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.

Uncovering dark biomarkers of glioblastoma survival through regulatory-deviation–based transcriptomic modeling

Yangyu Ning
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Zijun Qu
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Zehang Xie
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Shujun Wan
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Weiya Pei
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Xueqin Li
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Mengying Zhang
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Xiaolong Zhu
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Lei Li
The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University
Kun Lv
lvkun315@126.com

The First Affiliated Hospital of Wannan Medical College: Yijishan Hospital of Wannan Medical University

Research Article

Keywords: Glioblastoma, Transcriptomics, Regulatory deviation, Dark biomarkers, Prognostic biomarkers, Precision medicine

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9640328/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Glioblastoma (GBM) displays extreme survival heterogeneity that remains poorly resolved by conventional gene-level transcriptomic analyses, which often fail to identify robust survival-associated biomarkers. Here, we introduce mqTrans, a regulatory deviation modeling framework that shifts the analytical focus from absolute gene-level expression abundance to the deviation between observed expression and transcription-factor-inferred regulatory expectations. Applying mqTrans to two independent GBM cohorts (CGGA and CPTAC-GBM), we systematically defined and identified “dark biomarkers”—genes that show no significant expression changes in traditional differential expression analysis yet exhibit significant and reproducible regulatory deviations linked to survival. We uncovered 19 such dark biomarkers that consistently distinguished short-term from long-term survivors across both cohorts. These biomarkers converged on metabolic pathways (cholesterol homeostasis) and post-transcriptional regulatory programs (miRNA targets) and were enriched in genomically complex loci with overlapping non-coding transcripts, explaining their invisibility in abundance-based analyses. A LASSO-derived five-gene Cox model based on these dark biomarkers provided more stable and generalizable survival stratification than models built from conventional expression features. Our findings demonstrate that regulatory-deviation modeling can reveal hidden prognostic signals in GBM, offering a complementary regulatory-layer view of the transcriptome that captures regulatory perturbations that remain obscured under gene-expression abundance–based quantification.

Introduction

Glioblastoma (GBM) is one of the most common and aggressive primary malignant brain tumors in adults, characterized by diffuse cancer cell infiltration, vigorous angiogenesis, and extreme resistance to conventional drug therapies[1–4]. Even with aggressive multimodal treatment—including maximal safe resection, radiation therapy, and temozolomide chemotherapy—the clinical course of glioblastoma remains highly heterogeneous.[2, 5] Furthermore, there is a vast disparity in patient survival times: some patients experience extremely rapid disease progression, with only a few months elapsing between diagnosis and death, while others exhibit significantly slower progression, ranging up to several years in the case of a small number of long-term survivors[3, 6]. This marked prognostic difference—from short-term mortality to long-term survival—highlights key but not yet fully elucidated molecular determinants that govern tumor aggressiveness and the host response[4, 7]. Therefore, elucidating the molecular differences between GBM patients with short survival and those with long survival is crucial; this not only aids in refining prognostic stratification but also provides guidance for developing personalized treatment strategies[5, 8]. However, the precise molecular mechanisms driving these starkly different clinical outcomes remain inadequately elucidated by current analytical paradigms[7].

There are currently three main clinical strategies used to assess GBM prognosis and guide treatment decisions: neuroimaging assessments (e.g., evaluation of tumor burden and progression via contrast-enhanced MRI), histopathological classification and grading based on WHO guidelines, and limited molecular biomarker testing, most notably isocitrate dehydrogenase (IDH) mutation status and O^6-methylguanine-DNA methyltransferase (MGMT) promoter methylation[9–12]. Although these established methods provide a foundational framework for patient management, they have significant limitations in fully capturing the heterogeneity of GBM survival[10, 13]. Imaging and pathological assessments are inherently macroscopic or morphological in nature and often lag behind the molecular evolution of tumor dynamics[11, 14]. Furthermore, molecular assays for detecting MGMT methylation or IDH mutation status are typically resource-intensive, time-consuming, and susceptible to variations in experimental conditions, despite the clinical significance of these markers[12, 15]. Crucially, these traditional markers have limited ability to resolve the complex layers of post-transcriptional and post-translational regulation, which may profoundly influence tumor cell behavior and treatment sensitivity, leaving a significant portion of survival heterogeneity unexplained[15, 16].

The emergence and widespread adoption of high-throughput RNA sequencing (RNA-seq) have propelled computational methods to the forefront of GBM research, aiming to decipher this unresolved heterogeneity[17, 18]. Standard analytical workflows in this field extensively employ differential expression analysis (DEG), unsupervised clustering and dimension reduction techniques (such as PCA and t-SNE), as well as survival regression models (such as univariate and multivariate Cox proportional hazards regression)[18–20]. However, the vast majority of these computational methods are based on a critical implicit assumption: they model biology at the gene level, treating the total abundance of all transcripts produced at a gene locus as the primary functional variable[20, 21]. This perspective essentially posits that quantitative changes in total mRNA abundance are the primary drivers of phenotypic differentiation. However, a growing body of evidence suggests that relying solely on gene expression abundance is insufficient to fully characterize tumor-associated molecular abnormalities, and that complex transcriptional regulatory relationships and dysregulated networks may play a significant role in tumorigenesis and disease progression[22–24]. By merging different transcript isoforms with

potentially antagonistic functions into a single expression value, gene-level analyses risk systematic signal loss, thereby obscuring critical regulatory changes that occur even when overall gene expression abundance remains largely unchanged[21, 23, 25].

Here, we propose a regulatory deviation modeling framework (mqTrans) built on gene-level expression, which captures discrepancies between observed expression and transcription factor–inferred regulatory expectations.Unlike traditional abundance-centric approaches, mqTrans shifts the analytical focus from the absolute abundance of gene expression to the deviation between the observed expression levels and those predicted based on upstream transcriptional regulators[24–26]. This framework is used to identify potential biological signals that exhibit no significant changes in overall expression levels but have already undergone abnormalities in their transcriptional regulatory states. Based on this regulatory deviation analysis framework, we systematically defined and investigated a novel class of molecular features termed “dark biomarkers.” These biomarkers are operationally defined as genes that do not exhibit significant changes in traditional gene-level differential expression analyses but show significant differences under the mqTrans regulatory deviation perspective[25–27].

In this study, based on the mqTrans framework, we conducted a systematic analysis of two independent GBM cohorts (CGGA and CPTAC-GBM) and obtained the following key findings. First, we demonstrated that mqTrans modeling based on regulatory deviation can identify potential survival-related signals that are difficult to detect using traditional gene expression analysis, providing a new analytical perspective for elucidating prognostic heterogeneity in GBM. Second, this study defined and identified a class of “dark biomarkers” that are difficult to detect in traditional expression analysis but exhibit significant differences from the mqTrans perspective, suggesting that important molecular signals associated with survival may be hidden at the regulatory network level, which is difficult to capture in traditional expression analysis. Third, we further explored the potential regulatory mechanisms underlying the formation of these dark biomarkers, providing a basis for understanding their biological origins. Additionally, we validated the potential clinical utility of these dark biomarkers in patient prognostic stratification.

In addition to providing a new analytical perspective, this framework may have important translational implications. By capturing regulatory-deviation signals that are not detectable using conventional gene-expression approaches, mqTrans-derived biomarkers may improve patient stratification and prognostic assessment in glioblastoma. Furthermore, these dark biomarkers may represent potential therapeutic targets embedded within dysregulated regulatory networks, thereby offering new opportunities for precision medicine. Therefore, there is a critical need for analytical frameworks that can capture regulatory-layer perturbations beyond gene-expression abundance.

