A Glioma Stem Cell-Associated Signature Predicts Survival Across Adult and Pediatric High-Grade Gliomas and Reveals the FAM86B1/FAM86B2 Axis

Qiqi Xie [1], Bokai Wang [1, 2], Qin Ma [3, 4], Jia Shen [1, 5, 6, *]

[1] Medical Sciences Program, Indiana University School of Medicine, Bloomington, IN 47405, USA

[2] Cell, Molecular and Cancer Biology Graduate Program, Indiana University School of Medicine, Bloomington, IN 47405, USA

[3] Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center-The James, Columbus, OH 43210, USA

[4] Department of Biomedical Informatics College of Medicine, The Ohio State University, Columbus, OH 43210, USA

[5] Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN 46202, USA

[6] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[*] Corresponding author. E-mail address: shen17@iu.edu (J. Shen)

Abstract

High-grade gliomas (HGGs), including adult glioblastoma (GBM) and pediatric diffuse intrinsic pontine gliomas (DIPGs), are sustained by glioma stem cells (GSCs) that drive tumor initiation, therapeutic resistance, and recurrence. Although numerous prognostic models have been proposed, few are directly grounded in the core biology of GSCs across both adult and pediatric HGGs. In this study, we defined a GSC-associated gene signature by integrating transcriptomic profiles from patient-derived GSCs and their differentiated counterparts using in-house DIPG13 RNA-seq and the public GSE54791 dataset. The biological relevance of this signature was supported by functional enrichment and protein-protein interaction analyses. To assess its prognostic value, we applied machine learning-based modeling in a large training cohort (Chinese Glioma Genome Atlas, CGGA) and validated the resulting model across three independent datasets (Gravendeel, Rembrandt, and an integrated pediatric HGG cohort), demonstrating consistent predictive performance. To enhance clinical applicability, we developed a nomogram integrating the gene signature-derived risk score with key clinical factors, including age, race, and radiation therapy status, enabling individualized survival prediction. To further support the biological basis of the model, we experimentally examined *FAM86B1*, one of the five genes in the final signature and a gene not previously characterized in glioma biology, and found that the closely related FAM86B1/FAM86B2 axis was enriched in GSCs and overexpressed in GBM tissues, while suppression of this axis impaired GSC maintenance in both GBM and DIPG GSCs. Collectively, this study establishes a biologically grounded, GSC-centered prognostic model for HGG that improves patient stratification and may inform personalized therapeutic strategies.

Key words: HGG, GBM, DIPG, GSC, machine learning, prognostic model, FAM86B1/FAM86B2

62 # 1. Introduction

63 High-grade gliomas (HGGs) represent a devastating group of primary brain tumors that arise in both adult and pediatric
64 populations (1, 2). Classified by the World Health Organization (WHO) as grade 3 or 4 neoplasms, they include glioblastoma
65 (GBM) in adults and pediatric HGGs (pHGGs), with diffuse intrinsic pontine gliomas (DIPGs) representing the most lethal
66 subtype. For patients with GBM, median survival rarely exceeds 15 months (3), while children with DIPG, an inoperable
67 brainstem tumor typically diagnosed between the ages of 5 and 10, face a median survival of less than one year (4). The
68 uniformly poor outcomes associated with these malignancies underscore the urgent need for improved prognostic tools and
69 therapeutic strategies.

70 Glioma stem cells (GSCs) are a major driver of therapeutic resistance and recurrence in both adult and pediatric HGGs
71 (5, 6). These cells possess robust self-renewal capacity, initiate and sustain tumor growth, and exhibit resistance to
72 conventional therapies. Controlled by unique transcriptional programs, GSCs are molecularly and functionally distinct from
73 the bulk tumor population (7-9). Despite their recognized biological importance, however, the prognostic relevance of GSC-
74 associated molecular features remains insufficiently defined. Most existing prognostic models are constructed from bulk
75 tumor transcriptomes, where signals from the small yet clinically consequential GSC population are frequently masked (10-
76 12). Furthermore, models developed for specific subtypes such as GBM often fail to achieve robustness and generalizability
77 when applied across diverse HGG cohorts, particularly in extending from adult to pediatric cases, limiting their broader
78 clinical applicability.

79 To address these limitations, we developed a prognostic model grounded in a transcriptional signature that defines the GSC
80 state across adult and pediatric HGGs. Built from in-house and public datasets and validated in four independent cohorts,
81 the machine learning-derived model consistently outperformed existing signatures. Integrated with clinical variables into a
82 nomogram, it provides a robust tool for individualized survival prediction and underscores the central role of cancer
83 stemness in HGG outcomes, further supported by experimental validation of the FAM86B1/FAM86B2 axis in GSC
84 maintenance.

85

86 # 2. Results

87 **2.1. Identification of a core set of GSC–associated genes in HGGs**

88 To define a robust set of GSC-associated genes in HGGs, we integrated GSC transcriptomic data from both in-house
89 and public datasets (Fig. 1A). RNA-seq was performed on the pediatric HGG cell line DIPG13 cultured under stem cell–
90 maintaining versus differentiation conditions (Fig. 1B), and these data were analyzed together with an independent GBM
91 dataset (GSE54791) containing matched GSC and differentiated DGC pairs. Differential expression analysis identified
92 genes consistently upregulated in GSCs across both datasets, and their intersection yielded a core set of 603 GSC-associated
93 genes (Supplementary Table 1; Fig. 2A). Functional relevance of these genes was assessed through PPI network analysis,
94 which highlighted known stemness regulators such as SOX2 and OLIG1/2 as top hub genes (Fig. 2B and C), supporting the
95 validity of the approach. GO (Fig. 2D) and KEGG pathway analyses (Fig. 2E) further demonstrated that these genes are
96 enriched in key biological processes related to GSC biology, including neurogenesis, RNA splicing, and cancer signaling.

97 **2.2. Prognostic value of the GSC-associated gene signature**

98 To evaluate the prognostic value of the GSC-associated gene signature, we implemented a large-scale modeling strategy
99 that generated 135 machine learning pipelines within the CGGA training cohort (Fig. 3A). Each pipeline combined one of
100 ten feature selection methods with several survival learning algorithms, and model performance was assessed using the C-
101 index across the CGGA training set and three independent validation cohorts (Gravendeel, Rembrandt, and the integrated
102 pediatric HGG cohort). Among all models, the combination of forward stepwise Cox regression (StepCox[forward]) and
103 RSF achieved the highest and most consistent C-index across datasets. Using this model, patients were stratified into high-
104 and low-risk groups based on the median risk score derived from the training cohort. Kaplan–Meier survival analyses
105 showed that high-risk patients had significantly shorter overall survival than those in the low-risk group, with hazard ratios
106 ranging from 2.36 in the Gravendeel cohort (95% CI: 1.77–3.13, $p < 0.001$) to 4.27 in the CGGA cohort (95% CI: 3.40–
107 5.58, $p < 0.001$) (Fig. 3B). Time-dependent ROC analyses further confirmed the predictive capacity of the model, with AUC
108 values exceeding 0.80 in the CGGA cohort and above 0.70 in the Gravendeel, Rembrandt, and pediatric HGG cohorts (Fig.
109 3C and 4).

110 To further validate its robustness, the GSC-based prognostic model was compared with multiple previously published
111 glioma signatures across all four datasets. Using both AUC and C-index as evaluation metrics, our model consistently
112 ranked among the top performers, achieving the highest predictive accuracy in the CGGA cohort and comparable or superior
113 performance in the Gravendeel and pediatric HGG cohorts (Fig. 5 and 6). Time-dependent ROC analyses across all datasets
114 confirmed high discrimination for 1-, 3-, and 5-year overall survival predictions (Fig. 7A–C). Univariate Cox regression
115 analysis also demonstrated that the GSC-derived risk score (GSRScore) was a significant independent predictor of mortality

116 in CGGA (HR = 40.10, 95% CI: 26.67–60.27, $p < 0.001$), Gravendeel (HR = 2.65, 95% CI: 2.00–3.51, $p < 0.001$),
117 Rembrandt (HR = 2.14, 95% CI: 1.70–2.68, $p < 0.001$), and the pediatric HGG cohort (HR = 1.40, 95% CI: 1.36–1.76, $p <$
118 0.001) (Fig. 7D). A pooled meta-analysis yielded a summary HR of 3.08 (95% CI: 0.48–19.89) under the random-effects
119 model and 2.91 (95% CI: 2.50–3.40) under the fixed-effects model (Fig. 7D). Collectively, these results demonstrate that
120 the GSC-associated gene signature provides robust and consistent prognostic value across both adult and pediatric HGG
121 datasets.

