Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Nuclear morphology from H&E slides shows a small, composition- confounded association with hypoxia transcription in glioma: a cross-cohort benchmark

Kaihua Cao
Shanghai Jiao Tong University
Zhihua Cheng
Shanghai Jiao Tong University
Yu Zeng
Shanghai Jiao Tong University
Pengran Yu
Shanghai Jiao Tong University
Wenchuan Zhang
zhangwench88@sjtu.edu.cn
Shanghai Jiao Tong University

Article

Keywords: glioblastoma, hypoxia, computational pathology, whole-slide imaging, confounding, negative results

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11252688/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Nuclear morphology from H&E slides shows a small, composition-confounded association with hypoxia transcription in glioma: a cross-cohort benchmark

Kaihua Cao[1,†], Zhihua Cheng[1,†], Yu Zeng[1], Pengran Yu[1] & Wenchuan Zhang[1,*]

[1] Department of Neurosurgery, Shanghai Ninth People's Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200011, China.

[†] These authors contributed equally: Kaihua Cao, Zhihua Cheng.

[*] Correspondence: Wenchuan Zhang, zhangwench88@sjtu.edu.cn; zhangwench88@hotmail.com; Department of Neurosurgery, Shanghai Ninth People's Hospital, School of Medicine, Shanghai Jiao Tong University, No. 639 Zhizaoju Road, Shanghai 200011, China.

Abstract

Hypoxia is a hallmark of glioblastoma but is not measured in routine diagnosis, whereas hematoxylin and eosin (H&E) slides are universally available. We asked whether 31 interpretable nuclear and tissue-architecture features from H&E slides of TCGA-GBM, CPTAC-GBM and TCGA-LGG are associated with a bulk Hallmark hypoxia transcription score, and how far composition, IDH status, signature choice and technical confounding explain it. Hypotheses were defined post hoc, and only later-processed CPTAC patients (n = 60–61) formed a temporal hold-out. No feature was significant in TCGA-GBM alone (0/31), whereas 9/31 were significant in the pooled analysis. Three selected features had pooled $\rho \approx 0.19$ but were non-significant in the hold-out ($\rho = 0.05$–0.14; about 34% power). The score correlated strongly with immune, stromal and myeloid scores ($\rho = 0.35$–0.71), and composition predicted it better than morphology. Associations vanished after adjustment for IDH and grade in TCGA-LGG and for sequencing depth in Visium, and no independent prognostic value was found. In HCMI-CMDC, the only further open cohort with paired slides and RNA (n = 35; low power), the features did not replicate ($\rho = -0.15$ to -0.20; 0/31 BH), and the pooled estimate fell to $\rho \approx 0.15$. The findings are hypothesis-generating.

Keywords: glioblastoma; hypoxia; computational pathology; whole-slide imaging; confounding; negative results

Introduction

Hypoxia and the associated pseudopalisading necrosis are characteristic of glioblastoma (GBM) and are linked to angiogenesis, immune remodelling and therapy resistance[1,2]. Necrosis and microvascular proliferation are also used to classify and grade adult diffuse gliomas[3]. Transcriptional hypoxia signatures, such as the MSigDB Hallmark set and signatures derived from head-and-neck, pan-cancer and cell-line data, can be computed from bulk RNA[4–7], but RNA profiling is not part of routine diagnosis, whereas H&E staining is.

Computational pathology can infer molecular states, including mutations and gene expression, from H&E slides in many tumour types[8,9], typically with weakly supervised multiple-instance learning or foundation-model embeddings[10]. Interpretable feature pipelines have also been used[11]. In glioma, deep-learning pathomics has been applied to survival prediction, tissue-region segmentation and multimodal integration[12–16], and hypoxia-associated morphology has been modelled in breast cancer[17]. We make no claim of priority for glioma pathomics, for hypoxia inference from H&E, or for multiple-instance-learning models.

Our aim is narrower and deliberately conservative. We ask whether a small set of interpretable nuclear and tissue-architecture features is associated with bulk hypoxia transcription in GBM, whether the association is reproduced in an independent cohort, and how much of it is explained by tissue composition, IDH status and grade, signature choice, technical confounding (sequencing depth, stain, segmentation) and low within-patient reproducibility. Negative and attenuated results are reported in full, and exploratory analyses are distinguished from the single quasi-confirmatory test that the data allow.

Results

Study design, cohorts and analysis status

We analysed H&E diagnostic slides and matched bulk RNA from TCGA-GBM (93 slides with at least 10 usable tiles; 92 with the log2-scale score), CPTAC-GBM (n = 152) and TCGA-LGG (n = 488); a further open cohort, HCMI-CMDC (35 patients), was added on 1 October 2026 and is reported separately below. Thirty-one features were extracted per slide and related to the Hallmark hypoxia score (Fig. 1; Supplementary Table S19). The sequence of analyses matters and is stated in full in Methods. The 31-feature screen was exploratory. Three features (H1, hypocellular area fraction; H2, eosin-rich acellular fraction; H3, coefficient of variation of nearest-neighbour nuclear distance) were then defined post hoc, with knowledge of the screen in the first 91 CPTAC patients. Only the 60–61 CPTAC patients whose features were completed after the freeze time form a temporal hold-out. All other results, including pooled analyses that contain the patients used for feature selection, are exploratory and probably optimistic.

The Hallmark score tracks necrosis-adjacent tissue in IvyGAP but is composition-linked in bulk tumours

In IvyGAP, the Hallmark score separated pseudopalisading cells around necrosis (CTpan) from cellular tumour (CT; AUC = 0.979; Mann–Whitney $P = 2.0 \times 10^{-9}$), and the paired difference between necrosis-adjacent regions and cellular tumour was positive in all 8 tumours (Wilcoxon P = 0.0078). With Hallmark genes removed, separation persisted for the Winter signature (AUC = 0.772, $P = 6.7 \times 10^{-4}$; Fig. 2E) but not for Buffa (AUC = 0.557, P = 0.48; Fig. 2F). In bulk tumours, however, the Hallmark score correlated with composition (Fig. 2A–C): ρ with the immune score was 0.345

(TCGA-GBM, n = 154), 0.539 (CPTAC, n = 200) and 0.586 (TCGA-LGG, n = 516), with the stromal score 0.434, 0.589 and 0.708, and with the myeloid score 0.364, 0.483 and 0.492 (range 0.35–0.71). In TCGA-GBM, median-split differential expression (81 genes at FDR < 0.05 and |log2 fold change| > 1; 78 up, 3 down; Fig. 2G) and enrichment (Hallmark gene sets 44/50 at FDR < 0.05, 39/49 after removing the score genes; Fig. 2H) were dominated by inflammatory, TNF-α/NF-κB, epithelial–mesenchymal-transition and myeloid programmes, and the top protein–protein interaction hubs were *IL6*, *IL1B*, *CD44*, *FN1*, *CXCL8*, *TLR2*, *CCL2*, *ITGB2*, *COL1A1* and *CXCR4* (Fig. 2I; Supplementary Fig. S10). The bulk score is therefore entangled with myeloid and mesenchymal composition.

Exploratory screen and temporal hold-out

In TCGA-GBM alone (n = 90–92), none of the 31 features was significant after Benjamini–Hochberg (BH) correction (0/31; best max-T P = 0.76). In CPTAC alone, 8/31 were BH-significant (0/31 at max-T; best P = 0.088), and 9/31 were in the fixed-effect meta-analysis of the two cohorts (Fig. 3A). Feature-wise correlations were concordant between cohorts (ρ = 0.823; 26/31 same sign). In the pooled, cohort-stratified max-T analysis, 4/31 features reached a family-wise error rate below 0.05 (Fig. 3F,G): fraction of hypocellular tiles (P = 0.032), eosin-rich acellular fraction (P = 0.041), SD of the hypocellular fraction (P = 0.044) and its 90th percentile (P = 0.047). The features are highly inter-correlated (Supplementary Fig. S2).

H1–H3 had small positive correlations with the Hallmark score in every subset, but the evidence was weak (Fig. 3B–E; Table 1). Pooled ρ was 0.186, 0.193 and 0.188 (P = 0.0037, 0.0026 and 0.0034). In the hold-out (n = 60–61), the estimates were ρ = 0.135, 0.078 and 0.051 (P = 0.298, 0.550 and 0.699): same sign, non-significant, with about 34% power to detect ρ = 0.2. Of the three features, only H2 reached max-T family-wise significance (P = 0.041; H1 P = 0.056; H3 P = 0.053). The hold-out therefore did not confirm the association, and we interpret it as compatible with either a small true effect or none. Pooled P values include the patients used for feature selection and are not confirmatory.

Composition, signature choice and robustness

After adjustment for immune, stromal and myeloid scores and tile count, no feature was BH-significant in any cohort or in the meta-analysis (nominal P < 0.05: TCGA-GBM 0, CPTAC 4, meta 2; Fig. 4A). For H1–H3 (meta, adjusted), the eosin-rich fraction had ρ = 0.135 (95% CI 0.009–0.257; P = 0.036), the hypocellular fraction ρ = 0.112 (P = 0.084) and nearest-neighbour CV ρ = 0.066 (P = 0.306). In pooled ordinary least squares the eosin-rich coefficient was 0.125 (95% CI 0.007–0.243; P = 0.039; Fig. 4C). This single nominal result is not corrected for the number of signatures tested.

The association depended on the signature (Fig. 4B). Crude pooled correlations for H1–H3 were 0.15–0.16 for Winter and 0.14–0.16 for Elvidge but −0.002 to 0.04 for Buffa. With Hallmark-overlap genes removed, correlations were 0.08–0.12 for Winter (P = 0.065–0.24), 0.11–0.13 for Elvidge (P = 0.047–0.091) and −0.05 to −0.10 for Buffa (all non-significant); after composition adjustment they

were −0.08 to 0.06 (all $P \geq 0.24$). Part of the Hallmark association is therefore attributable to composition and to the gene content of the Hallmark set.

Within-patient reproducibility of slide-level features was low (median ICC(1) over the 31 features 0.106; H1 0.153, H2 −0.091, H3 0.098; Fig. 4G). Excluding five TCGA-GBM slides with failed stain estimation (n = 87) left the TCGA-GBM results unchanged (0/31 BH-significant), with the pooled analysis also unchanged at 9/31 BH-significant features (Fig. 4H shows the TCGA-GBM estimates; Supplementary Table S10). Averaging all TCGA-GBM slides per case (n = 154; same patients, not independent) strengthened crude associations (6/31 BH-significant) but none remained after composition adjustment (Supplementary Table S7). Nuclear-segmentation agreement with StarDist was moderate (slide-level density $\rho = 0.78$, n = 96) and the nearest-neighbour association was similar with either segmenter (Supplementary Fig. S4); no absolute accuracy claim is possible without expert annotation.

Cross-cohort prediction

A ridge model on the 31 morphology features trained on TCGA-GBM predicted the Hallmark score in CPTAC with $\rho = 0.255$ (95% CI 0.094–0.404; permutation $P \approx 0.001$; n_train = 90, n_test = 151), and the reverse direction gave $\rho = 0.191$ (−0.028 to 0.404; $P = 0.029$; Fig. 4D,E). An RNA-composition-only model (immune, stromal, myeloid) predicted the score better ($\rho = 0.484$ and 0.505). After residualising the outcome on composition, the incremental morphology signal was $\rho = 0.075$ ($P = 0.19$) and 0.091 ($P = 0.21$; Fig. 4F). ResNet-50 embeddings trained on TCGA-GBM predicted the CPTAC score with $\rho = 0.309$ ($P = 1.4 \times 10^{-4}$, n = 148) in an analysis without composition adjustment, whereas within TCGA-GBM, cross-validated prediction from ResNet-50 or morphology was not above chance (permutation P = 0.25–0.83).

