ecancermedicalscience

A cross-platform gene signature in glioblastoma

Anand Sharma[1,2a]

[1]Master of Science in Financial Engineering (MScFE), WorldQuant University, Washington, DC 20002, USA
[2]Masters in Renewable Energy and e-Mobility, Indian Institute of Technology Kanpur, Kanpur 208016, India
[a]https://orcid.org/0009-0004-7642-7737

Abstract

Glioblastoma (GBM) is a highly aggressive brain cancer with poor patient outcomes. This study uses a multi-step approach to identify reliable genetic markers for the disease. First, we analysed the GSE4290 dataset, comparing 77 tumour samples to 23 healthy brain controls to identify significantly overexpressed genes. To ensure these findings were not specific to a single study, we validated our results using the GSE147352 dataset. Finally, we correlated these gene expression levels with clinical outcomes using survival data from The Cancer Genome Atlas. Our results identified a core set of genes, including topoisomerase II alpha, maternal embryonic leucine zipper kinase, BIRC5 and abnormal spindle-like microcephaly (ASPM), which are consistently upregulated in tumours but do not significantly predict patient survival. These findings provide a validated molecular signature that defines the GBM transcriptome, though its clinical impact appears decoupled from overall survival, likely due to intensive treatment protocols or other dominant biological factors.

Keywords: *glioblastoma, gene expression profiling, biomarker validation, cross-platform validation, survival analysis*

Research

Introduction

Cancer remains one of the most significant public health challenges of the modern era. In the United States, it is estimated that one in two men and one in three women will be diagnosed with some form of malignancy during their lifetime [1]. At its core, cancer is a disease of uncontrolled cell division driven by the accumulation of genetic mutations. These mutations can be triggered by environmental factors such as ionising radiation and tobacco smoke or through internal errors during DNA replication.

Our research focuses specifically on glioblastoma (GBM), the most aggressive primary brain tumour in adults. Unlike many cancers that are staged by their physical spread to other organs, GBM rarely metastasizes outside the brain. Consequently, it is categorised by WHO Grade (1–4) based on its cellular aggressiveness; GBM is classified as Grade 4, the most invasive form. Traditional treatment involves a combination of surgery, radiotherapy and chemotherapy. Because clinicians prioritise the surgical removal of as much tumour mass as possible, early detection and precise characterisation are vital for improving survival. Despite maximal therapy, GBM remains a devastating diagnosis with a median survival of approximately 14–15 months and a 5-year survival rate of less than

Correspondence to: Anand Sharma
Email: anand.sharma9816@proton.me

ecancer 2026, **20**:2198
https://doi.org/10.3332/ecancer.2026.2198

Published: 27/08/2026
Received: 11/02/2026

Publication costs for this article were supported by *ecancer* (UK Charity number 1176307).

10% [2]. These statistics underscore the urgent need for reliable prognostic biomarkers to guide treatment decisions and improve patient outcomes.

Our work is situated within this diagnostic framework. By comparing gene expression profiles between GBM tumours and healthy brain tissue, we sought to identify a robust genetic signature of the disease. Using high-throughput genomic data, we identified a core set of genes - most notably topoisomerase II alpha (TOP2A), maternal embryonic leucine zipper kinase (MELK), BIRC5 and abnormal spindle-like microcephaly (ASPM) - which showed a consistent and substantial increase in expression across multiple independent datasets. These four genes were prioritised based on a dual rationale: they exhibited both the highest differential expression magnitude and statistical significance across discovery and validation, and they share a functionally coherent biological role as core regulators of cell cycle progression and mitosis (as expanded upon below). These biomarkers, therefore, hold significant potential for refining diagnostic accuracy and guiding the development of targeted molecular therapies, even as their independent impact on patient survival remains a subject of ongoing clinical investigation.

To understand the context of our work, we first review existing biomarker literature. While numerous studies have identified individual genes associated with poor prognosis in GBM, a critical limitation persists in the existing research: the historical reliance on single-platform analysis. Early genomic discovery was dominated by microarray technologies (e.g., Affymetrix), whereas modern research has transitioned to RNA-sequencing. This technological shift raises a fundamental question: do established biomarkers represent universal biological signals or merely platform-specific artefacts?

