Bayesian cell-type deconvolution and gene expression inference 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 complicated interactions between cells in their environment is a major challenge in genomics. Here we developed BayesPrism, a Bayesian method to jointly predict 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 1,412 bulk RNA-seq samples in primary glioblastoma, head and neck squamous cell carcinoma, and melanoma using single-cell datasets of 85 patients. We identified cell types correlated with clinical outcomes and explored spatial heterogeneity in tumor state and stromal composition. We refined 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 introduces a new lens that uses scRNA-seq to accurately infer cellular composition and expression in large cohorts of bulk data.

Introduction

Cells in an organism have complicated interactions with other cells in their environment. A quintessential example where cell-cell interactions have important ramifications in medicine is interactions between malignant tumor cells and functionally diverse non-malignant cell types known as stromal cells[1–3]. During the past two decades numerous studies have revealed interactions between malignant cells and the stroma that promote diverse 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 cancers.

Two layers of information are critical for understanding cell-cell interactions[26]: (1) the quantity of different cell types, and (2) systematic variation of gene expression in each cell type. Measurements of both cell type and expression can be made using single cell RNA sequencing (scRNA-seq)[27–32]. However, single cell transcriptomics are still costly and often require fresh tissues and hence scRNA-seq remains technically challenging to scale to large numbers of patient samples[33,34]. Additionally they are susceptible to confounding technical factors in capture efficiency that alter the composition of cell types. Other genomic studies have used bulk RNA-seq samples to infer cell type abundance using regression on a reference expression matrix constructed from a set of arbitrarily defined marker genes[17,18]. Although these pioneering studies have provided estimates of tumor infiltrating immune cells[17,18], they make strong assumptions about the invariant expression between the reference and the bulk mixture over the selected markers. This often results in a reduced accuracy when such assumptions are violated, especially in deconvolving datasets in which significant variation exists between the bulk and single cell reference, due to technical differences in sequencing platforms and/or heterogeneity in gene expression in the tumor and its microenvironment[35–37]. Critically, these cell type deconvolution studies were not able to learn gene expression in a heterogeneous population of tumor cells.

These studies leave open several foundational questions: How do malignant cells affect the composition of stromal cells? And which genes are correlated with these interactions? To answer these questions we need an accurate model for cell type-specific expression profiles in each bulk sample and cell type fraction that can accommodate uncertainty in the single cell reference. To address these issues, we devised a Bayesian model that infers both cell type composition and gene expression, called Bayesian cell Proportion Reconstruction Inferred using Statistical Marginalization (BayesPrism). BayesPrism infers the posterior distribution of cell types fractions and gene expression from bulk RNA-seq data using a scRNA-seq reference as prior information. We showed that by explicitly modeling the error distribution in single cell reference and marginalizing it out, BayesPrism significantly improves the inference of cell type fractions in both tumor and non-tumor settings.

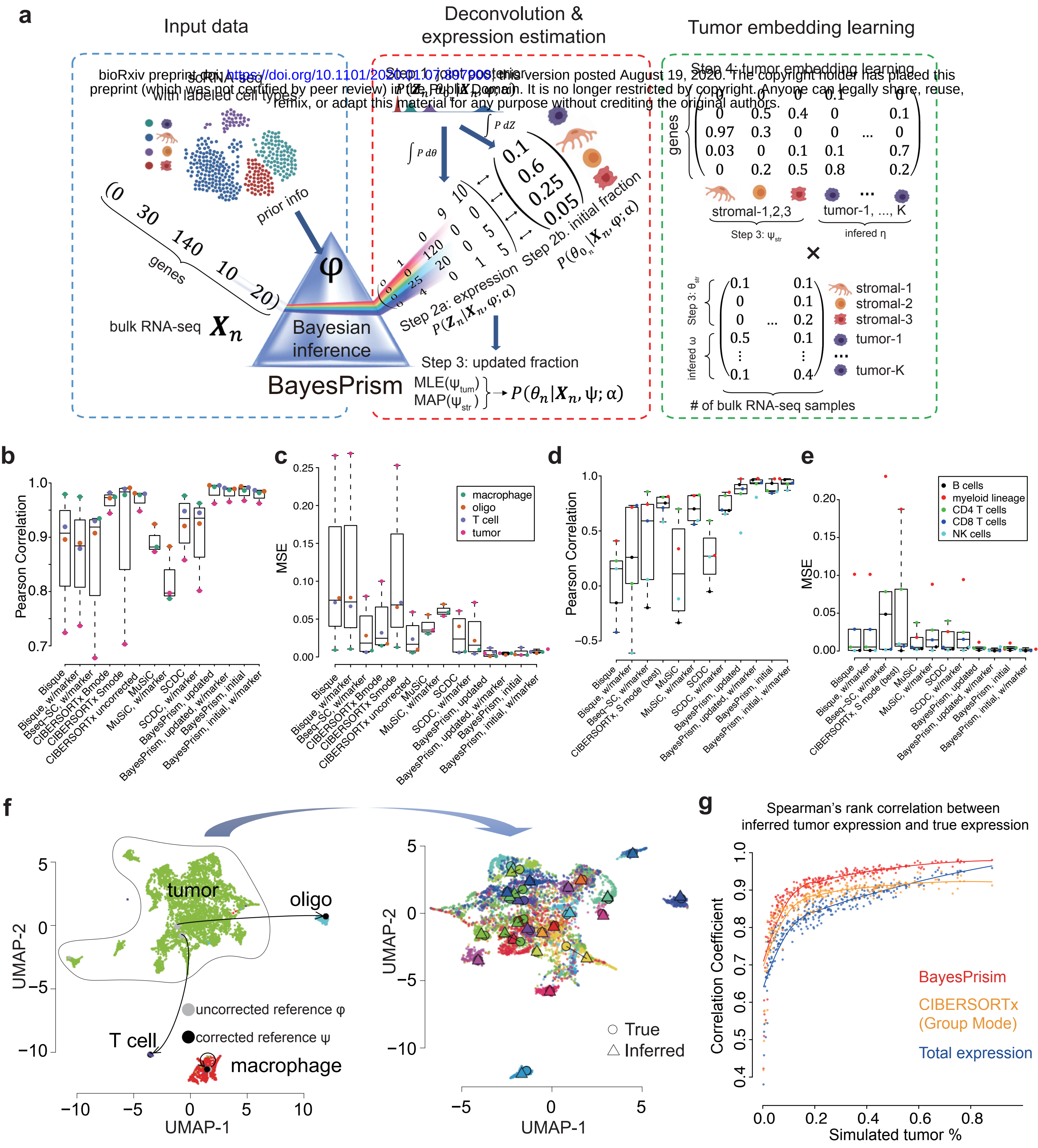


Fig. 1 | BayesPrism algorithm flow and performance validation. a) Algorithmist flow of BayesPrism. **b-e**) Boxplots show the cell type-level Pearson's correlation coefficient and MSE for deconvolution of pseudo bulks of GBM28 using refGBM8 (**b** and **c**), and bulk RNA-seq human whole blood samples with ground truth measured by flow-cytometry (**d** and **e**). Boxes mark the 25th percentile (bottom of box), median (central bar), and 75th percentile (top of box). Whiskers represent extreme values within 1.5 fold of the inter quartile range. **f**) UMAP visualization shows the expression of individual cells in GBM28. The expression profiles of stromal cells before (gray) and after (black) correction were projected onto the UMAP manifold of the scRNA-seq (left). Malignant cells in patients with greater than 10 malignant cells (N=27) were visualized on the zoomed-in UMAP (right), and are colored by patient. 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. **g**) Scatter plot shows Spearman's correlation between the average expression of malignant cells in pseudo-bulk and that estimated by BayesPrism (red), CIBERSORTx group mode (orange) or total bulk (blue), as a function of the fraction of malignant cells in a subsampled set (N=270).

Results

Bayesian inference of cell type composition and gene expression

BayesPrism uses scRNA-seq reference to infer two statistics of interest from each bulk RNA-seq sample: (i) the proportion of cell types and (ii) the expression level of genes in each cell type (**Fig. 1a, Supplementary Fig. 1, Supplementary Note 1**). The most challenging aspect of cellular deconvolution is accounting for various sources of uncertainty, including technical and biological batch variation, in gene expression between the bulk and scRNA-seq reference data. To account for uncertainty in the scRNA-seq reference, BayesPrism adopts a Baeysian strategy. BayesPrism models a prior distribution using scRNA-seq, and learns a posterior distribution of cell type proportion and gene expression in each cell type and sample conditional on each observed bulk after marginalizing uncertainty from the joint posterior. This strategy is implemented in an efficient algorithm (**Supplementary Note 2**). There are four major steps in the BayesPrism algorithm:

1. BayesPrism first infers a joint posterior distribution of the cell type proportion and gene expression, θ_{0n} and Z_n, conditional on the observed single cell reference φ and bulk expression X_n in the n_{th} bulk sample, i.e. $P(\theta_{0n}, Z_n \mid \varphi, X_n; \alpha)$, using Gibbs sampling, with α being a weak non-informative Bayesian hyper-parameter.
2. For each bulk sample n, BayesPrism estimates (2a) the gene expression matrix of each cell type, Z_n, and (2b) the proportion of each cell type, θ_{0n}, by marginalizing the joint posterior and reporting the posterior mean of the marginals. This strategy is highly robust to technical variation between the bulk and reference data, allowing BayesPrism to perform well despite substantial technical noise.
3. BayesPrism updates the reference matrix φ using information from Z to improve estimates of cell type fractions. The updated reference matrix, ψ, is the multinomial distribution parameters describing the distribution of Z. Two strategies are supported to infer ψ for use in distinct settings. First, BayesPrism can infer a maximum likelihood estimate (MLE) of cell type-specific gene expression that is unique to each bulk sample. This feature allows users to estimate gene expression in each bulk sample when there is substantial heterogeneity, as is generally the case with tumor cells from different patients[27–32]. Second, BayesPrism uses all bulk samples to summarize a maximum a posterior (MAP) estimate of cell type-specific expression. For most cell types, gene expression is reasonably similar in bulk samples[28,30,31], and in these cases it is appropriate to share information between samples by estimating ψ using all bulk data. BayesPrism then uses the updated prior distribution parameterized by ψ to re-sample the posterior marginal distribution of cell type composition for each bulk sample, θ, i.e. $P(\theta_n \mid \psi , X_n; \alpha)$. Sharing information across bulk samples results in a shrinkage property in the estimates, and provides higher accuracy for problems with batch effects.
4. Optionally, BayesPrism has an additional embedding learning module which can be used as factor analysis to explain the heterogeneous gene expression of one particular cell type across multiple bulk samples. This mode is particularly useful for analyzing common gene expression programs in malignant cells in bulk tumors, similar to the discovery of transcriptomic subtypes, after factoring out the confounding influence of stromal cells (**Fig. 1a, green**).