Materials and Methods

2.1 Datasets and Patient Selection

2.1.1 CGGA cohort

RNA-seq expression data and corresponding clinical annotations were obtained from the Chinese Glioma Genome Atlas (CGGA) cohort [28, 29], which included CGGA693 and CGGA325 samples. Samples were retained if they met the following criteria: (i) primary tumor samples; (ii) histologically annotated glioblastoma, WHO grade IV; and (iii) available overall survival (OS) information. Survival groups were defined to contrast the extremes of the survival spectrum. Short survival was defined as OS ≤ 360 days, and long survival was defined as OS ≥ 720 days. Samples with intermediate survival, defined as OS between 361 and 719 days, were excluded from the main analyses. After filtering, the CGGA cohort included 149 primary GBM samples and was used as the discovery cohort.

2.1.2 CPTAC-GBM cohort

An independent validation cohort was assembled from the CPTAC-GBM dataset [30]. The same sample selection criteria were applied, including primary GBM status, WHO grade IV annotation, and complete OS information. The same survival cutoffs were used to define the short-survival and long-survival groups, with OS ≤ 360 days classified as short survival and OS ≥ 720 days classified as long survival. Samples with intermediate survival were excluded. After filtering, the CPTAC-GBM cohort included 63 GBM samples and served as the independent validation cohort.

2.1.3 Expression matrix preprocessing

Gene-level TPM expression matrices from the CGGA and CPTAC-GBM cohorts were used for conventional expression-based downstream analyses. After sample filtering, the CGGA matrix contained 47,886 gene features across 149 samples, whereas the CPTAC-GBM matrix contained 10,959 gene features across 63 samples. Expression values were transformed as log2(TPM + 1) prior to modeling[16, 31]. To ensure cross-cohort comparability, downstream analyses were restricted to the 10,628 overlapping genes shared by the two cohorts. After missing or invalid value processing, 10,309 genes remained in the CGGA cohort and 9,269 genes remained in the CPTAC-GBM cohort. Batch effects were corrected before differential analysis, and available covariates were incorporated into the linear modeling framework[32, 33].

2.2 The mqTrans View of Transcriptomes

2.2.1 Construction of the TF-based prediction model

For each target gene, a transcription factor-based regression model was established to estimate its expected expression from the expression profiles of its candidate upstream regulators. Candidate transcription factors were obtained from the TRRUST regulatory interaction database [34], and their normalized expression values were used as model predictors.

For a given target gene j, the expected expression was modeled as:

$$\widehat{mRNA}_j = \beta_0 + \sum_k \beta_k TF_k + \epsilon \# (1)$$

where $\widehat{mRNA}_j$ denotes the predicted expression level of gene j, TF_k represents the expression level of the $k-th$ transcription factor, β_k is the estimated regression coefficient, and ϵ is the residual term.

The fitted model was then applied to each sample to obtain the expected expression value for each target gene. The mqTrans feature was defined as the absolute difference between the observed and predicted expression levels:

$$mqTrans_j = \left| mRNA_j - \widehat{mRNA}_j \right| \# (2)$$

This value quantifies the extent to which the observed expression of a gene deviates from the level expected from its transcription factor-based regulatory context.The models were pre-trained using transcriptomic data of healthy brain tissue samples from the UCSC Xena platform [5] (GTEx dataset, n = 1,157, tissue = brain). We selected healthy brain tissue because the core transcriptional regulatory relationships (e.g., TF→target gene) are expected to be relatively stable in non-malignant tissue. For each target gene, a linear regression model was fitted using the expression levels of its known TFs (from TRRUST) as predictors. The model parameters were estimated in the healthy samples using ordinary least squares with 5-fold cross-validation to guard against overfitting[35]. The trained models were then applied to each GBM sample to predict the ‘expected’ expression level of each target gene. This design assumes that deviations from the healthy regulatory baseline in tumor samples may reflect disease-associated dysregulation.

Unlike conventional transcriptomic analysis, which mainly evaluates absolute gene expression abundance, mqTrans focuses on regulatory deviation [36, 37]. This design allows the analysis to capture transcriptomic alterations that may not be evident from direct expression abundance alone. Thus, mqTrans provides a complementary view of transcriptome dysregulation by representing the discrepancy between observed expression and transcription factor-predicted expression.

mqTrans features were generated independently for the CGGA and CPTAC-GBM cohorts using the same computational procedure. The resulting mqTrans matrices were then used for downstream differential analysis and for the identification of dark biomarkers.

This framework assumes that transcriptional regulatory relationships learned from normal tissues provide a stable baseline, and that deviations from this baseline in tumor samples reflect disease-associated dysregulation.

Although transcriptional regulatory relationships may be partially altered in tumor contexts, a substantial proportion of core TF–target interactions are conserved across physiological and pathological states. Therefore, deviations from a normal regulatory baseline are more likely to reflect disease-associated dysregulation rather than random variation. This design allows mqTrans to capture biologically meaningful perturbations while maintaining a stable reference framework for cross-sample comparison.

2.3 Differential Analysis

To compare the mqTrans framework with conventional expression-based analysis, differential analyses were performed in parallel at two analytical levels: gene-level expression abundance and mqTrans-derived regulatory deviation. The same survival grouping strategy was used in both analyses, where samples were classified as short-survival or long-survival according to the predefined overall survival cutoffs. Differential testing was conducted separately in the CGGA and CPTAC-GBM cohorts.

For conventional expression-level analysis, each gene was evaluated independently using an ordinary least squares linear regression model. The processed gene expression value was used as the response variable, and the survival group indicator was included as the primary variable of interest. The model further incorporated available demographic, clinical, and technical covariates, including gender, age, OS days, OS event status, and batch, to account for potential confounding effects.

For mqTrans-level analysis, differential testing was performed on the mqTrans feature matrix. Each mqTrans feature quantified the regulatory deviation magnitude of a gene, defined as the discrepancy between its observed expression and the expression predicted from transcription factor-based regulation. Because mqTrans values represent deviation strength rather than direct expression abundance, differences between the short-survival and long-survival groups were assessed for each feature using a two-sided Mann–Whitney U test[38].

Multiple testing correction was performed separately for each cohort and each analytical level. For the gene-level expression analysis, nominal *P* values from the linear models were adjusted using Bonferroni correction[39], whereas *P* values from the mqTrans-level Mann–Whitney U tests were adjusted using the Benjamini–Hochberg procedure[40]. Features with adjusted *P* values below 0.05 were considered statistically significant.The threshold for minimal expression change was defined as |log2 fold change| < 0.2. Dark biomarkers were defined as genes that were significant at the mqTrans level in both the CGGA and CPTAC-GBM cohorts, but were not significant in conventional gene-level expression analysis in either cohort.

Results and Discussion

3.1 Conventional gene-expression abundance provides limited survival-associated signals in GBM

To evaluate whether conventional gene-level transcriptomic analysis could identify survival-associated molecular signals in glioblastoma, we first performed differential expression analysis between short-survival (< 1 year) and long-survival (> 2 years) patients in the CGGA and CPTAC-GBM cohorts based on normalized gene-level expression profiles.

In both cohorts, no genes remained significant after multiple testing correction, indicating that conventional gene-level differential expression analysis failed to identify robust survival-associated biomarkers[30, 41, 42]. To visualize potential expression trends, volcano plots based on nominal P-values were examined (Fig. 1A-B). In the CGGA cohort, most genes were symmetrically distributed around zero fold change with only a few genes showing weak differential trends. Similarly, although a limited number of genes in the CPTAC-GBM cohort exhibited nominal significance, these signals were sparse and lacked consistency, suggesting that survival-associated expression differences at the gene level were generally weak and unstable across datasets.

To further assess whether survival groups could be separated at the global transcriptomic level, principal component analysis (PCA) was performed using the top 20 genes ranked by nominal P-values in each cohort (Fig. 1C-D). In both the CGGA and CPTAC-GBM datasets, short-survival and long-survival samples showed substantial overlap in PCA space without clear clustering patterns. This result indicates that overall gene-expression patterns were insufficient to distinguish survival phenotypes, even when focusing on the most variable candidate genes[30, 43].