Several genes have emerged as potential markers for predicting GBM progression. TOP2A is an enzyme that manages DNA structure during cell division; its high expression in tumours is linked to more aggressive cell growth and shorter patient survival [3]. Similarly, ASPM associated) is a protein that promotes GBM cell growth and tumour progression [4,5]. MELK has also been identified as a key indicator of poor clinical outcomes when highly expressed [6]. Furthermore, BIRC5 (survivin) plays a critical role by preventing programmed cell death and driving continuous cell division in these aggressive Grade 4 tumours [7].

Despite these individual findings, the lack of systematic validation across disparate technological platforms remains a significant barrier to clinical translation. This gap in methodology limits the confidence required to move these biomarkers from the laboratory to the bedside. The present study addresses this gap by employing a three-tier validation strategy: (1) discovery of differentially expressed genes (DEGs) using microarray data (GSE4290), (2) technical validation using RNA-sequencing (GSE147352) and (3) prognostic evaluation using survival data from The Cancer Genome Atlas (TCGA). This framework enables systematic assessment of biomarker robustness across platforms, while critically evaluating whether this cross-platform stability translates into reliable prognostic value for the patient.

Methodology and data sources

Analytical pipeline overview: The analysis was conducted in three sequential stages: differential expression analysis (GSE4290), cross-platform validation (GSE147352) and survival analysis using TCGA-GBM data. Dataset-specific preprocessing, normalisation and statistical testing procedures are detailed below.

Discovery

For the discovery phase of this study, we utilised the GSE4290 transcriptomic dataset [8] and accessed it via the Gene Expression Omnibus database. The selection of this specific cohort was predicated on several rigorous criteria designed to ensure high statistical power and biological validity:

- Robust sample size: GSE4290 represents one of the most comprehensive bulk RNA expression cohorts available for neuro-oncology, comprising 180 total samples. This includes 81 pathologically confirmed GBM samples, providing sufficient statistical power ($n > 80$) to minimise Type I errors and identify high-confidence DEGs.
- High-quality control archetype: A recurring challenge in GBM research is the acquisition of non-malignant brain tissue. GSE4290 addresses this by utilising 'normal' brain tissue ($n = 23$) obtained from epilepsy surgery patients. Unlike post-mortem tissue or

peri-tumoural samples, which may exhibit degradation or field cancerisation effects, this cohort provides a biologically relevant baseline for identifying tumour-specific dysregulation.
- Standardised platform profiling: The data were generated using the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570). As a widely validated platform in transcriptomics, its widespread historical use facilitates seamless cross-study validation and ensures that probe-to-gene mapping is highly reliable and well-documented.
- Clinical heterogeneity: The inclusion of various glioma grades (Grades 2–4) within the same series allows for the filtration of genes that are specifically upregulated in the transition to highly aggressive GBM, rather than general markers of neural inflammation or low-grade gliosis.

Consequently, GSE4290 serves as an ideal 'Discovery Cohort' for identifying the molecular drivers of GBM. However, to ensure that the identified genetic signature was not an artefact of legacy microarray technology, a high-fidelity validation was required.

Validation

To confirm the technical and biological robustness of our findings, we utilised the GSE147352 dataset as our primary validation cohort [9]. This dataset was selected to provide a modern genomic contrast to our discovery phase based on the following factors:

- Technological shift to RNA-sequencing: While GSE4290 relies on hybridisation-based microarrays, GSE147352 utilises next-generation sequencing (Illumina HiSeq 2500). Validating our markers across these two fundamentally different technologies ensures that the observed differential expression is a true biological signal rather than a platform-specific bias.
- Replicated clinical architecture and selective filtering: Mirroring the composition of our discovery set, GSE147352 includes a spectrum of glioma grades. We applied the same rigorous exclusion criteria here, filtering out all Grade II and III samples to focus exclusively on the contrast between non-tumour tissue (n = 15) and Grade 4 GBM (n = 85). This consistent filtering strategy ensures that our validated signature specifically identifies the drivers of the most aggressive malignancy.
- Standardised processing pipeline: The use of raw count data in this cohort enabled the application of the PyDESeq2 framework - a Python-based implementation of the gold-standard DESeq2 algorithm. This facilitated sophisticated statistical modelling of gene-wise dispersion and shrinkage, providing a rigorous mathematical foundation for our validation.