The signature comparison was heterogeneous (Supplementary Table S20). Multivariate morphology predicted Winter-without-Hallmark-genes in one direction ($\rho = 0.192$, $P = 0.010$; reverse 0.139, P = 0.10) and Buffa-without-Hallmark-genes in the TCGA→CPTAC direction ($\rho = 0.151$, $P = 0.036$; reverse 0.031, P = 0.37), although single-feature Buffa correlations were near zero. The Buffa result is therefore more nuanced than “no association”, and the pattern is not consistent across analyses.

Confounding by IDH status and grade in TCGA-LGG and by sequencing depth in Visium

In TCGA-LGG (n = 488), 5/31 features were BH-significant with the Hallmark score in crude analysis (mainly nuclear elongation and eccentricity, $\rho \approx 0.29$), but none remained after adjustment for IDH status and grade (0/31; best q = 0.100; Fig. 5A). H1–H3 had crude $\rho = -0.013$, 0.018 and 0.036. Feature effects were not concordant with GBM ($\rho = -0.17$, $P = 0.37$; 12/31 same sign; Fig. 5E) and GBM-trained models did not transfer (morphology $\rho = 0.040$, $P = 0.38$; ResNet-50 $\rho = 0.007$, P =

0.88; Fig. 5D). The score was much higher in IDH-wild-type than in IDH-mutant tumours (medians 0.215 and −0.075; $P = 3.7 \times 10^{-23}$; Fig. 5B).

In the 22 Visium slides (9 patients), the uncorrected z-mean score was largely a readout of sequencing depth (median ρ with log UMI = 0.808; Fig. 5F). Its positive association with hematoxylin density (median ρ = 0.165; Wilcoxon P = 0.0083; patient-level P = 0.098) disappeared after depth correction (median ρ = −0.028, P = 0.406; patient-level mean −0.031, P = 0.65; Fig. 5G,I), as did the apparently lower score in pale zones (median difference −0.076, P = 0.013; after correction 0.005, P = 0.34; Fig. 5H). Spatial transcriptomics therefore provided no support for a local morphology–hypoxia relation.

Prognosis

The Hallmark score was not an independent prognostic factor (Fig. 6A). In TCGA-GBM, the hazard ratio (HR) per standard deviation was 1.267 (95% CI 1.05–1.53; P = 0.0130; n = 152, 121 events) univariably and 1.132 (0.922–1.388; P = 0.236; n = 148) after adjustment for age, sex and IDH. Adding the score raised the C-index from 0.605 to 0.618 (likelihood-ratio P = 0.235; bootstrap-corrected 0.596; calibration slope 0.816; Fig. 6D; Supplementary Fig. S8). In CPTAC-GBM the univariable HR was 1.17 (0.995–1.379; P = 0.058; n = 190) and 1.13 (0.80–1.59; P = 0.487; n = 98) after adjustment; with immune and myeloid scores added, the hypoxia HR was 1.03 (P = 0.856) whereas the myeloid HR was 3.47 (1.30–9.21; P = 0.013; Fig. 6H). In TCGA-LGG the univariable HR of 1.64 (1.43–1.87) fell to 1.127 (0.92–1.38; P = 0.249) after adjustment. Kaplan–Meier median splits were non-significant (TCGA-GBM n = 152, log-rank P = 0.364; image subset n = 95, P = 0.110; Fig. 6B,C), and out-of-fold time-dependent AUCs at 12, 18 and 24 months were 0.666, 0.543 and 0.571 for the score alone, 0.605, 0.563 and 0.592 for age, sex and IDH, and 0.664, 0.555 and 0.583 combined (Fig. 6F).

The ResNet-50 elastic-net Cox model failed cross-validation (out-of-fold C-index 0.435 ± 0.033; permutation P = 0.952; n = 96; apparent C-index 0.809, indicating overfitting), and the morphology-only model gave 0.486 (permutation P = 0.57; Fig. 6D,E). A Kaplan–Meier split by the ResNet out-of-fold risk had P = 0.028, but in the direction opposite to that expected (HR 0.60 for the “high-risk” group; Fig. 6G). We regard this as a cross-validation artefact, not a finding.

Public surrogates for pathologist annotation, immunohistochemistry and external cohorts

Because pathologist annotation, CA9/HIF1A immunohistochemistry (IHC) and a large independent image-bearing cohort were unavailable, we used public data as partial, post hoc substitutes. These are surrogates, not equivalents. H&E images of IvyGAP were analysed inside the polygons of the laser-microdissection regions with the same pipeline. The RNA score separated necrosis-adjacent wells (CTpan plus perinecrotic zone, n = 43) from cellular tumour (n = 95) with AUC = 0.959 ($P = 6.5 \times 10^{-18}$), whereas morphology separated them weakly (hypocellular fraction AUC = 0.628, paired P =

0.046; nuclear density AUC = 0.531, paired P = 0.58; Fig. 7A,B), and a morphology-only classifier reached a cross-validated AUC of 0.648 (permutation P = 0.0066). Across 194 wells, 8/31 features correlated with the score at BH < 0.05 and 4/31 within tumours, but 0/31 after adjustment for region and tumour (Fig. 7C); features therefore follow anatomical region (chiefly the hypocellular leading edge and infiltrating tumour) rather than hypoxia independently of region (Supplementary Fig. S12).

GDC records percent necrosis for 310 TCGA-GBM frozen slides from 154 cases. Seventeen of 31 features correlated with it at BH < 0.05 (hypocellular fraction $\rho = 0.290$, $P = 2 \times 10^{-7}$; Fig. 7D), and a patient-grouped ridge model predicted it with $\rho = 0.330$ (permutation P = 0.002; AUC for ≥ 10% necrosis 0.665; Fig. 7E). At case level the RNA score correlated with mean necrosis ($\rho = 0.426$, n = 154; partial $\rho = 0.355$ after composition adjustment). Features of the diagnostic slide did not significantly predict case-mean frozen-slide necrosis (0/31 BH; Supplementary Fig. S13). In CPTAC-GBM proteomics (n = 99), a 149-protein hypoxia composite correlated with the RNA score ($\rho = 0.662$, $P = 9 \times 10^{-14}$; 0.569 after adjustment; Fig. 7F) and so did CA9 protein ($\rho = 0.442$), but HIF1A protein was not detected, TCGA RPPA lacked CA9 and HIF1A antibodies, and Human Protein Atlas glioma CA9 IHC was negative in all 9 tumours (Supplementary Fig. S14).

In CGGA (n = 1,018; no images), the score's prognostic value in primary tumours attenuated from HR 1.71 (n = 626) to 1.13 after age, sex, IDH and grade (P = 0.032; n = 582) and to 1.10 (P = 0.23) with composition scores; in primary WHO IV GBM (n = 218) the adjusted HR was 1.09 (P = 0.23; Fig. 7G; Supplementary Fig. S15). In GLASS (355 RNA samples; no images), the score did not rise at first recurrence (paired change −0.029, 95% CI −0.103 to 0.046; Wilcoxon P = 0.577; n = 145 pairs; Fig. 7H). None of these analyses validates the morphology features, because CGGA and GLASS have no images.

An additional open cohort with paired slides and RNA: HCMI-CMDC

A survey of public repositories (Supplementary Note, 'Survey of public cohorts') found one further open human glioma cohort with both H&E slides and tumour RNA-seq: HCMI-CMDC. We analysed 35 patients (36 samples, 71 slides; 22 primary and 13 recurrent tumours in the primary analysis unit; 35/36 samples frozen-section), using the unchanged pipeline and an analysis plan written before any HCMI association was computed (Methods). The cohort is small: power was about 21% for $\rho = 0.2$, and 80% only for $|\rho| \geq 0.46$. In the 31-feature screen, no feature was significant after BH correction (0/31; best nominal feature nuclear eccentricity, $\rho = 0.345$, P = 0.042, q = 0.62). The three pre-specified features pointed in the opposite direction to TCGA-GBM and CPTAC: H1 $\rho = -0.155$ (95% CI −0.46 to 0.19; P = 0.37), H2 $\rho = -0.197$ (−0.50 to 0.15; P = 0.26) and H3 $\rho = -0.154$ (−0.46 to 0.19; P = 0.38; Supplementary Fig. S16). After adjustment for immune, stromal and myeloid scores the estimates were −0.25, −0.20 and −0.26 (P = 0.16–0.27), again with 0/31 BH-significant features. The sign was similar in all 36 samples ($\rho = -0.14$ to -0.17) and in the 22 primary tumours alone (H1

$\rho = -0.29$, P = 0.19; H2 $\rho = -0.13$, P = 0.57). The Hallmark score in HCMI correlated with immune ($\rho = 0.43$), stromal (0.41) and myeloid (0.39) scores, as in the other cohorts (Supplementary Fig. S16).

Across the 31 features, the HCMI correlations were not concordant with the TCGA/CPTAC mean (feature-wise $\rho = -0.43$, 12/31 same sign). Because the 31 features are strongly inter-correlated, this naive P value (0.016) is not valid; the permutation P value (5,000 permutations of the HCMI score) was 0.53 (Supplementary Fig. S16), so we interpret the result as no evidence of concordance rather than as significant discordance. Adding HCMI to the fixed-effect meta-analysis lowered the pooled estimates for H1–H3 from $\rho \approx 0.19$ to 0.146–0.149 (P = 0.014 for each; 95% CI lower bounds about 0.03), and the number of BH-significant features in the pooled analysis fell from 9/31 (two cohorts) to 0/31 (three cohorts; best q = 0.056; 9 features with $P < 0.05$). Heterogeneity across the three cohorts was moderate and not significant (Cochran Q P = 0.10–0.16; $I^2 = 0.45$–0.57), and the HCMI estimate differed nominally from the two-cohort pooled estimate for H2 (P = 0.036; H1 P = 0.066; H3 P = 0.065; uncorrected; Supplementary Fig. S16). The three-cohort pooled estimates are optimistic, because H1–H3 were chosen on TCGA-GBM and CPTAC.

Models trained on the 31 features in TCGA-GBM, in CPTAC or in both did not predict the Hallmark score in HCMI ($\rho = -0.10$, -0.08 and -0.07; 95% CI for the TCGA model -0.40 to 0.25; permutation $P \geq 0.65$ one-sided), and ResNet-50 embeddings transferred from TCGA also did not ($\rho = 0.07$, P = 0.70; Supplementary Fig. S16). With signatures from which Hallmark genes had been removed, results were inconsistent: the Buffa_ex score was predicted with $\rho \approx 0.43$–0.44 from all three training sets ($P \approx 0.009$ two-sided, uncorrected), whereas the Winter_ex score was predicted in the wrong direction by the CPTAC model ($\rho = -0.36$) and the Elvidge_ex score not at all. The single features H1–H3 correlated negatively with Buffa_ex ($\rho = -0.54$ to -0.59, $P \leq 0.001$) and with Winter_ex ($\rho =$ -0.40 to -0.47), and these negative associations persisted after composition adjustment (Supplementary Fig. S16). The sign of these associations is therefore opposite to that of the transfer result for Buffa_ex. We regard all alternative-signature results as nominal, uncorrected and descriptive (many tests in a small cohort), and they do not alter the conclusion that HCMI did not reproduce the TCGA/CPTAC association.