BayesPrism improves cell type deconvolution accuracy by accommodating uncertainty in the reference

The key innovation introduced in BayesPrism is the Bayesian inference strategy, which accommodates substantial variation in gene expression between the bulk sample and scRNA-seq reference. To benchmark the effect of measurement noise in cell-type deconvolution accuracy, we first multiplied the gene expression of the pseudo-bulk RNA-seq data of human peripheral blood mononuclear cells (PBMC) with log-normally distributed fold changes. We compared the Pearson correlation and mean squared error (MSE) between the ground truth and cell type proportions estimated using five different deconvolution methods[38–42]. BayesPrism was nearly invariant to the simulated noise and outperformed existing methods by up to an order of magnitude as noise increased (**Supplementary Fig. 2**). These data are consistent with our expectation that the Bayesian method outperforms existing methods by substantial margins when input data has noise.

To assess whether BayesPrism improved deconvolution performance in a more realistic setting, we next generated pseudo-bulk data by combining reads from similar samples analyzed using different scRNA-seq platforms. We benchmarked performance in three different settings: 1) technical batch effects with small amounts of biological variation using PBMCs and mouse cortex from different healthy subjects (**Supplementary Fig. 3**), 2) biological variation with small amounts of technical noise using leave-one-out test in datasets of two human cancer types generated by the same sequencing platforms (**Supplementary Fig. 4**), and 3) mixture of technical and biological variation using glioblastoma (GBM) datasets generated by different cohorts and sequencing platforms (**Fig. 1b-c; Supplementary Fig. 5**). When testing the effect of technical noise, we chose sequencing platforms that best recapitulate features common to bulk and scRNA-seq data modalities: full length SMART-seq2 data as a surrogate for bulk RNA-seq and 3' end enriched tag clusters obtained using 10X (for PBMCs), sci-RNA-seq (for mouse cortex) or a microwell-based platform (for GBM) as a reference scRNA-seq dataset. BayesPrism significantly outperformed all existing methods in all three settings ($p < 10^{-10}$, one-sided paired t-test). In the GBM dataset (the third setting), BayesPrism improved MSE over the next best performing method, CIBERSORTx, by ~4-7-fold (**Supplementary Fig. 4**). BayesPrism was particularly better than CIBERSORTx in estimating the proportion of tumor cells, in which gene expression was a poor match for the reference data, consistent with our expectation that the Bayesian method will provide the highest performance advantage in the presence of substantial gene expression variation between the bulk and reference data due to a mixture of biological and technical effects. BayesPrism was also robust to cell types that were missing from the scRNA-seq reference (**Supplementary Fig. 6; Supplementary Note 3**) and the number of cells and tumor patients collected by scRNA-seq (**Supplementary Fig. 7**).

As a final performance benchmark, we deconvolved real bulk RNA-seq data using ground truth obtained by orthogonal strategies. We obtained bulk RNA-seq data from 12 whole blood samples which were analyzed in parallel using flow cytometry[40]. Using a PBMC scRNA-seq data as a reference, BayesPrism obtained more accurate estimates of five cell types in the bulk sample than other deconvolution methods (**Fig. 1d-e**). BayesPrism also recovered the proportion of neutrophils in bulk RNA-seq data collected from bladder cancer that matched the neutrophil infiltration graded by a pathologist using H&E sections[17] (**Supplementary Fig. 8**). Taken together, benchmarks demonstrate that BayesPrism improves deconvolution performance in realistic data analysis tasks compared with existing deconvolution methods.

BayesPrism accurately estimates gene expression in heterogeneous cell types

We asked whether BayesPrism accurately recovered gene expression in heterogeneous cell types. We focused on the recovery of gene expression in tumor samples, in which scRNA-seq reference data is not able to accurately represent gene expression in new bulk samples due to substantial cross-patient heterogeneity[27–32]. We estimated cell types and gene expression in SMART-seq2 pseudo-bulk data from 28 GBMs, out of which 27 samples contain sufficient numbers of tumor cells, using a microwell-based scRNA-seq reference from 8 GBMs. We divided cell types into malignant and stromal groups. We computed a single estimator describing the average gene expression profile across all bulk samples for each stromal cell type (macrophages, oligodendrocytes and T cells) and a unique gene expression estimator for malignant cells in each sample. Gene expression estimates in both stromal and malignant cells were highly similar to the known ground truth (**Fig. 1f, Supplementary Fig. 9**). To determine how gene expression accuracy estimated by BayesPrism was affected by the proportion of tumor cells, we sampled random proportions of each cell type while controlling the tumor cells from an individual patient. BayesPrism estimated gene expression with correlations >0.95 for tumors with >50% purity (**Fig. 1g**) and accurately separated all tumors by patient (**supplementary Fig. 10**). Similar experiments on macrophages also accurately recovered subtle variation in gene expression in macrophage subclusters between simulated patients at moderate macrophage content (**Supplementary Fig. 11**). Gene expression estimates were substantially more accurate using BayesPrism than using either CIBERSORTx or the bulk tumor with no deconvolution (**Fig. 1g; Supplementary Fig. 12**). Taken together, BayesPrism accurately recovered gene expression in each cell type despite substantial differences between the bulk and scRNA-seq reference due to a mixture of batch effects and biological variation.

BayesPrism identifies tumor gene programs through embedding learning

Despite the heterogeneity in the gene expression of tumor cells, evolutionary pressure pushes tumors to optimize for different tasks that are essential for tumors to survive, which is done by modulating sets of co-expressed genes, known as gene programs[43]. The extent to which these gene programs are activated are often used to define molecular subtypes, using methods such as NMF and archetypal analysis. Although these methods provide a coarse grouping of genes and cancer samples, they may often reflect differential infiltration of stromal cell types rather than intrinsic tumor expression[44].

We developed a module in BayesPrism which recovered core tumor gene programs that best explain expression heterogeneity without contamination from non-malignant cell types (**Fig. 1a, green**). Motivated by recent observations that malignant cells in different tumors are heterogeneous mixtures of functionally distinct cell types[27,32,45], we modelled each patient as a linear combination of gene programs. BayesPrism infers the weights of each gene in each program and each program in each tumor using the expectation maximization (EM) algorithm, such that the linear combination of all gene programs most accurately approximates malignant cell expression in all patients (**Fig. 2a**). To evaluate whether BayesPrism learned subtypes that reflect intra-tumor heterogeneity, we identified four gene programs using the GBM28 pseudo-bulk

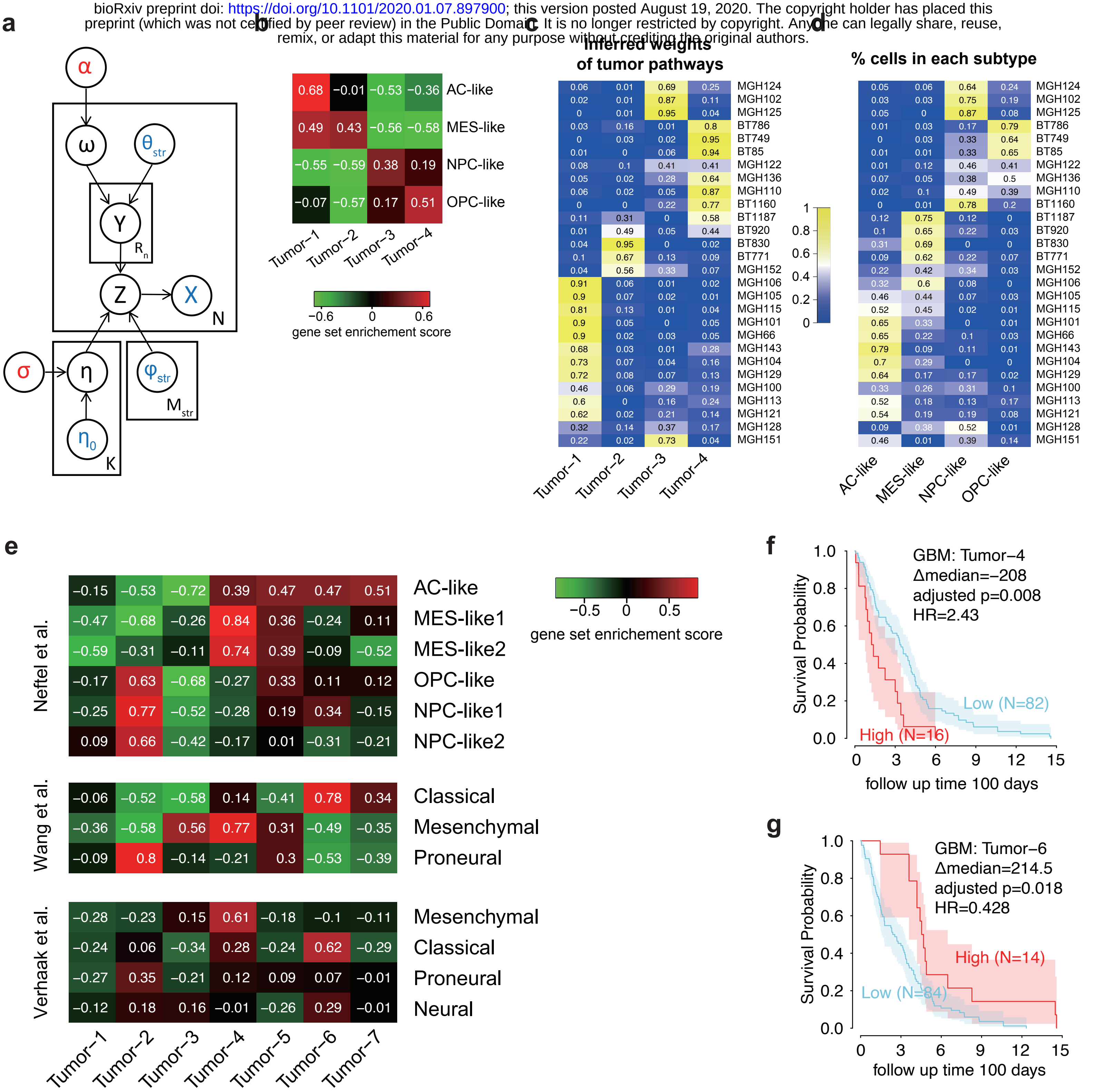


Fig. 2 | BayesPrism redefines GBM molecular subtypes after excluding expression in stromal cells. **a**) Graphical model illustrates the statistical dependencies and the generative process for the observed bulk RNA-seq data, X. Red text marks hyper-parameters; blue marks observed variables; black marks latent variables. **b**) Heatmap shows the gene set enrichment score for each tumor pathway from GBM28 inferred by BayesPrism. 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 BayesPrism 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.