By utilising GSE4290 for discovery and GSE147352 for validation, this study bridges two decades of neuro-oncology research, ensuring that our final gene signature is both historically consistent and technologically modern.

Prognosis

The final step of our methodology involves correlating the validated gene signature with clinical outcomes. While differential expression confirms a gene's presence in the tumour, prognostic validation determines its clinical significance.

- Dataset selection: We utilised TCGA GBM cohort (n = 135), accessed via the cBioPortal and TCGA Firehose Legacy databases. This dataset represents the definitive resource for clinical correlation, pairing high-depth RNA-sequencing data with long-term patient survival metadata.
- Survival analysis framework: To assess the impact of our markers on patient longevity, we employed Kaplan–Meier (KM) survival analysis. Patients were stratified into 'High' and 'Low' expression groups based on the median expression level of our target genes (TOP2A, MELK, ASPM and BIRC5).
- Statistical rigor: The divergence between survival curves was evaluated using the Log-rank (Mantel–Cox) test. A p-value of <0.05 was used to identify genes where overexpression significantly correlates with a shorter overall survival (OS).

By integrating this clinical tier, our study moves beyond mere observation. We evaluate whether these genes are not just markers of the tumour's existence, but active drivers of its lethality, making them high-priority targets for future therapeutic intervention.

Results and discussion

Data cleaning and verification

For the discovery phase, we analysed the GSE4290 transcriptomic dataset, as described. To ensure a high-confidence molecular signature, we applied a strict exclusion protocol: only samples with pathologically confirmed GBM (Grade 4) (n = 77) and non-tumour brain tissue from epilepsy surgery patients (n = 23) were included. We intentionally excluded lower-grade gliomas (astrocytomas and oligodendrogliomas) to isolate genes specific to the most aggressive form of the disease. Four samples were further removed due to incomplete clinical metadata, ensuring that the resulting 100-sample cohort provided a clean, verified comparison for differential expression analysis. Finally, the integrity of this dataset was bolstered by the absence of batch effects; all samples were collected and updated within the same timeframe, eliminating technical variance as a confounding factor in our analysis.

The expression profiles were derived from the Affymetrix HG-U133 Plus 2.0 platform, where numerical values represent normalised fluorescence intensities. To ensure cross-sample comparability, the data were scaled using a global scaling method, with a trimmed mean target intensity of 500 for each array. This linear-space normalisation accounts for technical variation in hybridisation efficiency, allowing for direct comparison between GBM and non-tumour control cohorts. For the subsequent differential expression analysis, these intensities were treated as proxies for mRNA abundance, with statistical significance determined by the magnitude of the log2 fold change (LFC), which represents the ratio of expression between the two groups on a logarithmic scale, and the Benjamini–Hochberg-adjusted p-value (a multiple test correction).

$$LFC = \log_2\left(\frac{\text{Mean Expression}_{GBM}}{\text{Mean Experssion}_{Control}}\right)$$

To verify the technical consistency of the dataset, we analysed the distribution of normalised intensities across all 100 samples. The resulting box plot, as seen in Figure 1 demonstrates that the global scaling method successfully aligned the expression levels, as shown by the uniform medians and interquartile ranges across the cohort. This uniformity confirms that the dataset is properly normalised, ensuring that the high expression of our target genes is a result of true biological activity in the tumour rather than technical error.