122 **2.3. Integrated nomogram for prognostic prediction**

123 To generate a clinically practical tool for individualized survival prediction, we first constructed a molecular signature
124 through a systematic multi-step feature selection process. Using univariate Cox regression, iterative LASSO Cox analysis
125 repeated 1,000 times, and forward stepwise selection, we identified five genes (*FAM86B1*, *IGFBP2*, *CA14*, *SCARA3*, and
126 *TSPAN13*) with the most stable and significant prognostic value. The time-dependent ROC curve analysis revealed that the
127 predictive performance of the model peaked when these five genes were included, achieving a 5-year AUC of 0.97 (Fig.
128 8A). Based on this model, a GSRScore was calculated for each patient. The prognostic accuracy of the GSRScore was
129 further validated by ROC analysis, yielding an AUC of 0.973 (Fig. 8B). Patients were then stratified into high- and low-risk
130 groups according to the median GSRScore, and Kaplan–Meier survival analysis confirmed a highly significant difference
131 in overall survival between the two groups (log-rank $p < 0.001$; Fig. 8C), with high-risk patients exhibiting markedly poorer
132 outcomes.

133 Building on this refined 5-gene signature, we next developed an integrated prognostic nomogram that combined the
134 GSRScore with established clinical variables, including patient age, sex, race, and radiation therapy status (Fig. 8D). Each
135 predictor was assigned a weighted point value proportional to its contribution to prognosis, and the total point score for an
136 individual patient corresponded to predicted 1-, 2-, and 3-year overall survival probabilities. Collectively, this integrated
137 nomogram provides a quantitative and user-friendly tool that improves prognostic accuracy and supports personalized risk
138 assessment in patients with high-grade glioma.

139 **2.4. FAM86B1-associated paralog axis supports glioma stemness**

140 Among the five genes comprising the final prognostic signature incorporated into the nomogram, *FAM86B1* emerged
141 as a particularly interesting candidate because of its limited prior characterization in glioma biology. Notably, FAM86B1
142 and FAM86B2 are closely related paralogs within the FAM86 family (13), a group of predicted methyltransferase-like

143 proteins that share substantial sequence homology. Compared with the more full-length FAM86 family members, *FAM86B1*
144 encodes a truncated FAM86 domain due to loss of a single exon, whereas *FAM86B2* more closely resembles the canonical
145 family structure (Fig. 9A). Consistent with this relatedness, analysis of the public GBM dataset GSE54791 and our in-house
146 DIPG13 RNA-seq dataset showed that FAM86B2, like FAM86B1, tended to be upregulated in GSCs relative to DGCs in
147 most models, with the exception of the GBM MGG4 pair (Fig. 9B). In addition, higher FAM86B2 expression, similar to
148 higher FAM86B1 expression, was associated with poorer GBM prognosis (Fig. 9C). Because the high sequence similarity
149 between FAM86B1 and FAM86B2 complicates their selective detection and perturbation, we next examined the
150 FAM86B1/2 axis across multiple patient-derived GSC models from both GBM and DIPG. Gene expression analysis
151 demonstrated that FAM86B1/2 expression was elevated in GSCs relative to DGCs across GBM GSC lines as well as DIPG
152 GSC models (Fig. 9D), supporting an association between this paralogous axis and the stem-like glioma state. To further
153 assess clinical relevance, we examined FAM86B1/2 protein expression in human tissues by immunohistochemistry and
154 found that GBM tumor specimens displayed higher expression than non-neoplastic tissues (Fig. 9E).

155 We then performed functional knockdown studies using two independent MISSION shRNAs (MilliporeSigma)
156 originally designed against FAM86B1. However, BLAST analysis revealed that both shRNAs also target FAM86B2,
157 indicating likely co-suppression of both transcripts. In GBM and DIPG GSC models, shRNA-mediated suppression of the
158 FAM86B1/2 axis reduced neurosphere formation capacity (Fig. 9F), decreased expression of the stemness regulators OLIG2
159 and SOX2 (Fig. 9G), and reduced self-renewal potential in limiting dilution assays (Fig. 9H), together indicating that the
160 FAM86B1/2 paralogous axis contributes to maintenance of core GSC stemness programs. Collectively, these results provide
161 experimental support for the biological relevance of a key gene family represented within our prognostic signature and
162 strengthen the link between the GSC-associated transcriptomic program and functional stemness phenotypes in high-grade
163 glioma.

164

165 ## 3. Discussion

166 By contrasting patient-derived GSCs with their differentiated counterparts from both adult and pediatric HGGs, we
167 defined a coherent set of 603 stemness-associated genes enriched in neurogenesis, RNA splicing, and oncogenic signaling
168 pathways, underscoring their direct relevance to GSC biology. Network analysis highlighted SOX2, OLIG1, and OLIG2 as
169 central regulators, consistent with their established roles in neural stem/progenitor cell identity and glioma plasticity (8, 14,

170 15). Accordingly, the risk score generated by our StepCox[forward] + RSF model likely quantifies the activity of this
171 stemness program, with high-risk tumors reflecting a stronger GSC phenotype and poorer prognosis.

172 A key innovation of this study is the translation of this gene signature into a clinically usable nomogram. Although
173 molecular risk scores are powerful, their direct adoption in practice is challenging. By integrating the GSRScore with clinical
174 factors such as age, sex, race, and radiation therapy status, the nomogram provides an individualized prediction of 1-, 2-,
175 and 3-year survival probabilities. This quantitative tool enables clinicians to move beyond population-level estimates and
176 engage in more precise, personalized discussions with patients, supporting both clinical decision-making and risk
177 stratification for future trials.

178 Beyond prognosis, our findings carry therapeutic implications. The GSC state is closely linked to resistance against
179 standard-of-care treatments such as temozolomide and radiation, largely through enhanced DNA damage repair (DDR) (7,
180 16). It is plausible that patients classified as "high-risk" by our model represent those whose tumors are most refractory to
181 conventional therapy. This opens opportunities for pairing the prognostic tool with therapeutic strategies. For instance,
182 enrolling high-risk patients into clinical trials evaluating GSC-targeting agents or DDR inhibitors. In this way, the model
183 could evolve into a predictive biomarker guiding precision therapy.

184 To further support the biological significance of the prognostic signature, we focused on *FAM86B1*, one of the five
185 genes in the final model. Our experimental data suggest that FAM86B1 and its closely related paralog FAM86B2 are linked
186 to the GSC state in both GBM and DIPG. Because the two shRNAs initially selected to target FAM86B1 were also found
187 to target FAM86B2, the functional phenotypes observed here are best interpreted as resulting from suppression of the
188 FAM86B1/2 axis rather than FAM86B1 alone. This interpretation is strengthened by the finding that FAM86B2, like
189 FAM86B1, was upregulated in GSCs and associated with poor prognosis. The observed reduction in neurosphere formation
190 together with decreased OLIG2 and SOX2 expression following FAM86B1/2 suppression supports a role for this paralogous
191 axis in maintaining glioma stemness. More broadly, the fact that the other genes in the signature, including *IGFBP2* (17),
192 *CA14* (18), *SCARA3* (19), and *TSPAN13* (20), have also been linked to poor prognosis in GBM further strengthens the
193 biological and clinical relevance of the model.

194 Nevertheless, several limitations warrant consideration. First, our analyses were retrospective, and prospective, multi-
195 center clinical trials will be essential to confirm real-world performance. Second, although the present study provides initial
196 functional validation of the FAM86B1/FAM86B2 axis, the individual contributions of FAM86B1 and FAM86B2 remain
197 to be fully resolved because of their high sequence similarity. Finally, tissue availability remains a barrier for HGG

198 diagnostics. Future efforts may explore liquid biopsy approaches, such as circulating tumor DNA or extracellular vesicles,
199 to detect the GSC signature non-invasively, which would greatly enhance clinical feasibility.