RNA-seq dataset. BayesPrism recovered gene programs that were highly correlated with those recently obtained by clustering 6,863 tumor cells from 28 patients[32] (**Fig. 2b**). Moreover, the weights of each gene program learned by BayesPrism were correlated with the fraction of cells in each tumor that represent each of the four major subtypes (**Fig. 2c-d**). Thus, tumor gene programs learned using BayesPrism can accurately approximate major tumor cell subpopulations, even when the expression of subtypes are not known from direct single cell measurements.

To characterize tumor heterogeneity in GBM using the most inclusive GBM cohort available to date, we next inferred gene programs from 169 TCGA bulk RNA-seq samples. We decomposed the TCGA dataset into seven gene programs, using the criterion that selected the number of gene programs, *K*, based on the degree of consensus clustering[46] (**Supplementary Fig. 13**). BayesPrism revealed several programs that were similar to those in previous studies[32,44,47], including program 2 (proneural, OPC, and NPC-like), 3 and 4 (mesenchymal), and 6 and 7 (classical and AC-like) (**Fig. 2e**). Two of the programs discovered using BayesPrism were correlated with clinical outcomes: program 4, similar to the mesenchymal subtype (HR = 2.43, p = 0.001; **Fig. 2f**; **Supplementary Fig. 14a**), and program 6, which bore similarities to the classical subtype (HR = 0.428, p = 0.005; **Fig. 2g**; **Supplementary Fig. 14a**). Notably, prior studies found no correlation between subtype and clinical outcomes in GBM, except when taking a subset of mesenchymal tumors[44]. In contrast, as BayesPrism naturally generates a continuum score of each gene program in each sample, and hence greatly facilitates the study between activation of gene programs and clinical covariates of interest in an unbiased way.

Cell type composition predicts clinical outcome in three tumor types

We analyzed the proportion of cell types in 1,142 TCGA samples from three tumor types: GBM, HNSCC, and melanoma[48–50]. 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. 3a**). Analysis using BayesPrism revealed that the majority of TCGA samples were >75% malignant cells in all three tumor types (**Fig. 3a**). Tumor estimates correlated with those obtained using CNVs and marker gene expression[51,52] (**Supplementary Fig. 15**). Across large cohorts of tumors, stromal cell types had a rich correlation structure with one another that mirrored several previously described observations (**Supplementary Fig. 16, Supplementary Note 4**).

We asked whether stromal cell types were correlated with patient survival. To avoid confounding our analysis with known genetic or clinical covariates, we accounted for clinical features known to strongly affect prognosis: GBM patients with the wild type IDH allele , and metastatic melanoma. In HNSCC, we studied all patients and found that results were robust when focusing on HPV-negative patients. The proportion of T cells was associated with better clinical outcomes in all three malignancies (hazard ratio [HR] = 0.416-0.604; **Fig. 3b-d; Supplementary Fig. 17 and 18**). 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. 3d and e**). Macrophages were significantly associated with survival in both GBM and melanoma, but not in HNSCC (**Fig. 3f-h**). Intriguingly, however, high macrophage infiltration had a poor prognosis in GBM (HR = 1.71; **Fig. 3f**), but a substantially better prognosis in melanoma (HR = 0.556; **Fig. 3h**), indicating substantial heterogeneity in the role of macrophages in different

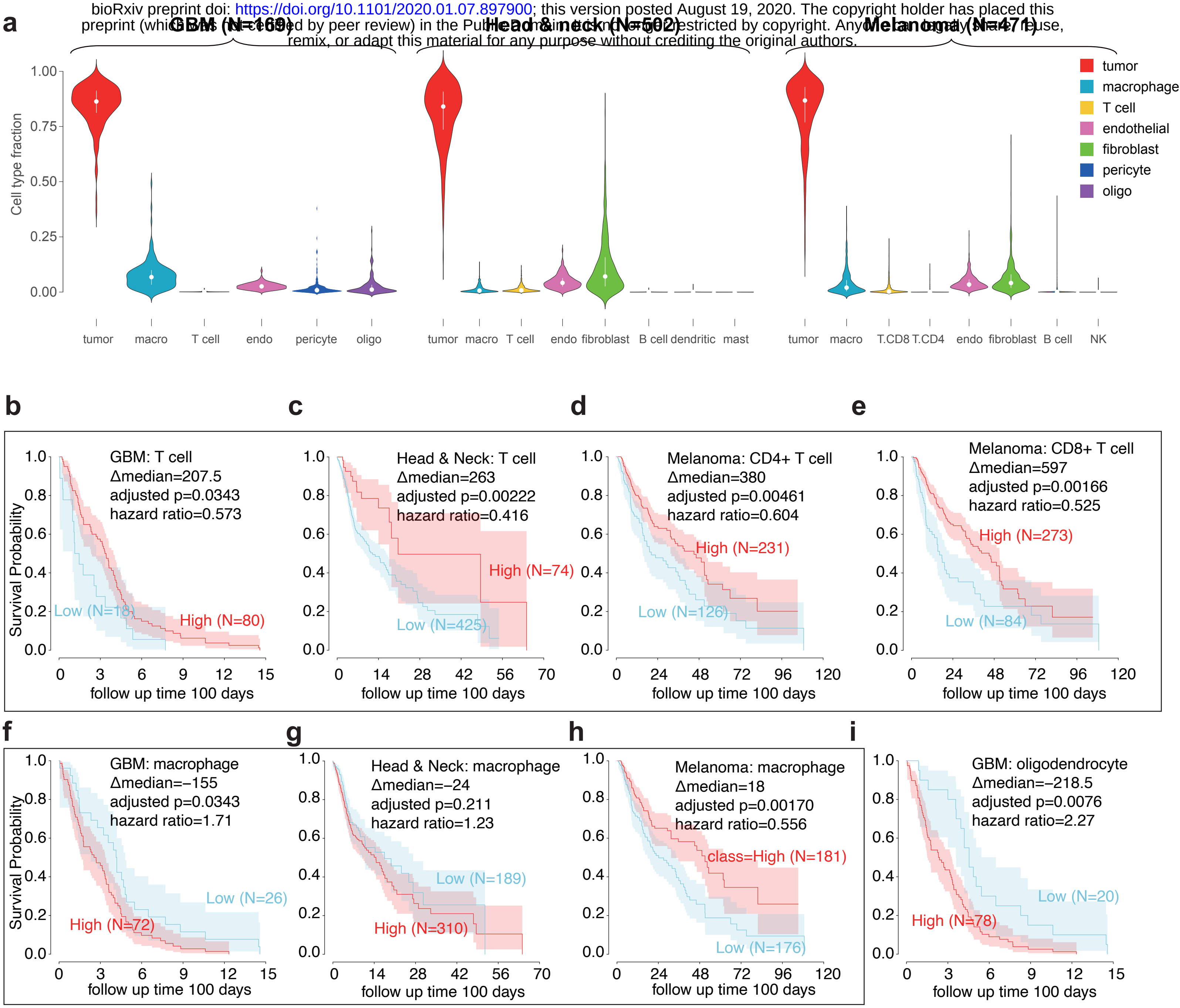


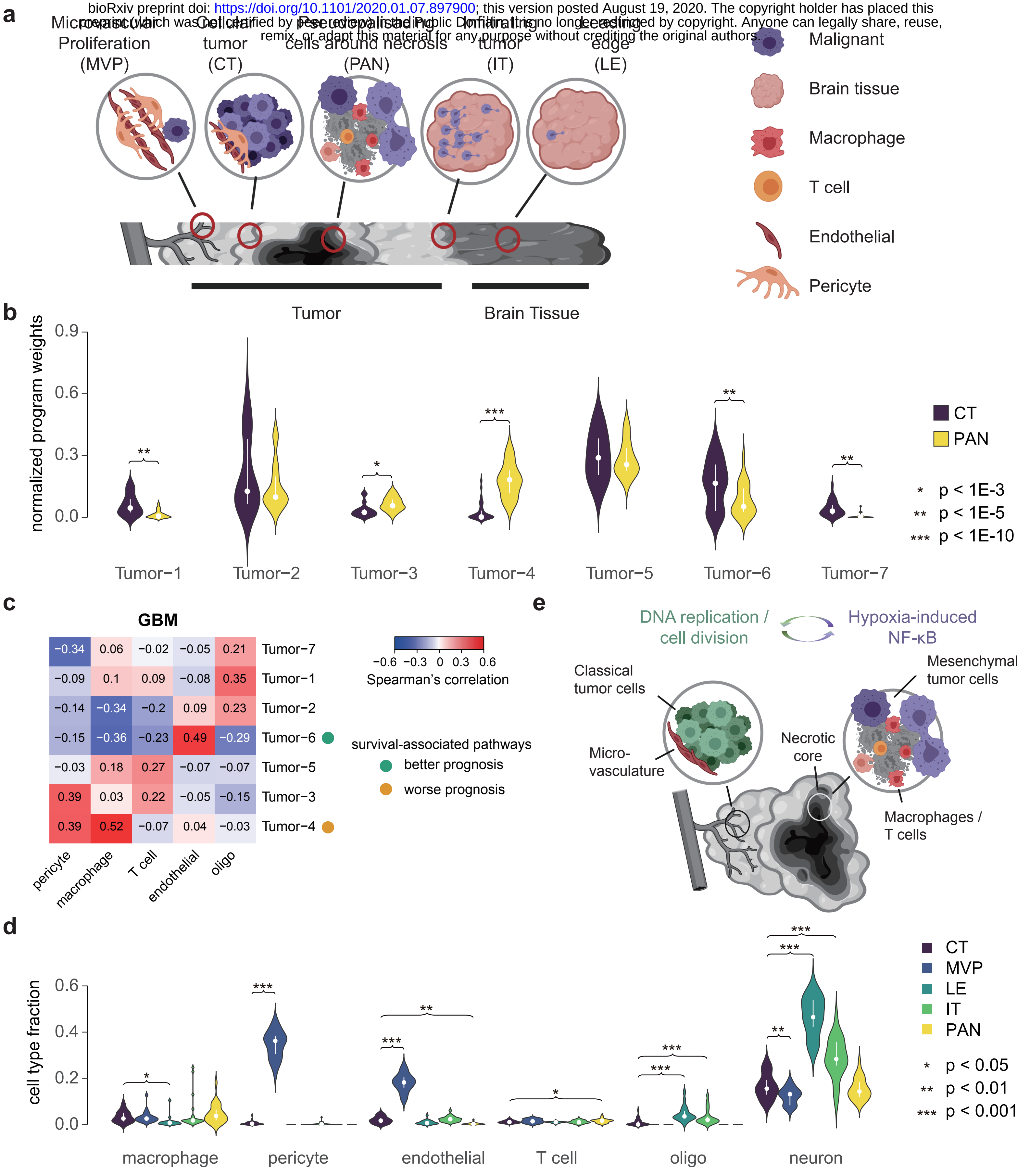
Fig. 3 | Cell type compositions in three tumor types. a) Violin plots show the distribution of cell type fractions in each tumor type. Median fractions are shown by white dots and upper/lower quartiles are shown by bars. **b-i**) KM plots show the survival associations with **b-e**) T cell infiltration, **f-h**) macrophage infiltration, and **i**) 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, and corrected using FDR. Hazard ratio is defined by high / low.

