. Low in Early Stage)	2.0	(1.2, 3.4)	0.01	
			Stage/(TmS x Stage)	-	-	-	
		PFI	Age	1.0	(0.98, 1.0)	0.8	
			TmS (High vs. Low in Early Stage)	6.0	(1.7, 20)	0.004	
			Stage/(TmS x Stage)	-	-	-	
Renal clear cell carcinoma	*n* = 291	OS	Age	1.0	(1.02, 1.06)	0.0002	
			TmS (High vs. Low)	2.3	(1.4, 4.0)	0.002	
			Stage (Advanced vs. Early)	4.5	(2.6, 7.7)	$5x10^{-8}$	
			TmS x Stage	-	-	-	
		PFI	Age	1.01	(0.99, 1.0)	0.3	
			TmS (High vs. Low)	2.6	(1.5, 4.3)	0.0004	
			Stage (Advanced vs. Early)	9.7	(5.5, 17)	$1x10^{-15}$	
			TmS x Stage	-	-	-	
Stomach adenocarcinoma	*n* = 238	OS	Age	1.0	(1.00, 1.05)	0.03	
			TmS (High vs. Low)	2.0	(0.85, 4.6)	0.1	
			Stage (Advanced vs. Early)	1.1	(0.23, 5.1)	0.9	
			TmS x Stage	2.0	(0.38, 10)	0.4	
		PFI	Age	0.99	(0.97, 1.01)	0.5	
			TmS (High vs. Low)	8.3	(1.1, 62)	0.04	
			Stage (Advanced vs. Early)	5.0	(0.59, 43)	0.1	
			TmS x Stage	0.41	(0.045, 3.7)	0.4	
Thyroid papillary carcinoma	*n* = 201	PFI	Age	1.0	(0.96, 1.06)	0.8	
			TmS (High vs. Low)	7.3	(0.88, 61)	0.07	
			Stage (Advanced vs. Early)	3.6	(0.32, 40)	0.3	
			TmS x Stage	1.2	(0.083, 16)	0.9	
Bladder urothelial carcinoma	*n* = 325	OS	Age	1.0	(1.0, 1.1)	0.04	Pattern II
			TmS (High vs. Low)	-	-	-	

			Stage/(TmS x Stage)	-	-	-
		PFI	Age	1.0	(0.99, 1.02)	0.4
			TmS (High vs. Low)	3.3	(0.79, 14)	0.1
			Stage (Advanced vs. Early)	9.1	(2.2, 37)	0.002
			TmS x Stage	0.18	(0.04, 0.79)	0.02
Head & neck squamous cell carcinoma (HPV-)	*n* = 193	OS	Age	1.0	(1.0, 1.1)	0.001
			TmS (High vs. Low)	4.4	(1.6, 12)	0.003
			Stage (Advanced vs. Early)	5.4	(2.3, 12)	$9x10^{-5}$
			TmS x Stage	0.14	(0.047, 0.40)	0.0003
		PFI	Age	1.0	(0.99, 1.0)	0.4
			TmS (High vs. Low)	0.55	(0.34, 0.91)	0.02
			Stage (Advanced vs. Early)	1.4	(0.75, 2.7)	0.3
			TmS x Stage	-	-	-
Lung squamous cell carcinoma	*n* = 420	OS	Age	1.0	(1.0, 1.03)	0.06
			TmS (High vs. Low)	2.8	(1.4, 5.8)	0.005
			Stage (Advanced vs. Early)	2.6	(1.7, 4.0)	$7x10^{-6}$
			TmS x Stage	0.16	(0.064, 0.42)	0.0002
		PFI	Age	1.0	(0.98, 1.0)	0.7
			TmS (High vs. Low)	2.2	(1.0, 4.5)	0.04
			Stage (Advanced vs. Early)	2.2	(1.5, 3.4)	0.0001
			TmS x Stage	-	-	-
ER-positive breast carcinoma	*n* = 624	OS	Age	1.1	(1.03, 1.07)	$6x10^{-8}$
			TmS (High vs. Low)	0.40	(0.14, 1.1)	0.09
			Stage (Advanced vs. Early)	2.7	(1.7, 4.4)	$6x10^{-5}$
			TmS x Stage	-	-	-
		PFI	Age	1.0	(0.99, 1.03)	0.2
			TmS (High vs. Low)	3.5	(1.4, 8.7)	0.007
			Stage (Advanced vs. Early)	2.4	(1.5, 4.0)	0.0007
			TmS x Stage	-	-	-
Triple-negative breast carcinoma†	*n* = 134	OS	Age	1.0	(0.97, 1.0)	0.9
			TmS + Stage	6.3	(2.6, 15)	$6x10^{-5}$
		PFI	Age	1.0	(0.96, 1.03)	0.5
			TmS + Stage	6.4	(2.7, 15)	$2x10^{-5}$
Renal papillary carcinoma†	*n* = 150	PFI	Age	0.97	(0.94, 1.0)	0.02
			TmS + Stage	7.6	(3.0, 19)	$1x10^{-5}$