Discussion

Interpretable H&E features showed a small, directionally consistent association with a bulk hypoxia transcription score across two GBM cohorts, in line with the expectation that hypocellular, eosin-rich (oedematous or necrotic) tissue and irregular nuclear spacing accompany hypoxic regions. The evidence is nevertheless weak and, in our judgement, does not justify calling H&E morphology a hypoxia biomarker in glioma. The association was (i) not significant in TCGA-GBM alone, (ii) not significant in the only temporal hold-out, which had limited power, (iii) largely shared with immune, stromal and myeloid composition, (iv) absent for the Buffa signature and weaker after removing

Hallmark genes, (v) undermined by low within-patient reproducibility, and (vi) absent in spatial transcriptomics once sequencing depth was controlled. In TCGA-LGG, apparent associations vanished after IDH and grade adjustment, a reminder of how easily morphology–expression correlations can reflect diagnostic category. The score was not an independent prognostic factor, and a deep-learning survival model failed cross-validation. In the one further open cohort with slides and RNA, HCMI-CMDC (n = 35), the association was not reproduced: point estimates were negative (ρ = −0.15 to −0.20), the confidence intervals included zero and, only just, the pooled value of about 0.19, and adding the cohort lowered the pooled estimate to 0.15 with no BH-significant feature. Because power was low (about 21% for $\rho = 0.2$) and the cohort differs from the others (mostly frozen sections, 13 of 35 recurrent tumours, a different consortium), this result neither refutes nor supports the association, but it removes the only independent test that was available and weakens any claim of generality.

The IvyGAP results show that the Hallmark score tracks pseudopalisading-necrosis regions, which supports its face validity as a necrosis-associated readout, but they do not validate the image features: the features follow anatomical region and separate necrosis-adjacent tissue only weakly (cross-validated AUC 0.65 versus 0.96 for RNA). Conversely, the differential-expression analysis shows that in bulk tumours the same score is dominated by inflammatory, mesenchymal and myeloid programmes, which likely explains why composition scores predict it better than morphology does. The public surrogates add coarse support: GDC-recorded necrosis on frozen sections was moderately associated with hypocellularity features ($\rho \approx 0.3$) and with the RNA score, and CPTAC proteomics showed that the score is reflected at protein level. However, these data are coarse percentages on frozen sections or different aliquots, HIF1A protein was undetected and HPA glioma CA9 IHC was negative, so no IHC-level confirmation could be obtained. Whether the Buffa signature is less composition-sensitive or simply noisier cannot be resolved here; its multivariate prediction in some analyses merits testing in independent data but may equally be noise given the number of analyses.

Deep-learning and multiple-instance pathomics studies in glioma, including a 2026 histomorphologic signature in TCGA gliomas[14], region-segmentation and RNA-correlation work in GBM[13] and multimodal survival models[12,15,16], report predictive models, and hypoxia-associated morphology has been modelled in breast cancer[17]. We make no claim of novelty in these areas. What this study adds is a transparent benchmark with a small interpretable feature set, explicit confounder analysis (composition, IDH and grade, signature, depth, stain, segmenter) and full reporting of negative and attenuated results. Foundation-model embeddings[10] were not tested and might carry stronger signal, so our ResNet-50 embeddings are a baseline only.

The study has limitations. (1) Hypotheses were defined post hoc, the only near-clean test set (n = 60–61) had about 34% power for $\rho = 0.2$, and pooled P values are optimistic. (2) Bulk RNA scores are confounded by composition and purity, and the ESTIMATE-like gene-set approximation and the ad

hoc 18-gene myeloid score are imperfect adjusters. (3) Within-patient reproducibility was low (ICC ≈ 0.1), single-slide sampling is noisy, and frozen sections carry artefacts. (4) We had no pathologist annotation and no CA9 or HIF1A IHC of our own, and the public surrogates are not equivalent to either. (5) TCGA and CPTAC differ in scanners, processing, RNA platforms and tile-sampling schemes, and stain normalisation was per slide only. (6) The Visium data comprise 9 patients without machine-readable region annotation, and IvyGAP polygons are microdissection regions. (7) Survival analyses are underpowered, and IDH-mutant GBM is rare in the cohorts. (8) Many signature, cohort and sensitivity analyses were run, so nominal P values outside the declared families are descriptive. (9) The only additional open cohort with slides and RNA that we could find (HCMI-CMDC, n = 35) was small (about 21% power for $\rho = 0.2$), mostly frozen-section, included recurrent tumours and a different consortium, was analysed after the main results were known (plan written before its association statistics were computed), and its alternative-signature results are nominal; it is not a sufficient validation, and no larger open cohort exists to our knowledge. (10) The TCGA-GBM score changed from 8 nominal hits (FPKM z-mean) to none (log2 z-mean), showing sensitivity to the score scale.

In conclusion, simple H&E nuclear-morphology features show only a small ($\rho \approx 0.2$), cross-cohort concordant but largely composition-explained and signature-dependent association with the Hallmark hypoxia programme in bulk RNA, no independent prognostic value and no depth-robust spatial support; the association was not reproduced in the one further open cohort with paired slides and RNA (HCMI-CMDC, n = 35, low power). Pathologist-annotated necrosis and hypocellular regions, CA9/HIF1A IHC, larger image-bearing independent cohorts and foundation-model comparisons would be needed before any diagnostic use is considered.

Methods

Study design and analysis status

This is a retrospective secondary analysis of public, de-identified data. The sequence of analyses matters for interpretation and is therefore stated explicitly. (i) A 31-feature panel was screened against the Hallmark hypoxia score in TCGA-GBM (n = 93, 0/31 BH-significant) and in the first 91 CPTAC-GBM patients (9/31 BH-significant) — *exploratory*. (ii) An internal, time-stamped analysis plan (Supplementary Note, ‘Analysis plans’; not a registered protocol) then defined three features (H1–H3) and the adjustment strategy. H1–H3 were chosen for biological rationale (acellular eosinophilic areas and irregular nuclear spacing as correlates of necrosis and hypoxia), but their selection was informed by the earlier CPTAC screen. Everything in this paper is therefore *defined post hoc*. (iii) CPTAC patients whose features were completed after the freeze time of the screen (29 September 2026, 16:45 China Standard Time; n = 60–61) are treated as a temporal hold-out and are the only quasi-confirmatory data. All other results, including cohort-level meta-analyses that contain the patients

used for feature selection, are exploratory and probably optimistic. The all-slide analyses and the stain-quality exclusion (see "Post hoc stain-quality sensitivity analysis") were performed after the main results were known and are labelled as post hoc sensitivity analyses. The public-surrogate analyses (IvyGAP images, GDC slide necrosis, CPTAC/RPPA/HPA protein evidence, CGGA and GLASS) were added on 30 September 2026, after all other results were known, and are entirely post hoc and exploratory.

Cohorts

TCGA-GBM[18]: diagnostic slide images (first diagnostic slide, "DX1"; 96 slides with matching RNA, of which 93 had ≥ 10 usable tiles and 92 had the log2-scale score) were obtained from the Genomic Data Commons[19] and are archived in TCIA[20]. Bulk RNA-seq, survival and clinical data came from GDC/UCSC Xena[21] and the public pan-glioma clinical annotation[22,23]. Gene-level expression from 154 primary tumours with slides was analysed as log2(x + 1). Sensitivity analyses used all slides per case (517 slides from 155 cases; median 3 per case, including additional diagnostic slides "DX2+", n = 117, and frozen-section slides "TS/BS", n = 309), averaged per case (n = 154).

CPTAC-GBM[24,25]: H&E slides were retrieved from the Imaging Data Commons[26]; RNA-seq (GDC CPTAC-3, log2 TPM) was available for 200 primary tumours, and 152 patients had slides with ≥ 10 usable tiles. Clinical variables came from GDC, the Proteomic Data Commons[27] and cBioPortal (age n = 179, sex n = 99, IDH1/2 mutation in 16 of 199 patients).

TCGA-LGG[28,29]: 488 diagnostic slides with ≥ 10 usable tiles; RNA (n = 516 with scores), IDH status and grade from the public annotation.

IvyGAP[30]: 270 RNA-seq samples from 37 tumours (104 blocks) with anatomic structure labels were obtained from the Allen Institute API (FPKM). The TCIA IvyGAP collection contains MRI only; H&E section images were instead obtained through the Allen Institute IvyGAP API (see "Public surrogate analyses"). RNA region analyses used the 122 reference-histology samples from 10 tumours (cellular tumour, CT, 30; microvascular proliferation, CTmvp, 25; pseudopalisading cells around necrosis, CTpan, 24; infiltrating tumour, IT, 24; leading edge, LE, 19).

CGGA[31]: mRNA-seq datasets 693 and 325 (RSEM, n = 693 and 325; no overlapping sample identifiers) with clinical data were downloaded from cgga.org.cn; no whole-slide images are available.

GLASS[32,33]: the Synapse repository requires login; the public cBioPortal datahub mirror of the study (difg_glass, release of 31 May 2022; RNA TPM) was used (355 RNA samples, 169 patients; 155 primary and 200 recurrent tumours; 46 samples derived from TCGA). No whole-slide images.

Visium[34,35]: 22 slides from 9 patients (GSE318562). The companion series GSE318564 is single-cell RNA-seq, not Visium. Core/edge/necrosis annotations exist only as figures in the source supplement

and were not available in machine-readable form; core-versus-edge comparisons were therefore not performed inferentially.

Image processing and feature extraction

Slides were read with OpenSlide[36]. Tissue tiles of 512 × 512 px were resampled to 0.5015 µm/px (40× slides, mpp ≈ 0.25, were down-sampled by a box filter). For TCGA-GBM DX1, up to 120 tiles per slide were taken on a regular grid (tissue fraction ≥ 0.10; median 41 tiles, interquartile range 28–59, among the 93 analysed slides); for CPTAC, TCGA-LGG and the additional TCGA slides, up to 60 tiles were sampled at random with a fixed seed (median 60). Large cohorts were streamed and the source images deleted after feature extraction; the 96 TCGA-GBM DX1 images are retained. Tiles with tissue fraction < 0.30 were excluded at aggregation, and slides with < 10 usable tiles were not analysed (counts in Fig. 1B refer to slides with ≥ 10 tiles).

Stain vectors were estimated per slide with the Macenko method[37] (no cross-slide stain normalisation was applied, although stain variation is a recognised confounder in computational pathology[38]), the hematoxylin concentration was normalised to the slide's 99th percentile and thresholded (0.35), and nuclei were separated by a distance-transform watershed (scikit-image[39]); objects of 40–1,500 px (≈ 10–380 µm²) were kept. Per-tile features were aggregated per slide (mean, 90th percentile, standard deviation, and fraction of tiles above a threshold), giving 31 features covering nuclear density, size, shape (elongation, eccentricity, solidity), hematoxylin and eosin intensity, hypocellular and hypercellular areas, eosin-rich acellular area, pyknotic-like nuclei, nearest-neighbour distance statistics and a palisade-orientation coherence measure. The three features of interest were defined as: **H1** hypocellular area fraction (fraction of tissue pixels with a Gaussian-smoothed nuclear-mask density < 0.03, $\sigma \approx 12$ µm); **H2** eosin-rich acellular fraction (normalised eosin concentration > 0.9 and nuclear density < 0.05); **H3** coefficient of variation of nearest-neighbour distance between nuclear centroids. A secondary representation was the mean-pooled 2,048-dimensional output of an ImageNet-pretrained ResNet-50[40] implemented in PyTorch.

Transcriptional scores

The primary outcome was the Hallmark HYPOXIA score (200 genes)[4], computed as the mean of within-cohort z-scores of $\log2(x + 1)$ expression (not ssGSEA). Sensitivity signatures were Winter (243 genes)[6], Buffa (50)[5], Elvidge (173)[7], Manalo (203)[41] and Harris (81; MSigDB set)[1], each also with Hallmark-overlap genes removed ("ex-Hallmark": Winter 170, Buffa 34, Elvidge 94 genes) to address circularity. Composition covariates were immune and stromal scores computed from ESTIMATE-derived gene sets (141 genes each; a gene-set approximation, not the full ESTIMATE algorithm)[42] and an ad hoc 18-gene myeloid score (CD68, CD163, CSF1R, AIF1, ITGAM, LYZ, MS4A6A, C1QA, C1QB, C1QC, TYROBP, CD14, FCGR3A, MSR1, MRC1, LAPTM5, CD86, SPI1). Expression units differ between cohorts (TCGA-GBM gene matrix, CPTAC TPM, TCGA-

LGG FPKM); scores were standardised within cohort. An earlier TCGA-GBM score calculated on the FPKM scale gave 8 nominal (0 BH) hits, whereas the log2-scale score used here gave none, so scale choice matters.