Transcriptomic filtering and signal enrichment

Following sample selection, we performed a gene-level filtering process to remove low-signal noise that could interfere with statistical sensitivity. We applied a threshold requiring a minimum intensity signal of ten units in at least 20% of the samples. This 'Independent Filtering' step ensures that the analysis focuses only on genes with a detectable presence across a significant portion of the cohort, thereby reducing the burden of multiple-testing corrections. Out of the original probe set, 54,217 high-confidence probes passed these criteria and were retained for the final differential expression model. This approach prioritises robust, reproducible biomarkers over stochastic background fluctuations.

Statistical analysis and gene mapping

To identify the molecular drivers of malignancy, we performed differential expression analysis using the PyDESeq2 framework. The magnitude of change for each gene was calculated as a LFC. Statistical significance was determined using the Wald test, comparing GBM (n = 77) against non-tumour controls (n = 23). The Wald test essentially evaluates whether the observed LFC is significantly different from zero; it divides the estimated change by its standard error to calculate a z-statistic. If this statistic is high enough, we can confidently reject the null hypothesis that the gene's expression is the same in both groups. The raw output is reported in Probe IDs (e.g., 201291_s_at).

To make these results biologically relevant, we mapped these identifiers to their official gene symbols (e.g., TOP2A). This 'translation' is essential, because while Probe IDs track technical signals on the DNA chip, gene symbols identify the actual proteins driving the tumour. Table 1 highlights the highest-ranking genes based on their adjusted p-values. In this model, an LFC of +1.0 represents a doubling of gene activity. Most of our top hits show massive increases; for example, the LFC of +6.99 for TOP2A indicates that the gene is approximately 127 times (2^6.99) more active in the tumour than in healthy tissue.

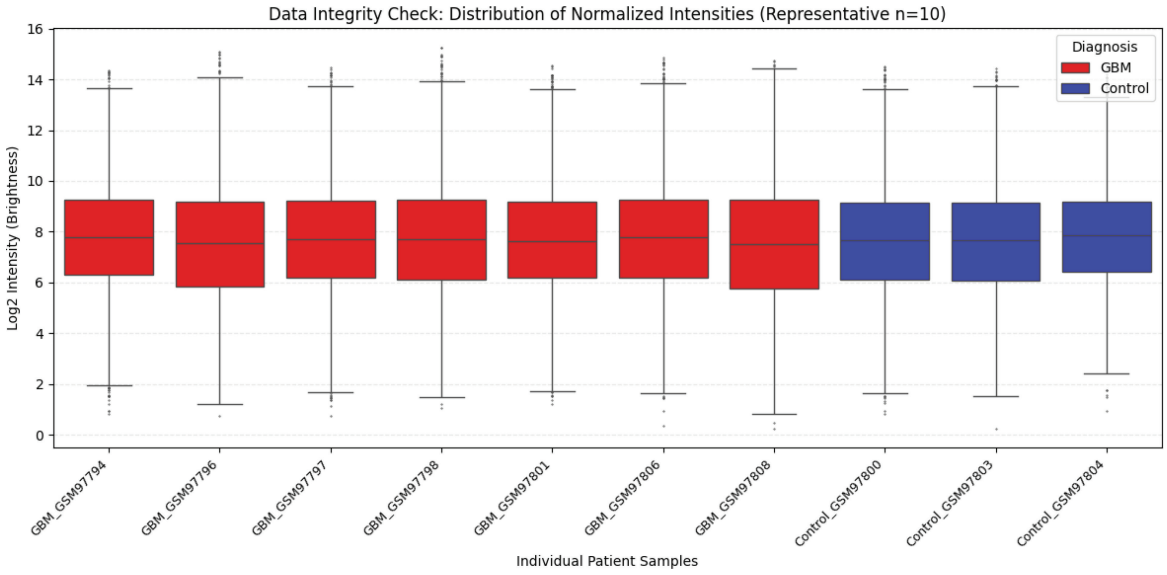

Figure 1. Box plot of normalised intensities.

Table 1. Top ten DEGs, discovery phase.