malignancies. BayesPrism 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. 3i**), 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 (all)], and 0.515 [melanoma]; **Supplementary Fig. 17 and 18**). Taken together, these analyses reveal new information about how heterogeneity in the microenvironment affects clinical outcomes in three cancer types.

Spatial heterogeneity of tumor gene program and stromal cell infiltration in GBM

We compared regional and inter-tumor heterogeneity in gene programs and stromal composition in GBM. To complement TCGA data, we deconvolved 122 bulk RNA-seq samples microdissected into five structures by IVY GAP[53] (**Fig. 4a and Supplementary Table 1a**). Different regions showed logical enrichments for stromal cell composition across different tumor regions (**Supplementary Note 5**). We examined which gene programs identified using TCGA (above) were enriched in cellular (CT) and necrotic (PAN) regions using IVY GAP data, whose microenvironments are known to differ in several respects, including blood supply, hypoxia, and local necrosis, which affect gene expression[54]. PAN regions were enriched for mesenchymal programs (especially programs 3 and 4), consistent with observations that tumors with higher necrosis were more likely to be mesenchymal GBMs[47] (**Fig. 4b**). We also discovered a novel association between CT and classical programs 6 and 7, and program-1 (which is not similar to previously discovered subtypes). To confirm the relationship between classical tumors and the microenvironment, we examined the correlation between stromal cells and each tumor subtype in TCGA samples. Consistent with the analysis of IVY GAP, Tumor-6 (classical) was correlated with endothelial cells (Spearman's rank correlation = 0.49), and Tumor-4 (mesenchymal) was not (Spearman's Rho = 0.04) (**Fig. 4c**). Gene ontology analysis for gene upregulated in these programs showed significant enrichment for biological process that echo prior knowledge[55,56], with program-4 enriched for NF-κB pathway and immune processes, and program-6 enriched for cell division and DNA replication ($p < 10^{-10}$, **Supplementary Table 2**). Our results implicate the balance between nutrient availability and hypoxia in establishing gene expression patterns that are characteristic of these subtypes.

Next we examined other stromal cell types which correlated with specific regulatory programs. BayesPrism recovered a correlation between Tumor-4 and macrophages as the strongest association, consistent with previous reports[44] (**Fig. 4c**). Weaker associations were discovered between pericytes and both mesenchymal-like programs (Tumor-3 and 4), which may reflect the differentiation of mesenchymal tumor propagating cells into pericytes[57]. Finally, two mesenchymal subtypes (Tumor-5 and Tumor-3) were associated with higher T cell infiltration (**Fig. 4c**). Several of these associations were confirmed by an analysis of IVY GAP data, including an enrichment of T cells in PAN (p=0.03) (**Fig. 4d**). These results are consistent with a model in which GBM tumor cells adopt a classical subtype which divides rapidly in a nutrient and oxygen-rich environment found near microvasculature, and a stress-induced mesenchymal subtype in hypoxia and resource-depleted necrotic regions, where T cells and macrophages are recruited (**Fig. 4e**)[32].



	pericyte	macrophage	T cell	endothelial	oligo
Tumor−7	−0.34	0.06	−0.02	−0.05	0.21
Tumor−1	−0.09	0.1	0.09	−0.08	0.35
Tumor−2	−0.14	−0.34	−0.2	0.09	0.23
Tumor−6 ●	−0.15	−0.36	−0.23	0.49	−0.29
Tumor−5	−0.03	0.18	0.27	−0.07	−0.07
Tumor−3	0.39	0.03	0.22	−0.05	−0.15
Tumor−4 ●	0.39	0.52	−0.07	0.04	−0.03

Fig. 4 | BayesPrism reveals spatial heterogeneity in GBMs. a) A graphical illustration of the anatomic structures of the IVY GAP samples. **b**) 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. **c**) Heatmaps show Spearman’s rank correlation between normalized weights of gene programs and the fraction of stromal cells in GBM. **d**) 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. **e**) A model depicting the interaction between tumor gene programs and microenvironment in GBM.

Our model provides mechanistic insight into disease progression. We suggest that the classical-like program-6 may reflect an earlier stage of cancer growth where blood supply is ample, and as the disease progresses the tumor outgrows the local nutrient supply resulting in a hypoxia and necrosis. This proposal explains why classical tumors recur as mesenchymal in longitudinal studies more frequently than the other direction[44,57]. Likewise, our survival analysis, described above, found that program-6 (classical) was associated with a better prognosis whereas program-4 (mesenchymal) was worse (**Fig. 2f-g**). Additionally, previous studies have had difficulty nailing down an association between mesenchymal composition and survival. Our results may suggest this difficulty reflects a proclivity for certain mesenchymal tumors to recruit T cells, which are generally associated with better outcomes. We also obtained broadly consistent results using several previously reported subtype definitions[32,44,47] (**Supplementary Fig. 19**). Collectively, our analysis relates known GBM subtypes to interpretable gene programs and their associated microenvironment, providing a mechanistic understanding of clinical prognosis.

Co-activation in gene programs across melanoma and HNSCC

We extended our analysis of GBM by learning gene programs in HNSCC and melanoma (**Fig. 5a-b**). Consensus clustering led us to divide each tumor type into five gene programs (**Supplementary Fig. 20 and 21**). As with GBM, several of these gene programs were associated with clinical outcomes (**Fig. 5c-g**). Both HNSCC and melanoma had an anti-angiogenic program (program 5 [HNSCC] and program 2 [melanoma]), which strongly and inversely correlated with endothelial cells, as well as a gene program that correlated with cancer associated fibroblasts (program 2 [HNSCC] and program 5 [melanoma]). We also noted several differences in tumor composition between HNSCC and melanoma. HNSCC had a single gene program which was highly immunogenic (program 4; **Fig. 5a**) and associated with extended survival (HR = 0.418; **Fig. 5c**). In melanoma, multiple gene programs correlated more weakly with immune infiltration (programs 3, 4, and 5). Interestingly, program 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. 5b**). We found this program was strongly associated with poor survival (**Fig. 5d**). Taken together these results indicate a strong correspondence between malignant cell expression and the tumor microenvironment.

BayesPrism identifies core genes involved in tumor-stroma interactions

For many applications, such as the identification of drug targets, prioritizing specific driver genes for tumor-stroma interactions is often needed. To identify such candidate genes interactions, we examined correlations between stromal cell type proportion and gene expression in malignant cells. BayesPrism reduced the correlations between stromal cell type and genes that were simply highly expressed in the same cell type (**Supplementary Fig. 22; Supplementary Note 6**), reducing the potential for false positives. To begin, we asked whether we could recover known positive regulators of macrophage infiltration in GBM[58,59]. Indeed, genes previously reported to have interactions all had statistically significant positive correlations with macrophage inflation, including *POSTN*, *ITGB1* and *LOX* (**Fig. 6a**). In addition, we identified numerous other correlations with a stronger magnitude, including *CASP5*, *GNG10*, *TNFAIP3, PI3, RIPK3*, and *PLB1*.