Statistical analysis

Associations used Spearman correlation with BH correction over the 31 features within cohort[43] (implemented with SciPy and statsmodels[44]). Family-wise control used a max-|ρ| permutation test over the 31 features (10,000 permutations; pooled analysis permuted within cohort). Cohorts were combined by fixed-effect inverse-variance (n − 3) pooling of Fisher-z values. Adjustment used partial Spearman correlation after rank-residualising feature and outcome on immune, stromal, myeloid scores and tile count (pooled ordinary least squares on ranks with cohort as covariate and HC3 errors as a check). Cross-cohort prediction trained ridge regression (standardised features, α selected by efficient leave-one-out cross-validation within the training cohort, α grid 10^0–10^4) in one cohort and tested it in the other (Spearman ρ; 1,000 bootstrap resamples for confidence intervals; 2,000 label permutations); incremental value was assessed after residualising the outcome on composition scores. Within-patient reproducibility used the one-way intraclass correlation ICC(1) across independent TCGA-GBM slides of the same patient[45].

Survival analyses used Cox proportional-hazards models (R survival, rms and lifelines)[46,47] with the score per standard deviation, adjusted for age, sex and IDH status (and MGMT methylation where available). To keep a single source, hazard ratios for TCGA-GBM are reported from R; lifelines gave near-identical values (Supplementary Table S9). The prognostic-model experiment used repeated (20 × 5-fold) nested cross-validation of an elastic-net Cox model (l1 ratio 0.5, univariate pre-filter and standardisation inside each training fold) using the coordinate-descent Coxnet algorithm[48] as implemented in scikit-survival (scikit-survival[49] with scikit-learn[50]), compared with label-permutation nulls. Post hoc power for a true correlation of 0.2 (two-sided α = 0.05) is 33–34% for n = 60–61, 48% for n = 92, 70% for n = 152 and 88% for n = 244 (Fisher z approximation).

IvyGAP and Visium

In IvyGAP, region separation used the area under the curve (AUC) for pseudopalisading cells around necrosis versus cellular tumour and a within-tumour paired comparison of the necrosis-adjacent regions (CTpan and perinecrotic zone, CTpnz) with cellular tumour (n = 8 tumours; Wilcoxon signed-rank).

For Visium, the Hallmark z-mean was strongly related to log UMI (median ρ = 0.81), so we regressed log UMI and log gene count (natural spline, 4 df) out of the score per slide and also used a UCell-style rank score based on the method of Andreatta and Carmona[51] (own implementation; top 2,000 genes per spot). Associations with hematoxylin density and with contrast between pale and other zones were

tested per slide (Wilcoxon signed-rank across 22 slides), at patient level (n = 9) and by patient-cluster bootstrap.

Segmentation reliability

The classical segmenter was compared with StarDist (pretrained 2D_versatile_he)[52] on 1,094 tiles from 96 slides and with Cellpose (cpsam)[53,54] on 60 tiles. Agreement was quantified by Spearman ρ, ICC and matched-nucleus F1 (IoU > 0.3). No pathologist ground truth was available.

Differential expression and enrichment

TCGA-GBM tumours were split at the median Hallmark score (n = 153; 77 high, 76 low); limma[55] was applied to 23,183 genes. Pre-ranked gene-set enrichment (GSEApy)[56,57] was run on the limma t-statistics with and without the 200 score genes; protein–protein interactions among the top 300 up-regulated genes were taken from STRING (score ≥ 0.7)[58].

Ethics approval

This study analyzed publicly deposited, de-identified data and collected no new participants or specimens. Ethics approvals and consent for the original studies are those reported by their investigators[18,24,30,34]. No new human-subject intervention was performed. No new participant consent was obtained because no new participants were enrolled.

Post hoc stain-quality sensitivity analysis

Inspection of example tiles revealed that per-slide Macenko estimation failed on five TCGA-GBM DX1 slides (atypical hematoxylin vector: TCGA-12-0616, TCGA-12-0618, TCGA-02-0047, TCGA-27-2526, TCGA-19-2624); in four of them nuclei were largely not segmented (hypocellular fraction 0.85–0.91). Because this can create artefactual "hypocellularity", the five slides were excluded in a sensitivity analysis (n = 87). CPTAC stain files were not available and could not be checked; three CPTAC slides have hypocellular fraction > 0.8 (C3L-00104, 4.4 nuclei/mm², suspicious; C3L-04036; C3N-00663, visually pale and probably genuine).

Public surrogate analyses (post hoc, added 30 September 2026)

Rationale and status. Pathologist-annotated necrosis or hypocellular regions, CA9/HIF1A IHC and large independent external cohorts were not available to us. We therefore used public data that partly substitute for them. These are **surrogates, not equivalents**, and all analyses below were performed after the main results were known.

IvyGAP images. The Allen Institute API lists 946 H&E section images; anatomic-structure polygons (SVG) were downloaded for the 720 annotated images. Polygon coordinates are given relative to the annotated image and were shifted by the section-image offset (verified against thumbnails, r = 0.70–0.84 versus ≈ 0 without the offset). All 270 RNA-seq laser-microdissection wells were matched to

their polygons. Within each well polygon, 512 × 512 px tiles were cut at native resolution (0.4945 μm/px) and passed through the same nuclear-segmentation pipeline as in "Image processing and feature extraction" (4,045 tiles; 31 wells had no complete tile; 194 wells with ≥ 5 tiles were analysed; median 16 tiles per well; 37 tumours). Region contrasts compared necrosis-adjacent regions (pseudopalisading cells around necrosis, CTpan, plus perinecrotic zone, CTpnz) with cellular tumour (CT), using Mann–Whitney AUC and within-tumour paired Wilcoxon tests (29 tumours with both). A morphology-only ridge/logistic classifier was evaluated with tumour-grouped cross-validation and label permutation. Associations between the 31 features and the Hallmark score used Spearman ρ, BH correction, within-tumour (tumour-demeaned) correlation, and rank regression with region and tumour fixed effects (tumour-clustered errors); sensitivity analyses adjusted for tile count and re-sampled a fixed 10 tiles per well (144 wells).

GDC slide-level necrosis. Fields percent_necrosis and percent_tumor_nuclei were retrieved from GDC biospecimen slide records (2,147 TCGA-GBM slides). They exist only for frozen-section slides (TS/BS), not for diagnostic (DX) slides, and are coarse percentages recorded by the contributing site. We correlated the 31 features (computed on the same frozen slides; 310 slides from 154 cases) with these values (Spearman, BH; 10-fold patient-grouped ridge prediction with permutation p), correlated case-mean necrosis with the RNA scores (with partial correlation on immune, stromal and myeloid scores), and, as an indirect test, correlated DX1 features with case-mean frozen-slide necrosis (different slides).

Protein-level surrogates for CA9/HIF1A IHC. CPTAC-GBM proteomics (cBioPortal gbm_cptac_2021, corresponding to PDC000204; 10,997 proteins, n = 99; TMT log2 ratios) and phosphosite data were related to the RNA Hallmark score (Spearman; BH; partial correlation on composition scores) and to a protein composite (mean z-score of 149 Hallmark-hypoxia proteins detected; and a 14-protein core set). Associations with the 31 morphology features were tested in the 96 patients with slides, with and without adjustment for composition scores and tile count. TCGA-GBM RPPA (cBioPortal gbm_tcga; 203 antibodies tested; n = 73 with RNA) was analysed likewise. HPA (v23 pathology.tsv) glioma IHC staining counts were tabulated for selected hypoxia-related proteins; HPA images and prognostic annotations were not analysed or used for inference.

CGGA and GLASS. The Hallmark score was computed as in "Transcriptional scores" within each dataset (183/184 of the 200 genes present in CGGA; 191 in GLASS; RSEM or TPM, log2(x + 1)). CGGA analyses: relation to IDH and grade, Cox models stratified by dataset (693 or 325) with adjustment for age, sex, IDH and grade (or MGMT, temozolomide, radiotherapy, and composition scores in primary GBM), and Kaplan–Meier median splits. GLASS analyses: paired primary-to-first-recurrence changes (Wilcoxon), sensitivity analyses without TCGA-derived samples and by IDH status, regression of the score change on composition changes, and Cox models for overall survival

from primary surgery and from recurrence. Because CGGA and GLASS have no images, only the score itself, not morphology, could be tested.

HCMI-CMDC cohort (post hoc, added 1 October 2026)

HCMI-CMDC (Human Cancer Models Initiative, Cancer Model Development Center) open-access data were obtained from the GDC and IDC (collection hcmi_cmdc; DICOM conversion of the GDC slides, CC BY 4.0[59]). We selected brain-primary patient tumour samples (Primary Tumor or Recurrent Tumor; next-generation cancer models were excluded) that had both an open-access H&E slide and an open-access STAR-Counts RNA-seq file with the same sample identifier: 71 slides from 36 samples from 35 patients (all frozen-section slides except one FFPE sample). One sample per patient (primary preferred) was the primary analysis unit (n = 35; 22 primary and 13 recurrent tumours), and slide features were averaged within sample. The unchanged pipeline was applied (up to 60 random tissue tiles per slide, 512 × 512 px at 0.50 μm/px, Macenko normalisation, the same 31 features and ResNet-50 embeddings); slides were streamed and the SVS files deleted after feature extraction (69/71 slides processed; 68 with ≥ 10 usable tiles; median 60 tiles). All 200 Hallmark genes were present; the score was the mean within-cohort z-score of log2(TPM + 1). An analysis plan listing the tests (Supplementary Note, 'Analysis plans'; internal, not a registered protocol) was written after cohort assembly and feature extraction had started, but before any HCMI feature was related to any HCMI score. It specified the following tests: H1–H3 and the 31-feature screen (Spearman, BH), feature-wise concordance with TCGA-GBM and CPTAC, composition-adjusted partial Spearman correlations, fixed-effect three-cohort meta-analysis with Cochran Q and I^2, cross-cohort ridge transfer with permutation P values and bootstrap confidence intervals, and sensitivity analyses (primary tumours only; all 36 samples; signatures without Hallmark genes). The concordance across features was tested by permuting the HCMI score (5,000 permutations), because the features are inter-correlated. Power was computed before analysis (≈ 21% for $\rho = 0.2$ at $n = 35$).

Statistics and reporting

All tests were two-tailed with $\alpha = 0.05$ unless stated, and exact P values are reported (Benjamini–Hochberg q values where stated). The word "significant" is used only together with a P value or q value and the multiplicity control applied. Nominal P values outside the declared families (the 31-feature screen, the three features H1–H3, and the Hallmark signature as primary outcome) are descriptive. In figures, P values are shown only where a test was pre-specified or where descriptive inference is sensible; they are omitted where a test would be uninformative (for example, comparisons of depth-corrected residuals that are centred on zero by construction are summarised by the per-slide Wilcoxon test only). Sample sizes are given for each analysis. No data were excluded except as described (tiles and slides below the quality thresholds, and the post hoc stain-quality sensitivity analysis).