Probe ID	Gene symbol	LFC	Adjusted *p*-value
201291_s_at	TOP2A	6.994330	1.429328e-72
201292_at	TOP2A	5.933276	1.883573e-72
204162_at	MELK	5.236609	2.860220e-63
202718_at	ASPM	5.479145	5.594756e-60
219918_s_at	ASPM	5.772027	1.076239e-59
204825_at	CDK1	4.679920	6.735471e-58
219978_s_at	BUB1B	5.002662	3.440197e-57
207035_at	SCG2	−4.114191	3.440197e-57
202095_s_at	BIRC5	4.599097	2.479962e-55
218678_at	COL4A2	3.256706	4.454844e-55

The statistical results show an overwhelming dominance of cell cycle and mitosis-related genes (TOP2A, CDK1 and BUB1B). This 'molecular signature' confirms that the GBM cells have completely hijacked the body's normal growth signals. Additionally, the significant downregulation of SCG2 (LFC = −4.11) suggests that these cells are losing their specialised neuronal identity as they become increasingly malignant.

The volcano plot, as seen in Figure 2, visually summarises these results. Each point represents a single gene; the *x*-axis indicates the magnitude of change, while the *y*-axis represents statistical significance. The distinct 'eruption' of red points in the top-right quadrant identifies a high-confidence cluster of upregulated oncogenes. These genes show both massive fold-changes and extreme statistical significance (adjusted *p* value < 0.05). Conversely, the blue points in the top-left quadrant represent genes that are significantly suppressed in the tumour environment. This visualisation clearly demonstrates that the GBM transcriptome is not merely drifting randomly, but is precisely re-wired to prioritise cell cycle progression and apoptosis evasion.

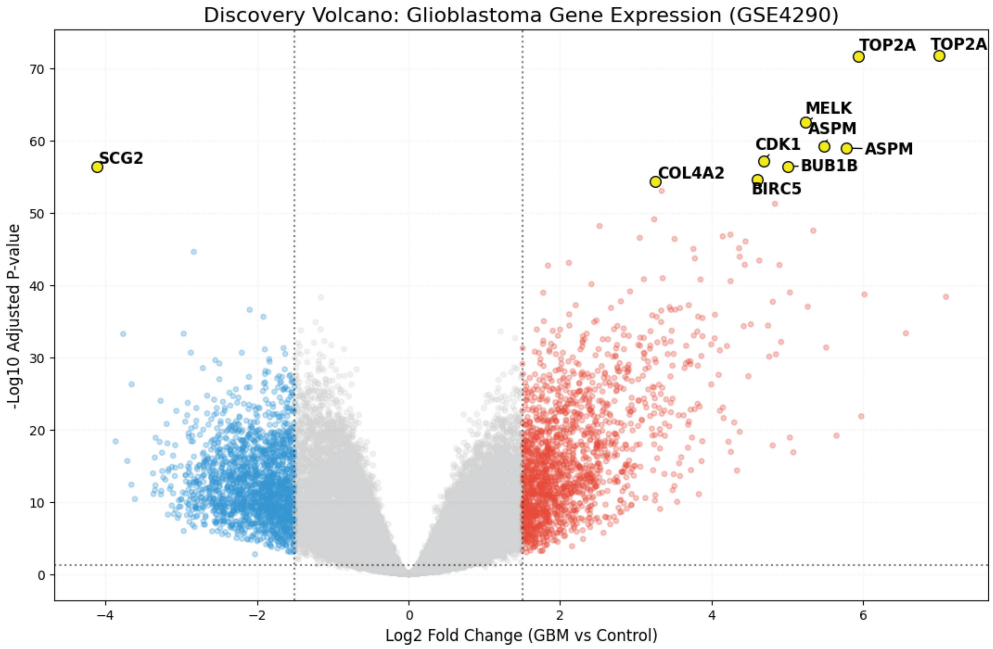

Figure 2. Volcano plot of DEGs, discovery phase.

Cross-platform validation strategy

To ensure the robustness of our molecular signature, we transitioned from microarray discovery to validation using the GSE147352 RNA-Seq dataset, as described. While microarrays rely on pre-defined probes to capture signals, RNA-Seq provides a more precise and unbiased quantification of the transcriptome. By verifying our top biomarkers across two distinct technologies and independent patient cohorts, we aim to minimise platform-specific bias and confirm that these genes are consistently overexpressed regardless of the detection method.