Pattern III

OS: Overall survival; PFI: Progression-free interval.

"-" stands for missing coefficients due to lack of events (death or progression) in the corresponding patient group.

For the cancer types with "†", hazard ratios of TmS and Stage cannot be estimated separately due to lack of events. Instead, a simplified model comparing "Early stage, Low TmS" and "Advanced stage, High TmS" was used and denoted as "TmS + Stage".

Table S4A | Summary of risk categories defined by TmS x Gleason score for TCGA prostate adenocarcinoma and ICGC-EOPC.

Risk category	TCGA prostate adenocarcinoma (*n*=259)			ICGC-EOPC (*n*=79)		
	No. of samples	Age at treatment (mean/sd)	Percentage of disease progression at 5 years (95% CI)	No. of samples	Age at treatment (mean/sd)	Percentage of disease progression at 5 years (95% CI)
Gleason score=6	22	59 (6.9)	0	10	46 (5.1)	0
Gleason score=7, Low TmS	59	60 (6.4)	4.3 (0-10)	32	47 (2.4)	13 (0-26)
Gleason score=7, High TmS	75	60 (7.5)	13 (1.6-23)	27	47 (3.2)	36 (14-52)
Gleason score>=8, Low TmS	77	62 (6.6)	34 (3.4-55)	5	48 (3.8)	40 (0-71)
Gleason score>=8, High TmS	26	62 (6.2)	57 (37-70)	5	48 (2.2)	80 (0-97)

Table S4B | Multivariate Cox proportional hazard models with Age, TmS, Gleason score and TmS x Gleason score as predictors for progression free-interval analysis of TCGA prostate adenocarcinoma and ICGC-EOPC.

TCGA prostate adenocarcinoma	Hazard ratio	95% CI	*P* value (Wald test)	Age and Gleason score model	Age and TmS x Gleason score model	Difference
Age	1.0	(0.97, 1.1)	0.5	median iAUC		d-iAUC (95% CI)
TmS (High vs. Low)	2.9	(0.58, 14)	0.2	0.74	0.78	(-0.0018, 0.077)
Gleason score (>=8 vs. 7)	7.5	(4.0, 70)	0.01	median IBS		d-IBS (95% CI)
TmS x Gleason score	0.78	(0.13, 4.7)	0.8	0.11	0.11	(-0.0080, 0.0033)
ICGC-EOPC	Hazard ratio	95% CI	*P* value (Wald test)	Age and Gleason score model	Age and TmS x Gleason score model	Difference
Age	0.94	(0.80, 1.1)	0.5	median iAUC		d-iAUC (95% CI)
TmS (High vs. Low)	3.9	(1.0, 14)	0.04	0.68	0.81	(0.029, 0.26)
Gleason score (>=8 vs. 7)	6.6	(1.1, 40)	0.04	median IBS		d-IBS (95% CI)
TmS x Gleason score	1.2	(0.13, 10.7)	0.9	0.19	0.18	(-0.013, 0.0054)

The “Age and Gleason score model” was trained on the TCGA prostate adenocarcinoma data with Gleason score and age as predictors.

The “Age and TmS x Gleason score model” was trained on the TCGA prostate adenocarcinoma data with age and subgroups defined by TmS and Gleason score as predictors.