Bayesian Inference of Cell Composition and Gene Expression Reveals Tumor-Microenvironment Interactions

Tinyi Chu[1,2,*] and Charles G. Danko[1,3,*]

[1] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
[2] Graduate field of Computational Biology, Cornell University, Ithaca, NY 14853.
[3] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

[*] **Address correspondence to:**
Charles G. Danko, Ph.D.
E-mail: dankoc@gmail.com

Tinyi Chu, Ph.D.
E-mail: tc532@cornell.edu

Abstract

Understanding the dynamic interactions between malignant cells and the tumor stroma is a major goal of cancer research. Here we developed a Bayesian model that jointly infers both cellular composition and gene expression in each cell type, including heterogeneous malignant cells, from bulk RNA-seq using scRNA-seq as prior information. We conducted an integrative analysis of 85 single-cell datasets and 1,412 bulk RNA-seq samples in primary human glioblastoma, head and neck squamous cell carcinoma, and melanoma. We identified cell types correlated with clinical outcomes and explored regional heterogeneity in tumor state and stromal composition. We redefined common molecular subtypes using gene expression in malignant cells, after excluding confounding non-malignant cell types. Finally, we identified genes whose expression in malignant cells correlated with infiltration of macrophages, T-cells, fibroblasts, and endothelial cells across multiple tumor types. Our work provides a new lens that we used to measure cellular composition and expression in a statistically powered cohort of three primary human malignancies.

Introduction

Tumors are complex mixtures comprised of malignant cells as well as functionally diverse non-malignant cell types and extracellular matrix proteins known as the tumor stroma[1–3]. The importance of the stroma was first recognized more than a century ago, beginning with observations that certain cancers metastasize exclusively to specific organs[1]. During the past two decades numerous studies have revealed interactions between malignant cells and the stroma that promote diverse tumor functions including angiogenesis[4,5], metastasis[6], and immunosuppression[7,8]. Stromal cells differ between patients and tumor types[9–15] and the abundance of certain stromal cell populations are used in the clinic as biomarkers[16–19] and therapeutic targets[20–25]. These studies motivate direct measurements of cell types and the interactions between them in human subjects.

Two layers of information are critical for understanding the tumor stroma and its interactions with malignant cells[26]: (1) the quantity of different cell types, and (2) each cell type can have systematic differences in gene expression pathways, also called cell "state". Measurements of both cell type and state can be made using single cell RNA sequencing (scRNA-seq)[27–32]. However, few scRNA-seq datasets are available and many important questions still require orders of magnitude larger sample sizes to have sufficient statistical power. Other genomic studies have used thousands of bulk RNA-seq samples in the cancer genome atlas (TCGA) and other data repositories to infer cell type abundance[17,18]. Although these pioneering studies have revealed new insights into tumor infiltrating immune cells[17,18], they have not taken advantage of recently released tumor-matched scRNA-seq datasets, limiting analysis to immune cell types which can be sorted from peripheral blood, and reducing deconvolution accuracy[33–35]. Critically, these cell type deconvolution studies were not able to measure changes in gene expression state in the different populations of tumor cells.

These studies leave open several foundational questions: How does malignant cell state affect the composition of stromal cells? And which genes are responsible for driving these interactions? To answer these questions we devised a Bayesian statistical model that infers both cell type composition and gene expression, called tumor and stromal environment deconvolution and expression inference (TED). TED infers cell types and states from bulk RNA-seq data using a scRNA-seq reference as prior information. We used TED to identify cell types correlated with clinical outcomes, explored regional heterogeneity in tumor state and stromal composition, and systematically identified gene expression pathways in malignant cells that were correlated with the tumor stroma. Our results provide numerous insights into how stromal cells interact within the tumor, and provides an accurate new tool for investigating bulk tissue heterogeneity.

Results

Bayesian inference of cell type composition and tumor expression

TED uses a scRNA-seq reference dataset to infer two parameters of interest from bulk RNA-seq data: (i) the proportion of cell types in the bulk population and (ii) the average expression profiles of each cell type. TED describes the proportion and expression profiles of each cell type as latent variables that it infers from the data (**Fig. 1a, Supplementary Fig. 1, Supplementary Note 1**). TED makes the key simplifying assumption that each non-malignant

cell type shares a common gene expression profile across patients, as observed in the cases analyzed to date[28,30,31]. Critically, each bulk RNA-seq sample is then assumed to have a unique tumor expression profile that we infer from the data.

Expression in the reference and bulk RNA-seq data often differ substantially due to batch effects or tumor heterogeneity. To account for uncertainty in the reference cell type expression matrix, TED implements a fully Bayesian inference of tumor composition. First, TED uses Gibbs sampling to estimate the posterior joint distribution of cell type composition, θ_0, and gene expression profiles, Z, i.e. $P(\theta_0, Z \mid \varphi, X; \alpha)$ (**Fig. 1a, red, top**). Second, to account for tumor cells, or other cell types which cannot be observed in the reference dataset, TED infers a maximum likelihood estimate (MLE) for tumor expression profiles, ψ_{tum} (**Fig. 1a, red, mid**). During this step TED also infers the maximum a posterior estimate (MAP) of the expression profile of non-malignant stromal cells, ψ_{str}, to correct for batch effects between bulk and single cell RNA-seq platforms. Last, TED uses the updated expression profile for each patient and cell type to re-sample the posterior distribution of cell type composition, θ, i.e. $P(\theta \mid \psi_{tum}, \psi_{str}, X; \alpha)$ (**Fig. 1a, red, bottom**). Optionally, TED has an additional mode which can be used to learn common patterns of expression heterogeneity of malignant cells, after factoring out the confounding influence of stromal cells (**Fig. 1a, green**).

TED performance evaluation on pseudo-bulk RNA-seq data

To evaluate the performance of TED in a realistic setting we analyzed two glioblastoma multiforme (GBM) datasets that mimic differences between bulk RNA-seq and scRNA-seq reference data. One scRNA-seq reference analyzed 23,793 cells from 8 patients using a microwell-based platform[31] (GBM8), which sequenced tag clusters near the 3' end of polyadenylated genes, similar to other high-throughput scRNA-seq methods (e.g., Drop-seq, 10x genomics, etc). A second scRNA-seq dataset was available which sequenced 7,930 cells from 28 patients using the SMART-Seq2 platform[32] (GBM28), which sequenced full length mRNA transcripts to a high read depth in each cell, similar to most bulk RNA-seq datasets.

We generated a "pseudo-bulk" RNA-seq dataset from GBM28 for which the proportion of each cell type was known by combining scRNA-seq reads in each patient. Despite substantial differences between the test and reference dataset, TED inferred initial estimates of cell type proportions (θ_0) that were reasonably accurate using GBM8 as a reference (**Supplementary Fig. 2a-e**). These estimates were substantially improved during the final step of TED after updating expression estimates of malignant and stromal cells (**Fig. 1b, and Supplementary Fig. 2f-j**). To test TED across a wider range of different tumor compositions, we simulated 1,400 pseudo bulk RNA-seq samples by sampling random proportions of each cell type using GBM28 (see **Online Methods**). TED accurately inferred the proportion of both malignant and stromal cells in this task (**Fig. 1c**). The primary limitation of TED was that it underestimated the fraction of T-cells by ~25% due to low coverage in the reference dataset (<0.3% of cells in the reference, N = 67 / 23793). Nevertheless, the relative fraction of T-cells was highly correlated with the ground truth, allowing a robust comparison between different patients.

We compared TED to CIBERSORTx[36], a recently published program for cell type estimation. As CIBERSORTx was substantially slower than TED in cell type deconvolution (**Supplementary Fig. 3**), we compared the performance of TED and CIBERSORTx on GBM28 (n = 28) and a subset of the 1,400 GBM28 simulated pseudo bulk RNA-seq samples (n = 280)

using GBM8 dataset as a reference (**Fig. 1d-e**; **Supplementary Fig. 4**). TED was substantially more accurate than CIBERSORTx regardless of the mode of batch correction (mean squared error [mse] = 0.0036 for TED, 0.024-0.087 for CIBERSORTx). CIBERSORTx systematically over-estimated the proportion of malignant cells and underestimated stromal cell types, especially at lower tumor purity, whereas estimates produced by TED were accurate across the full range (**Fig. 1d-e; Supplementary Fig. 4**). Additional benchmarking against CIBERSORT[37] and custom implementations of ordinary least squares and support vector regression (**Supplementary Fig. 5; Supplementary Note 2**), which serve as the basis for existing deconvolution strategies[17,36–38], suggest that TED performed significantly more accurately because of improvements in the reference dataset, as well as improvements in the way TED modeled gene expression in malignant cells and accommodated batch effects.

In addition to cell type composition TED was also designed to infer gene expression in both malignant and stromal cells. In stromal cells, TED recovered gene expression levels that were highly correlated with the ground truth, even when the scRNA-seq reference expression differed substantially due to batch effects (**Fig. 1f; Supplementary Fig. 6**). TED also accurately recovered expression in malignant cells (**Fig. 1g**). TED achieved correlations >0.95 for tumors with >50% purity, which most high-quality RNA-seq cancer datasets surpass[39] (**Fig. 1h**). Malignant cell expression profiles were separated correctly by hierarchical clustering in all tumors with >50% purity (**Fig. 1i**), indicating that expression estimates can be used for downstream genomic analysis. Gene expression estimates were substantially more accurate using TED than using either CIBERSORTx or the bulk tumor with no deconvolution (**Fig. 1h and Supplementary Fig. 7**).