During this validation, we implemented an independent filtering rule in Python to retain only genes with more than ten counts in at least 20 samples. Under these criteria, three candidates from our initial list - CDK1, BUB1B and SCG2 - were automatically filtered out, indicating low levels of expression in the newer dataset. Furthermore, our analysis revealed that COL4A2 failed to maintain statistical significance ($p > 0.05$) in the RNA-Seq environment, as visualised by its position near the baseline of the volcano plot (Figure 3). In contrast, TOP2A, ASPM, MELK and BIRC5 remained exceptionally significant and highly upregulated. By surviving this dual-platform audit and technical filtering, these four genes were confirmed as our definitive mitotic gene signature, providing a high-fidelity foundation for the subsequent clinical audit.

The attrition from ten candidate genes in discovery to four in validation underscores a critical methodological insight: single-platform studies risk overestimating biomarker reliability. Of our initial top candidates, three failed to meet expression thresholds in the RNA-seq dataset, while one lost statistical significance. This 40% failure rate highlights why cross-platform validation is essential: genes appearing robust on microarray may reflect platform-specific detection biases rather than true biological signals. The four genes that survived both technologies represent high-confidence markers whose differential expression is platform-independent, strengthening their candidacy for clinical translation, where RNA-sequencing is now the standard approach.

Prognosis

To evaluate the clinical relevance of the mitotic gene signature, we performed KM survival analysis on a curated TCGA-GBM cohort (n = 135). We ensured the integrity of the clinical correlation by analysing only primary solid tumours and excluding recurrent samples to maintain a consistent baseline for survival. RNA-seq expression data were then integrated with patient survival metadata using unique 12-character case IDs, while any patients missing genomic data or clinical follow-up were pruned to ensure a high-fidelity 'intersection' dataset. Contrary to our initial hypothesis, the mRNA expression levels of our gene signature did not serve as independent prognostic biomarkers for OS.

Statistical analysis revealed no significant divergence between the high- and low-expression groups for any of the four targets:

Table 2 shows the hazard ratios (HR) near 1.0 and the frequent intersection of the survival curves in Figure 4 indicate that, within this specific cohort, transcriptional activity of these genes alone does not sufficiently stratify patient risk.

While these genes are central to the mitotic machinery and are significantly upregulated in tumour tissue, their lack of independent prognostic power in the TCGA-GBM dataset may reflect several contributing factors, although these cannot be definitively established within this study:

- Treatment neutralisation: GBM patients in the TCGA cohort typically receive the 'Stupp Protocol' (aggressive surgery, radiation and Temozolomide). This standardised, intensive therapy may neutralise the biological advantage of lower mitotic gene expression, effectively 'levelling' the survival outcomes across the cohort.
- The ceiling effect: GBM is characterised by a remarkably narrow survival window. When median survival is highly compressed (typically 12–15 months), the mathematical opportunity to observe statistically significant divergence based on a single molecular variable is reduced compared to slower-progressing malignancies.
- Biological complexity: Our results suggest that a single gene's activity may be insufficient to predict a patient's outcome. In GBM, survival is likely decided by 'master' factors, such as how well the tumour can repair itself after chemotherapy or the underlying speed of the cancer type, which may overshadow the effects of the specific genes we studied.