Together, these findings demonstrate that conventional gene-level expression profiling provides limited discriminatory power for identifying survival-associated molecular signatures in GBM. The lack of reproducible differential expression signals across independent cohorts suggests that prognostic information may be concealed within regulatory perturbations not captured by absolute expression abundance alone, rather than being solely attributable to limited statistical power, a phenomenon frequently observed in highly heterogeneous tumors such as glioblastoma[44, 45]. These observations motivated the application of the mqTrans framework to detect hidden survival-associated signals from a regulatory deviation perspective.We further verified that the observed trends were robust under alternative survival grouping strategies, supporting the stability of the findings.

3.2 mqTrans-based regulatory deviation modeling reveals hidden survival-associated signals

To determine whether regulatory deviation–based transcriptomic modeling could reveal survival-associated signals that were undetectable at the conventional gene-expression level, we applied the mqTrans framework to the same short-survival and long-survival GBM cohorts.

Differential analysis based on mqTrans features identified multiple significantly altered genes in both the CGGA and CPTAC-GBM cohorts after multiple testing correction (adjusted $P < 0.05$), in contrast to the absence of significant signals in conventional gene-level analysis. Volcano plots demonstrated a substantially larger number of significant features with clearer distributions of differences in deviation magnitude between the two survival groups in both cohorts (Fig. 2A-B). These findings indicate that mqTrans captures a distinct layer of transcriptomic variation by quantifying deviations between observed expression and model-inferred regulatory expectations, thereby shifting the analytical focus from abundance-centric measurements to regulation-centric representations [37, 46, 47]. This framework therefore represents an alternative paradigm for transcriptomic analysis that complements and extends conventional gene-expression–based approaches.

To further evaluate the discriminatory ability of mqTrans-derived features, principal component analysis (PCA) was performed using the top 20 significant mqTrans biomarkers ranked by adjusted P-values. In the CGGA cohort, mqTrans features showed an evident trend toward separation between short-survival and long-survival samples, while in the CPTAC-GBM cohort the separation was more pronounced, with the two survival groups forming distinct clusters in PCA space (Fig. 2C-D).It should be noted that PCA was used here as a descriptive visualization tool rather than an independent classification method. Compared with the substantial overlap observed in gene-level PCA analysis, mqTrans-derived features demonstrated markedly improved discriminatory power for stratifying patient survival phenotypes[47, 48].

Taken together, these results demonstrate that regulatory deviation modeling reveals a distinct layer of survival-associated molecular information in GBM that is not captured by gene-expression abundance. By quantifying deviations between observed expression and model-inferred regulatory expectations, mqTrans detects regulatory perturbations that may underlie survival heterogeneity even in the absence of overt expression changes. This increased detectability of survival-associated deviation features provided the basis for defining dark biomarkers[37, 46, 49]. Conceptually, mqTrans can be interpreted as a regulatory deviation–based modeling framework that captures discrepancies between observed expression and inferred regulatory expectations.

3.3 Cross-cohort identification of 19 expression-invisible dark biomarkers

Building on the mqTrans-derived survival-associated signals identified above, we next sought to define a stable set of reproducible features across cohorts.To identify prognostic molecular signals that are not detectable by conventional differential expression analysis, we sought to rigorously define “dark biomarkers”. To this end, we imposed additional constraints beyond statistical non-significance at the gene-expression level. Specifically, dark biomarkers were defined as genes that (i) showed significant differences in mqTrans features (adjusted $P < 0.05$) in both cohorts, and (ii) exhibited minimal changes in gene-expression abundance, as indicated by minimal expression effect sizes (below a predefined threshold) and lack of consistent directional trends across cohorts.Conceptually, these dark biomarkers are analogous to hidden or latent regulatory signals described in network-based biomarker studies, which are not directly observable at the level of gene-expression abundance.Notably, the focus here is on identifying features that are reproducible across cohorts rather than further optimizing discriminatory performance.

Although no genes remained significant after multiple testing correction at the gene-expression level, a subset of genes exhibited weak but variable expression changes, with some showing nominal significance (unadjusted $P < 0.01$) and partially consistent fold-change directions across cohorts. Based on this stringent definition, dark biomarkers were subsequently identified as genes satisfying these criteria in both cohorts, ensuring that these biomarkers represent regulatory perturbations independent of expression abundance rather than artifacts of limited statistical power. Under this framework, we identified 19 consistently detected genes across the CGGA and CPTAC-GBM cohorts. Notably, none of these genes reached adjusted significance in conventional gene-

level analysis, confirming that they represent a class of regulatory-deviation signals that are not detectable by traditional abundance-focused approaches.

mqTrans-based differential analysis identified 115 candidate biomarkers in the CGGA cohort and 277 in the CPTAC-GBM cohort. Cross-cohort comparison revealed 19 overlapping genes, which were consistently detected in both datasets and were therefore defined as dark biomarkers (Fig. 3A). Notably, these biomarkers were not identified through conventional gene-level differential expression analysis, indicating that they represent hidden prognostic signals independent of absolute gene expression changes.

To evaluate whether these 19 dark biomarkers provided improved discriminative ability, we compared them with the top-ranked 19 genes selected by nominal P values from the raw-expression analysis. Clustering performance was assessed using the Silhouette score and Davies–Bouldin index. In both cohorts, the dark biomarker set achieved higher Silhouette scores and lower Davies–Bouldin indices than the top-ranked raw-expression genes (Fig. 3B), indicating that these features exhibit consistent clustering structure and support a more stable cross-cohort feature set.

Principal component analysis (PCA) further demonstrated that the 19 dark biomarkers enabled clearer separation between long-survival and short-survival patients in both the CGGA and CPTAC-GBM cohorts (Fig. 3C–D). In contrast, the top-ranked raw-expression genes showed weaker and less stable separation patterns. These results suggest that mqTrans-derived dark biomarkers provide a consistent feature space that supports stable separation patterns across independent cohorts.

Together, these findings demonstrate the existence of a class of latent regulatory biomarkers that are not detectable through conventional differential expression analysis but can robustly capture survival-associated molecular differences in glioblastoma, highlighting the importance of regulatory relationships beyond absolute expression abundance[50, 51].Importantly, the consistency of these signals across independent cohorts argues against insufficient statistical power as the primary explanation and instead supports a regulatory-origin mechanism.

Notably, several identified dark biomarkers have previously been implicated in glioblastoma biology. For example, PDPK1 is a key regulator of the PI3K/AKT signaling pathway and has been associated with tumor cell survival and therapeutic resistance.[52] SMARCA2, a chromatin remodeling factor, has been reported to influence transcriptional regulation and tumor progression[53]. SRPK2 plays a role in RNA splicing and has been linked to oncogenic processes in multiple cancers[54]. The identification of these genes further supports the biological relevance of mqTrans-derived dark biomarkers and suggests that regulatory-deviation signals may capture functionally important drivers of tumor behavior

Samples are colored by survival group (blue: long-survival; red: short-survival). mqTrans-derived dark biomarkers exhibit clearer separation between survival groups compared with conventional gene-expression features, indicating consistent separation patterns across cohorts.

3.4 Dark biomarkers converge on cholesterol homeostasis and miRNA-mediated regulatory programs

To investigate the biological relevance of the identified dark biomarkers, we performed functional enrichment analyses using curated pathway databases (MSigDB Hallmark) and miRNA target databases (miRTarBase), with statistical significance assessed using a hypergeometric test followed by Benjamini–Hochberg correction. Significant terms were selected based on an adjusted P value < 0.05, excluding single-gene matches to ensure robustness.

The enrichment results revealed that dark biomarkers were preferentially associated with metabolic regulatory pathways, particularly those related to lipid homeostasis (e.g., cholesterol homeostasis), suggesting that survival-associated transcriptional dysregulation may converge on metabolic reprogramming processes in glioblastoma[55–57].This is consistent with the known metabolic plasticity of glioblastoma, where lipid metabolism has been implicated in tumor growth and therapy resistance.