200 In conclusion, our integration of GSC transcriptomic profiling with machine learning, multi-cohort validation, and
201 experimental interrogation of the FAM86B1/FAM86B2 axis produced a robust prognostic model for HGGs. By distilling
202 complex biology into a clinically usable nomogram, we offer a practical tool for individualized survival prediction while
203 reinforcing glioma stemness as a key driver of disease behavior and a promising target for therapeutic intervention.

204

205 # 4. Materials and methods

206 **4.1. Cell culture**

207 DIPG cells (DIPG13 and BT245) were maintained in tumor stem cell medium consisting of Neurobasal-A and
208 DMEM/F-12 mixed 1:1 and supplemented with 10 mmol/L HEPES, sodium pyruvate, nonessential amino acids,
209 GlutaMAX-I, and B27 (all from Thermo Fisher Scientific). Growth factors included human FGF (20 ng/mL), human EGF
210 (20 ng/mL), and PDGF-AA and PDGF-BB (10 ng/mL each; FUJIFILM Biosciences), along with heparin (10 ng/mL;
211 Stemcell Technologies). Under these conditions, DIPG cells grew as neurospheres, consistent with GSC properties. Patient-
212 derived GBM GSCs, including GSC3565 and GSC23, were cultured according to established protocols (8). For
213 differentiation, GSCs were maintained in DMEM/F-12 supplemented with 10% fetal bovine serum (FBS) for at least 14
214 days, resulting in DGCs that grew adherently. Lenti-X 293T cells (Takara Bio, cat# 632180) were cultured in DMEM (Gibco,
215 cat# 11995065) supplemented with 10% FBS (Corning, cat# 35-015-CV).

216 **4.2. Lentiviral-mediated gene knockdown**

217 Lenti-X 293T cells were transfected using jetPRIME (Polyplus, cat# 101000001) with the packaging plasmids pMD2.G
218 and psPAX2, together with either control shRNA or FAM86B1/2-targeting shRNAs (#1, TRCN0000131177; #2,
219 TRCN0000130107) (Sigma-Aldrich). After 24 hours, the medium was replaced with GSC culture medium, and cells were
220 incubated for an additional 48 hours before viral supernatants were collected and stored at −80°C. GBM and DIPG GSCs
221 were infected with lentiviral supernatant for 18 hours, followed by medium replacement and incubation for an additional
222 48 hours before downstream analyses.

223 **4.3. Tumorsphere assay and extreme limiting dilution assay (ELDA)**

224 For the tumorsphere assay, GSCs expressing control or target shRNAs were dissociated into single cells and plated at

225 1 × 10^5 cells per well in 6-well plates under stem cell culture conditions. After 72 hours, tumorspheres were imaged using
226 an Eclipse Ts2-FL microscope (Nikon) and quantified. To assess self-renewal frequency, the ELDA was performed. GSCs
227 transduced with control or target shRNAs were dissociated into single cells and seeded into 96-well plates at serially diluted
228 cell densities in triplicate under stem cell culture conditions. After 10 days, wells containing tumorspheres were scored, and
229 stem cell frequency was estimated using the online ELDA tool (https://bioinf.wehi.edu.au/software/elda/).

230 **4.4. Quantitative PCR (qPCR)**

231 Total RNA was extracted using the RNeasy Plus Mini Kit (QIAGEN, cat# 74136) and reverse-transcribed using qScript
232 cDNA SuperMix (Quantabio, cat# 95048). qPCR was performed in triplicate using PowerUp SYBR Green Master Mix
233 (Applied Biosystems, cat# 25742) under the following cycling conditions: 95°C for 30 seconds, followed by 40 cycles of
234 95°C for 3 seconds and 60°C for 20 seconds. Primer sequences were as follows: FAM86B1/B2, forward 5′-
235 ACTGTGAGGCATCCTGTGTG-3′ and reverse 5′-TGATGGCTGTGCTCTTGGAG-3′; OLIG2, forward 5′-
236 TGGCTTCAAGTCATCCTCGTC-3′ and reverse 5′-ATGGCGATGTTGAGGTCGTG-3′; SOX2, forward 5′-
237 GCCGAGTGGAAACTTTTGTCG-3′ and reverse 5′-GGCAGCGTGTACTTATCCTTCT-3′; and ACTIN, forward 5′-
238 CATGTACGTTGCTATCCAGGC-3′ and reverse 5′-CTCCTTAATGTCACGCACGAT-3′.

239 **4.5. Immunohistochemistry**

240 Normal brain and GBM tissues (Supplementary Table 2) were obtained from the Enterprise Clinical Research
241 Operations (ECRO) Biorepository at Indiana University Health Methodist Hospital with Institutional Review Board (IRB)
242 approval for the collection and use of human biological materials. Tissues were processed as formalin-fixed, paraffin-
243 embedded (FFPE) specimens, and immunohistochemical staining was performed as previously described (21). The primary
244 antibody used for staining was anti-FAM86B1 (Thermo Fisher Scientific, cat# PA5-63164, 1:50). All slides were imaged
245 using a Motic EasyScan Pro 6 scanner (Motic).

246 **4.6. Data acquisition and preprocessing**

247 To identify genes associated with the GSC phenotype, we analyzed both in-house and publicly available datasets. The
248 in-house dataset was generated from DIPG13 cells, with total RNA extracted from GSC and DGC populations and
249 sequenced by Novogene Corporation. In parallel, a publicly available GBM dataset (GSE54791; (22)) containing matched
250 GSC and DGC pairs was included for comparative analysis.

For model construction and validation, one training cohort and three independent validation cohorts were assembled. The Chinese Glioma Genome Atlas (CGGA) RNA-seq dataset (http://www.cgga.org.cn/) served as the training cohort. Two well-established microarray datasets, Gravendeel (23) and Rembrandt (24), were used as validation cohorts. In addition, a pediatric HGG validation cohort was generated by integrating raw data from five high-quality studies: Hoffman (25), Paugh (26), Schwartzentruber (27), Sturm (28), and Witt (29). All data were uniformly curated and processed. Patients lacking complete survival information were excluded. RNA-seq data were normalized and log-transformed as log2(TPM+1) or log2(FPKM+1). Microarray data were processed from raw .CEL files using the Robust Multi-array Average (RMA) algorithm for background correction, normalization, and summarization. To ensure comparability across platforms, all expression values were standardized using z-scores prior to downstream analyses.

4.7. Identification of GSC–associated genes and pathway enrichment analysis

Differentially expressed genes (DEGs) between GSCs and DGCs were identified in both the in-house DIPG13 and GSE54791 datasets using the *limma* package (30) in R. Genes with |log2 fold change| > 1 and an adjusted p-value < 0.05 were considered significant. The intersection of DEGs from the two datasets yielded a set of 603 GSC–associated genes, which served as the candidate feature pool for model construction. Functional enrichment and network analyses were then performed to characterize these genes. A protein–protein interaction (PPI) network was constructed using STRING (https://string-db.org/) with a confidence score cutoff of 0.4 and visualized in Cytoscape (31). Hub modules were identified using the cytoHubba plugin (32). Gene Ontology (GO) (33) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (34) enrichment analyses were carried out with *clusterProfiler* (35), with terms and pathways considered significant at adjusted $p < 0.05$.

4.8. Machine learning model construction and selection

The HGG GSC–associated genes served as input features for prognostic model development in the CGGA training cohort. A total of 135 modeling pipelines were constructed by pairing 10 feature selection strategies, including LASSO and stepwise Cox regression, with multiple survival learning algorithms such as Random Survival Forest (RSF), CoxBoost, and Supervised Principal Components (SuperPC). Model performance was assessed using the concordance index (C-index), a standard measure of survival prediction accuracy. The model with the highest mean C-index across all four cohorts was selected, and patient-specific risk scores were subsequently calculated based on this model.