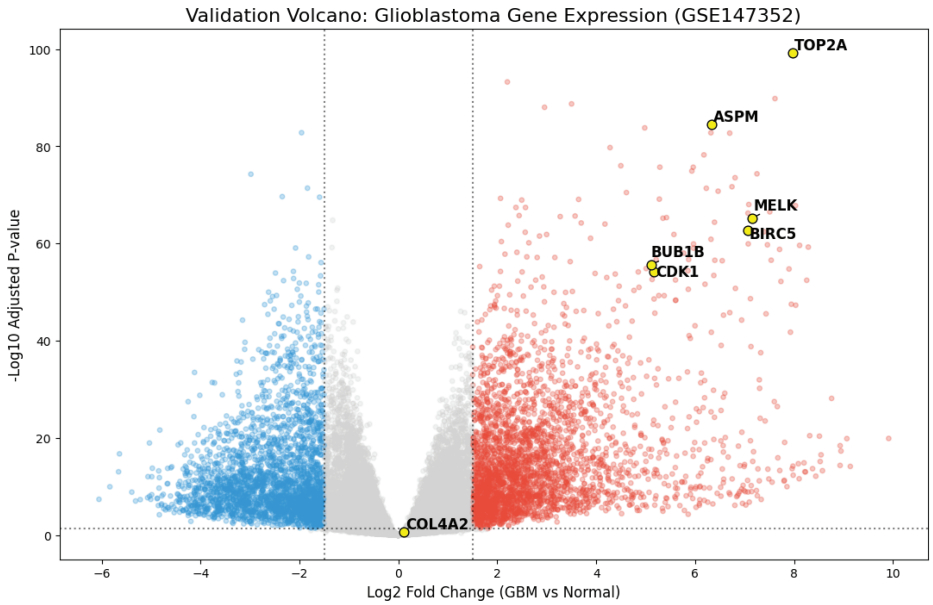

Figure 3. Volcano plot of DEGs, validation phase.

Research

Table 2. KM survival analysis results summary.

Gene	*p*-value	HR
TOP2A	0.33567	0.83
MELK	0.71685	1.07
BIRC5	0.83425	0.96
ASPM	0.75234	0.94

p-values & HR

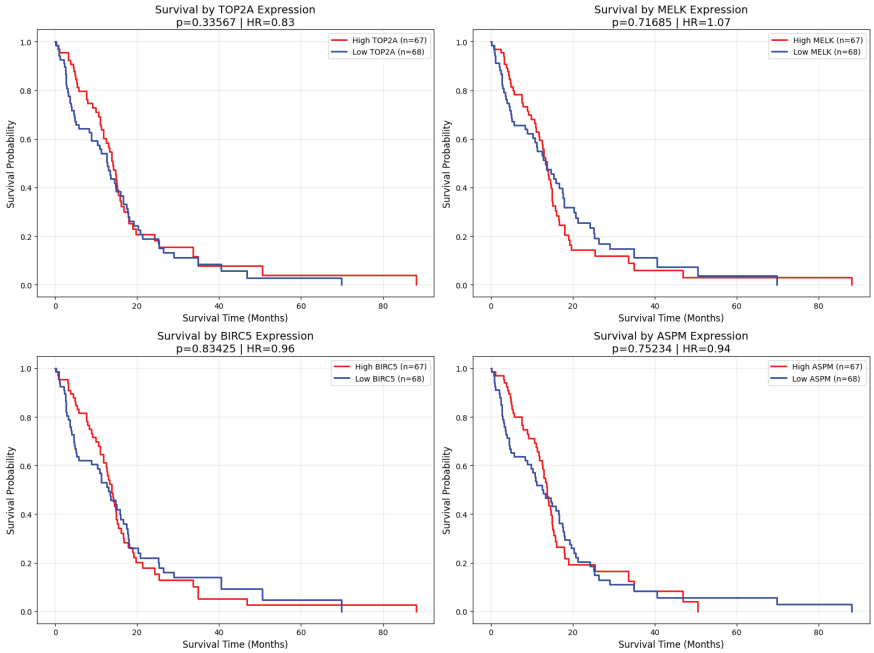

Figure 4. KM plots, across gene signature at high and low levels of expression.

Conclusion and future directions

In summary, while TOP2A, MELK, BIRC5 and ASPM are significantly upregulated in GBM and serve as robust markers of increased mitotic activity across technological platforms, they do not function as independent predictors of OS in our TCGA analysis. This finding contrasts with previous reports suggesting prognostic value for these genes [3,4,6], possibly reflecting differences in patient cohorts, treatment protocols or the limitations of mRNA-level measurements in predicting protein activity and clinical outcomes. Our results highlight the clinical challenge of GBM, where aggressive standard-of-care treatments and complex underlying molecular drivers may overshadow the prognostic impact of individual cell-cycle genes.