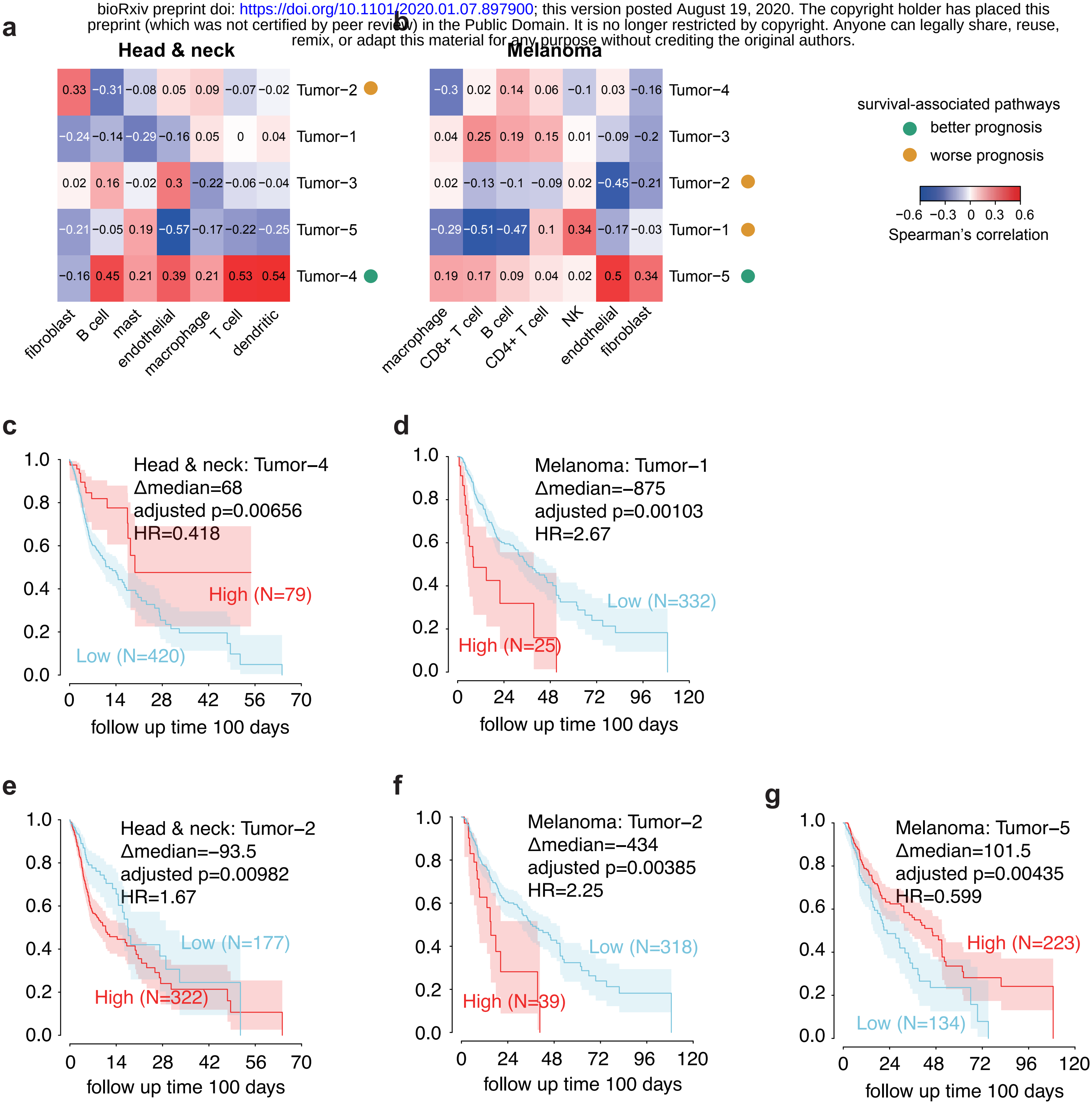


Fig. 5 | Tumor pathways correlate with stromal cell fractions. a-b) Heatmaps show Spearman's rank correlation between normalized weights of gene programs and the fraction of stromal cells in HNSCC and melanoma. **c-g**) KM plots show the survival duration for tumor pathways in HNSCC and melanoma. Δmedian: 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.

To validate correlations discovered using BayesPrism with independent data, we asked whether tumor regions expressing high levels of candidate 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[53]. We asked whether the proportion of macrophages estimated from RNA-seq using BayesPrism 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 BayesPrism identified correlations using TCGA that could be reproduced by the intratumoral spatial heterogeneity.

Having verified that BayesPrism can identify correlations observed between tumor expression and the proportion of stromal cells that could be validated using independent sources of data, we next analyzed candidate pan-cancer tumor-stroma interactions across three tumor types. Surprisingly, 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[60]; **Fig. 6e**). We examined the genes that were most enriched in multiple tumor types (**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, a small metabolite which activates T regulatory cells and myeloid-derived suppressor cells[61]. Endothelial cells were correlated with several genes known to be involved in angiogenesis, including *ANGPT4*, *CP*, and *VASH2*[62,63]. The top gene correlating with fibroblasts was *LOXL2*, which is a factor secreted by tumor cells that promotes proliferation of fibroblasts[64]. 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 BayesPrism recovers genes known to interact with stromal cells in cancer.

We found that CD4+ and CD8+ T cells were correlated with 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; $p < 7.5\times10^{-30}$; Fisher's exact test). Tissue stiffness affects a variety of T cell responses[65,66], 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. Consistent with distinct mechanisms affecting the deposition of CD4+ and CD8+ T cells, we also noted two separate submodules of immune cells in melanoma, one consisting of CD4+ T cells and NK cells, and the other consisting of CD8+ T cells, macrophages, and B cells (**Supplementary Fig. 16c**). Taken together these results support similar genes that interact with stromal cell types between different tumors, and reveal differential modes of interaction between melanomas and different T cell subsets.

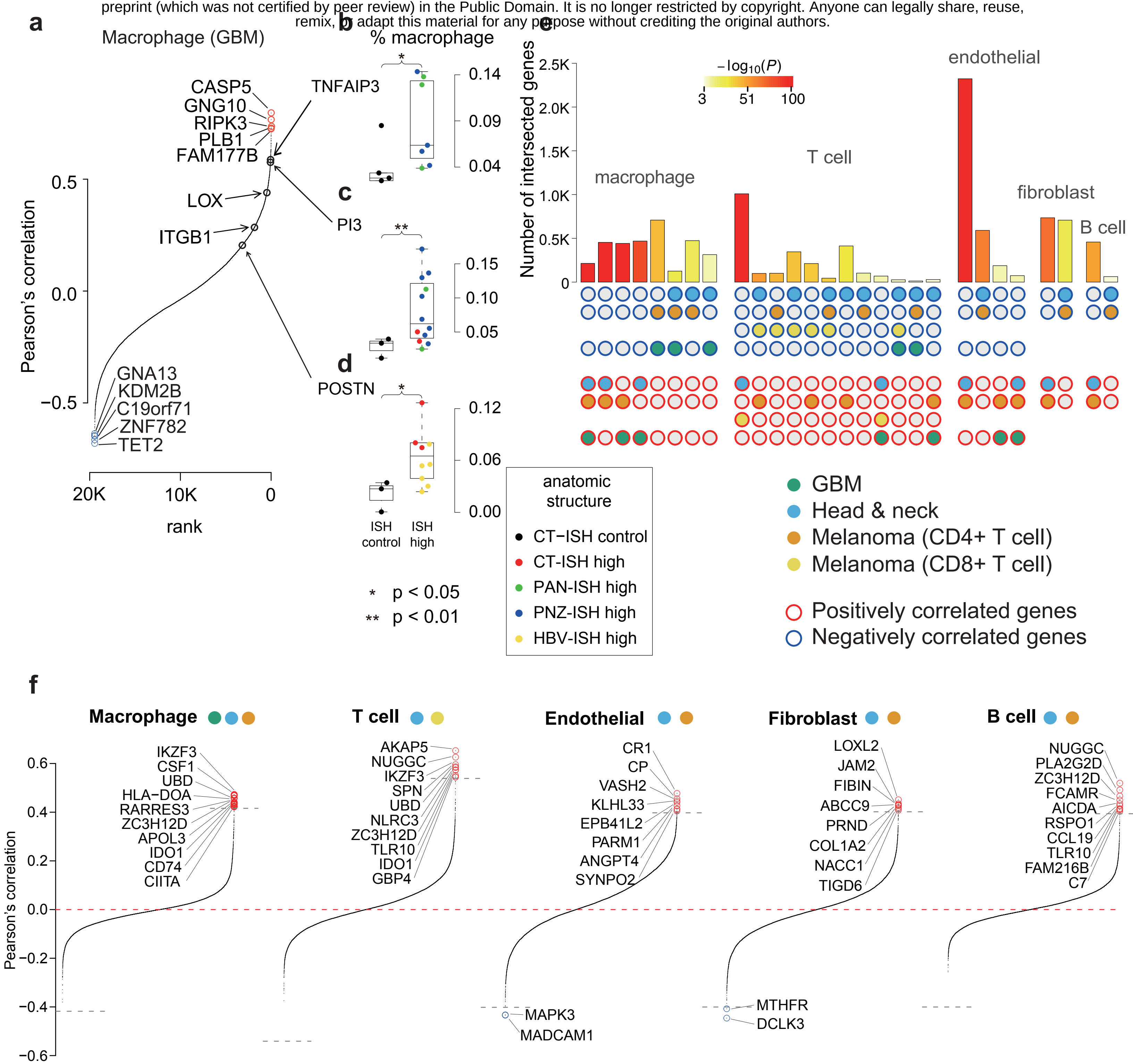


Fig. 6 | Correlation between malignant cell gene expression and stromal cell fraction. a) Rank-ordered plot shows Pearson's correlation between malignant cell gene expression inferred by BayesPrism and macrophage fraction in the TCGA GBM dataset. Positively correlated outlier genes are marked in red; negative correlations are marked in blue. Black circles highlight experimentally validated regulators of macrophage infiltration in GBM, or genes whose expression correlates with macrophage infiltration in IVY GAP. **b-d**) Boxplots show the BayesPrism 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**) Bars show the number of genes whose malignant cell expression level was correlated with the indicated cell types in the indicated tumor type. Bars are 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 BayesPrism inferred gene expression in malignant cells 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.

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[48–50]. 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 programs 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 gene programs 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[19,67], 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 genomics toolbox. BayesPrism 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 BayesPrism is not just a deconvolution algorithm - it jointly models cell types and their average expression, which was crucial for analyses reported herein. Thus BayesPrism 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 cell-cell interactions.

Acknowledgements

We thank Peter Sims for sharing the cell type annotation, Xin Bing for discussions on statistical inference, and Mariano Viapiano for 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. Some figures were created with BioRender.com.

Author Contributions

The project was conceived by TC with input and advice by CGD. TC developed and implemented BayesPrism, 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 BayesPrism

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

BayesPrism 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, BayesPrism first obtains an initial estimate of the fraction of each cell type. BayesPrism 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, BayesPrism estimates gene expression in all cell types. BayesPrism 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 BayesPrism assumes that the stromal cells share the same expression profiles across patients, allowing it to pool the statistical strength across bulk RNA-seq samples. BayesPrism 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 BayesPrism was designed to identify gene expression patterns that arise commonly among bulk RNA-seq samples after removing stromal cells influtrating the tumor. BayesPrism 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[46]. BayesPrism then uses EM to determine the gene programs whose linear combination best estimates gene expression in the observed bulk RNA-seq malignant cells. In the E step BayesPrism 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 BayesPrism 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, BayesPrism runs a final Gibbs sampling to derive the distribution of ω under the maximum a posterior estimates of η, and uses its mean to get a point estimator.