Code availability

Analysis scripts, including the scripts that regenerate every figure in this paper, are available from the corresponding author on reasonable request and will be deposited in a public repository with a persistent identifier upon acceptance. Software: Python (NumPy, SciPy, statsmodels, scikit-learn, scikit-survival, lifelines, scikit-image, matplotlib), R (survival, rms, limma), OpenSlide, PyTorch and GSEApy (see references).

Data availability

The primary datasets analysed in this study are publicly available from the following resources: TCGA-GBM and TCGA-LGG whole-slide images, transcriptomic data and clinical annotations from the Genomic Data Commons (GDC), The Cancer Imaging Archive (TCIA) and UCSC Xena; CPTAC-GBM whole-slide images, transcriptomic data and clinical annotations from the Imaging Data Commons (IDC), GDC (project CPTAC-3) and Proteomic Data Commons (PDC); Ivy Glioblastoma Atlas Project (IvyGAP) RNA-sequencing data, H&E images and anatomical-region polygons from the Allen Institute API; slide-level biospecimen annotations from GDC; CPTAC-GBM proteomic data from cBioPortal (study gbm_cptac_2021, corresponding to PDC000204) and TCGA-GBM reverse-phase protein array data from cBioPortal (study gbm_tcga); Human Protein Atlas version 23; Chinese Glioma Genome Atlas (CGGA) mRNA-seq datasets 693 and 325 from https://www.cgga.org.cn/; GLASS data from the public cBioPortal DataHub mirror (study difg_glass, release dated 31 May 2022); HCMI-CMDC whole-slide images and RNA-sequencing data from the open-access GDC resource and IDC collection hcmi_cmdc; and spatial transcriptomic data from NCBI Gene Expression Omnibus under accession GSE318562 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE318562). Further dataset descriptions and source references are provided in the Methods.

Derived per-slide feature tables, analysis tables and figures are available from the corresponding author, Wenchuan Zhang (zhangwench88@sjtu.edu.cn), on reasonable request and will be deposited in a public repository with a persistent identifier upon acceptance. Original whole-slide images are not redistributed by the authors and should be accessed through the source repositories listed above. This study used de-identified, publicly deposited data and collected no new participant data or specimens.

References

1. Harris, A. L. Hypoxia–a key regulatory factor in tumour growth. *Nat. Rev. Cancer* **2**, 38–47 (2002). https://doi.org/10.1038/nrc704

2. Brat, D. J. et al. Pseudopalisades in glioblastoma are hypoxic, express extracellular matrix proteases, and are formed by an actively migrating cell population. *Cancer Res.* **64**, 920–927 (2004). https://doi.org/10.1158/0008-5472.CAN-03-2073

3. Louis, D. N. et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro. Oncol.* **23**, 1231–1251 (2021). https://doi.org/10.1093/neuonc/noab106

4. Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015). https://doi.org/10.1016/j.cels.2015.12.004

5. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010). https://doi.org/10.1038/sj.bjc.6605450

6. Winter, S. C. et al. Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res.* **67**, 3441–3449 (2007). https://doi.org/10.1158/0008-5472.CAN-06-3322

7. Elvidge, G. P. et al. Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1alpha, HIF-2alpha, and other pathways. *J. Biol. Chem.* **281**, 15215–15226 (2006). https://doi.org/10.1074/jbc.M511408200

8. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020). https://doi.org/10.1038/s43018-020-0087-6

9. Schmauch, B. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 3877 (2020). https://doi.org/10.1038/s41467-020-17678-4

10. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3

11. Diao, J. A. et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat. Commun.* **12**, 1613 (2021). https://doi.org/10.1038/s41467-021-21896-9

12. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970-E2979 (2018). https://doi.org/10.1073/pnas.1717139115

13. Zadeh Shirazi, A. et al. A deep convolutional neural network for segmentation of whole-slide pathology images identifies novel tumour cell-perivascular niche interactions that are associated with poor survival in glioblastoma. *Br. J. Cancer* **125**, 337–350 (2021). https://doi.org/10.1038/s41416-021-01394-x

14. Wu, X. et al. A synaptogenesis-associated histomorphologic signature from H&E whole-slide images predicts glioma prognosis and identifies EFNB2-positive malignant cells as a candidate neuro-glioma communication hub. *Int. J. Mol. Sci.* **27**, 4300 (2026). https://doi.org/10.3390/ijms27104300

15. Dai, Y. et al. Integrating computational pathology and multi-transcriptomics to characterize glioblastoma heterogeneity and identify prognostic biomarkers. *Hum. Pathol.* **166**, 105982 (2025). https://doi.org/10.1016/j.humpath.2025.105982

16. Zadeh Shirazi, A. & Gomez, G. A. Deep learning-based multimodal fusion of whole-slide images and RNA sequencing identifies survival-relevant glioblastoma clusters. *Cancer Med.* **15**, e72182 (2026). https://doi.org/10.1002/cam4.72182

17. Manescu, P., Geradts, J. & Fernandez-Reyes, D. Computational pathology detection of hypoxia-induced morphologic changes in breast cancer. *Am. J. Pathol.* **195**, 663–670 (2025). https://doi.org/10.1016/j.ajpath.2024.10.023

18. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008). https://doi.org/10.1038/nature07385

19. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016). https://doi.org/10.1056/NEJMp1607591

20. Scarpace, L. et al. The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM). The Cancer Imaging Archive https://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9 (2016).

21. Goldman, M. J. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020). https://doi.org/10.1038/s41587-020-0546-8

22. Ceccarelli, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016). https://doi.org/10.1016/j.cell.2015.12.028

23. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013). https://doi.org/10.1126/scisignal.2004088

24. Wang, L. B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509-528.e20 (2021). https://doi.org/10.1016/j.ccell.2021.01.006

25. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme Collection (CPTAC-GBM). The Cancer Imaging Archive https://doi.org/10.7937/K9/TCIA.2018.3RJE41Q1 (2018).

26. Fedorov, A. et al. NCI imaging data commons. *Cancer Res.* **81**, 4188–4193 (2021). https://doi.org/10.1158/0008-5472.CAN-21-0950

27. Thangudu, R. R. et al. NCI's proteomic data commons: a cloud-based proteomics repository empowering comprehensive cancer analysis through cross-referencing with genomic and imaging data. *Cancer Res. Commun.* **4**, 2480–2488 (2024). https://doi.org/10.1158/2767-9764.CRC-24-0243

28. The Cancer Genome Atlas Research Network et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015). https://doi.org/10.1056/NEJMoa1402121

29. Pedano, N. et al. The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG). The Cancer Imaging Archive https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK (2016).

30. Puchalski, R. B. et al. An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018). https://doi.org/10.1126/science.aaf2666

31. Zhao, Z. et al. Chinese glioma genome atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. *Genomics Proteomics Bioinformatics* **19**, 1–12 (2021). https://doi.org/10.1016/j.gpb.2020.10.005

32. Barthel, F. P. et al. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019). https://doi.org/10.1038/s41586-019-1775-1

33. Varn, F. S. et al. Glioma progression is shaped by genetic evolution and microenvironment interactions. *Cell* **185**, 2184-2199.e16 (2022). https://doi.org/10.1016/j.cell.2022.04.038

34. Khan, S. M. et al. Mapping the spatial architecture of glioblastoma from core to edge delineates niche-specific tumor cell states and intercellular interactions. Preprint at https://doi.org/10.1101/2025.04.04.647096 (2025).

35. Gene Expression Omnibus. GSE318562: Mapping the spatial architecture of glioblastoma from core to edge delineates niche-specific tumor cell states and intercellular interactions [Spatial Transcriptomics]. NCBI GEO https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE318562 (2026).

36. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013). https://doi.org/10.4103/2153-3539.119005

37. Macenko, M. et al. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107–1110 (IEEE, 2009). https://doi.org/10.1109/ISBI.2009.5193250

38. Bejnordi, B. E. et al. Stain specific standardization of whole-slide histopathological images. *IEEE Trans. Med. Imaging* **35**, 404–415 (2016). https://doi.org/10.1109/TMI.2015.2476509

39. van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014). https://doi.org/10.7717/peerj.453

40. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). https://doi.org/10.1109/CVPR.2016.90

41. Manalo, D. J. et al. Transcriptional regulation of vascular endothelial cell responses to hypoxia by HIF-1. *Blood* **105**, 659–669 (2005). https://doi.org/10.1182/blood-2004-07-2958

42. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013). https://doi.org/10.1038/ncomms3612

43. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

44. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

45. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979). https://doi.org/10.1037/0033-2909.86.2.420

46. Therneau, T. M. & Grambsch, P. M. *Modeling survival data: extending the Cox model* (Springer New York, 2000). https://doi.org/10.1007/978-1-4757-3294-8

47. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019). https://doi.org/10.21105/joss.01317

48. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011). https://doi.org/10.18637/jss.v039.i05

49. Pölsterl, S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020). https://jmlr.org/papers/v21/20-729.html

50. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). https://jmlr.org/papers/v12/pedregosa11a.html

51. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021). https://doi.org/10.1016/j.csbj.2021.06.043

52. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 265–273 (Springer, 2018). https://doi.org/10.1007/978-3-030-00934-2_30

53. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021). https://doi.org/10.1038/s41592-020-01018-x

54. Pachitariu, M., Rariden, M. & Stringer, C. Cellpose-SAM: superhuman generalization for cellular segmentation. Preprint at https://doi.org/10.1101/2025.04.28.651001 (2025).

55. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015). https://doi.org/10.1093/nar/gkv007

56. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023). https://doi.org/10.1093/bioinformatics/btac757

57. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005). https://doi.org/10.1073/pnas.0506580102

58. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638-D646 (2023). https://doi.org/10.1093/nar/gkac1000

59. Clunie, D., Clifford, W. & Fedorov, A. HCMI-CMDC: DICOM converted whole slide images from the Human Cancer Models Initiative (HCMI) Cancer Model Development Center (CMDC). Zenodo https://doi.org/10.5281/zenodo.17381442 (2026).

60. Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017). https://doi.org/10.1126/science.aan2507

Acknowledgements

We thank the participants and investigators of TCGA, CPTAC, IvyGAP, CGGA, GLASS and the Visium study, and the maintainers of the GDC, TCIA, IDC, PDC, GEO, cBioPortal, CGGA, GLASS, Human Protein Atlas and Allen Institute resources.

Author contributions

Conceptualization, K.C. and W.Z.; methodology, K.C.; formal analysis, K.C.; data curation, Z.C.; validation, Z.C., Y.Z. and P.Y.; visualization, K.C.; clinical interpretation, Y.Z., P.Y. and W.Z.; writing—original draft preparation, K.C. and Z.C.; writing—review and editing, Y.Z., P.Y. and W.Z.; supervision, project administration and resources, W.Z. K.C. and Z.C. contributed equally. All authors have read and agreed to the manuscript.

Funding

This research received no external funding.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.Z. **Supplementary information** accompanies this paper: Supplementary Figs. S1–S16 (Supplementary Information file), Supplementary Tables S1–S22 (Excel workbook) and Supplementary Notes (Supplementary Information file).

Figure legends

Statistical tests are two-tailed and all analyses are exploratory unless stated (hypotheses were defined post hoc; see Methods). P values are displayed inside the panels only where a test was pre-specified or where descriptive inference is sensible; the test is named on the panel or here. Colours are consistent across figures: TCGA-GBM blue, CPTAC-GBM orange, TCGA-LGG green, positive or significant red, negative blue-grey, non-significant grey.