We extended the validation experiments to two other tumor types: melanoma and head and neck squamous cell carcinoma (HNSCC)[29,30]. Since only one scRNA-seq dataset was available for both cancers, we used a leave-one-out test, in which we generated a “pseudo-bulk” RNA-seq dataset from one patient, and asked how accurately TED deconvolved expression using the remaining datasets as a reference. As observed with GBM, TED consistently estimated cellular proportions that were similar to the true values (**Supplementary Fig. 8a-d**). TED and CIBERSORTx were more equally matched in leave-one-out test data which did not have batch effects and in which the number of tumor reference samples was large (#HNSCC=31; #melanoma=18), supporting our speculation that performance improvements in TED partially reflect more effective modeling of the uncertainty in the reference, including the batch effect and tumor heterogeneity. TED also produced gene expression estimates that were highly correlated with malignant cells (**Supplementary Fig. 8e-h**). Thus, we conclude that TED accurately infers cell type composition and expression in multiple cancers.

Cell type composition predicts clinical outcome in three malignancies

We analyzed the proportion of stromal cell types in 1,142 TCGA samples from three tumor types: GBM, HNSCC, and melanoma[40–42]. To maintain the highest possible accuracy for cell type proportions, we used the scRNA-seq reference from the same tumor type[29–31]. Using these reference datasets provided estimates of 6 cell types for GBM, 8 for HNSCC, and 8 for melanoma (**Fig. 2a**). Analysis using TED revealed that the majority of TCGA samples were comprised of >75% malignant cells in all three tumor types (**Fig. 2a**). GBM had, on average, the highest purity, consistent with previous estimates[39].

To determine how stromal cell types co-varied with each other, we examined the pairwise correlations between each cell type in the TCGA cohort. In GBM, the strongest correlation was between pericytes and endothelial cells (Spearman's rank correlation [ρ] = 0.35; **Fig. 2b**), consistent with their combined presence in vascular structures[43]. However, correlations were weaker overall in GBM than in other cancer types. In HNSCC, the proportion of immune cell types were highly correlated with each other (**Fig. 2c**). We also noted a high correlation between most immune cell types and endothelial cells. In melanoma, endothelial cells had a relatively strong positive association with fibroblasts (ρ = 0.6; **Fig. 2d**), which may be consistent with reports that fibroblast ECM remodelling promotes angiogenesis in melanoma[44]. We also noted two separate submodules of highly correlated immune cells, one consisting of CD4+ T-cells and NK cells, and the other consisting of CD8+ T-cells, macrophages, and B-cells (**Fig. 2d**). This finding suggests that melanoma patients may have two distinct types of immune response to varying degrees between patients.

The proportion of certain cell types were strongly associated with survival in all three cancer types, consistent with an important role for the microenvironment in clinical presentation. The proportion of T-cells was associated with better clinical outcomes in all three malignancies (hazard ratio [HR] = 0.416-0.604; **Fig. 2e-h**). In melanoma, where CD4+ and CD8+ cells were annotated separately in the reference scRNA-seq dataset, we found that CD8+ T-cells had a stronger correlation with survival (**Fig. 2g and h**). Macrophages were significantly associated with survival in both GBM and melanoma, but not in HNSCC (**Fig. 2i-k**). Intriguingly, however, high macrophage infiltration had a poor prognosis in GBM (HR = 1.71; **Fig. 2i**), but a substantially better prognosis in melanoma (HR = 0.556; **Fig. 2k**), indicating substantial heterogeneity in the role of macrophages in different malignancies.

TED also revealed substantial information about stromal cell types that have not been explored in prior deconvolution studies. First, the strongest association with survival in GBM was with oligodendrocytes, which mark poor clinical outcomes (HR = 2.27; **Fig. 2l**), suggesting that the presence of oligodendrocytes in GBM may interact with malignant cells in some way. Second, endothelial cells were marginally associated with better clinical outcomes in all three tumors (HR = 0.444 [GBM], 0.665 [HNSCC], and 0.515 [melanoma]; **Supplementary Fig. 9**). Taken together, these analyses reveal new information about heterogeneity in the microenvironment of three cancer types, and how these patterns correlate with clinical variables.

Spatial heterogeneity in cell type composition in GBM

To understand the spatial distribution of cells within a tumor, we applied TED to 122 bulk RNA-seq samples from the Ivy Glioblastoma Atlas Project (IVY-GAP) that interrogate laser microdissected tissue from GBM[45]. Data was available for ten tumors microdissected into five structures: leading edge (LE), infiltrating tumor (IT), cellular tumor (CT), microvascular proliferation (MVP) and pseudopalisading cells around necrosis (PAN) (**Fig. 3a and Supplementary Table 1a**). As we expected normal brain cells, including neurons, at high abundance in the leading edge and infiltrating tumor structures[45], we deconvolved all samples using a reference scRNA-seq dataset which includes GBM8 and adult neurons[46] (see Methods).

TED revealed several striking features of GBM regional cellular heterogeneity (**Fig. 3b**). First, pericytes and endothelial cells were significantly enriched in regions of microvascular

proliferation (p < 1e-4, linear mixed model, see Methods), and comprised nearly 60% of cells in these regions. Second, oligodendrocytes and adult neurons were enriched in the leading edge and infiltrating tumor (p < 1e-4, linear mixed model), with a relative magnitude that matches H&E stained sections from these same patients[45]. Third, pseudopalisading cells around necrosis showed a depletion of endothelial cells (p = 0.0057, linear mixed model) and enrichment for T cell infiltration (p = 0.0311, linear mixed model). Fourth, macrophages were higher in the cellular tumor than in the leading edge (p = 0.0381, linear mixed model).

To confirm these observations using an independent dataset, we analyzed 169 GBMs in the TCGA dataset. To identify tumors in the TCGA cohort enriched for anatomical features microdissected by IVY-GAP, we identified genes that were differentially expressed in malignant cells in PAN relative to CT and examined how these genes correlated with stromal cell proportions. We focused on validating differences between PAN and CT, because malignant cells were extremely rare in other anatomical regions (as low as 1%, based on H&E). As observed in the IVY-GAP dataset, we noted that genes up-regulated in PAN trended to have positive correlations with macrophages and T-cells, and negative correlations with endothelial cells (**Fig. 3c**), in agreement with the analysis using IVY-GAP data. Taken together, TED revealed a rich and highly heterogeneous picture of cell type composition, where immune cells accumulate in necrotic regions of the tumor depleted of the tumor microvasculature.

Tumor pathway embeddings identify new distinctions between molecular subtypes

We developed a module in TED which recovered core tumor pathways that best describe expression heterogeneity without contamination from non-malignant cell types (**Fig. 4a**). To accommodate recent observations that malignant cells in different tumors are heterogeneous mixtures of functionally distinct cell types[27,32,47], we modelled each patient as a linear combination of different pathways comprised of genes whose expression covaries in a similar manner. TED selected genes in each pathway and the weights for each tumor using the expectation maximization (EM) algorithm, such that the linear combination of all pathways most accurately approximates malignant cell expression in all patients.

To evaluate whether TED learned subtypes that reflect intra-tumor heterogeneity, we identified four pathways using the GBM28 pseudo-bulk RNA-seq dataset. TED recovered pathways that were highly correlated with those recently obtained by clustering 7,930 single cells from 28 patients[32] (**Fig. 4b**). Moreover, the weights of each pathway learned by TED were correlated with the fraction of cells in each tumor that represent each of the four subtypes (**Fig. 4c-d**). Thus, tumor pathways learned using TED can accurately identify major tumor cell subpopulations, even when the expression of subtypes are not known from direct single cell measurements.

To understand tumor heterogeneity in GBM using the most inclusive GBM cohort available to date, we next inferred pathways from 169 TCGA bulk RNA-seq samples. We decomposed the TCGA dataset into seven pathways, using the criterion that selected the pathway number, K, with the highest degree of consensus clustering[48] (**Supplementary Fig. 10**). TED revealed several pathways that were similar to those in previous studies[32,49,50], including pathway 2 (proneural, OPC, and NPC-like), 3 and 4 (mesenchymal), and 6 and 7 (classical and AC-like) (**Fig. 4e**). Notably, prior studies found no correlation between subtype and clinical outcomes in GBM, except when taking a subset of mesenchymal tumors[49]. Two of

the pathways discovered using TED were correlated with clinical outcomes: pathway 4, similar to the mesenchymal subtype (HR = 2.43, p = 0.001; **Fig. 4f**; **Supplementary Fig. 11a**), and pathway 6, which bore similarities to the classical subtype (HR = 0.428, p = 0.005; **Fig. 4g**; **Supplementary Fig. 11a**). Thus pathways identified by TED correlate with clinical outcomes better than existing molecular subtypes.

Comparison with IVY-GAP data revealed substantial regional heterogeneity in tumor pathways. Pathways-3 and 4 (both mesenchymal) were enriched for PAN regions, suggesting that malignant cells with a mesenchymal phenotype tend to accumulate in necrotic regions (**Fig. 4h**). Pathways-1 (not similar to previously discovered subtypes), 6, and 7 (classical) were enriched in cellular tumor. Previous studies have demonstrated plasticity between GBM cell subtypes[32], and we speculate that regional differences in subtype reflect interactions between GBM cells and local features of the tumor microenvironment.