4.9. Model validation and performance benchmarking

The final model was validated across all cohorts. Patients were stratified into high- and low-risk groups according to the median risk score derived from the training cohort. Survival differences between groups were assessed using Kaplan–Meier analysis with the log-rank test. Model discrimination was further evaluated using time-dependent Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) values at 1-, 3-, and 5-year survival endpoints. To benchmark performance, the C-index and AUCs of our model were systematically compared against a broad panel of previously published glioma prognostic signatures across all four cohorts.

4.10. Meta-analysis

To evaluate the overall prognostic significance and generalizability of the model-derived risk score, a meta-analysis was performed across the four cohorts. Within each cohort, the risk score as a continuous variable was assessed using univariate Cox proportional hazards regression. Hazard ratios (HRs) with 95% confidence intervals (CIs) were pooled using a random-effects model based on the DerSimonian and Laird method (36), yielding a summary HR. This approach provided a robust estimate of prognostic value while accounting for heterogeneity among cohorts.

4.11. Construction and validation of a prognostic nomogram

A prognostic nomogram was constructed to provide individualized survival prediction by integrating molecular and clinical factors. A multivariable Cox proportional hazards regression model (37) was fitted using the cph function from the *rms* and *survival* packages in R (version 4.4.1), with overall survival time and status as dependent variables and the gene signature risk score (GSRScore), age, sex, race, and radiation therapy status as independent covariates. Based on the regression coefficients, a visual nomogram was generated using the *regplot* package, where each predictor was assigned a point value proportional to its contribution to prognosis. The total point score for each patient corresponded to predicted 1-, 2-, and 3-year survival probabilities. Model performance was assessed by discrimination and calibration, where discrimination was quantified by the C-index (0.5 indicates no discrimination and 1.0 indicates perfect discrimination), and calibration was evaluated by comparing predicted and observed survival probabilities using calibration curves generated from 1,000 bootstrap resamples to reduce overfitting.

4.12. Statistical analysis

All statistical analyses were performed in R. Survival analyses were conducted using the *survival* (https://CRAN.R-project.org/package=survival) and *survminer* (https://CRAN.R-project.org/package=survminer) packages. Time-dependent

304 ROC analyses were carried out with the *timeROC* package (38), and meta-analyses were implemented using the *meta*
305 package (39). A two-sided p value < 0.05 was considered statistically significant.

306 **4.13. Data availability**

307 RNA-seq data for DIPG13 cells cultured under stem cell-maintaining and differentiation conditions will be deposited
308 in the NCBI Gene Expression Omnibus (GEO). All other data supporting the findings of this study are available from the
309 corresponding authors upon reasonable request.

310

311 ## Author contributions

312 QX: Investigation, Visualization, Data curation, writing–original draft. BW: Investigation, Visualization, Data curation,
313 writing–original draft. QM: Writing–review and editing. JS: Conceptualization, Methodology, Funding acquisition,
314 Supervision, Project administration, Writing–original draft, Writing–review and editing.

315

316 ## Acknowledgments

317 This study was supported by the ACS-IRG Grant Mechanism (22-147-37; J. Shen), the Schwarz Family and Friends
318 Cancer Research Fund (J. Shen), the Indiana University Seed Grant (J. Shen), the IUSCCC Early Career Investigator Pilot
319 Fund (J. Shen), and the IUSM LAMP PLUS Scholars Project Fund (J. Shen). The authors thank Dr. Jeremy Rich at the
320 University of North Carolina for providing the GBM GSC lines. The authors also thank Dr. Stefanie Galban and Dr. Sriram
321 Venneti at the University of Michigan, Dr. Q. Richard Lu at the University of Cincinnati, and Dr. Michelle Monje at Stanford
322 Medicine for providing the DIPG cell lines.

323

324

325

326

327

328 ## References

329 1. Poorva, P., Mast, J., Cao, B., Shah, M.V., Pollok, K.E., and Shen, J. 2025. Killing the killers: Natural killer cell
330 therapy targeting glioma stem cells in high-grade glioma. *Mol Ther* 33:2462-2478.
331 2. Sturm, D., Bender, S., Jones, D.T., Lichter, P., Grill, J., Becher, O., Hawkins, C., Majewski, J., Jones, C., Costello,
332 J.F., et al. 2014. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer* 14:92-
333 107.
334 3. Vredenburgh, J.J., Desjardins, A., Herndon, J.E., 2nd, Dowell, J.M., Reardon, D.A., Quinn, J.A., Rich, J.N.,
335 Sathornsumetee, S., Gururangan, S., Wagner, M., et al. 2007. Phase II trial of bevacizumab and irinotecan in
336 recurrent malignant glioma. *Clin Cancer Res* 13:1253-1259.
337 4. Monje, M., Cooney, T., Glod, J., Huang, J., Peer, C.J., Faury, D., Baxter, P., Kramer, K., Lenzen, A., Robison, N.J.,
338 et al. 2023. Phase I trial of panobinostat in children with diffuse intrinsic pontine glioma: A report from the Pediatric
339 Brain Tumor Consortium (PBTC-047). *Neuro Oncol* 25:2262-2272.
340 5. Prager, B.C., Bhargava, S., Mahadev, V., Hubert, C.G., and Rich, J.N. 2020. Glioblastoma Stem Cells: Driving
341 Resilience through Chaos. *Trends Cancer* 6:223-235.
342 6. Suva, M.L., and Tirosh, I. 2020. The Glioma Stem Cell Model in the Era of Single-Cell Genomics. *Cancer Cell*
343 37:630-636.
344 7. Bao, S., Wu, Q., McLendon, R.E., Hao, Y., Shi, Q., Hjelmeland, A.B., Dewhirst, M.W., Bigner, D.D., and Rich,
345 J.N. 2006. Glioma stem cells promote radioresistance by preferential activation of the DNA damage response.
346 *Nature* 444:756-760.
347 8. Li, C., Xie, Q., Ghosh, S., Cao, B., Du, Y., Vo, G.V., Huang, T.Y., Spruck, C., Carpenter, R.L., Wang, Y.A., et al.
348 2025. SUV39H1 maintains cancer stem cell chromatin state and properties in glioblastoma. *JCI Insight* 10.
349 9. Surowiec, R.K., Ferris, S.F., Apfelbaum, A., Espinoza, C., Mehta, R.K., Monchamp, K., Sirihorachai, V.R., Bedi,
350 K., Ljungman, M., and Galban, S. 2021. Transcriptomic Analysis of Diffuse Intrinsic Pontine Glioma (DIPG)
351 Identifies a Targetable ALDH-Positive Subset of Highly Tumorigenic Cancer Stem-like Cells. *Mol Cancer Res*
352 19:223-239.
353 10. Niu, W., Yu, H., Fan, X., Li, S., Sun, S., Gong, M., Zhang, S., Bi, W., Chen, X., and Fang, Z. 2024. Development
354 of stemness-related signature to optimize prognosis prediction and identify XMD8-85 as a novel therapeutic
355 compound for glioma. *Cell Signal* 120:111231.