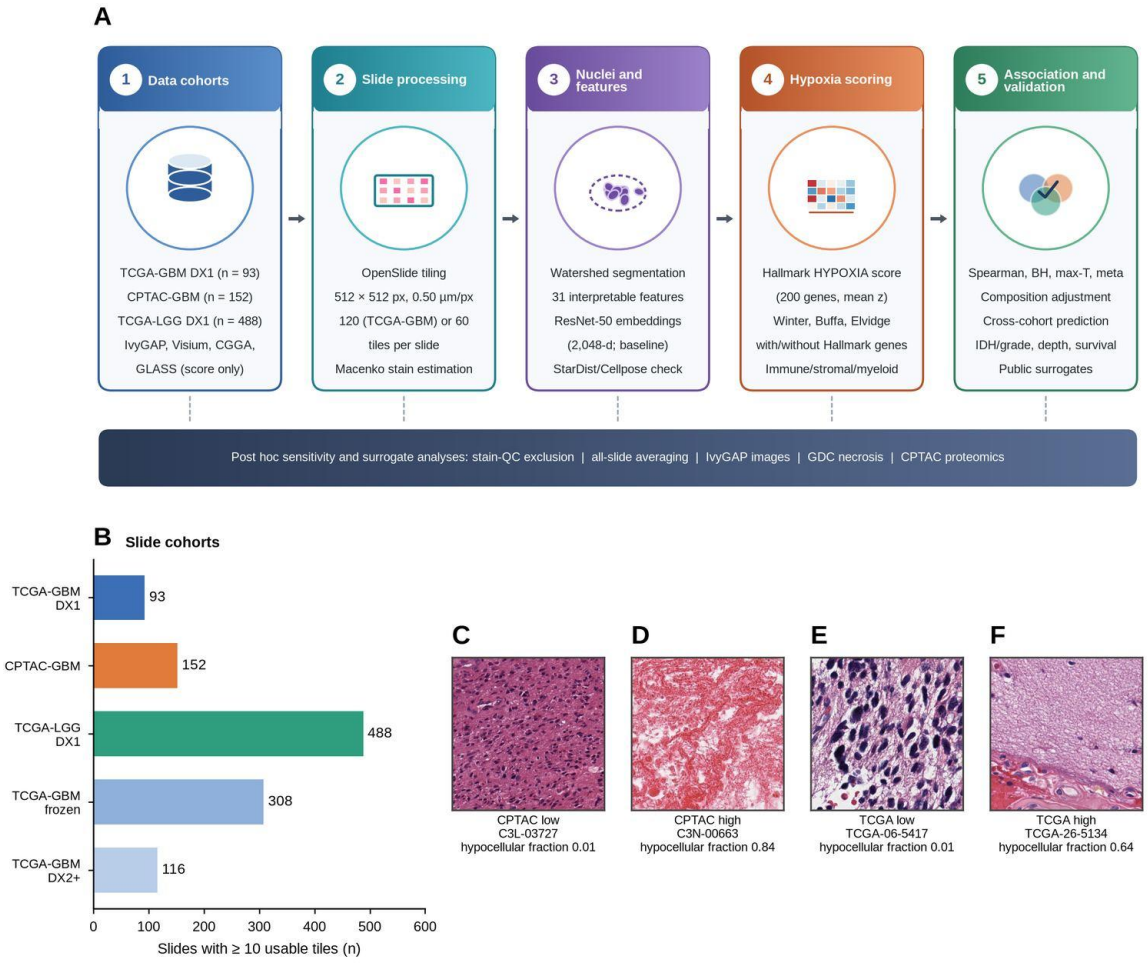
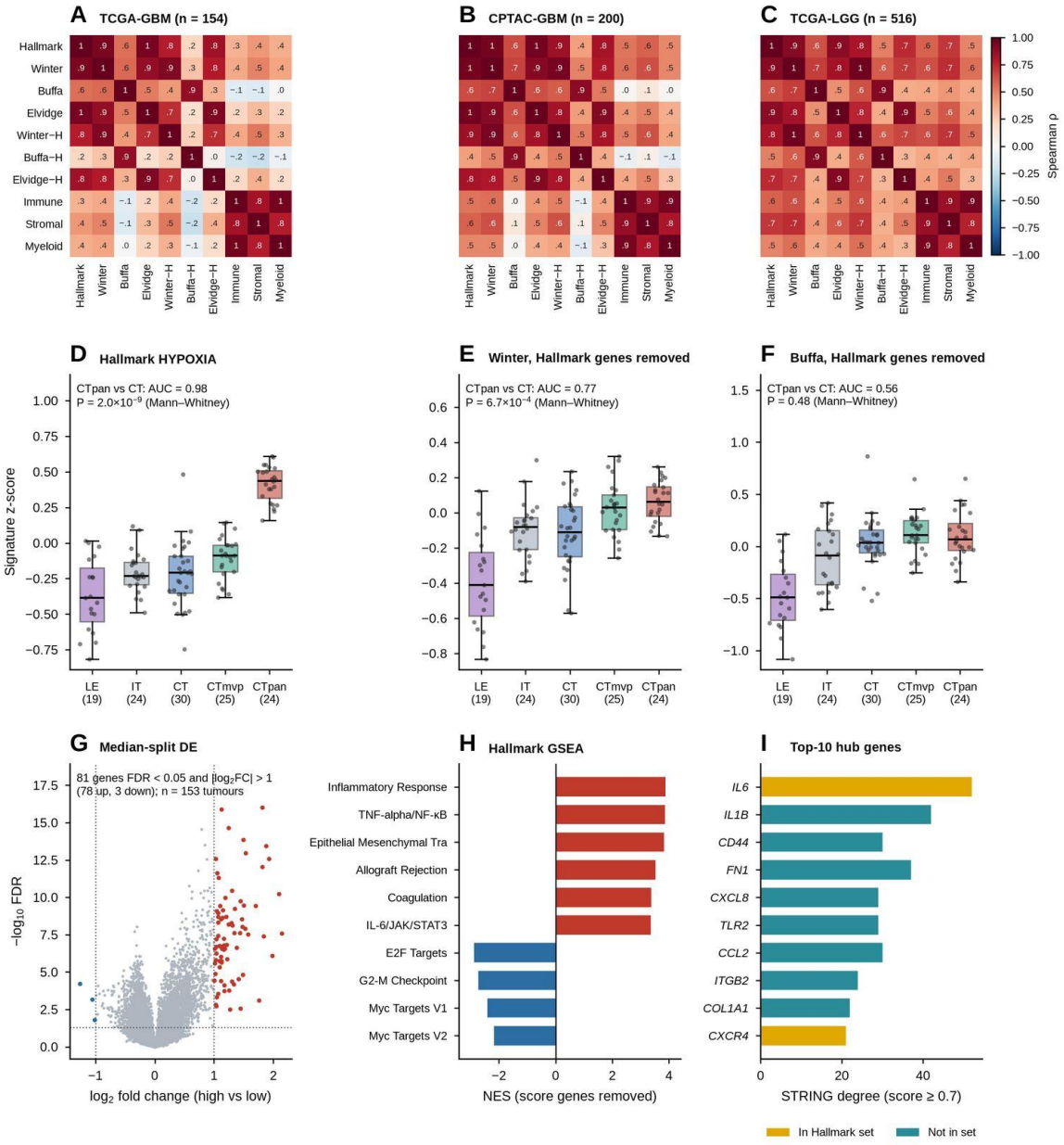


Fig. 1. Study design and cohorts. (A) Five-stage workflow: data cohorts; slide processing (OpenSlide tiling, 512 × 512 px at 0.50 µm/px, Macenko stain estimation); nuclear segmentation, 31 interpretable features and ResNet-50 embeddings; hypoxia scoring (Hallmark HYPOXIA and alternative signatures, with and without Hallmark genes, and immune, stromal and myeloid scores); and association and validation analyses. (B) Number of slides with at least 10 usable tiles per cohort (TCGA-GBM DX1 93, CPTAC-GBM 152, TCGA-LGG 488, TCGA-GBM frozen 308, DX2+ 116). (C,D) Example CPTAC-GBM tiles from slides with low (C3L-03727) and high (C3N-00663) hypocellular fraction. (E,F) Example TCGA-GBM tiles from typical-stain slides with low (TCGA-06-5417) and high (TCGA-26-5134) hypocellular fraction; the tile nearest the slide-level value is shown. Numbers above each tile give the slide-level hypocellular fraction.



744

745 **Fig. 2. The Hallmark score tracks necrosis-adjacent tissue in IvyGAP but is composition-linked**
746 **in bulk tumours.** (A–C) Spearman correlations among hypoxia signatures (with "−H": Hallmark
747 genes removed) and immune, stromal and myeloid scores in TCGA-GBM (n = 154), CPTAC-GBM
748 (n = 200) and TCGA-LGG (n = 516). (D–F) IvyGAP RNA signature by anatomic region (LE, leading
749 edge; IT, infiltrating tumour; CT, cellular tumour; CTmvp, microvascular proliferation; CTpan,
750 pseudopalisading cells around necrosis; 122 reference-histology samples from 10 tumours). Boxes
751 show median and interquartile range; inset text gives the AUC and Mann–Whitney P for CTpan
752 versus CT. (G) Volcano plot of median-split differential expression in TCGA-GBM (n = 153; limma;
753 red up, blue down; FDR < 0.05 and |log2 fold change| > 1). (H) Hallmark gene-set enrichment
754 (normalised enrichment score, NES) with the 200 score genes removed (six highest positive and four

lowest negative NES). (I) Top ten STRING hub genes among the top 300 up-regulated genes (degree at interaction score ≥ 0.7; gold, gene in Hallmark hypoxia set).