At the post-transcriptional level, dark biomarkers exhibited significant enrichment in multiple miRNA-mediated regulatory modules, including targets of hsa-miR-16-5p and hsa-miR-95-5p[58]. This indicates that these genes are embedded within coordinated post-transcriptional regulatory networks rather than acting as independent expression units.

Notably, such enrichment patterns were not observed in conventional Gene Ontology categories, highlighting that the functional convergence of dark biomarkers is not reflected in standard gene-centric annotations but instead emerges at the level of regulatory

interactions[59].This pattern is consistent with the concept that regulatory convergence may arise at the network level rather than through individual gene annotations, further supporting the view that mqTrans captures a distinct layer of biological organization beyond absolute gene-expression changes.

Collectively, these findings demonstrate that dark biomarkers converge on metabolic and post-transcriptional regulatory programs that are closely linked to glioblastoma progression and survival heterogeneity, reinforcing their role as regulatory-network-level biomarkers. Detailed enrichment results are provided in Supplementary Table S4.

Term	Source	Ratio	P.value	adj.P.value
Cholesterol Homeostasis	MSigDB_Hallmark_2020	2/74	2.22E-03	1.33E-02
hsa-miR-16-5p	miRTarBase_2017	8/1555	4.55E-05	2.59E-02
mmu-miR-3074-2-3p	miRTarBase_2017	2/22	1.95E-04	4.37E-02
hsa-miR-95-5p	miRTarBase_2017	3/128	2.30E-04	4.37E-02

3.5 Dark biomarker loci overlap with non-coding transcripts, suggesting transcriptomic structural complexity

Having established that dark biomarkers converge on regulatory programs, we next examined whether their invisbility at the gene-expression level could be partially explained by transcriptomic structural complexity.To explore the structural basis underlying the hidden regulatory signals captured by mqTrans, we examined whether dark biomarkers are preferentially located in transcriptionally complex genomic regions. Using GENCODE annotation, we identified multiple non-coding transcripts overlapping dark biomarker loci, including antisense, intronic, TEC, and other non-coding transcript types[60, 61].

These overlapping transcriptional elements indicate that many dark biomarkers reside in loci with complex transcriptional architectures.These overlapping transcripts are not limited to long non-coding RNAs but include diverse classes of non-coding transcriptional elements. In such regions, RNA-seq reads may originate from multiple partially overlapping transcripts, leading to ambiguity in read assignment during conventional gene-level quantification.

Importantly, inspection of representative loci using genome browser data revealed RNA-seq coverage spanning overlapping transcript boundaries, supporting the active transcription of these non-coding elements rather than annotation artifacts. This suggests that transcriptional overlap introduces systematic uncertainty in gene-level expression estimates, consistent with transcriptional interference and isoform ambiguity reported in complex genomic loci.

In this context, conventional abundance-based analyses may obscure biologically relevant regulatory perturbations arising from transcriptomic complexity, as signal contributions from overlapping transcripts cannot be disentangled. By contrast, mqTrans captures deviations in transcriptional relationships, enabling the detection of regulatory perturbations that arise from such structurally complex regions.

Notably, several key dark biomarkers, including PDPK1, SRPK2, and SMARCA2, were associated with multiple overlapping non-coding transcripts, further supporting the prevalence of transcriptional interference and regulatory complexity at these loci.

Together, these findings suggest that dark biomarkers are enriched in transcriptionally complex genomic regions, where overlapping non-coding transcription introduces quantification ambiguity. This provides a mechanistic explanation for why such regulatory signals remain obscured in conventional analyses while becoming detectable under a regulatory-deviation framework.More broadly, this highlights a fundamental limitation of gene-centric quantification in resolving transcriptional complexity.

Dark Biomarker	chr_id	gene_name	lncRNA_ensemble_id	lnc_strand	lnc_gene_type	lnc_gene_name
ENSG00000137449	chr4	CPEB2	ENSG00000249252	-	antisense	RP11-665G4.1
ENSG00000177706	chr7	FAM20C	ENSG00000240093	-	antisense	AC093627.12
ENSG00000177706	chr7	FAM20C	ENSG00000249852	-	TEC	AC145676.2
ENSG00000138757	chr4	G3BP2	ENSG00000201644	-	misc_RNA	Y_RNA
ENSG00000125734	chr19	GPR108	ENSG00000283950	-	miRNA	MIR6791
ENSG00000173110	chr1	HSPA6	ENSG00000273112	+	processed_transcript	RP11-25K21.6
ENSG00000140992	chr16	PDPK1	ENSG00000261613	-	antisense	RP11-20I23.13
ENSG00000140992	chr16	PDPK1	ENSG00000261288	+	sense_intronic	RP11-20I23.11
ENSG00000140992	chr16	PDPK1	ENSG00000269937	-	antisense	RP11-20I23.8
ENSG00000140992	chr16	PDPK1	ENSG00000261140	-	antisense	RP11-20I23.6
ENSG00000140992	chr16	PDPK1	ENSG00000260436	-	antisense	RP11-20I23.7
ENSG00000140992	chr16	PDPK1	ENSG00000279568	-	TEC	RP11-20I23.5
ENSG00000140992	chr16	PDPK1	ENSG00000280402	+	TEC	RP11-20I23.10
ENSG00000140992	chr16	PDPK1	ENSG00000279162	+	TEC	CTD-3126B10.2
ENSG00000140992	chr16	PDPK1	ENSG00000261093	-	antisense	CTD-3126B10.1
ENSG00000080503	chr9	SMARCA2	ENSG00000236199	-	antisense	RP11-264I13.2
ENSG00000080503	chr9	SMARCA2	ENSG00000222973	-	snRNA	RNU2-25P
ENSG00000135250	chr7	SRPK2	ENSG00000270764	+	processed_pseudogene	CTB-152G17.5
ENSG00000135250	chr7	SRPK2	ENSG00000271482	+	processed_pseudogene	CTB-152G17.4
ENSG00000135250	chr7	SRPK2	ENSG00000213361	+	processed_pseudogene	RP4-778K6.1
ENSG00000135250	chr7	SRPK2	ENSG00000244490	+	processed_pseudogene	RWDD4P1
ENSG00000135250	chr7	SRPK2	ENSG00000242154	+	antisense	RP4-778K6.3
ENSG00000135250	chr7	SRPK2	ENSG00000201179	+	snRNA	RNU6-1322P
ENSG00000062716	chr17	VMP1	ENSG00000267637	+	sense_intronic	RP11-619I22.1
ENSG00000062716	chr17	VMP1	ENSG00000284190	+	miRNA	MIR21
ENSG00000167962	chr16	ZNF598	ENSG00000260107	+	lincRNA	AC005606.15

3.6 Reduced mqTrans-derived biomarker models show survival-stratification potential

To evaluate the prognostic value of mqTrans-derived dark biomarkers, multigene Cox proportional hazards models were constructed and assessed in both the CGGA and CPTAC-GBM cohorts[35, 62].

Using the full set of 19 dark biomarkers, the model achieved significant survival stratification in the CGGA cohort, as evidenced by clearly separated Kaplan–Meier curves and a significant hazard ratio. However, when applied to the CPTAC-GBM cohort, the model exhibited extreme hazard ratios and unstable confidence intervals, indicating severe overfitting likely due to the imbalance between feature dimensionality and sample size[63, 64].

To mitigate overfitting and improve model robustness, feature selection was performed using a regularization-based Cox regression framework (LASSO), with the optimal penalty parameter (λ) determined via 10-fold cross-validation by minimizing the partial

likelihood deviance.To avoid information leakage, model training and parameter tuning were performed within the training set only. This procedure yielded a reduced set of five dark biomarkers for model construction[62, 65].

A risk score was calculated for each patient as a linear combination of selected features weighted by their Cox regression coefficients. Patients were stratified into high- and low-risk groups based on the median risk score. The resulting five-gene model demonstrated stable and consistent survival stratification across both the CGGA and CPTAC-GBM cohorts, with well-separated Kaplan–Meier curves and more reasonable hazard ratios[64, 66].