Deconvolution of bulk RNA-seq using BayesPrism

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 [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[68] 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 many of the non-coding genes in TCGA were close to zero across all patients, and hence we subset the inference on protein-coding genes when deconvolving TCGA data to speed up downstream analysis. Deconvolution over all genes generated almost identical results (data not shown). 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 BayesPrism 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 BayesPrism. We set the default parameters of BayesPrism 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.

Embedding learning analysis

To initialize the tumor gene programs (bases), we used the NMF R package[46] to learn a linear combination that best approximates the normalized tumor expression inferred by the deconvolution module of BayesPrism (res$ Zkg.tum.norm). We optimized the number of tumor bases using from 2 to 12, and chose the K that yields the best cophenetic score before a significant drop begins. We then fix the K and randomly initialize the algorithm 200 times and choose the bases that yielded the minimal residuals. The tumor bases optimized by NMF are then used as inputs for the embedding learning module of BayesPrism. Although BayesPrism does not necessarily require the use of input bases learned using external algorithms such as NMF, and can use multiple replicates of an averaged tumor expression as inputs when no user-defined input is used, we do find the use of NMF learned bases significantly speeds up the convergence of EM, and also helps in selecting the number of gene programs K when no prior information is provided.

Performance Benchmarks

Benchmarks against other deconvolution tools. We benchmarked BayesPrism against CIBEROSRTx, Bseq-SC, Bisque, SCDC, and MuSiC. Marker genes are required for CIBERSORT-based methods including Bseq−SC and CIBERSORTx (all modes), while they are optional for all other methods including BayesPrism. For CIBERSORTx we used the online portal of CIBERSORTx (https://cibersortx.stanford.edu) to perform all the benchmarking. All parameters were used at their default values, except for the "Min. Expression", which was set to 0 for single cell references to generate a signature matrix, following the author's recommendations for droplet-based platforms. Quantile normalization is disabled by default following the author's recommendations for RNA-seq. For all other methods, we used their R packages. In all non-tumor sample deconvolutions, we used a single batch of scRNA-seq dataset as the reference for Bseq-SC, SCDC, SCDC, and split the scRNA-seq dataset equally into two batches for each cell type for Bisque and MuSiC, as multiple batches of single cell references are required by these methods. When benchmarking Bisque when used the "ReferenceBasedDecomposition" and disabled "use.overlap", as we do not have samples with matched scRNA-seq and bulk RNA-seq. For tumor deconvolution (simulated GBM28), we leveraged the information of individual patients from scRNA-seq reference (GBM8) to label biological replicates whenever possible. This includes Bseq-SC, Bisque and MuSiC. Tumor cells in each patient are used as a single cell type cluster. When deconvolving genes without markers, ribosomal and mitochondrial genes, and genes on chromosome X and Y are removed in all benchmarks. To speed up computation, we removed lowly expressed genes, by subsetting genes expressed in at least 5 cells. In addition, outlier genes, defined as genes that show >1% of total reads (or normalized reads if only TPM data is available) in any of the bulk samples are removed, unless otherwise specified below. In all pseudo-bulk analysis, we defined the ground truth as fractions of total reads over all annotated genes in each cell type, which is modeled by BayesPrisim. As the cell type level correlation is not affected by the difference in total expression, we benchmarked this metric against other deconvolution tools.

To ensure a fair comparison between all methods, we use the same set of markers generated by CIBERSORTx when applicable (**Fig. 1b-e; Supplementary Fig. 2,3**) and labeled as "method name, w/ marker". The BayesPrism package also provides options for the use of marker genes to combat cases where significant batch effects exist, such as when using ribosomal-depleted RNA or statistical assumptions are possibly violated, such as when using references collected from unmatched samples. The implementation is based on the findMarker function implemented by the scran package. Briefly, single cell reference is normalized by median of library size followed by $\log_2(X+0.1)$ - log(0.1) transformation. There are two types of marker defined by the findMarker function, the “all marker” (genes that are significantly differentially transcribed between one cell type and all other cell types) and “any marker” (those that are significantly differentially transcribed between one cell type and any other cell type). Significance is calculated based on t test, and only genes upregulated are used to define markers.

Linear multiplicative noise model (Supplementary Fig. 2). We used the scRNA-seq of PBMC collected from the first donor using 10x Chromium (v2) A[69], as the reference. The expression profile of each cell type was first collapsed to a vector of length equal to the number of genes by summing up reads for all cells within each cell type. The collapsed expression profile was then normalized to sum to one using the “norm.to.one” function in the BayesPrism package, such that the added pseudo-count equal to 10^{-8} after normalization.

We simulated 200 pseudo-bulk RNA-seq samples from the same dataset used to build the reference. The cell type fractions were drawn from a uniform dirichlet distribution (α=1), and the cell numbers of each cell type were sampled from a multinomial distribution parameterized by the cell type fractions with the total cell number equal to the cell number in the original batch (N=3222). Cells were then sampled with replacement according to the simulated cell number of each cell type, and then collapsed by summing up the reads over the sampled single cells to make pseudo-bulks. In the simulation, no outliers from the bulk were removed.

To generate noise, we simulated a zero-centered log normally distributed fold change at one particular σ independently and identically distributed for each gene, which generated a vector of length equal to the number of genes. To mimic the real biological batch effects, we penalized extreme fold changes that result in unrealistic expression values, which is particularly frequent at high σ levels. This is done by sampling the fold change vector 10000 times and choosing the one that induced the minimal change to the total expression as measured by elemental-wise multiplying the reference expression with the fold change vector. The chosen fold change vector was then elemental-wise multiplied with the pseudo-bulks which were then rounded up to the nearest integers.

Cross-platform deconvolution using pseudo-bulk RNA-seq from non-tumor samples (Supplementary Fig. 3). For PBMC data, we used the 10x Chromium (v2) dataset collected from the second donor as the reference to deconvolve pseudo-bulks generated by the Smart-seq2 from the first donor in the original paper [69]. For mouse cortex data we used the sci-RNA-seq dataset collected from the second mouse as the reference to deconvolve pseudo-bulks generated by the Smart-seq2 from the first mouse in the original paper[69]. The choice of these dataset is to represent the strongest batch effect based on the correlation shown in the Supplementary Figure 4 of the paper by Ding et al.[69].

Single-platform leave-one-out deconvolution using pseudo-bulk RNA-seq from HNSCC and melanoma (Supplementary Fig. 4). 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 BayesPrism deconvolved expression using the remaining datasets as a reference. All parameters of BayesPrism and CIBERSORTx were default. Batch effect correction is disabled for leave-one-out tests. Two cell types, “-Fibroblast” and “myocyte”, in the HNSCC scRNA-seq dataset are of very low cell number <20 cells, and only show up in a small subset of patients (N=4), which may lead to unreliable estimates of correlation coefficients. Therefore, we excluded them during the leave-one-out test. Benchmarking the expression inference cannot be done, as CIBERSORTx requires the number of mixtures to be greater than the number of reference components.

As observed with GBM, BayesPrism consistently estimated cellular proportions that were more accurate to the true values than CIBERSORTx (**Supplementary Fig. 4a-d**). As the leave-one-out test data were generated from the same sequencing platforms and processed by a uniform pipeline, they represent minimal technical batch effects, and hence the superiority in performance of BayesPrism mainly reflects its ability to account for the uncertainty in the reference caused by biological variations in the tumor environment.

Cross-platform deconvolution using pseudo-bulk RNA-seq from GBM samples (Fig.1b-c; Supplementary Fig. 5, 9, 10, 11). We analyzed two glioblastoma multiforme (GBM) datasets collected from different patients using different scRNA-seq platforms to represent a mixture of technical and biological variations. 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, and hence also mimics differences between single cell and bulk.

We generated “pseudo-bulk” RNA-seq datasets from GBM28 by 1) adding up scRNA-seq counts for each patients (N=28), and 2) 1,350 pseudo bulk RNA-seq samples by sampling random proportions of each cell type using GBM28 while controlling the tumor cells from an individual patient for each pseudo-bulk to test BayesPrism across a wider range of different tumor compositions. Specifically for the second pseudo-bulk simulation, for each patient among the 27 out of 28 GBM patients we simulated 50 pseudo-bulk RNA-seq samples (the tumor cells in one sample BT1187 were excluded due to only having 8 tumor cells). 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.

When benchmarking CBIERSORTx, 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 subtypes and 5 non-malignant cell types. Same as BayesPrism, the total tumor fraction was computed by summing up the fractions of 60 tumor phenotypes.

As S mode and B mode batch correction worsened the performance, we diabled batch correction in imputing expression. For gene expression imputation in **Fig. 1g** and **Supplementary Fig. 12**, 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 the high resolution mode (by excluding the “1” and "NA" values in the

CIBERSORTxHiRes_job1_PJ0XX-tumor-X_Window140.txt). The spearman's rank correlation coefficients in **Supplementary Fig. 12** were computed on these 53 genes for four different approaches.

To show BayesPrism also infers the expression of stromal cells in addition to tumors, we generated an additional set of references and pseudo-bulks using GBM8 and GBM28, by taking the heterogeneities in the macrophages into the simulation (**Supplementary Fig. 11**). We first clustered the macrophages found in GBM28 and GBM8. We processed the scRNA-seq of macrophages in each dataset as follows. As GBM28 data is TPM normalized, we skipped the normalization step, and $\log_2$(X+0.1) transformed the data followed by removal of ribosomal mitochondrial genes, and genes on chromosome Y. We then filtered out genes expressed in less than 10 cells. We performed dimension reduction using the rsvd package, using parameters k= 20, p= 15, q= 3. Phenograph was then used to cluster the macrophages over the 20 dimensions imputed by SVD using the default parameter at K=30. Phenograph yielded 10 clusters for macrophages in GBM28. Similarly for macrophages in GBM8, we performed transformation, gene filtering, dimension reduction and clustering using the same methods and parameters, while adding the medium library size normalization step for the raw count UMI data before all these steps. Phenograph yielded 11 clusters for macrophages in GBM8. For each patient (N=27) and each macrophage cluster (N=10) in GBM28, we simulated 5 pseudo-bulk, also constituting 1,350 samples in total. The parameters for simulation are the same as mentioned above. When deconvolving pseudo-bulks, we treated each 60 tumor sub-clusters and 11 macrophage clusters as individual cell types, and summed the inferred expression over all macrophage clusters to represent the total macrophage expression profile in each bulk. The cluster purity was calculated using the "purity" function from the NMF package.