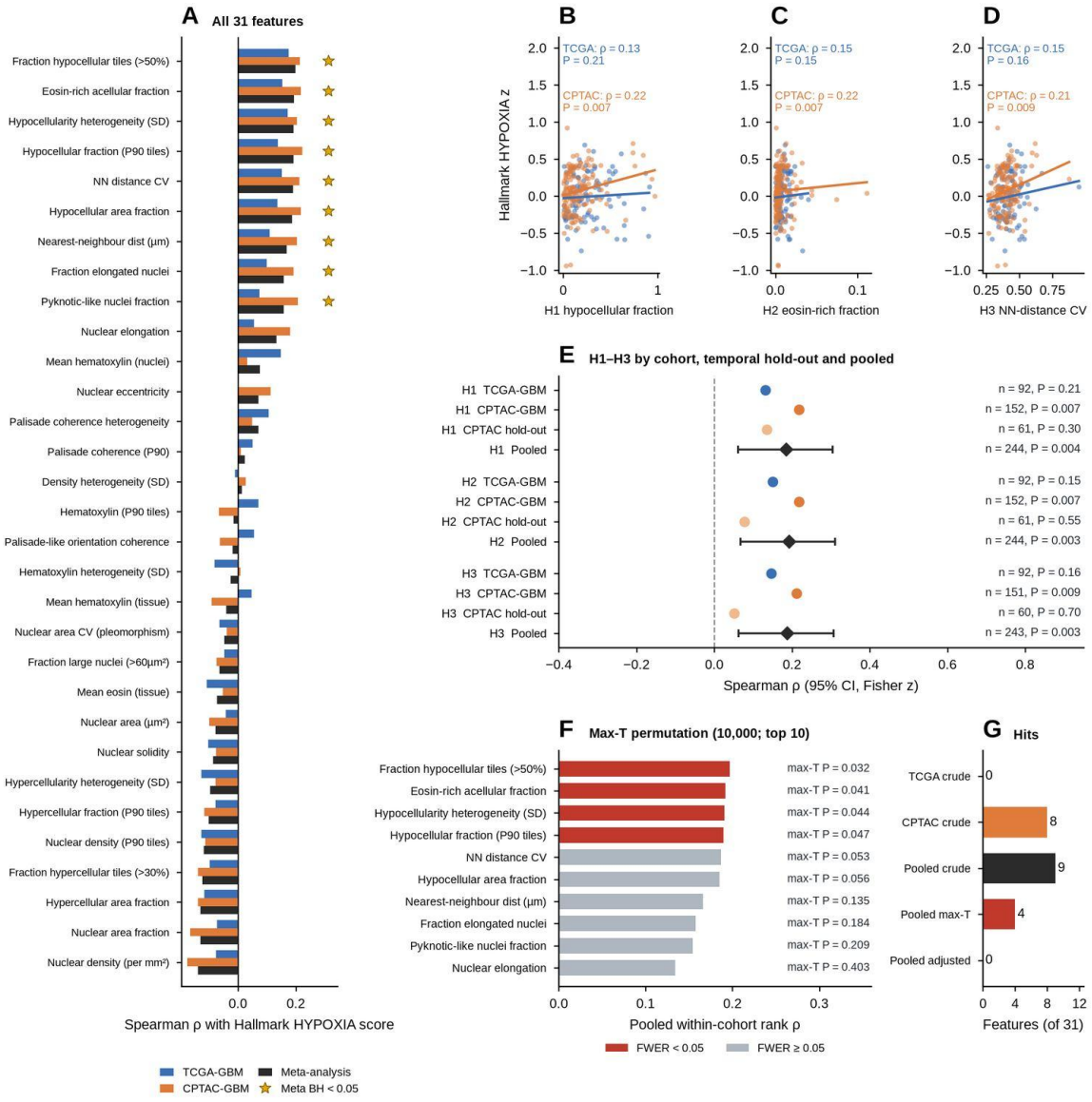


Fig. 3. Morphology–hypoxia associations (exploratory; hypotheses defined post hoc). (A) Spearman ρ of all 31 features with the Hallmark score in TCGA-GBM (n = 90–92), CPTAC-GBM (n = 151–152) and the fixed-effect meta-analysis (n = 241–244). Stars mark features with meta-analysis BH q < 0.05 (legend in panel). (B–D) Scatter plots with regression lines for H1, H2 and H3 by cohort; Spearman ρ and P are shown per cohort. (E) Forest plot of ρ with 95% confidence intervals (Fisher z) for TCGA-GBM, CPTAC-GBM, the CPTAC temporal hold-out (n = 60–61) and the pooled analysis. (F) Max-T permutation adjusted P values for the ten strongest features (10,000 permutations, cohort-stratified; red, family-wise error rate < 0.05). (G) Number of features (of 31) with BH q < 0.05 (crude TCGA-GBM, crude CPTAC-GBM, pooled crude, pooled max-T family-wise error rate < 0.05, pooled composition-adjusted).

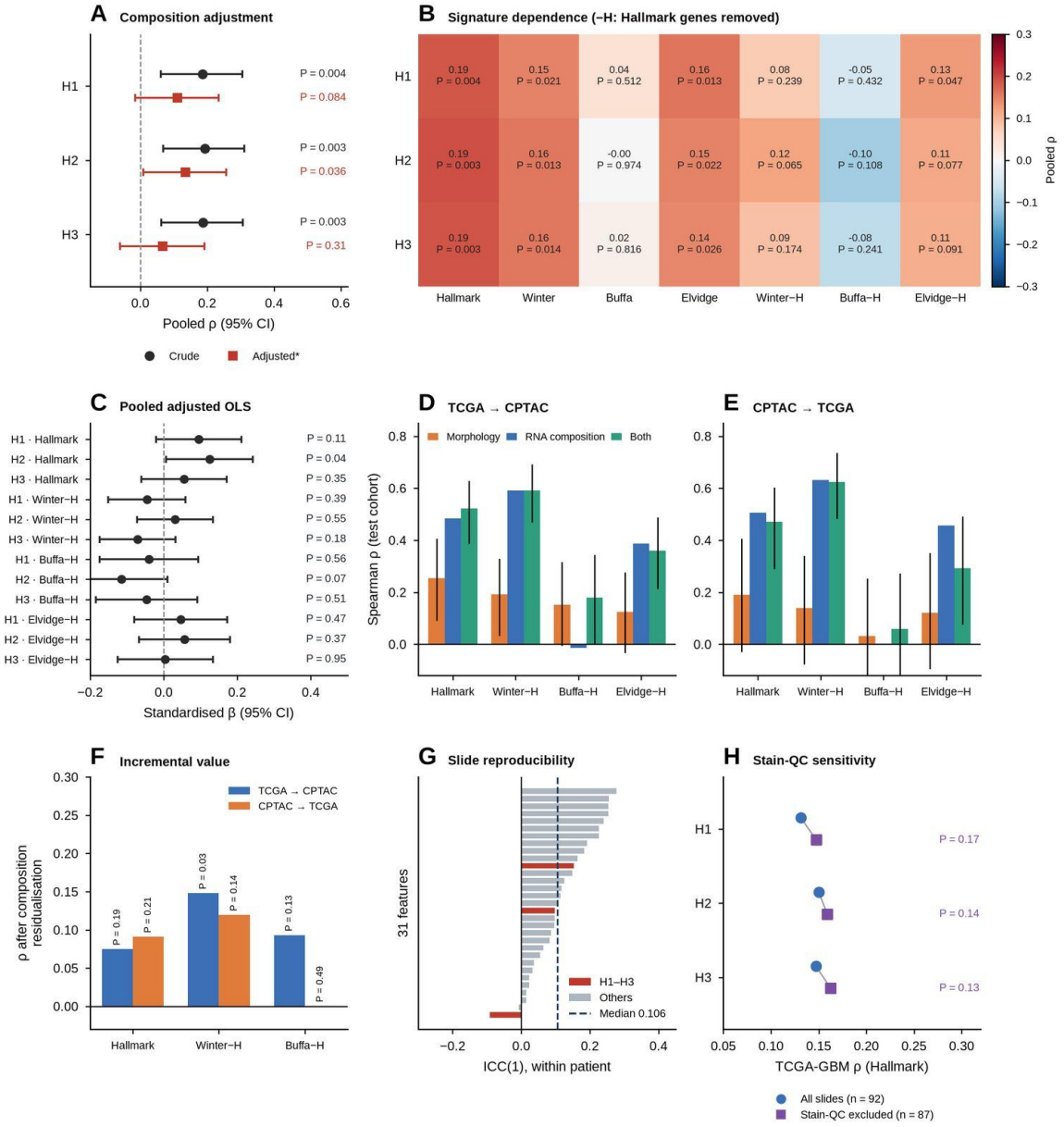
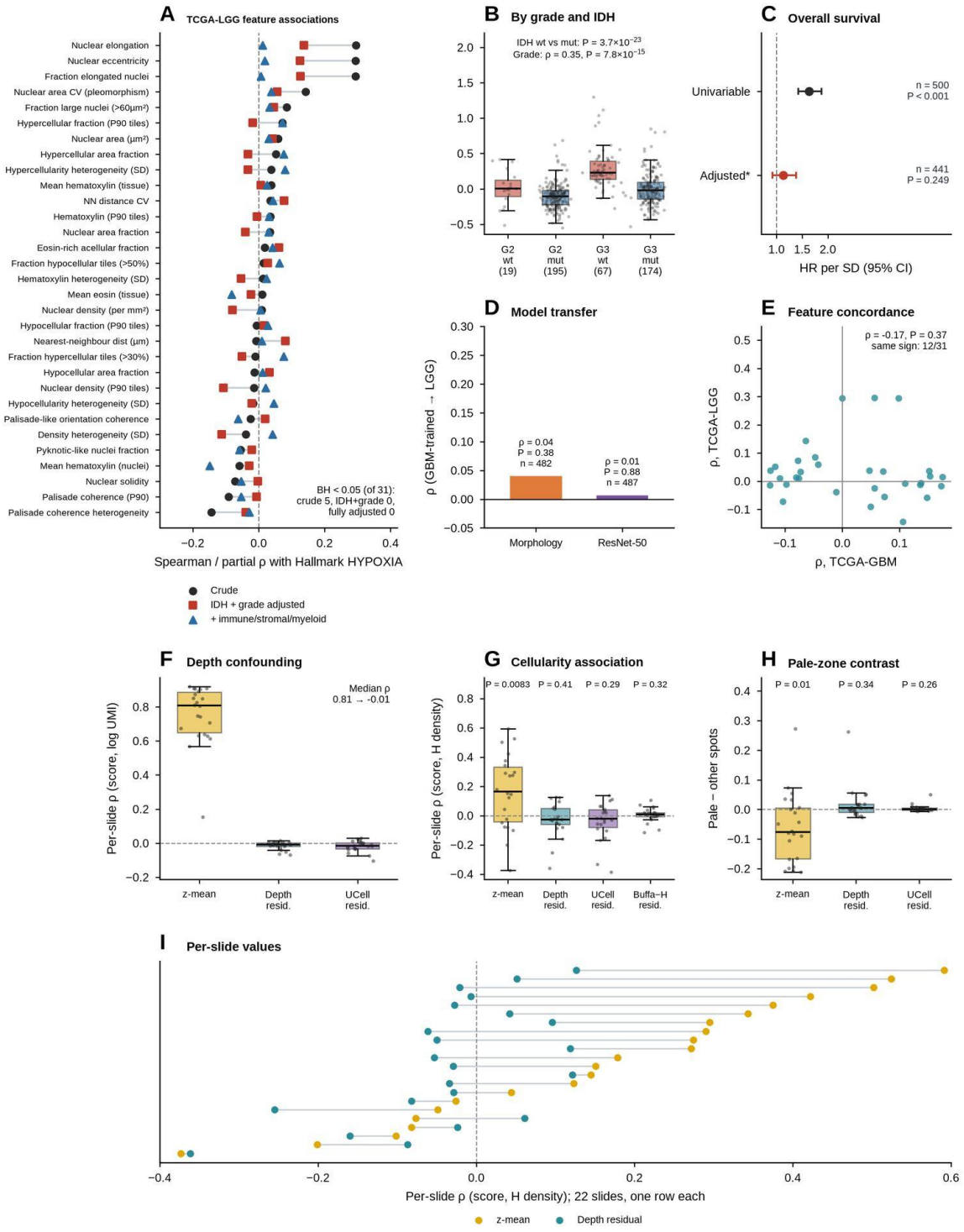


Fig. 4. Robustness: composition, signature choice, cross-cohort prediction and reproducibility. (A) Crude and composition-adjusted pooled ρ (95% CI) for H1–H3 with P values (adjusted for immune, stromal and myeloid scores and tile count). (B) Dependence of pooled ρ on the hypoxia signature. (C) Pooled adjusted ordinary-least-squares coefficients by signature. (D,E) Cross-cohort prediction (ridge regression; ρ in the test cohort with bootstrap 95% CI) for TCGA→CPTAC and CPTAC→TCGA using morphology, RNA composition only, or both. (F) Incremental value of morphology after residualising the outcome on composition (permutation P above bars). (G) Within-patient ICC(1) of the 31 features across independent TCGA-GBM slides (red, H1–H3; grey, other features; dashed line, median). (H) Stain-quality sensitivity analysis in TCGA-GBM: ρ of H1–H3 with the Hallmark score in all slides (n = 92) and after excluding five slides with failed stain estimation (n = 87); P values refer to the latter.



780

781 **Fig. 5. Confounding by IDH status and grade in TCGA-LGG, and by sequencing depth in**
782 **Visium.** (A) Correlation of each of 31 features with the Hallmark score in TCGA-LGG: crude, IDH
783 and grade adjusted, and fully adjusted (legend below the panel; count of BH q < 0.05 inside). (B)
784 Hallmark score by grade and IDH status (wt, wild-type; mut, mutant; n in brackets); the Mann–
785 Whitney P for IDH and the Spearman correlation with grade are given inside. (C) Cox hazard ratios
786 per SD in TCGA-LGG, univariable and adjusted (age, sex, IDH, grade). (D) Transfer of GBM-trained
787 morphology and ResNet-50 models to LGG. (E) Concordance of feature correlations between TCGA-

GBM and TCGA-LGG. (F) Per-slide correlation of the score with log UMI (22 Visium slides): raw z-mean, depth-residual and UCell-style residual. (G,H) Per-slide association with hematoxylin density and pale-zone contrast; P values are per-slide Wilcoxon signed-rank tests of the median per-slide value against zero for each score type. (I) Per-slide values for z-mean and depth-residual scores (one row per slide).