Interactions between tumor pathways and stromal cell types

Motivated by our speculation that tumor pathways may be driven in part by interactions with the tumor microenvironment, we examined how tumor pathways learned by TED correlate with the composition of stromal cells. We focused on the 7 pathways identified in GBM, in which the only correlation reported is with macrophage infiltration in mesenchymal tumors[49]. TED recovered this known association as the strongest correlation observed, between pathway 4 (mesenchymal-like) and macrophages (**Fig. 5a**). We also noted a strong correlation between pathway 6 (classical-like) and endothelial cells, consistent with the IVY-GAP analysis (above) that implicates the classical molecular subtype as the most highly correlated with angiogenesis. Weaker associations were discovered between pericytes and both mesenchymal pathways (pathways 3 and 4). These associations may reflect the reported differentiation of mesenchymal tumor propagating cells into pericytes[51]. We also obtained broadly consistent results using several previously reported subtype definitions[32,49,50] (**Supplementary Fig. 12**).

We extended our analysis of GBM by learning pathway embeddings in HNSCC and melanoma (**Fig. 5b-c**). Consensus clustering led us to divide each tumor type into five pathways (**Supplementary Fig. 13 and 14**). As with GBM, several of these pathways were associated with clinical outcomes (**Fig. 5d-h**). Both HNSCC and melanoma had an anti-angiogenic pathway (pathway 5 [HNSCC] and pathway 2 [melanoma]), which strongly and inversely correlated with endothelial cells, as well as a pathway that correlated with cancer associated fibroblasts (pathway 2 [HNSCC] and pathway 5 [melanoma]). We also noted several differences in tumor composition between HNSCC and melanoma. HNSCC had a single pathway which was highly immunogenic (pathway 4; **Fig. 5b**) and associated with extended survival (HR = 0.418; **Fig. 5d**). In melanoma, multiple pathways correlated more weakly with immune infiltration (pathways 3, 4, and 5). Interestingly, pathway 1 was strongly and inversely correlated with infiltration of CD8+ T-cells, B-cells, and to a lesser extent with macrophages, but was positively correlated with NK cells (**Fig. 5c**). We found this pathway was strongly associated with poor survival (**Fig. 5e**). Taken together these results indicate a strong correspondence between malignant cell expression and the tumor microenvironment.

TED identifies genes involved in tumor-stroma interactions

To identify individual genes that may be involved in either regulating or responding to the stromal composition of tumors, we examined correlations between cell type proportions and gene expression in malignant cells. Correlations between gene expression and cell type composition in bulk data are prone to false positives when a gene is expressed highly and specifically in the cell type under investigation. We reasoned that TED recovered tumor cell expression accurately enough to avoid this type of false positive. To test this hypothesis, we examined the distribution of marker genes that were specifically expressed in each cell type based on independently derived data. Whereas marker genes had systematically higher correlations in the bulk data, these were reduced to a median near 0 when using the estimates of expression in malignant cells produced by TED (**Supplementary Fig. 15**). This suggests that TED expression estimates are effective in removing many trivial false positives.

Next we examined genes correlated with macrophage infiltration in GBM. Several known positive regulators of macrophage infiltration have been reported in GBM[52,53], and these all had statistically significant positive correlations with macrophage inflation, including *POSTN*, *ITGB1* and *LOX* (**Fig. 6a**). However, we identified numerous other correlations with a stronger magnitude, including *CASP5*, *GNG10*, *RIPK3*, and *PLB1*. To validate additional correlations using independent data, we asked whether tumor regions expressing high levels of positively correlated genes tended to have higher macrophage infiltration. We analyzed 148 bulk RNA-seq samples from 34 GBMs that were collected adjacent to sections analyzed by in situ hybridization (ISH) for tumor propagating cell markers[45]. We asked whether the proportion of macrophages estimated from RNA-seq using TED was higher in ISH positive regions of the tumor compared to ISH negative regions. Despite low power in the IVY-GAP dataset, we observed significantly higher macrophage content in ISH positive sections for three of the five genes analyzed, *PI3*, *TNFAIP3* and *POSTN* (**Fig. 6b-d**, **Supplementary Table 1b**). Thus TED identified correlations using TCGA that could be reproduced by the regional heterogeneity within a tumor using independent estimates of gene expression.

Consistent gene stroma interactions in three tumors

We hypothesized that similar genes are involved in stromal cell type interactions across malignancies. We asked whether similar genes correlated with cell proportions in GBM, HNSCC, and melanoma using statistical tools to evaluate the enrichment of intersections between all tumor types[54]. Genes that had a statistically significant positive correlation in one tumor were strongly enriched for positive correlations in one or both of the other two tumor types, and the same was observed for genes with negative correlations ($p < 0.001$, super exact test; **Fig. 6e**).

For most stromal cell types, there were no significant intersections that had both positive and negative correlations. However, there was one interesting exception to this rule: CD4+ and CD8+ T-cells were correlated with somewhat different sets of genes. In melanoma, there was a statistically significant intersection consisting of 212 genes which had a negative correlation with CD8+ cells and a positive correlation with CD4+ cells ($p < 0.001$, super exact test; **Fig. 6e**). Thirty-eight of these 212 (18%) were enriched in keratinization pathways (17-fold enrichment; $p < 7.5e-30$; Fisher's exact test). Tissue stiffness affects a variety of T-cell responses[55,56], and thus one interpretation of our results is that keratinization by malignant cells affects tumor stiffness and has different effects on CD4+ and CD8+ T-cells. Taken together these results

support similar genes that interact with stromal cell types between different tumors, and reveals differential modes of interaction between melanomas and different T-cell subsets.

Next we examined the genes in the overlaps that were most enriched between different tumors (**Fig. 6f**). Notably, *UBD* and *IDO1* were positively correlated with both macrophages and T-cells. *IDO1* encodes an enzyme that catalyzes the conversion of tryptophan into kynurenine, resulting in the activation of T regulatory cells and myeloid-derived suppressor cells[57]. Endothelial cells were correlated with several genes known to be involved in angiogenesis, including *ANGPT4*, *CP*, and *VASH2*[58,59]. The top gene correlating with fibroblasts was *LOXL2*, which is a factor secreted by tumor cells that promotes proliferation of fibroblasts[60]. Several other extracellular proteins, including *JAM2*, *PRND*, and *FIBIN* were correlated with fibroblasts which have not, to our knowledge, been directly implicated in fibroblast deposition in tumors. Taken together with previous literature, these results show that TED recovers genes known to interact with stromal cells in cancer.

Discussion

A large body of literature now provides numerous examples of stromal influence on malignant cell function[25,26], confirming more than a century of speculation about the critical role of the stroma[1]. However, our knowledge remains largely anecdotal and based mostly on work in animal models rather than human subjects. scRNA-seq has recently made it possible to measure both cell types present in the tumor and their gene expression states in a systematic manner[26]. Although scRNA-seq provides the right data modality to chart the various ways in which tumor-stroma interactions occur, current studies are underpowered to address these questions in a statistically rigorous manner. In parallel, thousands of bulk RNA-seq datasets are now available that provide weak information about the entire cellular milieu in a variety of malignancies[40–42]. Here we leveraged these advancements in genomic resources by developing a rigorous statistical modeling strategy and using it to integrate scRNA-seq data from 37 thousand cells over 85 patients and 1,412 bulk RNA-seq samples, providing a new lens into both the cell type and expression in three human cancers.

Our analysis revealed numerous examples in which systematic differences in malignant cell gene expression pathways correlated with the presence of specific stromal cell types. Although different tumor types have unique somatic mutations and transcription states, we identified substantial overlap in the genes that were correlated with stromal cell types, suggesting that a few key pathways are used to control malignant and stromal cell communication. Our findings suggest that therapies targeting a few key genes could have broad impact in manipulating tumor-microenvironment interactions in multiple tumor types.

Many stromal cell types and tumor expression pathways correlate with clinical outcomes, highlighting how tumor-stroma interactions affect tumor phenotype. T-cell infiltration was associated with a better prognosis in all three of the tumors we examined. This was consistent with prior reports in melanoma and HNSCC[19,61], but to our knowledge this is a novel finding in GBM that was likely missed by previous studies because T-cells are so rare in GBM.

Our modeling approach fills several critical needs in the cancer genomics toolbox. TED more accurately deconvolves bulk RNA-seq into the proportion of cell types than previous approaches thanks in part to the Bayesian statistical model which allows the scRNA-seq reference to have substantial expression differences from the bulk data. Most importantly TED

is not just a deconvolution algorithm - it jointly infers cell types and their average expression, which was crucial for analyses reported herein. Thus TED provides a new type of lens for integrating new scRNA-seq data with the statistically powered cohorts of bulk RNA-seq data, allowing insights into interactions between malignant and stromal cells.

Acknowledgements

We thank Peter Sims for sharing the cell type annotation, Xin Bing for discussions on the statistical inference, and Mariano Viapiano for critical discussions and biological insights. Work in this publication was supported by R01-HG009309 (NHGRI) to CGD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Author Contributions

The project was conceived by TC with input and advice by CGD. TC developed and implemented TED, conducted all analyses, and wrote the first draft of the manuscript. CGD supervised all work and edited the paper.

Competing financial interests

The authors declare no competing financial interests.

Methods

Overview of TED

A complete mathematical description and justification of TED is included in **Supplementary Note 1**. Here we provide a brief summary of TED and its use in this manuscript. The R package of TED can be downloaded at https://github.com/Danko-Lab/TED.git**.**

TED is comprised of two functional modules (**Fig. 1a**): (1) a module that infers the cell type fractions, denoted by θ, and gene expression of each cell type in each bulk RNA-seq sample, denoted by Z, and (2) a module designed to identify commonly occurring subtype clusters after removing gene expression in stromal cells that are influtrating the tumor. Both modules take as input a reference matrix, φ, that describes gene expression in each cell type that is constructed from scRNA-seq data, and a matrix, X, representing gene expression in all available bulk RNA-seq samples. The second module additionally depends on the output of the deconvolution module.