Real bulk RNA-seq human whole blood with ground truth measured by flow-cytometry (Fig.1d-e). To test the performance of BayesPrism on real bulk RNA-seq, we deconvolved 12 human whole blood samples for which the cell type composition was known using flow-cytometry. We used the same PBMC RNA-seq dataset from the CIBERSROTx paper as the reference, which was obtained from a patient with non-small cell lung cancer (NSCLC) using 10x Genomics Chromium v2 (3′ assay)[40]. The bulk PBMC dataset and scRNA-seq reference were mismatched, as the bulk RNA-seq reference was performed on whole blood and the scRNA-seq reference with PBMCs. Neutrophils present in high abundance in the whole blood sample were not represented in the reference because neutrophils are polynucleated and do not isolate with PBMCs. Missing neutrophils may inflate the fraction of other myeloid[70] cell types that have similar expression. Thus, we inferred the proportion over a combined myeloid population and used the ground truth as the combined fraction of monocytes and neutrophils in all analyses. The bulk RNA-seq of human whole blood and the scRNA-seq reference of PBMCs from non-small cell lung cancer patients were downloaded from the CIBERSORTx website at: https://cibersortx.stanford.edu/download.php. As only the S mode of CIBERSORTx produced accurate results, as shown by the authors, we did not benchmark against the uncorrected and B mode.

Validation of inferred neutrophil fractions over TCGA bladder cancer dataset based on H&E grouping. (Supplementary Fig. 8). We benchmarked BayesPrism using haematoxylin and eosin

(H&E) histopathology estimates published with TIMER. In the TIMER dataset, H&E slides from TCGA bladder cancer samples were scored by a pathologist for the level of infiltration of neutrophils. The pathologist divided samples into High, Medium, and Low neutrophil groups. The pathological estimation of neutrophil levels for TCGA bladder cancer samples was downloaded from the http://cistrome.org/TIMER/download.html. Reference scRNA-seq data was limited for bladder cancer: Only one individual was available[71], and neutrophils were not annotated as a separate cell type from other myeloid cells. Nevertheless, deconvolution with BayesPrism revealed a myeloid cell fraction that was correlated with the pathologist designation (low, medium, and high; see the revised **Supplementary Fig. 8a**). To separate neutrophils from other myeloid cell types, we also tried a separate analysis by adding bulk RNA-seq data in purified blood cell populations[72] to the bladder cancer scRNA-seq reference. BayesPrism estimated neutrophil cell fractions that were correlated with the pathologist designation (**Supplementary Fig. 8b**). Moreover, the two samples selected for display in the TIMER paper were correctly estimated by BayesPrism as having either very low or very high infiltration of myeloid or neutrophil cells in both analyses (**Supplementary Fig. 8c**). Statistical significance was computed using a two-sided wilcoxon test. This analysis, especially when combined with both new and existing analyses, supports the use of BayesPrism in deconvolving cell type fractions in bulk cancer data.

Gene set enrichment analysis

The gene set enrichment scores shown in Fig. 2b and e were computed using the GSVA R package[73], using the marker genes of each subtype as the gene set. The GO analysis for inferred TCGA tumor gene programs was done using the topGO R package with the gene set of "biological process". Genes highly upregulated (one-sided t test, p<0.01) in each gene program were used as input.

Choosing cell type marker genes for correlation analysis

In **Supplementary Fig. 22**, 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[72], by assigning each gene to the cell type with the maximum expression value. Markers for oligodendrocytes, endothelial cells, pericytes, and microglias in normal brains were derived from the gene list generated by Lake et al. using normal brain scRNA-seq[74]. 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[75] 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[76] 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 BayesPrism 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

To avoid known clinical or genetic factors that have a strong influence on patient survival from confounding our survival analysis, we focused on the largest homogenous population of patients available for each cancer type. These included IDH-1 wildtype tumors for GBM and metastatic melanoma. We also attempted to control for HPV status in HNSCC. Although only 72 of 500 samples were annotated for HPV, we nevertheless reproduced trends (FDR adjusted $p < 0.1$) observed for T cells and endothelial cells in a small cohort of 56 HPV negative patients (**Supplementary Fig. 18**). We divided patients into high and low groups based on the feature of interest, e.g. weights of tumor gene programs 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	Normalization	# of cells	# of patient (scRNA-seq) / # of samples (bulk)	Clustered tumor cells	Accession ID
Tumor scRNA-seq					
refGBM8	Raw (UMI)	23793	8	YES	GSE103224
GBM28	TPM	7930	28	NO	GSE131928
scMel	Raw	6879	31	NO	GSE115978
scHNSCC	TPM	5902	18	NO	GSE103322

scBladder	TPM(UMI)	2075	1	NO	GSM4307111
Normal scRNA-seq					
PBMC 1 - 10x Chromium (v2) A	Raw (UMI)	3222	1	NA	GSE132044
PBMC 2 - 10x Chromium (v2)	Raw (UMI)	3362	1	NA	GSE132044
PBMC 1 - Smart-seq2	Raw	253	1	NA	GSE132044
PBMC from NSCLC patients	Raw (UMI) (converted from CPM)	1054	1	NA	https://cibersortx.stanford.edu/download.php
Normal mouse cortex single nucleus-seq					
Cortex2 sci-RNA-seq	Raw (UMI)	3791	1	NA	GSE132044
Cortex1 Smart-seq2	Raw	295	1	NA	GSE132044
TCGA bulk					
TCGA-GBM	Raw	NA	169	NA	https://portal.gdc.cancer.gov
TCGA-SKCM	Raw	NA	471	NA	
TCGA-HNSC	Raw	NA	502	NA	
TCGA-BLCA	Raw	NA	414	NA	
Normal bulk					
Whole blood bulk (with flow-sorted ground truth)	TPM	NA	12	NA	https://cibersortx.stanford.edu/download.php

Flow-sorted blood cells (from health controls)	Raw	NA	4	NA	GEO60424
IVY GAP					
IVY Anatomic Structures RNA-Seq	Raw	NA	122 samples across 10 tumors	NA	https://glioblastoma.alleninstitute.org
IVY Cancer Stem Cells RNA-Seq	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. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).

34. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).

35. 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).

36. 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).

37. Zheng, S. Benchmarking: contexts and details matter. *Genome Biol.* **18**, 129 (2017).

38. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).

39. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346–360.e4 (2016).

40. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

41. 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).

42. Dong, M. *et al.* SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. doi:10.1101/743591.

43. Hart, Y. *et al.* Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods* **12**, 233–5, 3 p following 235 (2015).

44. 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).

45. Hovestadt, V. *et al.* Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).

46. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).

47. 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).

48. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).

49. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).

50. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
51. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
52. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
53. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).
54. Joseph, J. V. *et al.* Hypoxia enhances migration and invasion in glioblastoma by promoting a mesenchymal shift mediated by the HIF1α-ZEB1 axis. *Cancer Lett.* **359**, 107–116 (2015).
55. Bhat, K. P. L. *et al.* Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. *Cancer Cell* **24**, 331–346 (2013).
56. Chu, T. *et al.* Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat. Genet.* **50**, 1553–1564 (2018).
57. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat. Genet.* (2016) doi:10.1038/ng.3590.
58. 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).
59. Chen, P. *et al.* Symbiotic Macrophage-Glioma Cell Interactions Reveal Synthetic Lethality in PTEN-Null Glioma. *Cancer Cell* **35**, 868–884.e6 (2019).
60. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5**, 16923 (2015).
61. Liu, M. *et al.* Targeting the IDO1 pathway in cancer: from bench to bedside. *Journal of Hematology & Oncology* vol. 11 (2018).
62. Raju, K. S., Alessandri, G., Ziche, M. & Gullino, P. M. Ceruloplasmin, copper ions, and angiogenesis. *J. Natl. Cancer Inst.* **69**, 1183–1188 (1982).

63. Xue, X. *et al.* Vasohibin 2 is transcriptionally activated and promotes angiogenesis in hepatocellular carcinoma. *Oncogene* **32**, 1724–1734 (2013).

64. Mahjour, F. *et al.* Mechanism for oral tumor cell lysyl oxidase like-2 in cancer development: synergy with PDGF-AB. *Oncogenesis* **8**, 34 (2019).

65. Judokusumo, E., Tabdanov, E., Kumari, S., Dustin, M. L. & Kam, L. C. Mechanosensing in T lymphocyte activation. *Biophys. J.* **102**, L5–7 (2012).

66. Saitakis, M. *et al.* Different TCR-induced T lymphocyte responses are potentiated by stiffness with variable sensitivity. *Elife* **6**, (2017).

67. 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).

68. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

69. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).

70. Schafflick, D. *et al.* Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nat. Commun.* **11**, 247 (2020).

71. Lee, H. W. *et al.* Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. *Genome Medicine* vol. 12 (2020).

72. 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).

73. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

74. 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).

75. 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).

76. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. (2019).

Main figures

Fig. 1 | BayesPrism algorithm flow and performance validation. a) Algorithmist flow of BayesPrism. **b**-**e**) Boxplots show the cell type-level Pearson's correlation coefficient and MSE for deconvolution of pseudo bulks of GBM28 using refGBM8 (**b** and **c**), and bulk RNA-seq human whole blood samples with ground truth measured by flow-cytometry (**d** and **e**). Boxes mark the 25th percentile (bottom of box), median (central bar), and 75th percentile (top of box). Whiskers represent extreme values within 1.5 fold of the inter quartile range. **f**) UMAP visualization shows the expression of individual cells in GBM28. The expression profiles of stromal cells before (gray) and after (black) correction were projected onto the UMAP manifold of the scRNA-seq (left). Malignant cells in patients with greater than 10 malignant cells (N=27) were visualized on the zoomed-in UMAP (right), and are colored by patient. 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.**g**) Scatter plot shows Spearman's correlation between the average expression of malignant cells in pseudo-bulk and that estimated by BayesPrism (red), CIBERSORTx group mode (orange) or total bulk (blue), as a function of the fraction of malignant cells in a subsampled set (N=270).

Fig. 2 | BayesPrism 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 BayesPrism. 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 BayesPrism 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.

Fig. 3 | 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-i**) KM plots show the survival associations with **b-e**) T cell infiltration, **f-h**) macrophage infiltration, and **i**) 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, and corrected using FDR. Hazard ratio is defined by high / low.