This study demonstrates several important limitations. First, our survival analysis focused solely on OS without stratification by TCGA molecular subtypes (Proneural, Classical, Mesenchymal and Neural), which may mask subtype-specific prognostic effects. Second, mRNA expression may not fully reflect protein-level activity or post-translational modifications that directly influence patient outcomes. Third, the standardised intensive treatment protocols in the TCGA cohort may have neutralised any survival differences that would otherwise be observable in untreated or minimally treated populations.

Future research should investigate these targets not as solo biomarkers, but as part of multi-gene risk scores, in the context of specific molecular subtypes, or at the protein level using immunohistochemistry. Additionally, while these genes may lack independent prognostic value, they remain biologically relevant as potential therapeutic targets given their consistent overexpression and central role in tumour cell proliferation.

Acknowledgments

The author expresses his sincere gratitude to the editors and anonymous reviewers for their invaluable comments and constructive feedback, which significantly contributed to the enhancement of this paper. The author also acknowledges the use of Large Language Models (LLMs) for assistance with code generation and writing review; however, the intellectual endeavour, analysis and conclusions are solely that of the author.

Conflicts of interest

The author declares no conflicts of interest.

Funding

The author received no financial support for the research and authorship of this article, though publication costs were generously covered by ecancermedicalscience.

References

1. American Cancer Society (2024) *Cancer Facts & Figures 2024* (American Cancer Society) [https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2024.html]
2. Omuro A and Deangelis LM (2013) **Glioblastoma and other malignant gliomas: a clinical review** *JAMA* **310**(17) 1842–1850 https://doi.org/10.1001/jama.2013.280319 PMID: 24193082
3. Zhou T, Wang Y, and Qian D, *et al* (2018) **Over-expression of TOP2A as a prognostic biomarker in patients with glioma** *Int J Clin Exp Pathol* **11**(3) 1228–1237 [https://pmc.ncbi.nlm.nih.gov/articles/PMC6958105/]
4. Chen X, Huang L, and Yang Y, *et al* (2020) **ASPM promotes glioblastoma growth by regulating G1 restriction point progression and Wnt-β-catenin signaling** *Aging (Albany NY)* **12**(1) 224–241 https://doi.org/10.18632/aging.102612 PMID: 31905171 PMCID: 6977704
5. Horvath S, Zhang B, and Carlson M, *et al* (2006) **Analysis of oncogenic signaling networks in glioblastoma may identify candidate therapeutic targets** *Proc Nat Acad Sci (PNAS)* **103**(46) 17402–17407 https://doi.org/10.1073/pnas.0608396103 PMID: 17090670 PMCID: 1635024
6. Yang H, Zhou H, and Wang G, *et al* (2022) **MELK is a prognostic biomarker and correlated with immune infiltration in glioma** *Front Neurol* **13** 977180 https://doi.org/10.3389/fneur.2022.977180 PMID: 36353126 PMCID: 9637824
7. Fäldt Beding A, Larsson P, and Helou K, *et al* (2022) **Pan-cancer analysis identifies BIRC5 as a prognostic biomarker** *BMC Cancer* **22** 322 https://doi.org/10.1186/s12885-022-09371-0 PMID: 35331169 PMCID: 8953143
8. Sun L, Hui AM, and Su Q, *et al* (2006) **Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain** *Cancer Cell* **9**(4) 287–300 https://doi.org/10.1016/j.ccr.2006.03.003 PMID: 16616334
9. Huang T, Yang Y, and Song X, *et al* (2021) **PRMT6 methylation of RCC1 regulates mitosis, tumorigenicity, and radiation response of glioblastoma stem cells** *Mol Cell* **81**(6) 1276–1291 https://doi.org/10.1016/j.molcel.2021.01.015 PMID: 33539787 PMCID: 7979509