In the deconvolution module, TED first obtains an initial estimate of the fraction of each cell type. TED uses the Gibbs sampling, a method for Markov chain Monte Carlo (MCMC) estimation, to approximate the joint posterior distribution of cell type proportion and gene expression, and then takes the mean over Gibbs samples to estimate the posterior distribution of Z. Next, TED estimates gene expression in all cell types. TED assumes the gene expression profiles for each cell type follow the multinomial distribution. It infers multinomial parameters of the tumor expression profiles ψ_{tum} in each bulk RNA-seq sample using maximum likelihood estimation. As TED assumes that the stromal cells share the same expression profiles across patients, allowing it to pool the statistical strength across bulk RNA-seq samples. TED infers a maximum a posterior estimator for the parameters of the multinomial distribution that control the expression profiles of stromal cells across all bulk RNA-seq samples ψ_{str}. The cell type fractions are then updated by re-sampling θ conditional on ψ_{tum} , ψ_{str} and X.

The second module of TED was designed to identify gene expression patterns that arise commonly among bulk RNA-seq samples after removing stromal cells influtrating the tumor. TED learns a series of latent embeddings, called tumor bases (denoted by η), chosen such that their linear combination best approximates gene expression levels in malignant cells. The learning module takes the input K vectors of tumor bases η_0 as an initialization, and uses the expectation-maximization (EM) algorithm to optimize the tumor bases by maximizing the log of the posterior of η, conditional on X and the cell type proportions and expression of stromal cells, i.e. θ_{str} and ψ_{str} inferred by the deconvolution module. We used the non-negative matrix factorization approach followed by consensus clustering on ψ_{tum} to approximate η_0 and selected the number of clusters, K, that yields the most consensus structure[48]. TED then uses EM to determine the pathways whose linear combination best estimates gene expression in the observed bulk RNA-seq malignant cells. In the E step TED uses the Gibbs sampling to approximate the posterior distribution of the cell type expression Z and the weights ω associated with each tumor basis. In the M step TED uses the conjugate gradient method to optimize the expectation of the log posterior of η with respect to the distribution of Z and ω that were approximated in the E step. At convergence, TED runs a final Gibbs sampling to derive the distribution of ω under the maximum a posterior estimates of η, and use its mean to get a point estimator.

Deconvolution of bulk RNA-seq using TED

Generating the reference expression profiles from scRNA-seq data. We used reference expression profiles generated from scRNA-seq data to deconvolve the bulk RNA-seq data of the corresponding tumor type. We collapsed, i.e. summed up, the raw read counts whenever count data is available (for ref[30,31]). For data where only TPM normalized data is available (scHNSCC), we collapsed TPM normalized reads. To generate the reference profiles of stromal cells, we collapsed read count in each cell type across all patients. To account for the heterogeneities in malignant cells, to generate tumor expression references, we collapsed expression of each subcluster of tumor cells generated by PhenoGraph[62] in each individual patient, whenever tumor cells are clustered (refGBM8, 60 subclusters in total for 8 patients). For datasets where tumor cells were not clustered by the original paper (scHNSCC and scMel), we collapsed expression of tumor cells in each patient. We found the expression of some of the non-coding genes in TCGA were close to zero across all patients, but were consistently highly expressed in refGBM8. Such systematic batch effects that violate the relative expression between cell types are unlikely to be corrected by TED, and hence we subset the inference on protein-coding genes when deconvolving TCGA data. In addition genes on the Y chromosome are also excluded in the reference to avoid sex-specific transcriptions. The collapsed expression profiles were normalized by the total count across each cell. To avoid exact zeros in the reference profile, we added a customized pseudo count to each cell type (provided as the norm.to.one function by the TED package), such that genes with zero expression have the same small value (default=10^{-8}) across all cell types after normalization.

Choice of hyper-parameters and retrieving the output from TED. We set the default parameters of TED to: the standard deviation of the log-normal distribution σ=2, and sparse dirichlet prior $\alpha=10^{-8}$ and used these defaults throughout the present study. We used the default setting for Gibbs sampling as follows: length of chain = 150, burn in = first 50 and thinning = 2 (i.e., we ran an MCMC chain of 150 samples, discarded the first 50, and used every other sample to estimate parameters of interest). The maximum number of iterations of conjugate gradient method was set to 10^5. All cell type fractions used were the updated θ.

Statistical tests for cell type fractions. When comparison is done for two groups, we used the two-sided Wilcoxon test. For comparisons between multiple groups, we used one-way ANOVA using the built-in function “aov” in R. For ANOVA F test statistics that passed alpha level (p value < 0.05), we used the function TukeyHSD to perform multiple testing-corrected pairwise tests based on the studentized range statistics.

Simulating pseudo-bulk RNA-seq data for GBM

For each patient among the 28 GBM patients[32] we simulated 50 pseudo-bulk RNA-seq samples, totalling up to 1,400 samples. The cell type fractions were drawn from a uniform dirichlet distribution (α=1). The tumor cells in each simulated sample were drawn from one particular patient with replacement, while the stromal cells were drawn from pooled patients with replacement. As raw data were TPM normalized, we rounded up the counts after summing them up across each cell.

Benchmarking CIBERSORTx

We used the online portal of CIBERSORTx (https://cibersortx.stanford.edu) to perform all the benchmarking. To create the signature matrix, we provided the same single cell reference matrix file used to prepare the refGBM8 of TED. Each column of the single cell reference matrix denotes a cell phenotype, which corresponds to either a sub-cluster of tumor cells in a particular patient or a non-malignant cell type. Same as in the refGBM8, there are 60 tumor cell phenotypes and 5 non-malignant cell phenotypes. All parameters were used at their default values, except for the "Min. Expression", which was set to 0 for refGBM8, following the author recommendations for droplet-based platforms. To impute the cell fractions and expression, we used all default setting, and the quantile normalization is disabled by default. S mode failed for the GBM280 dataset, due to memory overflow. As S mode and B mode batch correction worsened the performance in GBM28 (S and B mode) and GBM280 (B mode), we diabled batch correction in imputing expression. To compute the expression of tumor cells in each simulated sample using the group mode, we matrix-multiplied the unfiltered GEPs.txt and Weights.txt over 60 tumor sub-clusters. As the high-resolution mode of CIBERSORTx only allows the imputation for maximum of 1000 genes, we selected the top 1000 most variable genes across the merged tumor expressions of 28 GBM patients after variance-stabilizing transformation. We added up the imputed expression values across 60 tumor sub-clusters to get the tumor expression in each sample. Only 53 genes are imputable across all tumor sub-cluster references (by excluding the "1" and "NA" values in the CIBERSORTxHiRes_job1_PJ0XX-tumor-X_Window140.txt). The spearman's rank correlation coefficients in **Supplementary Figure 7** were computed on these 53 genes for four different approaches.

For leave-one-out tests on HNSCC and melanoma, all parameters were under the default settings. Batch effect correction is disabled for leave-one-out tests. Expression inference cannot be done, as CIBERSORTx requires the number of mixtures to be greater than the number of reference components.

Choosing cell type marker genes for correlation analysis

In Supplementary Figure 14. We computed the Pearson's correlation coefficient between the variance-stabilized transformed expression over a set of marker genes with the cell type fractions. Marker genes for CD4+ and CD8+ T cell and monocytes were derived from the LM6 matrix from the CIBERSORT website (https://cibersort.stanford.edu/download.php), which were based on based on GSE60424[63], by assigning each gene to the cell type with the maximum expression value. Markers for oligodedrocytes, endothelial cells, pericyted, and microglia were derived from the gene list generated by Lake et al. using normal brain scRNA-seq[64]. Only marker genes that are uniquely assigned to each cell type were used for the plot.

Analysis of anatomically resolved transcriptomics data from IVY GAP

Anonymized BAM files for each sample were downloaded from glioblastoma.alleninstitute.org, and raw counts for each gene were obtained using featureCounts[65] using the GENCODE annotation v24lift37.

To test the statistical significance in the mean of cell type fractions across multiple anatomic structures while taking account of the multiple biological replicates of each patient, we fit a linear mixed model using the lme function from the R package nlme[66] with random intersect. We modeled anatomic structures as the fixed effects and patient IDs as random effects. The ground level was set to the cellular tumor (CT). We maximized the log-likelihood function by setting the method as “ML”. We used “optim” as the optimizer.

To quantify the level of differential transcription between PAN and CT. We first estimated the tumor expression profile by TED using GBM8[31] as the reference. The estimated expression profile was rounded up to the closest integer for differential expression analysis by DESeq2. To account for the patient-specific means in the transcription level, we incorporated patient IDs as an independent variable in the model, which resulted in a design formula of design= ~patient ID + anatomic structure. We used the adjusted Wald test statistics to define genes that were differentially transcribed ($p<0.01$) or unchanged ($p>0.5$).

Survival analysis

We divided patients into high and low groups based on the feature of interest, e.g. weights of tumor pathways or stromal cell fractions, and then computed the hazard ratio by fitting a Cox proportional hazards regression model for survival time of patients in these two groups. We used two approaches to define a cutoff. First we reported the hazard ratio at the threshold between 0.1 quantile and 0.9 quantile that gave the lowest two-sided p-value between survival times using a Chi-squared test. This ensured that we reported the largest possible difference in survival time for each individual feature. As this scanning threshold method may suffer from inflated false positives due to multiple testing, we also used a second approach which was dividing patients into the upper and lower 20% quantiles, which ensures that all genes were fit for the regression model using a roughly balanced number of patients. When applying the decision rule in testing the null hypothesis, we took the results from both approaches into consideration.