356 11. Tan, J., Zhu, H., Tang, G., Liu, H., Wanggou, S., Cao, Y., Xin, Z., Zhou, Q., Zhan, C., Wu, Z., et al. 2021. Molecular
357 Subtypes Based on the Stemness Index Predict Prognosis in Glioma Patients. *Front Genet* 12:616507.
358 12. Wang, Z., Wang, Y., Yang, T., Xing, H., Gao, L., Guo, X., Xing, B., and Ma, W. 2021. Machine learning revealed
359 stemness features and a novel stemness-based classification with appealing implications in discriminating the
360 prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Brief Bioinform* 22.
361 13. Francis, J.W., Shao, Z., Narkhede, P., Trinh, A.T., Lu, J., Song, J., and Gozani, O. 2023. The FAM86 domain of
362 FAM86A confers substrate specificity to promote EEF2-Lys525 methylation. *J Biol Chem* 299:104842.
363 14. Ligon, K.L., Alberta, J.A., Kho, A.T., Weiss, J., Kwaan, M.R., Nutt, C.L., Louis, D.N., Stiles, C.D., and Rowitch,
364 D.H. 2004. The oligodendroglial lineage marker OLIG2 is universally expressed in diffuse gliomas. *J Neuropathol*
365 *Exp Neurol* 63:499-509.
366 15. Balasubramaniyan, V., Timmer, N., Kust, B., Boddeke, E., and Copray, S. 2004. Transient expression of Olig1
367 initiates the differentiation of neural stem cells into oligodendrocyte progenitor cells. *Stem Cells* 22:878-882.
368 16. Li, Y., Wang, T., Wan, Q., Wang, Q., Chen, Z., Gao, Y., Ye, Y., Lin, J., Zhao, B., Wang, H., et al. 2022. TRAF4
369 Maintains Deubiquitination of Caveolin-1 to Drive Glioblastoma Stemness and Temozolomide Resistance. *Cancer*
370 *Res* 82:3573-3587.
371 17. Qian, C., Xiufu, W., Jianxun, T., Zihao, C., Wenjie, S., Jingfeng, T., Kahlert, U.D., and Renfei, D. 2022. A Novel
372 Extracellular Matrix Gene-Based Prognostic Model to Predict Overall Survive in Patients With Glioblastoma. *Front*
373 *Genet* 13:851427.
374 18. Qin, G., Pang, G., Wu, S., Bi, S., Lan, S., Tang, X., Hu, B., Zhou, J., Shi, F., and Qin, C. 2024. Construction and
375 validation of a novel prognostic model with palmitoylation-related genes for glioblastoma. *Transl Cancer Res*
376 13:6117-6135.
377 19. Chen, X., Xu, Y., Wang, M., and Ren, C. 2023. Development of Prognostic Indicator Based on AU-Rich Elements-
378 Related Genes in Glioblastoma. *World Neurosurg* 175:e601-e613.
379 20. Wang, H., Liu, Z., Peng, Z., Lv, P., Fu, P., and Jiang, X. 2025. Identification and validation of TSPAN13 as a novel
380 temozolomide resistance-related gene prognostic biomarker in glioblastoma. *PLoS One* 20:e0316552.
381 21. Xie, Q., Du, Y., Ghosh, S., Rajendran, S., Cohen-Gadol, A.A., Baizabal, J.M., Nephew, K.P., Han, L., and Shen, J.
382 2025. Multi-omics analysis identifies glioblastoma dependency on H3K9me3 methyltransferase activity. *NPJ*
383 *Precis Oncol* 9:78.

384 22. Suva, M.L., Rheinbay, E., Gillespie, S.M., Patel, A.P., Wakimoto, H., Rabkin, S.D., Riggi, N., Chi, A.S., Cahill,
385 D.P., Nahed, B.V., et al. 2014. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma
386 stem-like cells. *Cell* 157:580-594.
387 23. Gravendeel, L.A., Kouwenhoven, M.C., Gevaert, O., de Rooi, J.J., Stubbs, A.P., Duijm, J.E., Daemen, A., Bleeker,
388 F.E., Bralten, L.B., Kloosterhof, N.K., et al. 2009. Intrinsic gene expression profiles of gliomas are a better predictor
389 of survival than histology. *Cancer Res* 69:9065-9072.
390 24. Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.C., Fine, H., and Madhavan, S. 2018. The REMBRANDT
391 study, a large collection of genomic data from brain cancer patients. *Sci Data* 5:180158.
392 25. Hoffman, L.M., Veldhuijzen van Zanten, S.E.M., Colditz, N., Baugh, J., Chaney, B., Hoffmann, M., Lane, A.,
393 Fuller, C., Miles, L., Hawkins, C., et al. 2018. Clinical, Radiologic, Pathologic, and Molecular Characteristics of
394 Long-Term Survivors of Diffuse Intrinsic Pontine Glioma (DIPG): A Collaborative Report From the International
395 and European Society for Pediatric Oncology DIPG Registries. *J Clin Oncol* 36:1963-1972.
396 26. Paugh, B.S., Broniscer, A., Qu, C., Miller, C.P., Zhang, J., Tatevossian, R.G., Olson, J.M., Geyer, J.R., Chi, S.N.,
397 da Silva, N.S., et al. 2011. Genome-wide analyses identify recurrent amplifications of receptor tyrosine kinases and
398 cell-cycle regulatory genes in diffuse intrinsic pontine glioma. *J Clin Oncol* 29:3999-4006.
399 27. Schwartzentruber, J., Korshunov, A., Liu, X.Y., Jones, D.T., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A.M.,
400 Quang, D.A., Tonjes, M., et al. 2012. Driver mutations in histone H3.3 and chromatin remodelling genes in
401 paediatric glioblastoma. *Nature* 482:226-231.
402 28. Sturm, D., Capper, D., Andreiuolo, F., Gessi, M., Kolsche, C., Reinhardt, A., Sievers, P., Wefers, A.K., Ebrahimi,
403 A., Suwala, A.K., et al. 2023. Multiomic neuropathology improves diagnostic accuracy in pediatric neuro-oncology.
404 *Nat Med* 29:917-926.
405 29. Mueller, S., Fangusaro, J., Thomas, A.O., Jacques, T.S., Bandopadhayay, P., de Blank, P., Packer, R.J., Fouladi,
406 M., van Meeteren, A.S., Jones, D., et al. 2024. Consensus framework for conducting phase I/II clinical trials for
407 children, adolescents, and young adults with pediatric low-grade glioma: Guidelines established by the International
408 Pediatric Low-Grade Glioma Coalition Clinical Trial Working Group. *Neuro Oncol* 26:407-416.
409 30. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. 2015. limma powers differential
410 expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47.

411 31. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker,
412 T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome*
413 *Res* 13:2498-2504.
414 32. Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T., and Lin, C.Y. 2014. cytoHubba: identifying hub objects
415 and sub-networks from complex interactome. *BMC Syst Biol* 8 Suppl 4:S11.
416 33. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight,
417 S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
418 *Nat Genet* 25:25-29.
419 34. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of
420 Genes and Genomes. *Nucleic Acids Res* 27:29-34.
421 35. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. 2021. clusterProfiler
422 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2:100141.
423 36. DerSimonian, R., and Laird, N. 1986. Meta-analysis in clinical trials. *Control Clin Trials* 7:177-188.
424 37. Balachandran, V.P., Gonen, M., Smith, J.J., and DeMatteo, R.P. 2015. Nomograms in oncology: more than meets
425 the eye. *Lancet Oncol* 16:e173-180.
426 38. Blanche, P., Dartigues, J.F., and Jacqmin-Gadda, H. 2013. Estimating and comparing time-dependent areas under
427 receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 32:5381-5397.
428 39. Balduzzi, S., Rucker, G., and Schwarzer, G. 2019. How to perform a meta-analysis with R: a practical tutorial. *Evid*
429 *Based Ment Health* 22:153-160.
430
431