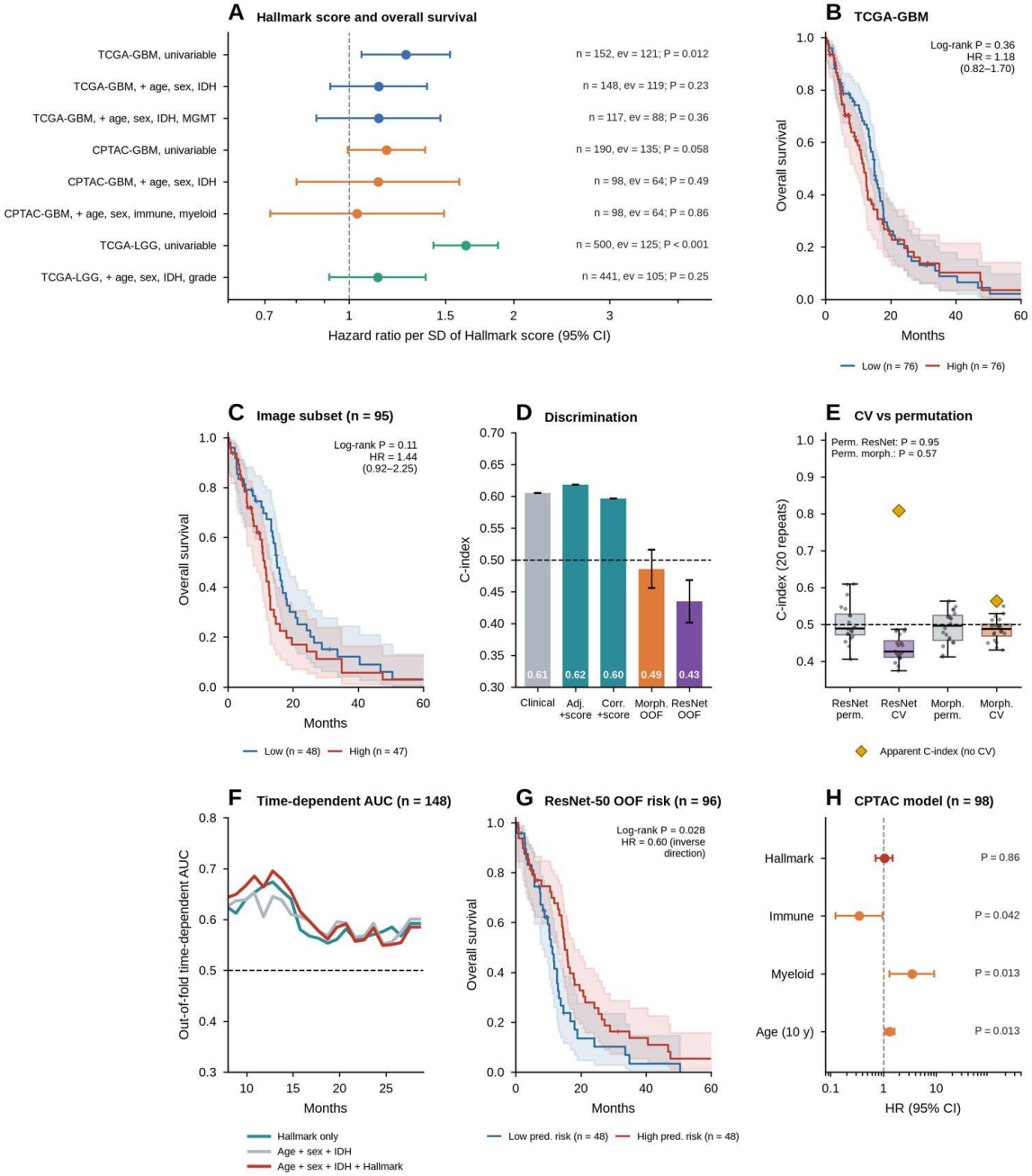


Fig. 6. Prognostic analyses. (A) Hazard ratios per SD of the Hallmark score in TCGA-GBM (univariable; + age, sex, IDH; + age, sex, IDH, MGMT), CPTAC-GBM (univariable; + age, sex, IDH; + age, sex, immune and myeloid scores) and TCGA-LGG (univariable; + age, sex, IDH, grade) with n and P. (B,C) Kaplan–Meier curves by median split of the score: TCGA-GBM (n = 152; log-rank P = 0.36; HR 1.18, 95% CI 0.82–1.70) and the image subset (n = 95; P = 0.11; HR 1.44, 0.92–2.25). (D)

C-index of a clinical model (age, sex, IDH), with the score (apparent and bootstrap-corrected), and of morphology and ResNet-50 elastic-net Cox models (out-of-fold, OOF). (E) Repeated cross-validated C-index versus label-permutation null for the ResNet-50 and morphology models; gold diamonds, apparent C-index. (F) Out-of-fold time-dependent AUC curves (n = 148). (G) Kaplan–Meier split by ResNet-50 OOF risk (n = 96; log-rank P = 0.028). The direction is opposite to that expected (HR 0.60 for the "high-risk" group) and is a cross-validation artefact, not a finding. (H) CPTAC-GBM Cox model with immune and myeloid scores (n = 98).

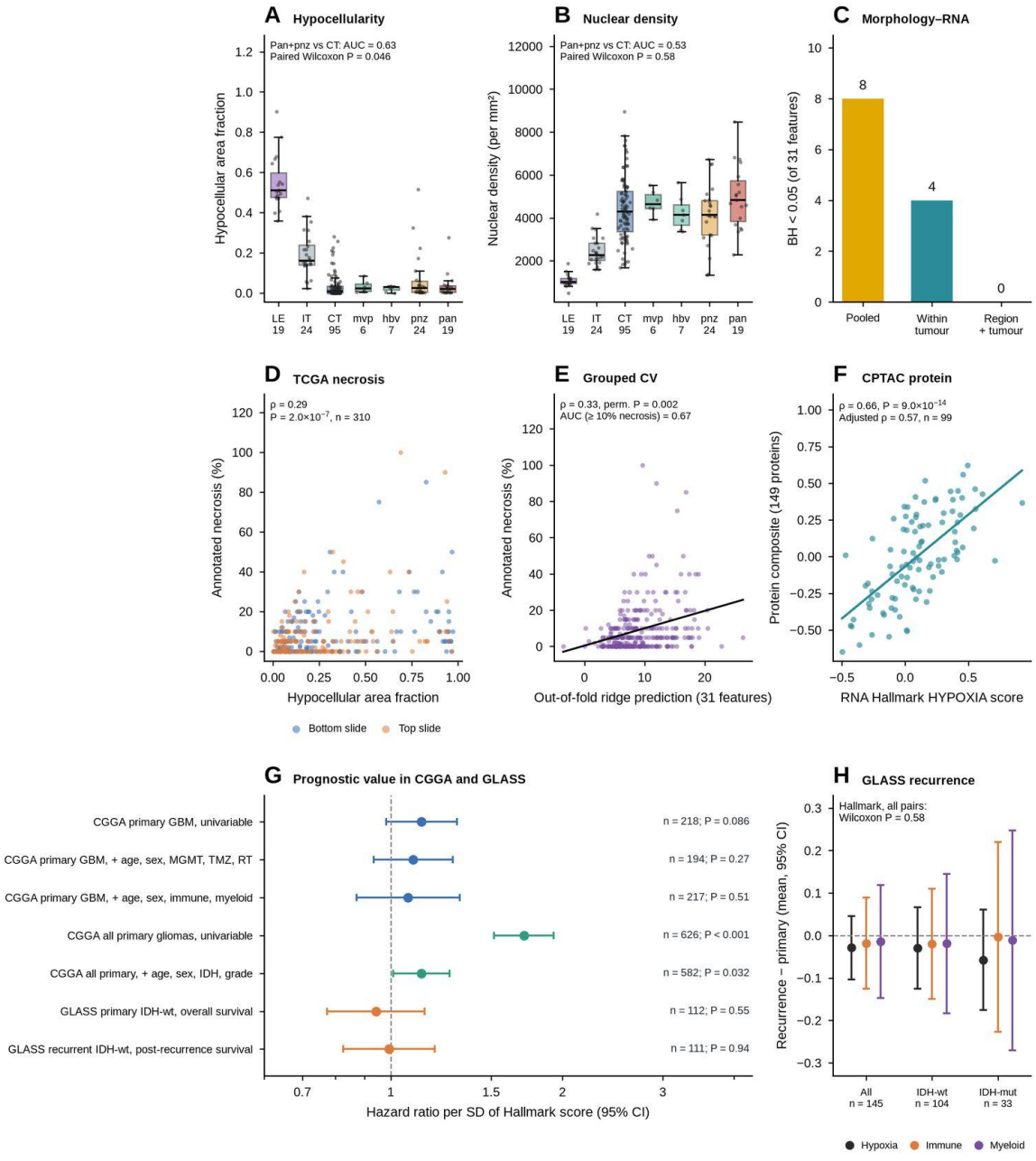


Fig. 7. Public surrogates for pathologist annotation, IHC and external cohorts (post hoc). (A,B) IvyGAP hypocellular fraction and nuclear density by region inside polygons of RNA-sampled wells, with AUC and paired Wilcoxon P for necrosis-adjacent regions versus cellular tumour. (C) Number of

BH-significant features (of 31) for the Hallmark score in IvyGAP under three models (pooled, within-tumour, region and tumour fixed effects). (D) TCGA-GBM frozen-slide hypocellular fraction versus GDC percent necrosis (310 slides; $\rho = 0.29$). (E) Patient-grouped cross-validated ridge prediction of percent necrosis ($\rho = 0.33$; permutation $P = 0.002$; AUC for $\geq 10\%$ necrosis 0.67). (F) CPTAC-GBM 149-protein hypoxia composite versus RNA Hallmark score (n = 99). (G) Hazard ratios per SD for CGGA and GLASS. (H) GLASS paired recurrence-minus-primary change in hypoxia, immune and myeloid scores by subgroup (mean and 95% CI; Wilcoxon P for the Hallmark score, all pairs). These are surrogates; none is equivalent to pathologist annotation or IHC, and CGGA and GLASS have no images.

Tables

Table 1. Hypotheses H1–H3 versus the Hallmark score: Spearman ρ (P) by subset.

Feature	TCGA-GBM DX1 (n = 92)	CPTAC (n = 151–152)	CPTAC hold-out (n = 60–61)	Pooled (n = 243–244)	Pooled, composition-adjusted	Max-T P (pooled)
H1 hypocellular area fraction	0.131 (0.213)	0.218 (0.007)	0.135 (0.298)	0.186 (0.0037)	0.112 (0.084)	0.056
H2 eosin-rich acellular fraction	0.150 (0.153)	0.218 (0.007)	0.078 (0.550)	0.193 (0.0026)	0.135 (0.036)	0.041
H3 nearest-neighbour distance CV	0.147 (0.162)	0.212 (0.009)	0.051 (0.699)	0.188 (0.0034)	0.066 (0.306)	0.053

Pooled: fixed-effect meta-analysis of TCGA-GBM and CPTAC (includes the hold-out patients and the patients used for feature selection; exploratory). Adjusted for immune, stromal and myeloid scores and tile count. Two-tailed nominal P values, not corrected for multiplicity.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.zip