Dataset used

Dataset name	Annotation	Normalization	# of cells	# of patients	Clustered tumor cells	Accession ID
refGBM8	GENCODE V24	Raw (UMI)	23793	8	YES	GSE103224
GBM28	UCSC Refgene	TPM	7930	28	NO	GSE131928
scMel	UCSC Refgene	Raw	6879	31	NO	GSE115978
scHNSCC	UCSC Refgene	TPM	5902	18	NO	GSE103322
TCGA bulk						

TCGA-GBM	GENCODE V22	Raw	NA	169	NA	https://portal.gdc.cancer.gov
TCGA-SKCM	GENCODE V22	Raw	NA	471	NA	
TCGA-HNSC	GENCODE V22	Raw	NA	502	NA	
IVY GAP						
IVY Anatomic Structures RNA-Seq	GENCODE v24lift37	Raw	NA	122 samples across 10 tumors	NA	https://glioblastoma.alleninstitute.org
IVY Cancer Stem Cells RNA-Seq	GENCODE v24lift37	Raw	NA	148 samples across 34 tumors	NA	

References

1. Paget, S. THE DISTRIBUTION OF SECONDARY GROWTHS IN CANCER OF THE BREAST. *The Lancet* vol. 133 571–573 (1889).
2. Greene, H. S. & Harvey, E. K. THE RELATIONSHIP BETWEEN THE DISSEMINATION OF TUMOR CELLS AND THE DISTRIBUTION OF METASTASES. *Cancer Res.* **24**, 799–811 (1964).
3. Auerbach, R. *et al.* Specificity of adhesion between murine tumor cells and capillary endothelium: an in vitro correlate of preferential metastasis in vivo. *Cancer Res.* **47**, 1492–1496 (1987).
4. Crawford, Y. *et al.* PDGF-C mediates the angiogenic and tumorigenic properties of fibroblasts associated with tumors refractory to anti-VEGF treatment. *Cancer Cell* **15**, 21–34 (2009).
5. Kobayashi, H. *et al.* Cancer-associated fibroblasts in gastrointestinal cancer. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 282–295 (2019).
6. Murgai, M. *et al.* KLF4-dependent perivascular cell plasticity mediates pre-metastatic niche formation and metastasis. *Nat. Med.* **23**, 1176–1190 (2017).
7. Sena, I. F. G. *et al.* Glioblastoma-activated pericytes support tumor growth via immunosuppression. *Cancer Medicine* vol. 7 1232–1239 (2018).
8. Paiva, A. E. *et al.* Pericytes in the Premetastatic Niche. *Cancer Res.* **78**, 2779–2786 (2018).
9. Brubaker, D. B. & Whiteside, T. L. Localization of human T lymphocytes in tissue sections by a rosetting technique. *Am. J. Pathol.* **88**, 323–332 (1977).
10. Richters, A. & Kaspersky, C. L. Surface immunoglobulin positive lymphocytes in human breast cancer tissue and homolateral axillary lymph nodes. *Cancer* **35**, 129–133 (1975).
11. Hersh, E. M., Mavligit, G. M., Gutterman, J. U. & Barsales, P. B. Mononuclear cell content of human solid tumors. *Med. Pediatr. Oncol.* **2**, 1–9 (1976).
12. Russell, S. W., Doe, W. F. & Cochrane, C. G. Number of macrophages and distribution of mitotic activity in regressing and progressing Moloney sarcomas. *The Journal of Immunology* (1976).
13. Folkman, J., Merler, E., Abernathy, C. & Williams, G. Isolation of a tumor factor responsible for angiogenesis. *J. Exp. Med.* **133**, 275–288 (1971).
14. Sidky, Y. A. & Auerbach, R. Lymphocyte-induced angiogenesis in tumor-bearing mice. *Science* **192**, 1237–1238 (1976).
15. Schor, S. L. *et al.* Occurrence of a fetal fibroblast phenotype in familial breast cancer. *Int. J. Cancer* **37**, 831–836 (1986).
16. Cao, Y. *et al.* Pericyte coverage of differentiated vessels inside tumor vasculature is an independent unfavorable prognostic factor for patients with clear cell renal cell carcinoma. *Cancer* **119**, 313–324 (2013).
17. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
18. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812–830.e14 (2018).
19. Fu, Q. *et al.* Prognostic value of tumor-infiltrating lymphocytes in melanoma: a systematic review and meta-analysis. *Oncoimmunology* **8**, 1593806 (2019).
20. Sandler, A. *et al.* Paclitaxel–Carboplatin Alone or with Bevacizumab for Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **355**, 2542–2550 (2006).
21. Ostermann, E. *et al.* Effective immunoconjugate therapy in cancer models targeting a serine protease of tumor fibroblasts. *Clin. Cancer Res.* **14**, 4584–4592 (2008).
22. Massoud, R. V., Vivian Massoud, R., Solimando, D. A. & Aubrey Waddell, J. Leucovorin, Fluorouracil, and Irinotecan (FOLFIRI) plus Bevacizumab for Metastatic Colorectal Cancer. *Hospital Pharmacy* vol. 46 748–754 (2011).
23. Waldhauer, I. *et al.* Novel tumor-targeted, engineered IL-2 variant (IL2v)-based

immunocytokines for immunotherapy of cancer. (2013).
24. Xu, C. *et al.* Interferon-α-secreting mesenchymal stem cells exert potent antitumor effect in vivo. *Oncogene* **33**, 5047–5052 (2014).
25. Hinshaw, D. C. & Shevde, L. A. The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Res.* **79**, 4557–4566 (2019).
26. Suvà, M. L. & Tirosh, I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Mol. Cell* **75**, 7–12 (2019).
27. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
28. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
29. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611–1624.e24 (2017).
30. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984–997.e24 (2018).
31. Yuan, J. *et al.* Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Medicine* vol. 10 (2018).
32. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835–849.e21 (2019).
33. Newman, A. M., Gentles, A. J., Liu, C. L., Diehn, M. & Alizadeh, A. A. Data normalization considerations for digital tumor dissection. *Genome biology* vol. 18 128 (2017).
34. Li, B., Liu, J. S. & Liu, X. S. Revisit linear regression-based deconvolution methods for tumor gene expression data. *Genome biology* vol. 18 127 (2017).
35. Zheng, S. Benchmarking: contexts and details matter. *Genome Biol.* **18**, 129 (2017).
36. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
37. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
38. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
39. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
40. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
41. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).
42. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
43. Bergers, G. & Song, S. The role of pericytes in blood-vessel formation and maintenance. *Neuro. Oncol.* **7**, 452–464 (2005).
44. Goldstein, L. J., Chen, H., Bauer, R. J., Bauer, S. M. & Velazquez, O. C. Normal human fibroblasts enable melanoma cells to induce angiogenesis in type I collagen. *Surgery* **138**, 439–449 (2005).
45. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).
46. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–7290 (2015).
47. Hovestadt, V. *et al.* Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).
48. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).

49. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42–56.e6 (2017).
50. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
51. Cheng, L. *et al.* Glioblastoma stem cells generate vascular pericytes to support vessel function and tumor growth. *Cell* **153**, 139–152 (2013).
52. Zhou, W. *et al.* Periostin secreted by glioblastoma stem cells recruits M2 tumour-associated macrophages and promotes malignant growth. *Nat. Cell Biol.* **17**, 170–182 (2015).
53. Chen, P. *et al.* Symbiotic Macrophage-Glioma Cell Interactions Reveal Synthetic Lethality in PTEN-Null Glioma. *Cancer Cell* **35**, 868–884.e6 (2019).
54. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5**, 16923 (2015).
55. Judokusumo, E., Tabdanov, E., Kumari, S., Dustin, M. L. & Kam, L. C. Mechanosensing in T lymphocyte activation. *Biophys. J.* **102**, L5–7 (2012).
56. Saitakis, M. *et al.* Different TCR-induced T lymphocyte responses are potentiated by stiffness with variable sensitivity. *Elife* **6**, (2017).
57. Liu, M. *et al.* Targeting the IDO1 pathway in cancer: from bench to bedside. *Journal of Hematology & Oncology* vol. 11 (2018).
58. Raju, K. S., Alessandri, G., Ziche, M. & Gullino, P. M. Ceruloplasmin, copper ions, and angiogenesis. *J. Natl. Cancer Inst.* **69**, 1183–1188 (1982).
59. Xue, X. *et al.* Vasohibin 2 is transcriptionally activated and promotes angiogenesis in hepatocellular carcinoma. *Oncogene* **32**, 1724–1734 (2013).
60. Mahjour, F. *et al.* Mechanism for oral tumor cell lysyl oxidase like-2 in cancer development: synergy with PDGF-AB. *Oncogenesis* **8**, 34 (2019).
61. Schneider, K. *et al.* Immune cell infiltration in head and neck squamous cell carcinoma and patient outcome: a retrospective study. *Acta Oncol.* **57**, 1165–1172 (2018).
62. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
63. Linsley, P. S., Speake, C., Whalen, E. & Chaussabel, D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS One* **9**, e109760 (2014).
64. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
65. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
66. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. (2019).