In contrast, models constructed using the top-ranked 19 genes derived from conventional gene-expression analysis showed weaker and less consistent stratification performance, particularly in the CPTAC-GBM cohort[67].

Collectively, these results indicate that mqTrans-derived dark biomarkers showed more favorable survival-stratification patterns than abundance-based features in the analyzed cohorts. Importantly, controlling feature dimensionality is essential for achieving stable cross-cohort performance in survival modeling[35, 64].

Hazard ratios (HR) with 95% confidence intervals and log-rank P values are indicated in each panel. High-risk and low-risk groups are shown in blue and orange, respectively. The mqTrans-derived dark biomarker model provides improved and more generalizable prognostic stratification compared with conventional gene-expression–based models.

Conclusion

In this study, we developed mqTrans, a regulatory deviation modeling framework built on gene-level expression designed to detect regulatory perturbations that are not captured by conventional abundance-based transcriptomic analyses. By quantifying the deviation between observed gene expression and model-inferred regulatory expectations, mqTrans provides a regulatory-layer view of the transcriptome that emphasizes regulatory dysregulation rather than absolute expression abundance[68–70].

Applying this framework to two independent Chinese Glioma Genome Atlas and Clinical Proteomic Tumor Analysis Consortium glioblastoma cohorts, we identified a set of dark biomarkers that were reproducibly associated with patient survival despite showing no significant differences in conventional gene-level expression analyses[69, 71]. These biomarkers demonstrated stronger discriminatory power for survival stratification than abundance-based features, suggesting that clinically relevant prognostic information may reside in hidden regulatory signals that are overlooked by traditional transcriptomic methods[70, 72].

Functional analyses further indicated that these dark biomarkers converge on post-transcriptional regulatory programs, while structural analyses revealed their enrichment in transcriptionally complex genomic loci, particularly regions overlapping with non-coding transcripts[69, 73]. These findings provide a plausible mechanistic explanation for why important prognostic signals may remain undetected in standard gene-level analyses and support the view that regulatory-layer perturbation represents an important layer of glioblastoma heterogeneity[73, 74].Importantly, these findings also suggest potential translational applications in clinical prognostic assessment and precision medicine.

Importantly, the mqTrans-derived biomarkers retained prognostic relevance across independent cohorts and provided improved survival stratification compared with conventional expression-based models, indicating their potential utility for molecular prognostic assessment in Glioblastoma[69, 75]. More broadly, this work suggests that regulatory-layer transcriptomic modeling can reveal biologically and clinically meaningful signals beyond gene-expression abundance alone[70, 76].

Several limitations should be acknowledged. The current framework was evaluated retrospectively in public cohorts with limited sample sizes, and the biological functions of the identified dark biomarkers remain to be experimentally validated. Future studies incorporating larger prospective cohorts and functional validation experiments will be necessary to determine the mechanistic roles and translational applicability of these biomarkers.

In summary, our findings demonstrate that regulatory deviation modeling can uncover hidden survival-associated signals that are not captured by conventional gene-expression analysis, providing a new computational strategy for identifying prognostic biomarkers and improving the understanding of molecular heterogeneity in aggressive cancers.

Declarations

Ethics approval and consent to participate

This study is a secondary analysis of publicly available, de-identified transcriptomic datasets. The original studies (CGGA and CPTAC-GBM) from which the data were obtained each received ethical approval from their respective institutional review boards, and informed consent was obtained from all participants. No additional ethics approval was required for this study.

Consent for publication

Not applicable. This study contains no individual person's data in any form.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Natural Science Foundation of Anhui Province (Grant No. 2408085JX011), the Program for Excellent Scitech Innovation Teams of Universities in Anhui Province (Grant No. 2022AH010074), the Excellent Young Teacher Training Program of the Department of Education of Anhui Province (Grant No. YQZD202403), the Anhui Provincial Department of Education Higher Education Research Project (Grant No. 2025AHGXZK31534), the Open Research Fund of Anhui Province Key Laboratory of Non-coding RNA Basic and Clinical Transformation (Grant No. NcRNA202509), the Major Projects of Natural Science Research of Universities in Anhui Province (Grant No. 2025AHGXZK20259), and the General Projects of Natural Science Research of the Health Commission of Anhui Province (Grant Nos. AHWJ2024Aa20054 and AHWJ2024Aa20251).

Authors' contributions

Y.N. conceived the study, developed the methodology, performed all computational analyses, and wrote the manuscript. All other authors contributed through discussion, data interpretation, and manuscript review. K.L. supervised the project and acquired funding. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge the Chinese Glioma Genome Atlas (CGGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) for providing publicly available data. We also thank the TRRUST database and the UCSC Xena platform for their open-access resources.

Availability of data and material

The datasets analyzed in this study are available in the following public repositories: CGGA (http://www.cgga.org.cn) and CPTAC-GBM (https://proteomics.cancer.gov/programs/cptac). Additional reference data were obtained from the UCSC Xena platform (GTEx dataset, https://xena.ucsc.edu). Transcription factor–target interactions were retrieved from the TRRUST database (https://www.grnpedia.org/trrust). GENCODE annotation was obtained from https://www.gencodegenes.org. All analysis code and processed data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

1. Louis DN et al (2021) The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol 23(8):1231–1251
2. Tan AC et al (2020) Management of glioblastoma: State of the art and future directions. CA Cancer J Clin 70(4):299–312
3. Wen PY, Kesari S (2008) Malignant gliomas in adults. N Engl J Med 359(5):492–507

4. Ostrom QT et al (2023) CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2016–2020. Neuro Oncol 25(12 Suppl 2):iv1–iv99
5. Stupp R et al (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med 352(10):987–996
6. Weller M et al (2021) EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat Rev Clin Oncol 18(3):170–186
7. Brennan CW et al (2013) The somatic genomic landscape of glioblastoma. Cell 155(2):462–477
8. Phillips HS et al (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9(3):157–173
9. Eckel-Passow JE et al (2015) *Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors.* N Engl J Med, 372(26): pp. 2499 – 508
10. Hegi ME et al (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med 352(10):997–1003
11. Ceccarelli M et al (2016) Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell 164(3):550–563
12. Parsons DW et al (2008) An integrated genomic analysis of human glioblastoma multiforme. Science 321(5897):1807–1812
13. Verhaak RG et al (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17(1):98–110
14. Reifenberger G et al (2017) Advances in the molecular genetics of gliomas - implications for classification and therapy. Nat Rev Clin Oncol 14(7):434–452
15. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, (2008) 455(7216): p. 1061–1068
16. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550
17. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140
18. Ringnér M (2008) What is principal component analysis? Nat Biotechnol 26(3):303–304
19. Trapnell C et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578
20. Pertea M et al (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11(9):1650–1667
21. Wang ET et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456(7221):470–476
22. Pan Q et al (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40(12):1413–1415
23. Katz Y et al (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7(12):1009–1015
24. Soneson C, Love MI, Robinson MD (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4:1521
25. Robert C, Watson M (2015) Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol 16(1):177
26. Marbach D et al (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9(8):796–804
27. Alvarez MJ et al (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet 48(8):838–847
28. Zhao Z et al (2021) Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. Genom Proteom Bioinform 19(1):1–12
29. Zhao Z et al (2017) Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. Sci data 4(1):170024
30. Wang LB et al (2021) Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell 39(4):509–528e20
31. Conesa A et al (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13

32. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–127
33. Leek JT et al (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11(10):733–739
34. Han H et al (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 46(D1):D380–D386
35. Simon N et al (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw 39(5):1–13
36. Duan M et al (2023) Orchestrating information across tissues via a novel multitask GAT framework to improve quantitative gene regulation relation modeling for survival analysis. Brief Bioinform 24(4):bbad238
37. Li K et al Generating the Transcriptional Regulation View of Transcriptomic Features for Prediction Task and Dark Biomarker Detection on Small Datasets. J Vis Exp, 2024(205).
38. Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 4:1–39
39. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310(6973):170
40. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc: Ser B (Methodol) 57(1):289–300
41. Neftel C et al (2019) An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell 178(4):835–849e21
42. Singh S et al (2025) Glioblastoma at the crossroads: current understanding and future therapeutic horizons. Signal Transduct Target Therapy 10(1):213
43. Couturier CP et al (2020) Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nat Commun 11(1):3406
44. Yan Y et al (2024) Suppression of ITPKB degradation by Trim25 confers TMZ resistance in glioblastoma through ROS homeostasis. Signal Transduct Target Ther 9(1):58
45. Darmanis S et al (2017) Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. Cell Rep 21(5):1399–1410
46. Han J et al (2022) Multilayered control of splicing regulatory networks by DAP3 leads to widespread alternative splicing changes in cancer. Nat Commun 13(1):1793
47. Glinos DA et al (2022) Transcriptome variation in human tissues revealed by long-read sequencing. Nature 608(7922):353–359
48. Stuart T et al (2019) Comprehensive Integration of Single-Cell Data. Cell 177(7):1888 –1902.e21
49. Vaquero-Garcia J et al (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife 5:e11752
50. Zhang Y et al (2022) Identifying network biomarkers of cancer by sample-specific differential network. BMC Bioinformatics 23(1):230
51. Tang S, Yuan K, Chen L (2022) Molecular biomarkers, network biomarkers, and dynamic network biomarkers for diagnosis and prediction of rare diseases. Fundam Res 2(6):894–902
52. Migliozzi S et al (2023) Integrative multi-omics networks identify PKCδ and DNA-PK as master kinases of glioblastoma subtypes and guide targeted cancer therapy. Nat Cancer 4(2):181–202
53. Harling JD, Tinworth CP (2023) A two-faced selectivity solution to target SMARCA2 for cancer therapy. Nat Commun 14(1):515
54. Wu Y et al (2024) LTR retrotransposon-derived LncRNA LINC01446 promotes hepatocellular carcinoma progression and angiogenesis by regulating the SRPK2/SRSF1/VEGF axis. Cancer Lett 598:217088
55. Venneti S, Thompson CB (2017) Metabolic Reprogramming in Brain Tumors. Annu Rev Pathol 12:515–545
56. Miska J, Chandel NS (2023) Targeting fatty acid metabolism in glioblastoma. J Clin Invest, 133(1)
57. Kao TJ et al (2023) Dysregulated lipid metabolism in TMZ-resistant glioblastoma: pathways, proteins, metabolites and therapeutic opportunities. Lipids Health Dis 22(1):114
58. Chen M, Medarova Z, Moore A (2021) Role of microRNAs in glioblastoma. Oncotarget 12(17):1707–1723
59. Liberzon A et al (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1(6):417–425

60. Statello L et al (2021) Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 22(2):96–118
61. Ransohoff JD, Wei Y, Khavari PA (2018) The functions and unique features of long intergenic non-coding RNA. Nat Rev Mol Cell Biol 19(3):143–157
62. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16(4):385–395
63. Sauerbrei W, Royston P, Binder H (2007) Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 26(30):5512–5528
64. Waldron L et al (2014) Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst, 106(5)
65. Zhang HH, Lu W (2007) Adaptive Lasso for Cox's proportional hazards model. Biometrika 94(3):691–703
66. Bøvelstad HM et al (2007) Predicting survival from microarray data–a comparative study. Bioinformatics 23(16):2080–2087
67. Geeleher P, Cox NJ, Huang RS (2014) Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol 15(3):R47
68. Aibar S et al (2017) SCENIC: single-cell regulatory network inference and clustering. Nat Methods 14(11):1083–1086
69. Xu Z et al (2025) Integration of Single-Cell RNA and Bulk RNA Sequencing Reveals Cellular Heterogeneity and Identifies Survival-Associated Regulatory Networks in Glioblastoma. IET Syst Biol 19(1):e70025
70. Zheng Y et al (2023) Spatial cellular architecture predicts prognosis in glioblastoma. Nat Commun 14(1):4122
71. Ahmed YB et al (2024) *Identification of Hypoxia Prognostic Signature in Glioblastoma Multiforme Based on Bulk and Single-Cell RNA-Seq.* Cancers. 16(3):633
72. Ruffle JK et al (2023) Brain tumour genetic network signatures of survival. Brain 146(11):4736–4754
73. Ni B et al (2023) The short isoform of MS4A7 is a novel player in glioblastoma microenvironment, M2 macrophage polarization, and tumor progression. J Neuroinflammation 20(1):80
74. Hoogstrate Y et al (2023) Transcriptome analysis reveals tumor microenvironment changes in glioblastoma. Cancer Cell 41(4):678–692 .e7
75. Wan Z et al (2023) Identification of angiogenesis-related genes signature for predicting survival and its regulatory network in glioblastoma. Cancer Med 12(16):17445–17467
76. Oliveira FD et al (2024) Expression of key unfolded protein response genes predicts patient survival and an immunosuppressive microenvironment in glioblastoma. Translational Med Commun 9(1):5

Figures

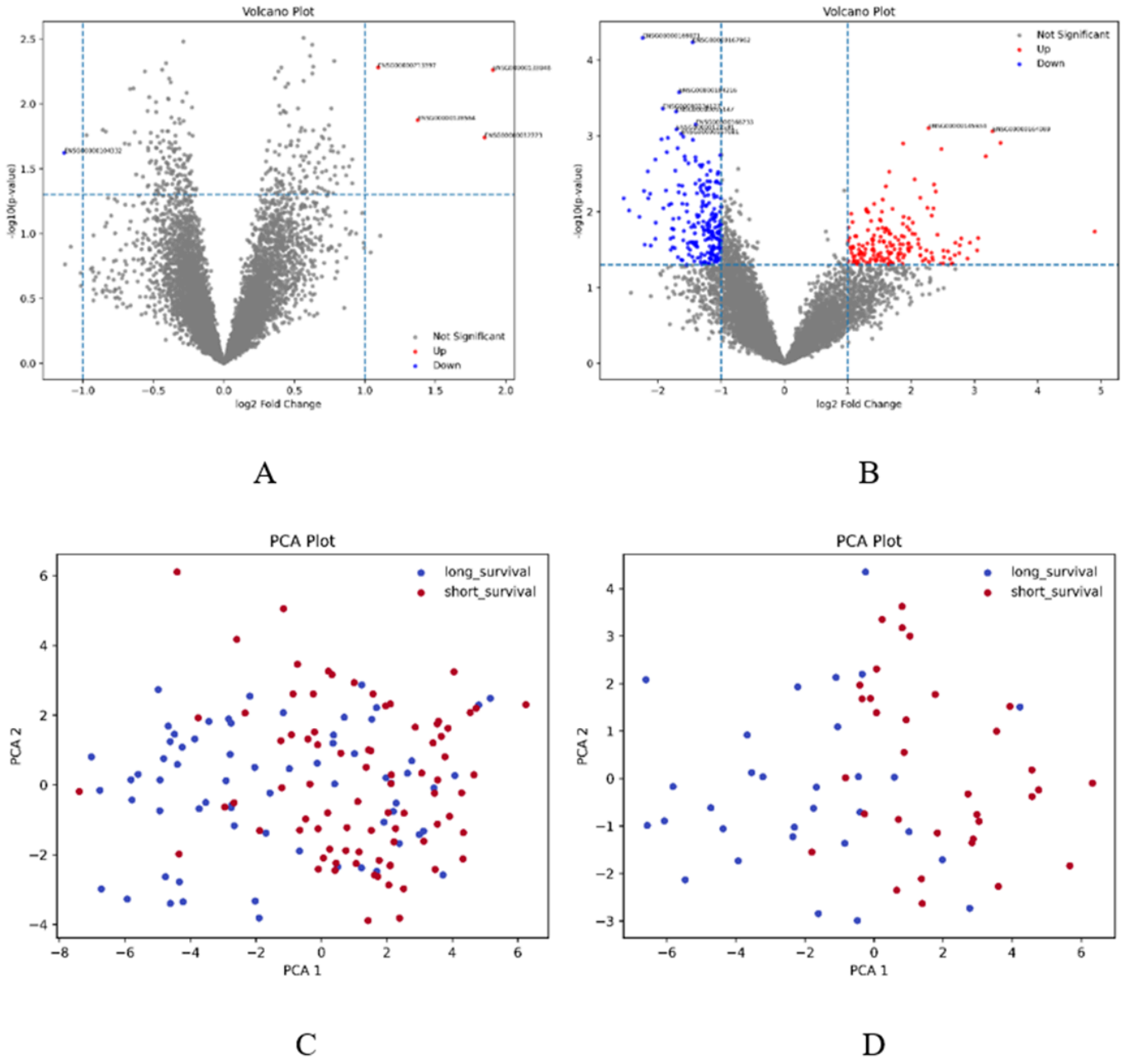


Figure 1

Conventional gene-level expression analysis fails to distinguish survival groups in GBM.