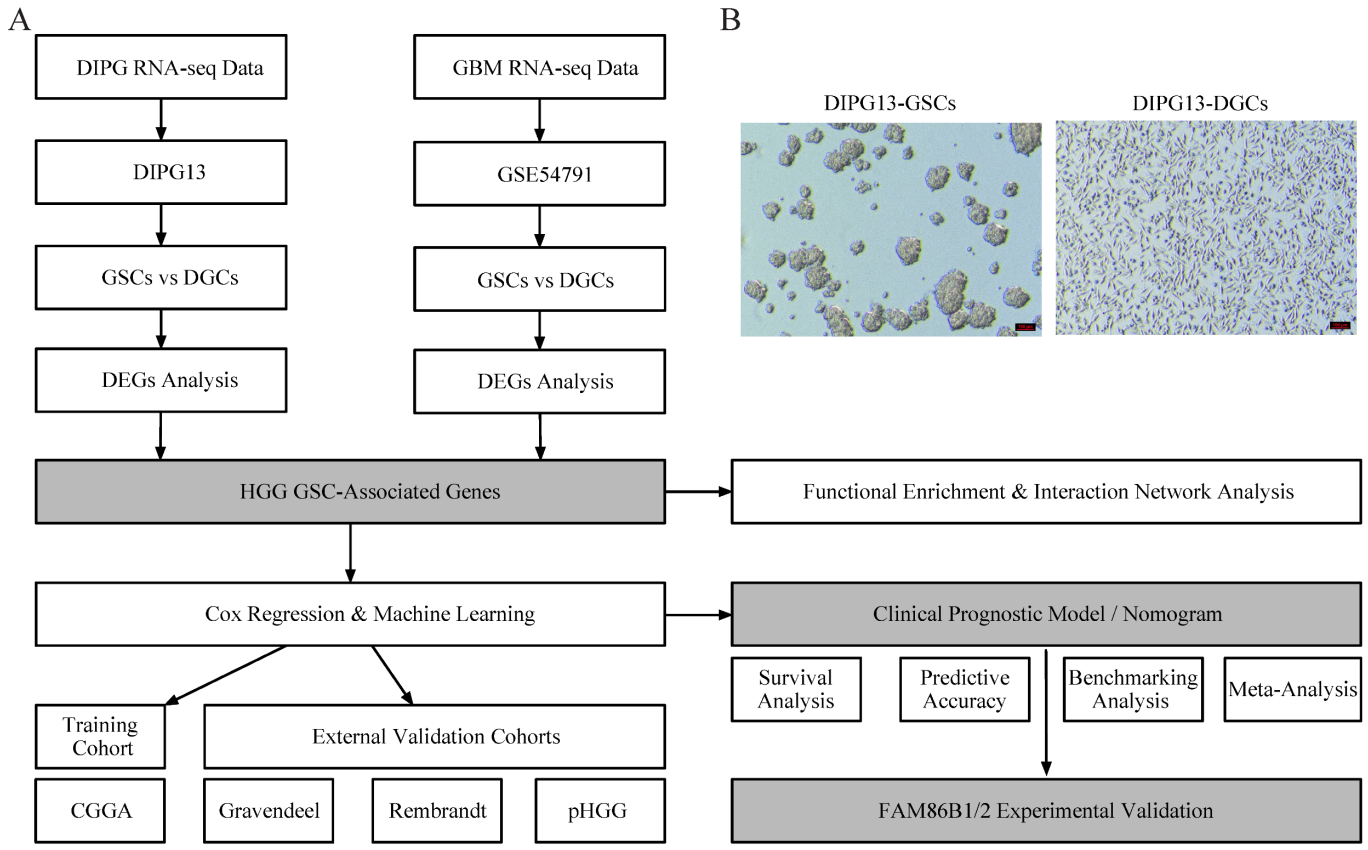


Fig. 1. Schematic overview of the study design. (A) RNA-seq datasets from DIPG13 (in-house, pediatric HGG) and GSE54791 (adult GBM, HGG) were analyzed to identify differentially expressed genes (DEGs) between glioma stem cells (GSCs) and differentiated glioma cells (DGCs). The intersecting DEGs defined a core set of HGG GSC-associated genes, which were subjected to functional enrichment, network analysis, Cox regression, and 135 machine learning algorithms to construct prognostic models. Model performance was trained in the CGGA cohort and validated in three independent cohorts (Gravendeel, Rembrandt, and pHGG). Subsequent survival, predictive accuracy, benchmarking, and meta-analyses established a clinically applicable prognostic model and nomogram. The final model also guided experimental validation of the FAM86B1/2 axis in GSC maintenance. (B) Representative images of DIPG13 cells under stem cell conditions (left) and following serum-induced differentiation (10% FBS, right). DIPG13 cells displayed neurosphere morphology in stem cell medium and adopted an adherent morphology upon differentiation. Scale bars, 100 μm.

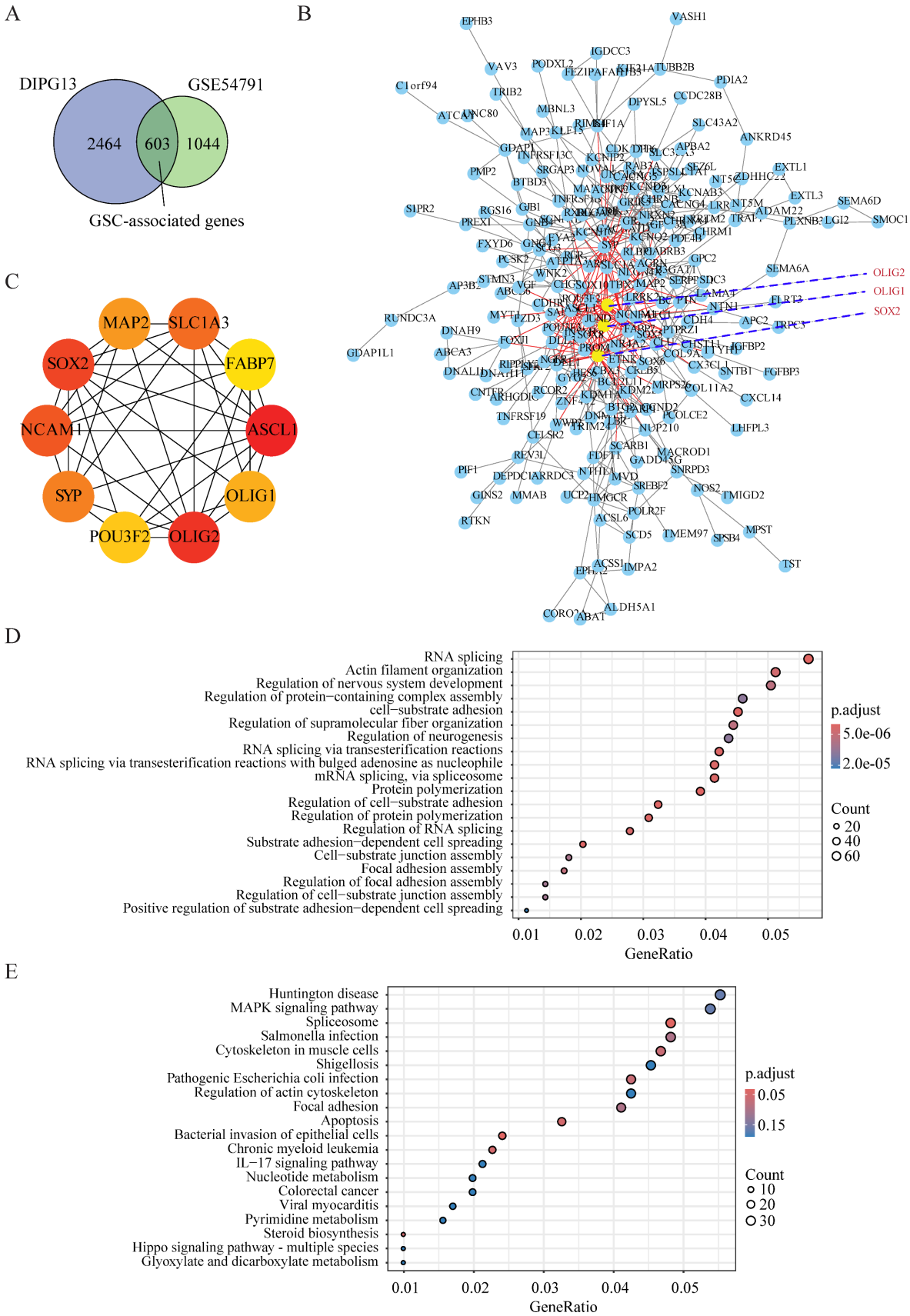


Fig. 2. Functional characterization of HGG GSC-associated genes. (A) Venn diagram showing overlap of DEGs between GSCs and DGCs in DIPG13 and GSE54791 datasets. The 603 shared DEGs were defined as the core HGG GSC-associated genes. (B) PPI network of these 603 genes generated using STRING, with key regulators such as SOX2 and OLIG1/2 highlighted. (C) Top functional module from the PPI network identified by MCODE, highlighting 10 hub genes. (D) GO enrichment analysis for biological processes. (E) KEGG pathway enrichment analysis. In (D) and (E), dot size reflects gene counts and color indicates adjusted p values.