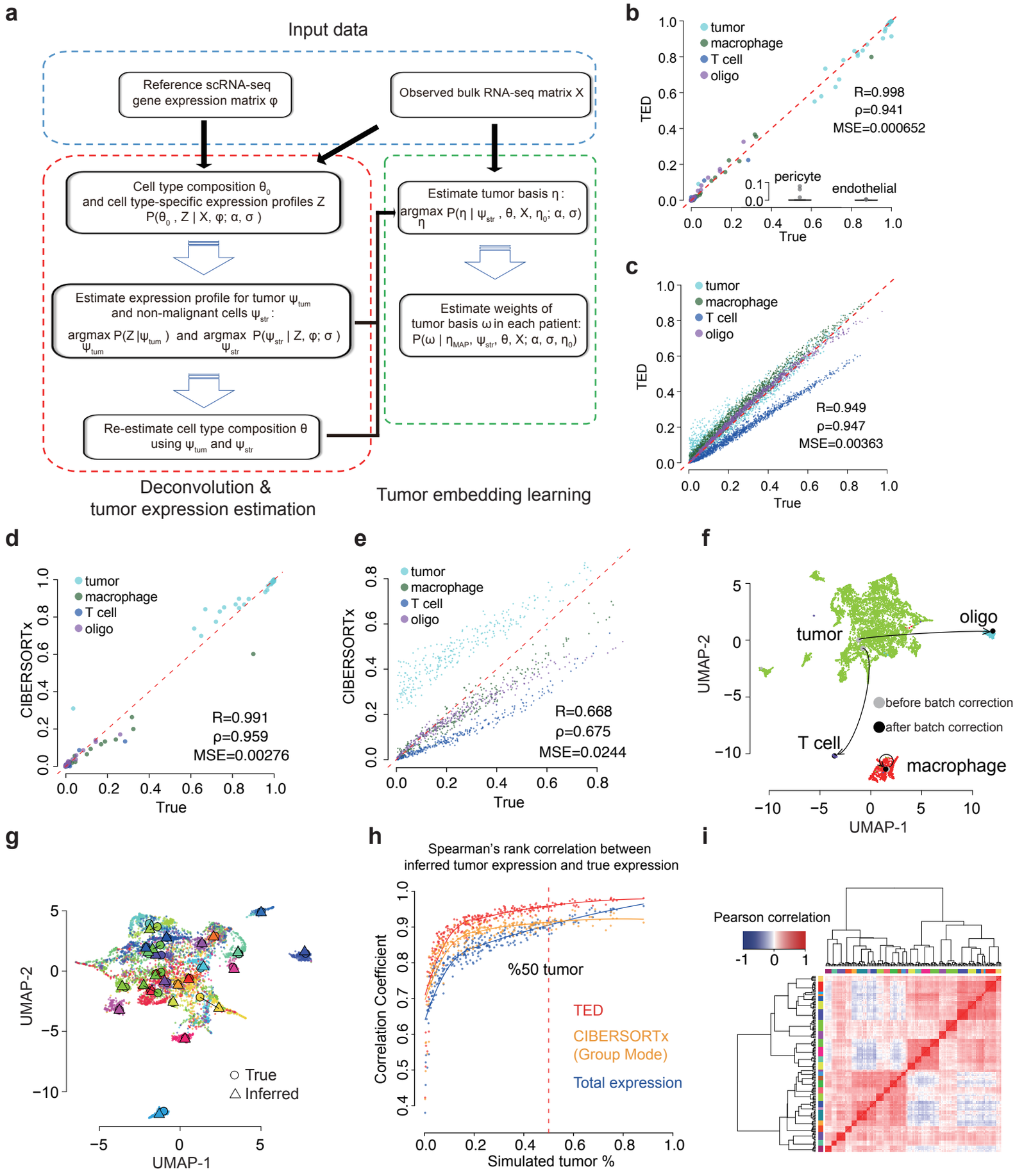

Fig. 1 | TED algorithm flow and performance validation. a) Algorithmist flow of TED. **b**) Scatter plot shows the cell type fraction predicted by TED versus the true value of the pseudo-bulk GBM28 dataset. R denotes Pearson's correlation coefficient; ρ denotes Spearman's rank correlation coefficient; MSE stands for mean squared error. Boxplot (insert) shows predicted cell type fractions for cell types not annotated in GBM28. Gray dots mark outliers defined by as >1.5 times the interquartile range. **c**) Scatter plot shows the cell type fraction predicted by TED versus the true value of the 1400 pseudo-bulk RNA-seq datasets simulated by drawing from GBM28 dataset. **d**) Scatter plot shows the cell type fraction predicted by CIBERSORTx and the true value of the pseudo-bulk GBM28 dataset and **e**) the down-sampled 280 out of the 1400 boot-strapped pseudo-bulk GBM28 dataset. **f**) UMAP visualization shows the expression of individual cells in GBM28 colored by their cell type. The expression profiles of stromal cells before (gray) and after (black) batch effect correction were projected onto the UMAP manifold of the scRNA-seq. **g**) UMAP shows the expression profile of individual malignant cells in GBM28 colored by patient. Patients with greater than 10 malignant cells (N=27) are shown. The inferred expression profile, shown as △, and the averaged expression profile from scRNA-seq for each patient, shown as ○ , are projected onto the UMAP manifold. **h**) Scatter plot shows Spearman's correlation between gene expression estimated by TED (red), CIBERSORTx group mode (orange) or total bulk (blue) and the average expression from malignant cells in scRNA-seq as a function of the fraction of malignant cells in the dataset containing the 280 simulated samples. **i**) Heatmap shows the correlation matrix between TED estimated gene expression of simulated pseudo-bulk RNA-seq with greater than 50% tumor cells. The heatmap is ordered by hierarchical clustering as shown by the dendrogram. The color bars indicate patient IDs.

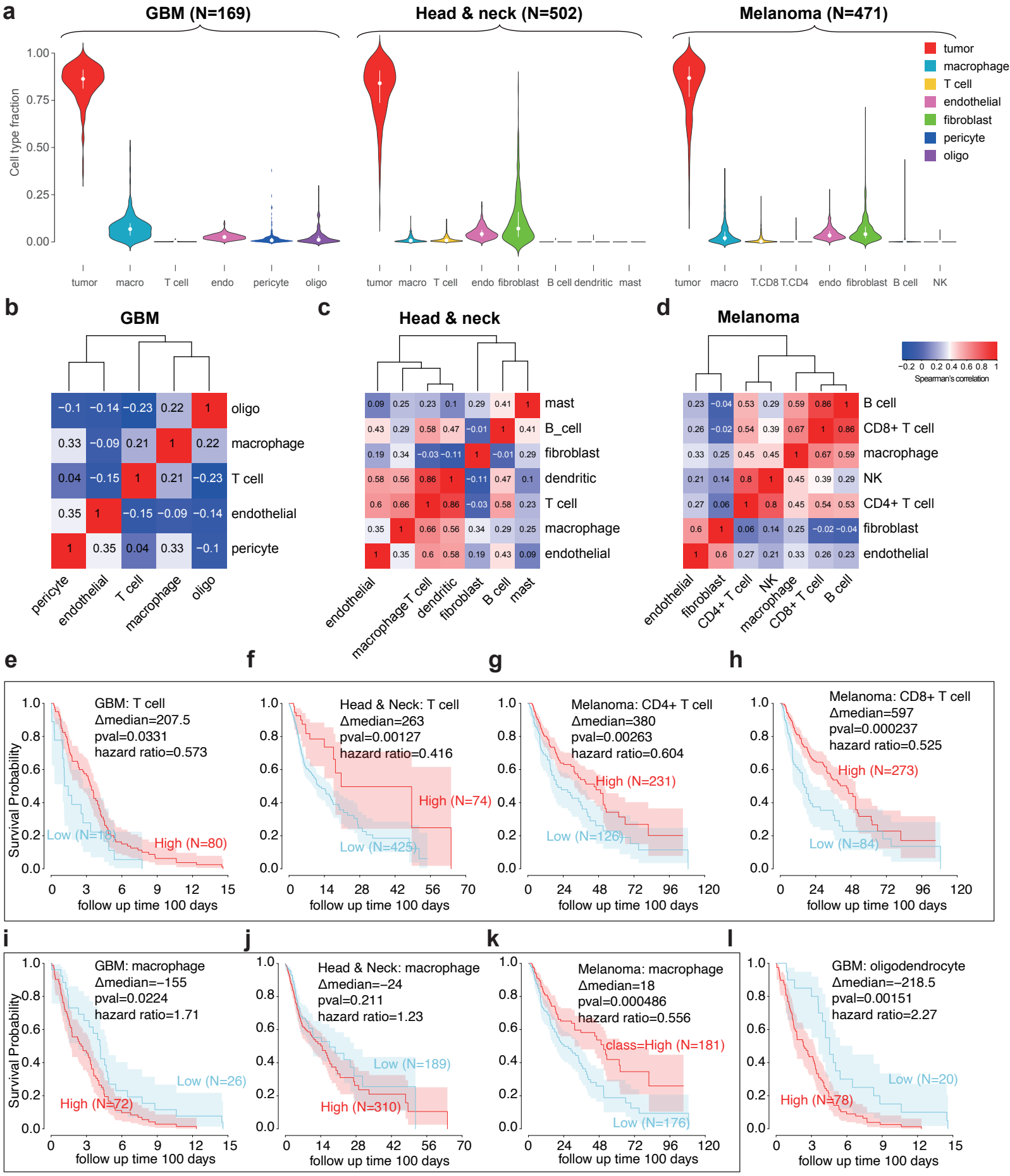

Fig. 2 | Cell type compositions in three tumor types. a) Violin plots show the distribution of cell type fractions in each tumor type. Median fractions are shown by white dots and upper/lower quartiles are shown by bars. **b**-**d**) Heatmaps show the Spearman's rank correlation between stromal cells in each tumor type. **e**-**l**) KM plots show the survival associations with **e**-**h**) T cell infiltration, **i**-**k**) macrophage infiltration, and **l**) oligodendrocytes. Δmedian: median survival time in the high group - median survival time in the low group. P values were computed using the log-rank test. Hazard ratio is defined by high / low.