(A-B) Volcano plots of differential gene expression between short-survival and long-survival GBM patients in the CGGA and CPTAC-GBM cohorts based on gene-level expression profiles. No genes remained significant after multiple testing correction; nominal P-values are shown to illustrate overall expression trends.

(C-D) PCA plots based on the top 20 genes ranked by nominal P-values in the CGGA and CPTAC-GBM cohorts. Samples from short-survival and long-survival groups showed substantial overlap without clear separation.

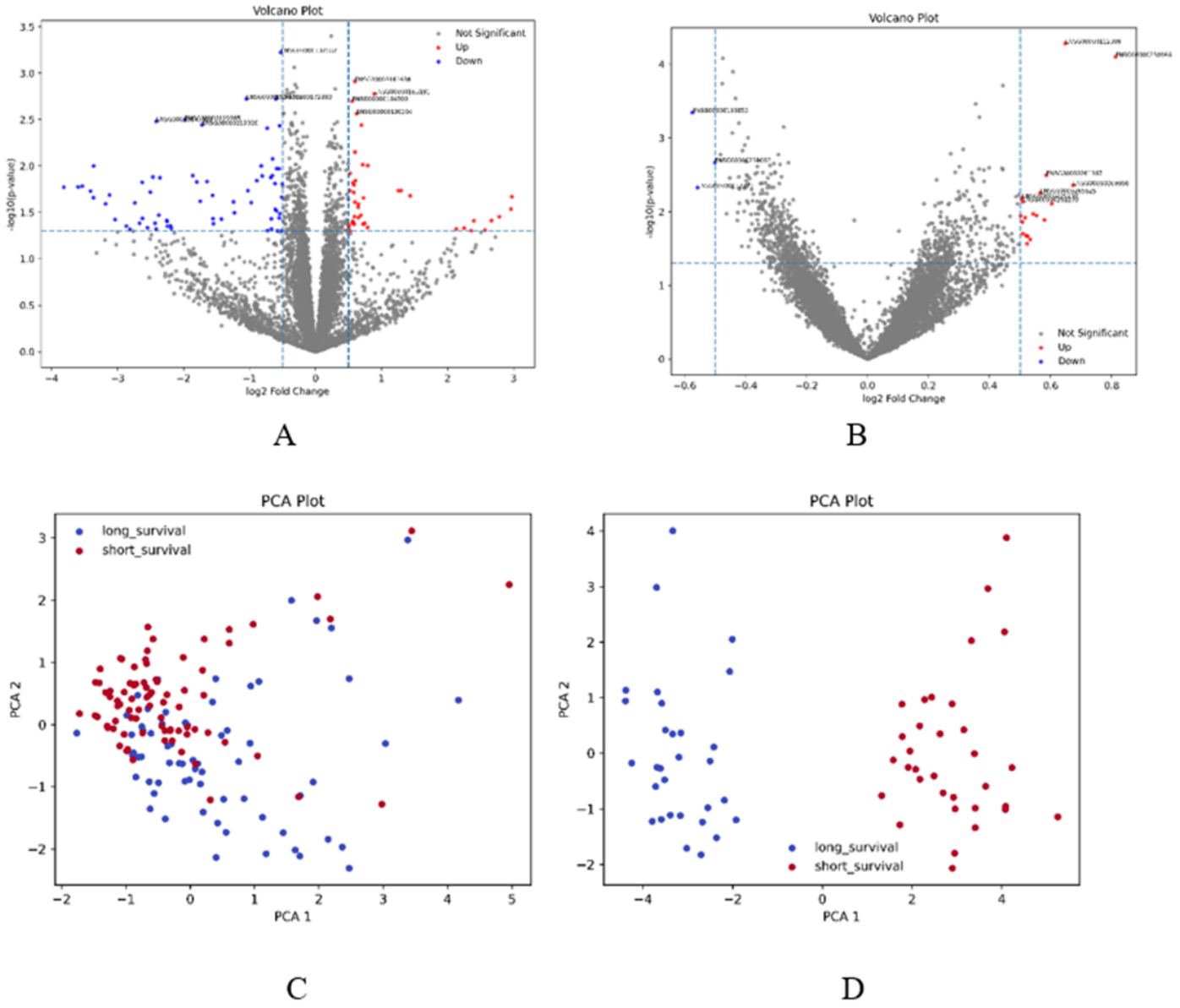


Figure 2

mqTrans-based regulatory deviation modeling reveals survival-associated signals in GBM

(A-B) Volcano plots of differential mqTrans features between short-survival and long-survival GBM patients in the CGGA and CPTAC-GBM cohorts. Significant mqTrans biomarkers remained detectable after multiple testing correction (adjusted P < 0.05), demonstrating stronger survival-associated signals compared with conventional gene-level analysis.

(C-D) PCA plots based on the top 20 significant mqTrans biomarkers in the CGGA and CPTAC-GBM cohorts. mqTrans-derived features improved separation between short-survival and long-survival groups, particularly in the CPTAC-GBM cohort.