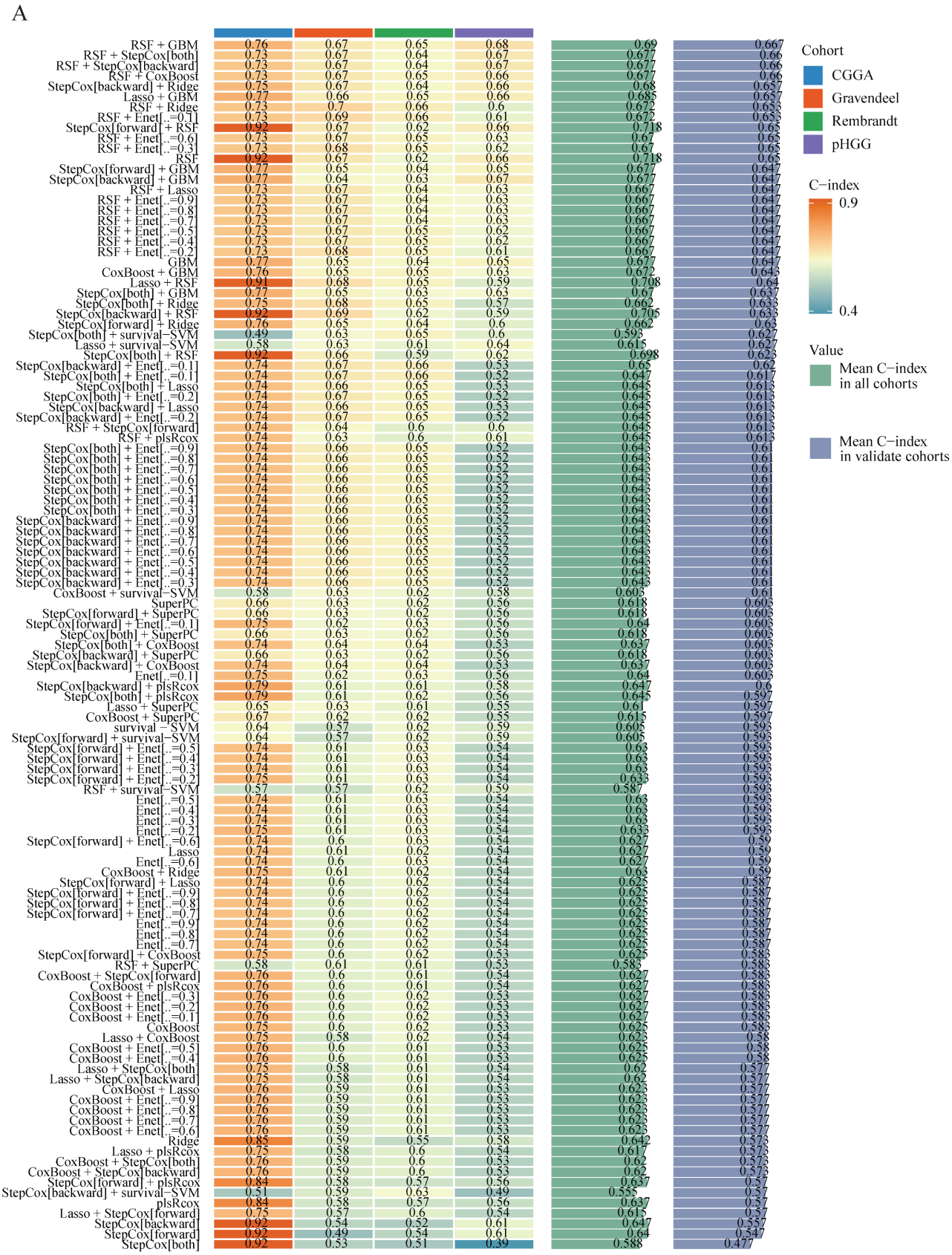
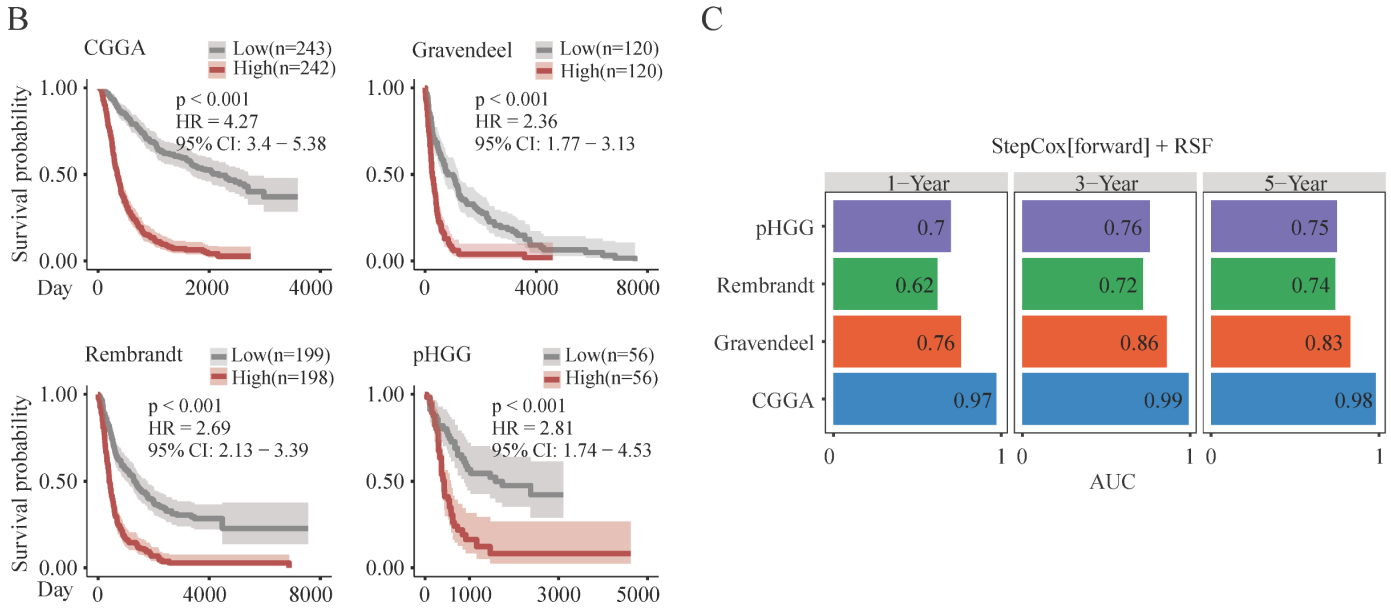


Fig. 3. Development of the optimal machine learning–based prognostic model. (A) Heatmap summarizing the performance of 135 machine learning algorithm combinations across four HGG cohorts (CGGA, Gravendeel, Rembrandt, and pHGG), measured by C-index. Bars on the right indicate mean C-index across all cohorts and within validation cohorts. (B) Kaplan–Meier survival curves for the RSF model, showing significantly worse outcomes in high-risk versus low-risk patients (log-rank $p < 0.001$). Hazard ratios (HR) with 95% confidence intervals (CI) are indicated. (C) Time-dependent AUCs at 1-, 3-, and 5-years for the StepCox[forward] + RSF model.



Continued - Fig. 3. Development of the optimal machine learning–based prognostic model. (A) Heatmap summarizing the performance of 135 machine learning algorithm combinations across four HGG cohorts (CGGA, Gravendeel, Rembrandt, and pHGG), measured by C-index. Bars on the right indicate mean C-index across all cohorts and within validation cohorts. (B) Kaplan–Meier survival curves for the RSF model, showing significantly worse outcomes in high-risk versus low-risk patients (log-rank $p < 0.001$). HR with 95% CI are indicated. (C) Time-dependent AUCs at 1-, 3-, and 5-years for the StepCox[forward] + RSF model.