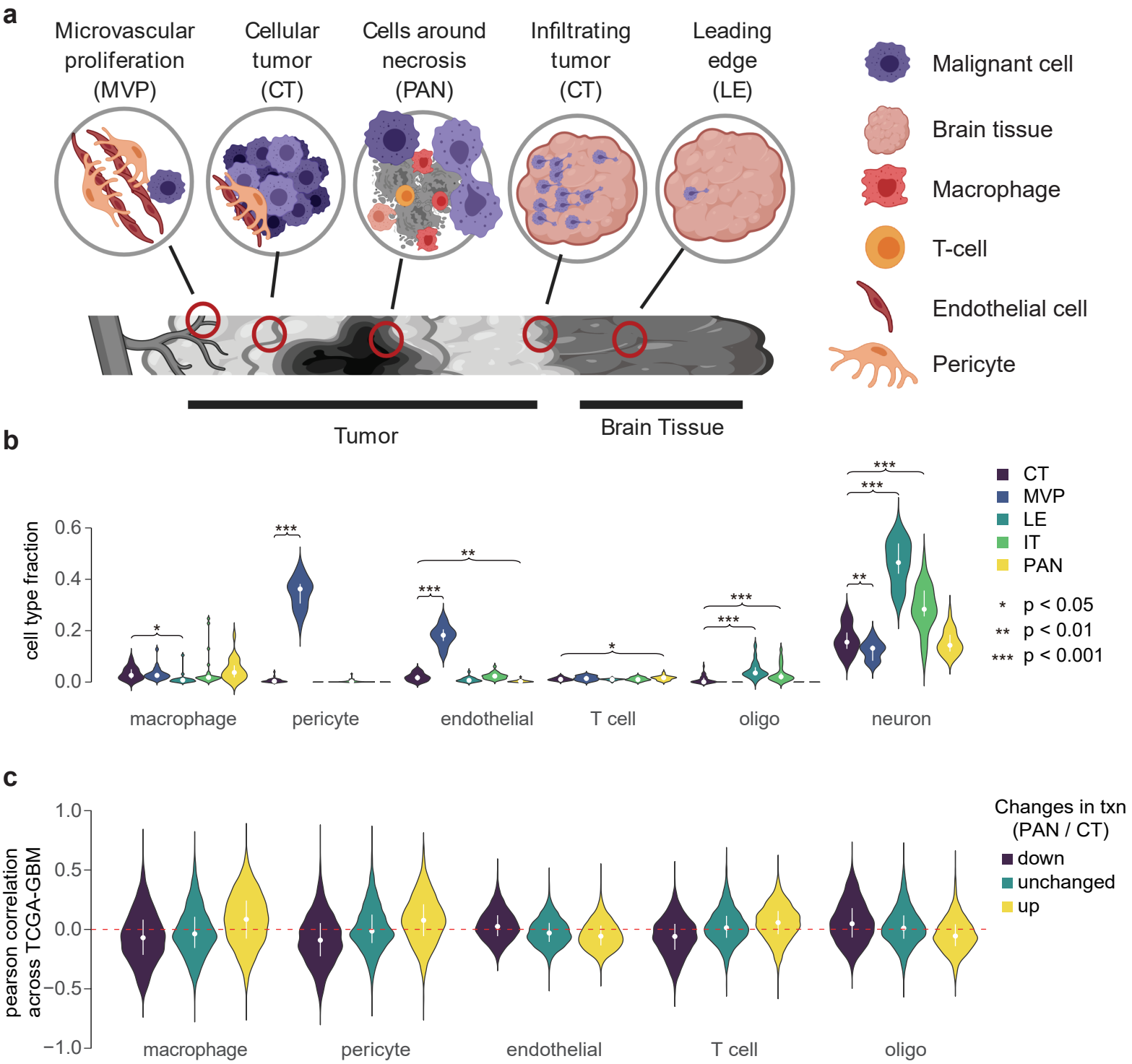

Fig. 3 | TED reveals spatial heterogeneity in GBMs. a) A graphical illustration of the anatomic structures of the IVY GAP samples based on microscopy. **b**) Violin plot shows the distribution of cell type fractions in each anatomic structure over 122 IVY GAP samples. Median fractions are shown by white dots and upper/lower quartiles are shown by bars. Asterisks mark the significant differences between CT and other anatomic structures based on a linear mixed model. **c**) Violin plot shows the distribution of Pearson correlation between genes differentially transcribed between PAN versus CT and the fraction of stromal cells over 169 TCGA-GBM samples. Median fractions are shown by white dots and upper/lower quartiles are shown by bars. All cell types show statistical significance between up/down/unchanged genes by Wilcoxon test ($p < 0.001$).

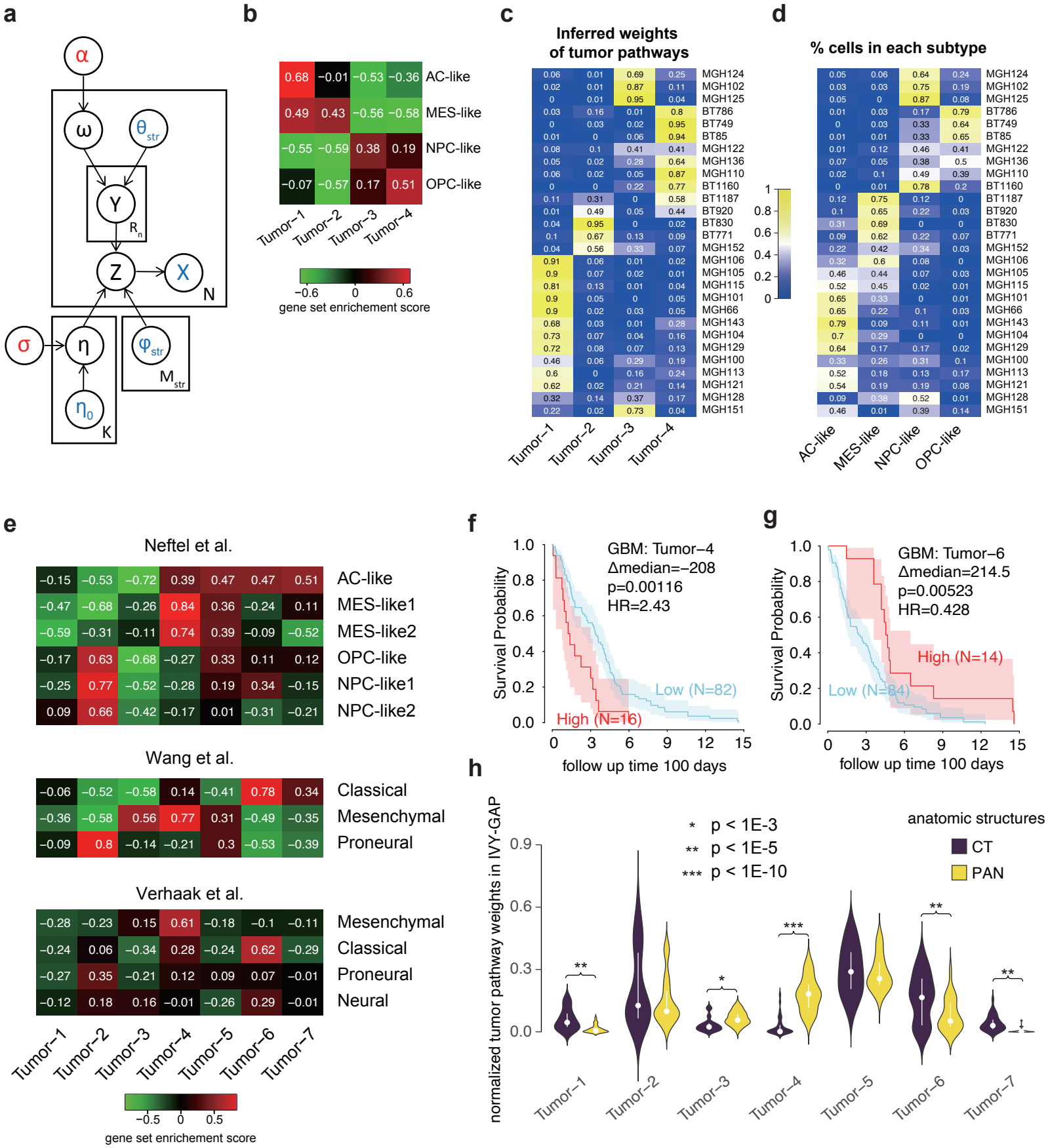

Fig. 4 | TED redefines GBM molecular subtypes after excluding expression in stromal cells. a) Graphical model illustrates the statistical dependencies and the generative process for the observed bulk RNA-seq data, X. Red text marks hyper-parameters; blue marks observed variables; black marks latent variables. **b**) Heatmap shows the gene set enrichment score for each tumor pathway from GBM28 inferred by TED. Marker genes in each cluster reported by Neftel et al. (2019) are used as the gene sets. **c**) Heatmap shows the inferred weights of each pathway in GBM28. **d**) Heatmap shows the fraction of tumor cells assigned to each cluster in GBM28. **e**) Heatmap shows the gene set enrichment score for each tumor pathway inferred by TED from TCGA-GBM. Three sets of subtype classification schemes and their marker genes are used for computing the enrichment scores. **f-g**) KM plots show the survival duration for tumor pathways in GBM. Δmedian: median survival time in the high group - median survival time in the low group. P values were computed using the log-rank test. Hazard ratio is defined by high / low. **h**) Violin plot shows the distribution of inferred weights of tumor pathways normalized to one for each sample over CT and PAN regions of the IVY-GAP samples. Asterisks mark the significant differences between CT and PAN based on a linear mixed model.