Fig. 4 | BayesPrism reveals spatial heterogeneity in GBMs. a) A graphical illustration of the anatomic structures of the IVY GAP samples. **b**) 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. **c**) Heatmaps show Spearman's rank correlation between normalized weights of gene programs and the fraction of stromal cells in GBM. **d**) 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. **e**) A model depicting the interaction between tumor gene programs and microenvironment in GBM.

Fig. 5 | Tumor pathways correlate with stromal cell fractions. a-b) Heatmaps show Spearman's rank correlation between normalized weights of gene programs and the fraction of stromal cells in HNSCC and melanoma. **c-g**) 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 BayesPrism 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 correlates with macrophage infiltration in IVY GAP. **b-d**) Boxplots show the BayesPrism 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**) Bars show the number of genes whose malignant cell expression level was correlated with the indicated cell types in the indicated tumor type. Bars are colored by -log10 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 BayesPrism inferred gene expression in malignant cells 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.

Supplementary figures

Fig. S1 | A detailed algorithmist flow of BayesPrism. Gray grids show the dimension of the variables used or inferred in each step.

Fig. S2 | Comparison between BayesPrism and other deconvolution methods using simulated noise. Line plots show the cell type-level Pearson's correlation coefficient (left) and MSE (right) as a function of the noise level.

Fig. S3 | Comparison between BayesPrism and other deconvolution methods on pseudo-bulks across different sequencing platforms and biological samples. Boxplots show the cell type-level Pearson's correlation coefficient and MSE for the deconvolutions of pseudo-bulk human PBMC scRNA-seq (**a** and **b**) and mouse cortex single nucleus-seq (**c** and **d**). Boxes mark the 25th percentile (bottom of box), median (central bar), and 75th percentile (top of box). Whiskers represent extreme values within 1.5 fold of the inter quartile range.

Fig. S4 | Performance of BayesPrism in inferring cell type composition and tumor gene expression on the leave-one-out pseudo-bulk data of HNSCC and melanoma. a-b) Scatter plots show the initial estimates of cell type fraction in **a**) HNSCC and **b**) melanoma versus the ground truth in pseudo-bulk. **c-d**) Scatter plots show the CIBERSORTx inferred cell type fraction in **c**) HNSCC and **d**) melanoma versus the ground truth in pseudo-bulk. **e-f**) UMAP shows the expression profile of individual tumor cells in scRNA-seq of **e**) HNSCC and **f**) melanoma colored by patient ID. Patients with >50 tumor cells and cells with reads detected for >3000 genes 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. **g-h**) Scatter plot shows the Pearson's correlation coefficient between inferred expression and that of the averaged expression from malignant cells in scRNA-seq of **g**) HNSCC and **h**) melanoma as a function of the fraction of tumors in each simulated data. The correlation coefficient was computed on DESeq2 variance-stabilized transformed values. Red marks the correlation inferred by BayesPrism, while blue marks that of total expression of the simulated data.

Fig. S5 | Comparison between BayesPrism and various modes of CIBERSORTx. Scatter plots show the inferred cell type fraction in the pseudo-bulk GBM28 (**a-d**), and simulated 270 pseudo-bulk dataset (**e-h**). (**i-p**) Scatter plots show the performance of individual cell types in **e** and **f**.

Fig. S6 | BayesPrism is robust to missing cell types in the reference. Line plots show the cell type-level Pearson's correlation coefficient (left) and MSE (right) for the deconvolution of simulated GBM28 (N=1350) using refGBM8 with T cells removed as the reference. The X axis marks the midpoints of each 10% width bin. Lines are colored by cell type. Solid lines represent the initial fractions, while dashed lines represent updated fractions. Vertical bars mark the average observed T cell fraction in GBM8 and GBM28.

Fig. S7 | BayesPrism is robust to downsampled reference. a) Boxplots show the distribution of cell type-level Pearson's correlation coefficient as a function of the number of downsampled patients in refGBM8, in which tumor cells are excluded from patients that are not sampled. Boxes mark the 25th percentile (bottom of box), median (central bar), and 75th percentile (top of box). Whiskers represent extreme values within 1.5 fold of the inter quartile range. **b-e**) Line plots show the cell type-level Pearson's correlation coefficient (left) and MSE (right) for the deconvolution of simulated GBM28 (N=1350) by refGBM8 (**b** and **c**) and HNSCC leave-one-out test using reference with downsampled single cells (**d** and **e**). The X axis marks the maximum number of cells in each cell type (or subclones for tumor) in the reference. Lines are colored by cell type. Dashed vertical lines mark the observed number of cells in each cell type in the original reference.

Fig. S8 | Validation of inferred neutrophil fractions in TCGA bladder cancer using H&E staining. a-b) Boxplot shows distribution of the **a**) estimated fractions of myeloid lineage cells using macrophage in the scRNA-seq bladder cancer as the reference, and **b**) estimated fractions of neutrophils using neutrophil from purified bulk as the reference. P values are computed using a two-sided wilcoxon test. **c**) Samples selected for representation in the TIMER paper were compared to BayesPrism.

Fig. S9 | BayesPrism improves estimates of gene expression in stromal cells. Scatter plots show log2 gene expression in stromal cells before and after batch correction versus that of true expression. Genes with zero expression counts in the reference are colored in red, and those with non-zero expression counts are colored in blue.

Fig. S10 | Heatmap shows the correlation matrix computed on tumor gene expression in pseudo-bulks estimated by total expression (the left column) and BayesPrism (the right column), at various thresholds of tumor cell fraction.

Fig. S11 | BayesPrism accurately recovers the heterogeneity in the expression of macrophage. a) Scatter plot shows Spearman's correlation between the average expression from macrophages in the pseudo-bulk and gene expression estimated by BayesPrism (red) or total bulk (blue) as a function of the fraction of macrophages in the simulated pseudo-bulk (N=1350). **b**) UMAP visualization shows the expression of individual macrophages in GBM28. 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. **c-f**) Heatmap shows the correlation matrix computed on macrophage gene expression in pseudo-bulks estimated by total expression (**c** and **e**) and BayesPrism (**d** and **f**), at various thresholds of macrophage fraction.

Fig. S12 | Comparison between BayesPrism and the two different modes of expression inference by CIBERSORTx. Scatter plot shows Spearman's correlation between gene expression estimated by BayesPrism (red), total bulk (blue), CIBERSORTx group mode (orange) or CIBERSORTx high resolution mode (purple) and the average expression from malignant cells in scRNA-seq as a function of the fraction of malignant cells in the dataset containing the 270 simulated samples. The correlation coefficient was calculated on 53 imputable genes by the high-resolution mode out of the top 1000 most variable genes in tumor cells.

Fig. S13 | Choosing number of pathways K and initializing tumor basis η_0 for TCGA-GBM using NMF factorization and consensus clustering. a) Line plots show various metrics on consensus clustering as a function of K. **b**) Heatmaps show the consensus clustering matrix of different choices of K.

Fig. S14 | KM plots of tumor pathway weights inferred from three tumor types across TCGA samples. KM plots show the survival associations of all tumor pathways in **a**) GBM, **b**) HNSCC, and **c**) melanoma. The left column of each panel shows the survival curves at the cutoff between 0.1 and 0.9 quantile that yields the minimum p value. The right columns of each panel shows the survival curve at the cutoff using the 0.2 and 0.8 quantile.

Fig. S15 | Comparison between tumor purity inferred by BayesPrism, ABSOLUTE and ESTIMATE. a-c) Scatter plots show the correlation between tumor fractions inferred by each method. Dashed lines mark the y=x and regression fit. Rectangular boxes mark the outliers predicted as 1 by ABSOLUTE. **d-e**) Scatter plots show the correlation between tumor fractions inferred by each method with outliers in **a** and **b** removed. **f**) Violin plot shows the distribution of library size of tumor and non-tumor expression in GBM8. The higher total expression of tumor cells explains the linear shift in estimating tumor fraction by BayesPrism, as it estimates the total fraction of reads rather than cell number for each cell type.

Fig. S16 | Heatmaps show the Spearman's rank correlation between stromal cells in each tumor type.

Fig. S17 | KM plots of all cell types inferred from three tumor types across TCGA samples. KM plots show the survival association of all stromal cell types in **a**) GBM, **b**) HNSCC, and **c**) melanoma. The left column of each panel shows the survival curves at the cutoff between 0.1 and 0.9 quantile that yields the minimum p value. The right columns of each panel shows the survival curve at the cutoff using the 0.2 and 0.8 quantile.

Fig. S18 | KM plots of all cell types inferred from HPV-negative HNSCC TCGA samples. The left column of each panel shows the survival curves at the cutoff between 0.1 and 0.9 quantile that yields the minimum p value. The right columns of each panel shows the survival curve at the cutoff using the 0.2 and 0.8 quantile.

Fig. S19 | Stromal cell distribution in previously reported GBM subtypes. Asterisks mark the significant differences between subtypes based on one-way ANOVA and studentized range statistics.

Fig. S20 | Choosing the number of pathways, K, and initializing tumor basis η_0 for TCGA-HNSC (HNSCC) using NMF factorization and consensus clustering. a) Line plots show various metrics on consensus clustering as a function of K. **b**) Heatmaps show the consensus clustering matrix of different choices of K.

Fig. S21 | Choosing the number of pathways, K, and initializing tumor basis η_0 for TCGA-SKCM (melanoma) using NMF factorization and consensus clustering. a) Line plots show various metrics on consensus clustering as a function of K. **b**) Heatmaps show the consensus clustering matrix of different choices of K.

Fig. S22 | BayesPrism removes false positive correlates from cell type marker genes. Violin plot shows the distribution of Pearson's correlation between gene expression in malignant cells inferred by BayesPrism (left violins) or total gene expression of bulk RNA-seq (right violins) and BayesPrism predicted fractions of each cell type on their corresponding marker genes over TCGA-GBM. Median correlations are shown by white dots and upper/lower quartiles are shown by bars. Braces on the horizontal direction label the cell type fraction on which the correlations were computed. Color indicates the cell type of which the marker genes are curated from independent datasets.

Table S1 | Anatomic structures in IVY GAP dataset. Table shows the abbreviations of anatomic structures and their associated features in **a**) the anatomic structures RNA-Seq study and **b**) the cancer stem cells RNA-Seq study.

Table S2 | Gene ontology analysis of tumor gene programs learned in three cancer types.