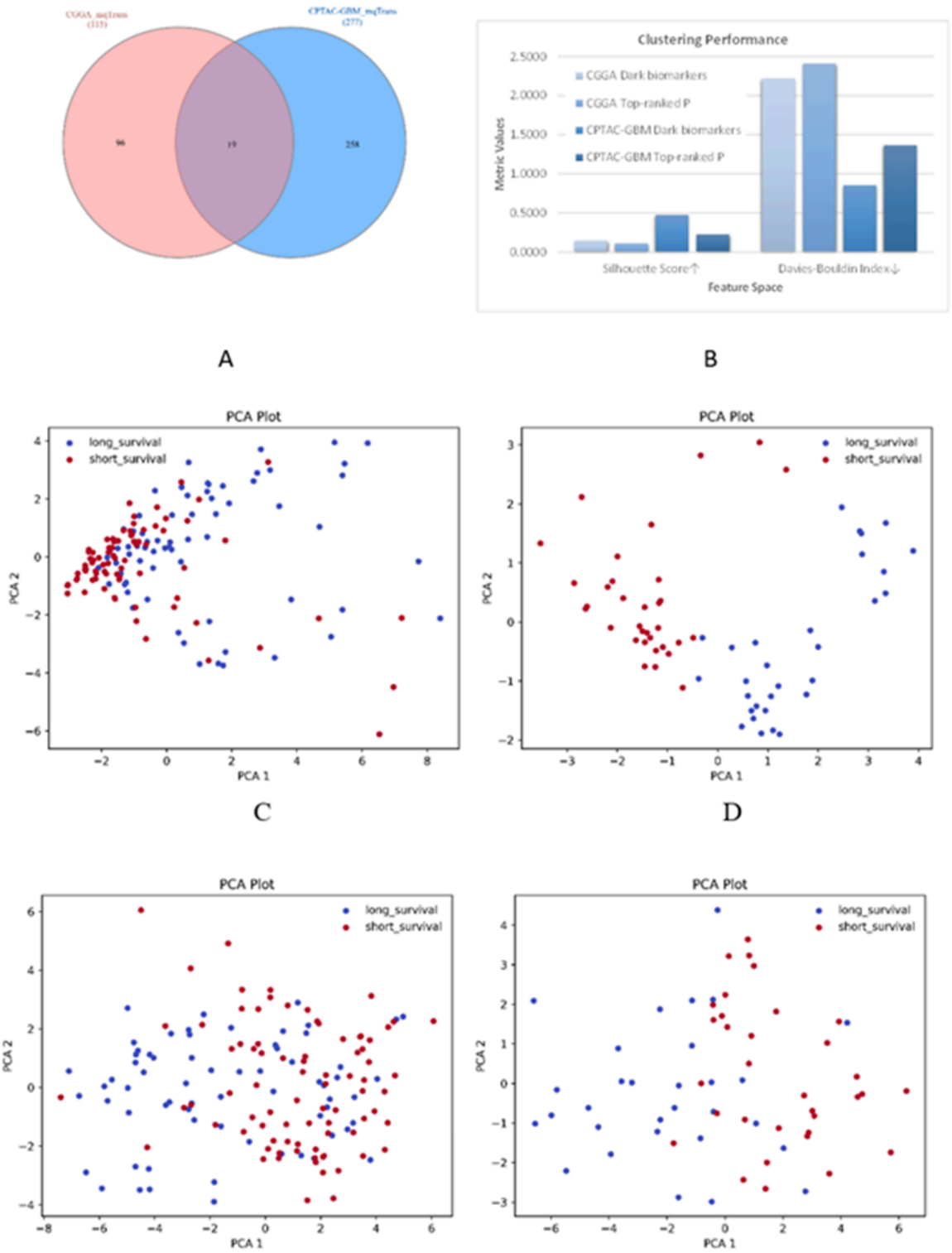


Figure 3

Identification and comparative evaluation of mqTrans-derived dark biomarkers in GBM.

(A) Venn diagram showing the overlap of significant mqTrans-derived biomarkers identified in the CGGA (n = 115) and CPTAC-GBM (n = 277) cohorts. A total of 19 shared genes were identified and defined as “dark biomarkers”, representing regulatory signals detected by mqTrans but not by conventional gene-expression analysis.

(B) Comparison of clustering performance between the 19 dark biomarkers and the top-ranked 19 genes selected based on nominal P values from gene-level expression analysis. Clustering quality was evaluated using Silhouette score (higher is better) and Davies–

Bouldin index (lower is better) in both cohorts, demonstrating improved group separation and compactness for dark biomarkers.

(C–F) Principal component analysis (PCA) plots comparing the discriminative ability of different feature sets.

(C) CGGA cohort using dark biomarkers.

(D) CPTAC-GBM cohort using dark biomarkers.

(E) CGGA cohort using top-ranked gene-expression features.

(F) CPTAC-GBM cohort using top-ranked gene-expression features.

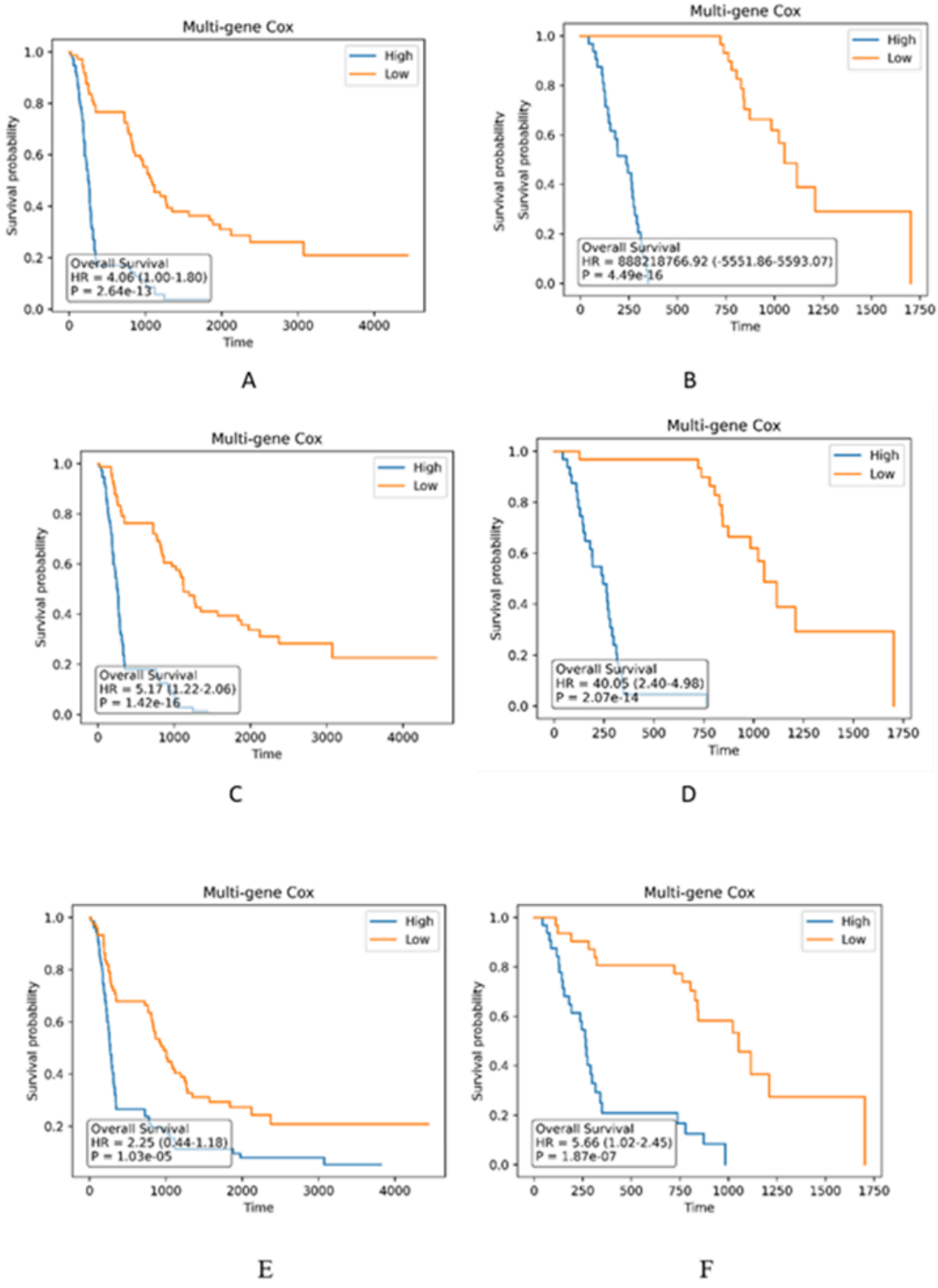


Figure 4

Survival stratification performance of mqTrans-derived dark biomarker–based Cox models.

Kaplan–Meier survival curves comparing high-risk and low-risk groups defined by multigene Cox proportional hazards models in the CGGA and CPTAC-GBM cohorts. Patients were stratified based on the median risk score calculated from model coefficients.

(A–B) Survival curves based on the full set of 19 dark biomarkers in the CGGA (A) and CPTAC-GBM (B) cohorts. While significant stratification was observed in the CGGA cohort, the CPTAC-GBM cohort exhibited extreme hazard ratios and unstable confidence intervals, indicating overfitting.

(C–D) Survival curves based on the LASSO-selected five-gene model in the CGGA (C) and CPTAC-GBM (D) cohorts. The reduced model demonstrated improved robustness, with more stable hazard ratios and consistent survival separation across cohorts.

(E–F) Survival curves based on the top-ranked 19 genes derived from conventional gene-expression analysis in the CGGA (E) and CPTAC-GBM (F) cohorts. These models showed weaker and less consistent stratification performance compared with mqTrans-derived features.