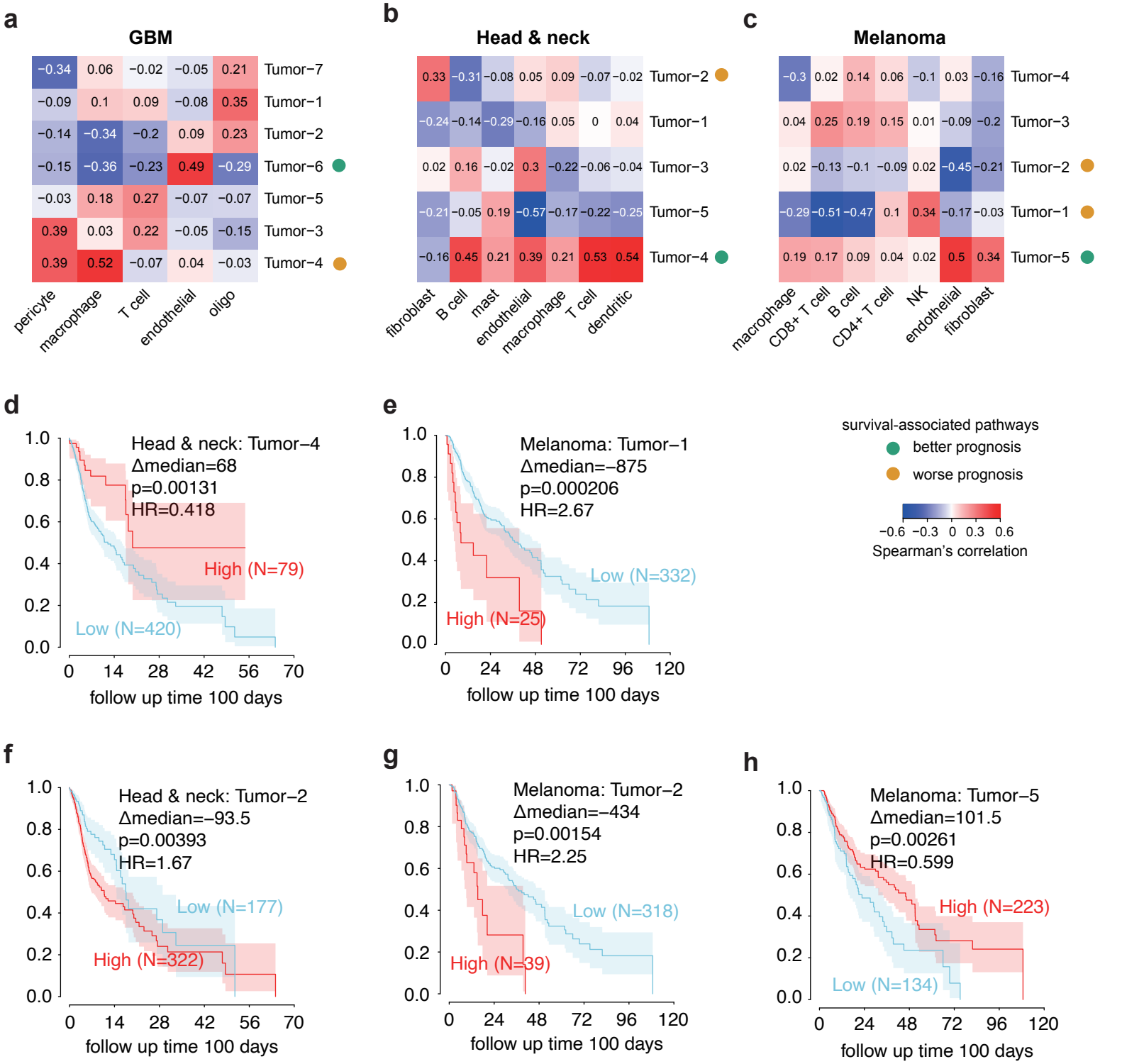

Fig. 5 | Tumor pathways correlate with stromal cell fractions. a-c) Heatmaps show Spearman's rank correlation between normalized weights of each tumor pathway and the fraction of stromal cells in each tumor type. **d-h**) KM plots show the survival duration for tumor pathways in HNSCC and melanoma. Δmedian: median survival time in the high group - median survival time in the low group. P values were computed using the log-rank test. Hazard ratio is defined by high / low.

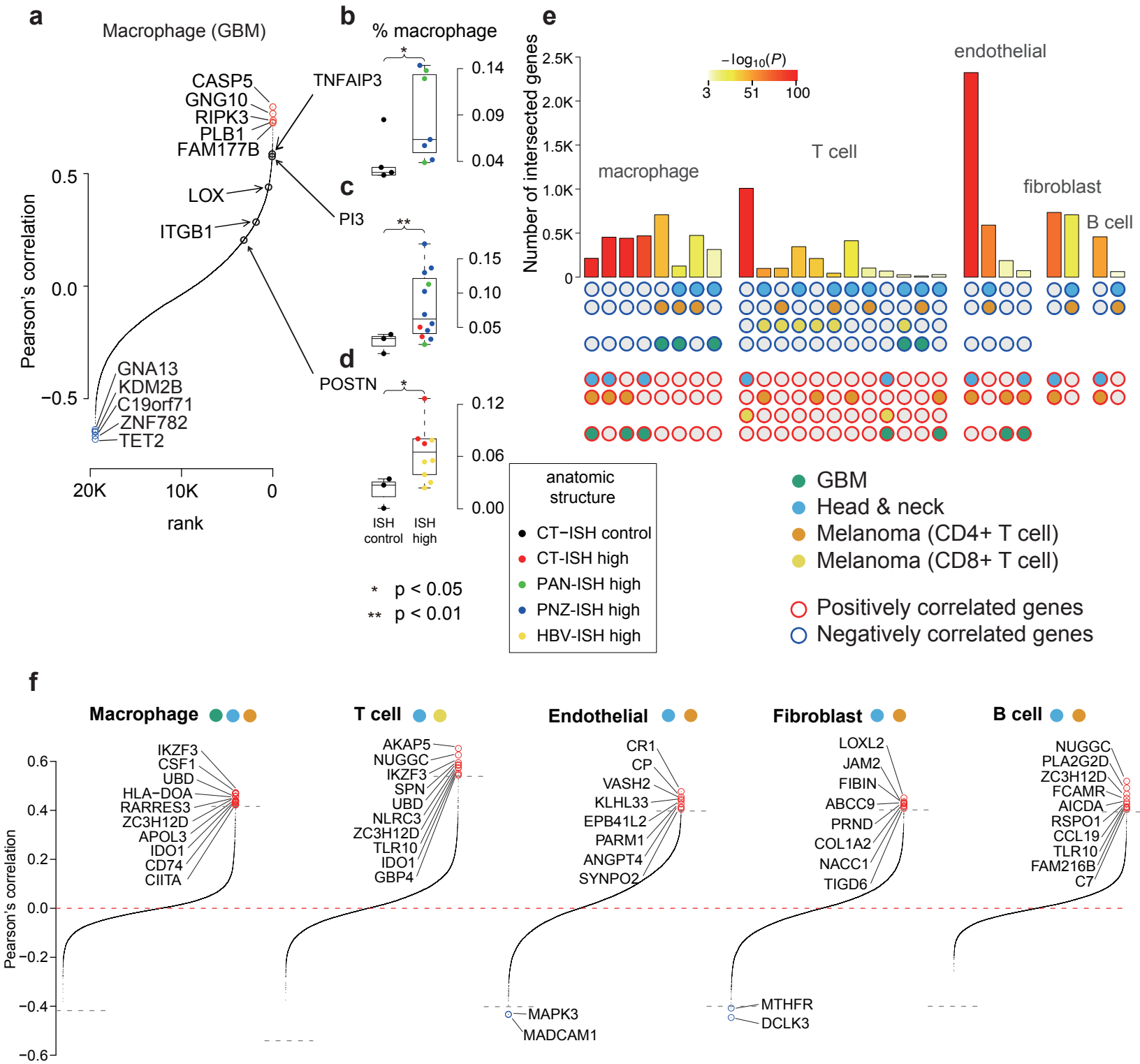

Fig. 6 | Correlation between malignant cell gene expression and stromal cell fraction. a) Rank-ordered plot shows Pearson's correlation between malignant cell gene expression inferred by TED and macrophage fraction in the TCGA GBM dataset. Positively correlated outlier genes are marked in red; negative correlations are marked in blue. Black circles highlight experimentally validated regulators of macrophage infiltration in GBM, or genes whose expression correlated with macrophage infiltration in IVY-GAP. **b**-**d**) Boxplots show the TED inferred fraction of macrophage infiltration for regions with low (ISH-control) or high (ISH-high) expression of three target genes. Color indicates anatomic structures associated with the ISH experiments. Asterisks mark significant differences as shown by a Wilcoxon test. **e**) Histogram shows the number of intersected genes in each category colored by -$\log_{10}$ p-value computed using the super-exact test. Only intersections with p<10-3 are shown. Circles below the histogram indicate the set of intersections. Only genes with significant association with cell type fractions (p<0.001, t-test) are used for the intersection study. **f**) Rank-ordered plots show the minimum absolute value of Pearson's correlation between TED inferred malignant cell gene expression and macrophage fraction over the tumor types in the most significant intersections shown by **e**). Positively correlated outlier genes are red; negative correlations are blue.