436

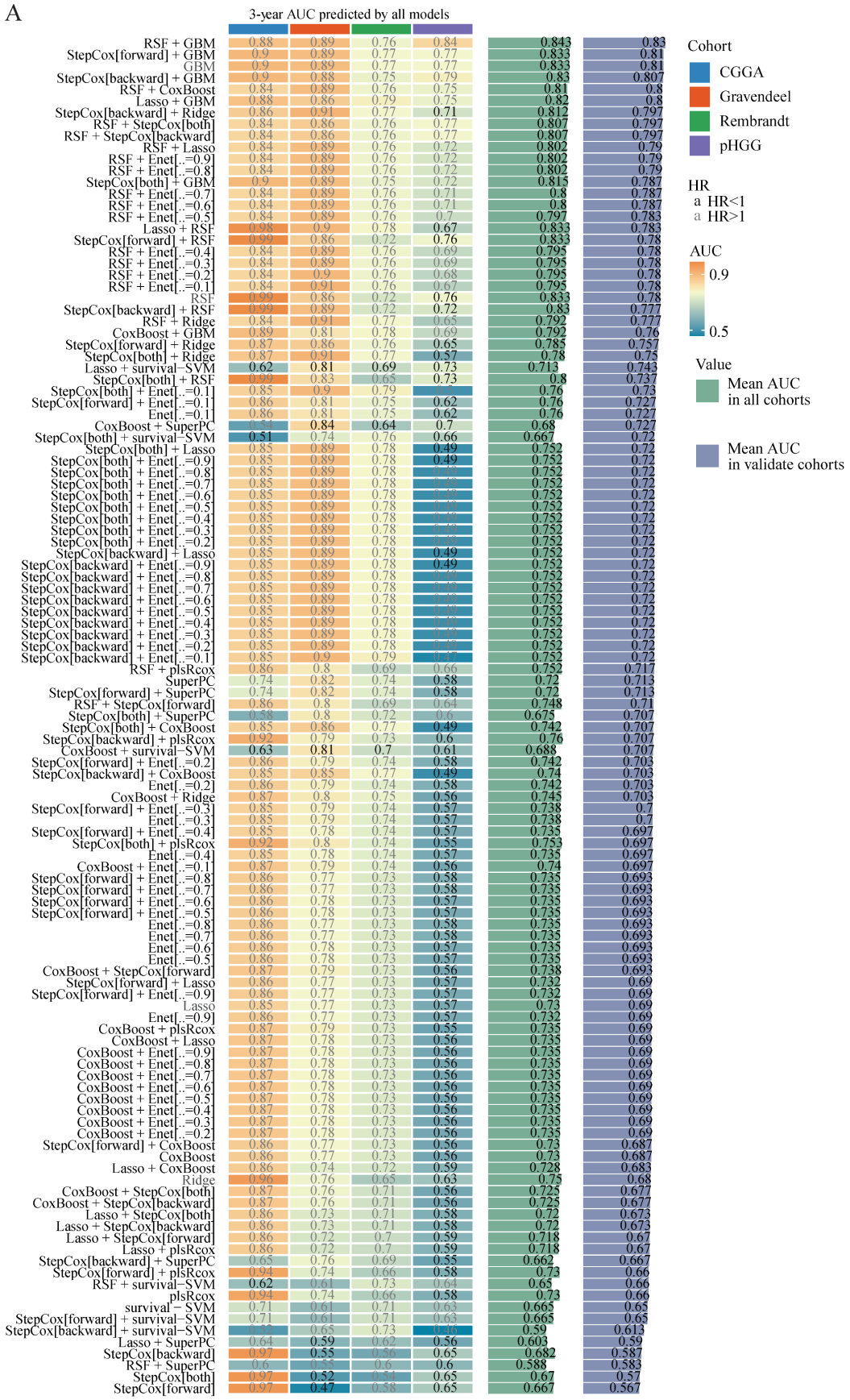
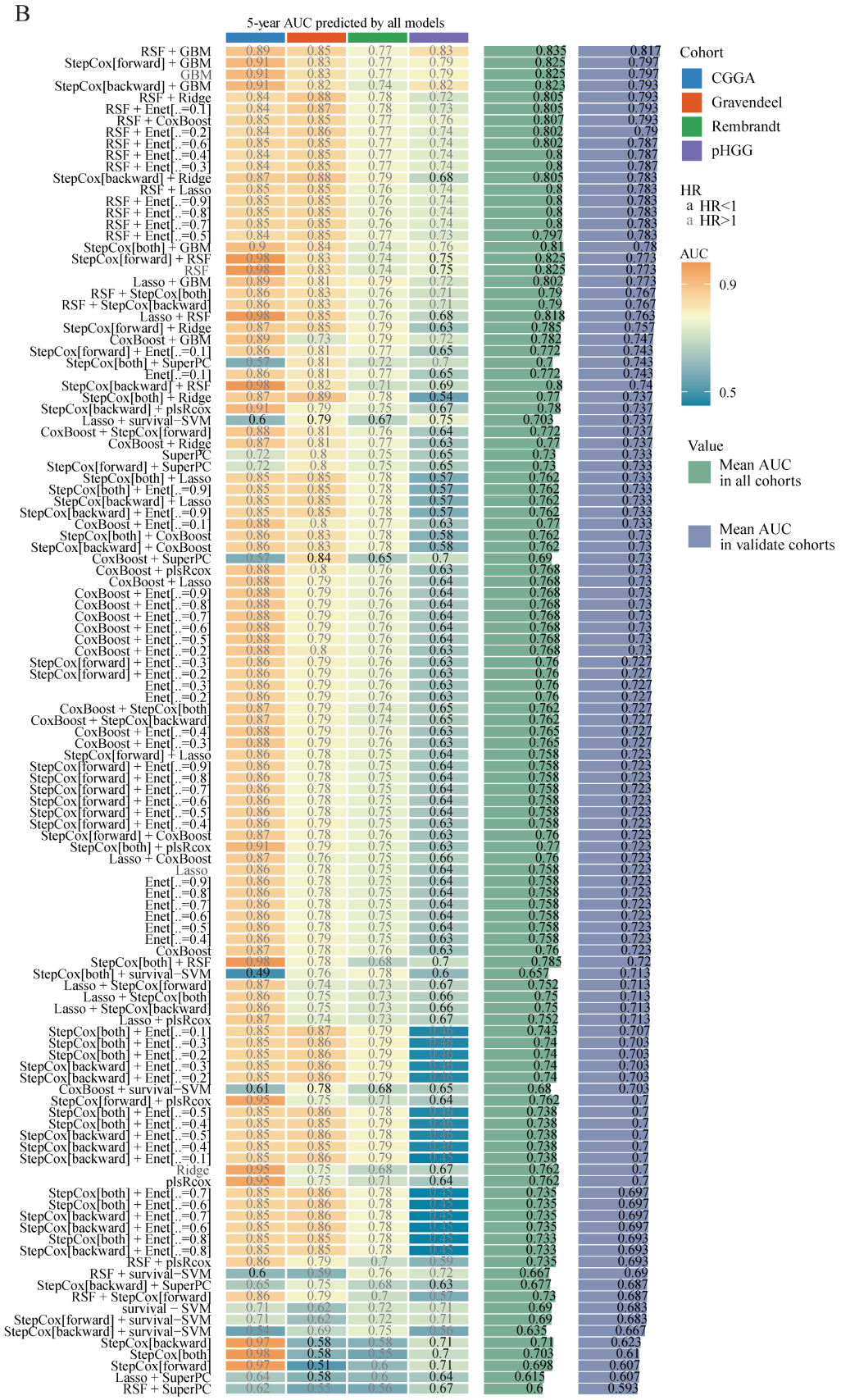


Fig. 4. Comparative evaluation of predictive accuracy across candidate models. (A–B) Heatmaps comparing the predictive performance of all 135 candidate models across four HGG cohorts (CGGA, Gravendeel, Rembrandt, and pHGG). Performance was evaluated using time-dependent AUC at (A) 3 years and (B) 5 years. Each heatmap column represents an individual model, displaying its corresponding HR, cohort-specific AUC values, and mean AUC across cohorts. Right-side bars summarize the mean AUC values within all cohorts and specifically within the validation cohorts.



Continued - Fig. 4. Comparative evaluation of predictive accuracy across candidate models. (A–B) Heatmaps comparing the predictive performance of all 135 candidate models across four HGG cohorts (CGGA, Gravendeel, Rembrandt, and pHGG). Performance was evaluated using time-dependent AUC at (A) 3 years and (B) 5 years. Each heatmap column represents an individual model, displaying its corresponding HR, cohort-specific AUC values, and mean AUC across cohorts. Right-side bars summarize the mean AUC values within all cohorts and specifically within the validation cohorts.

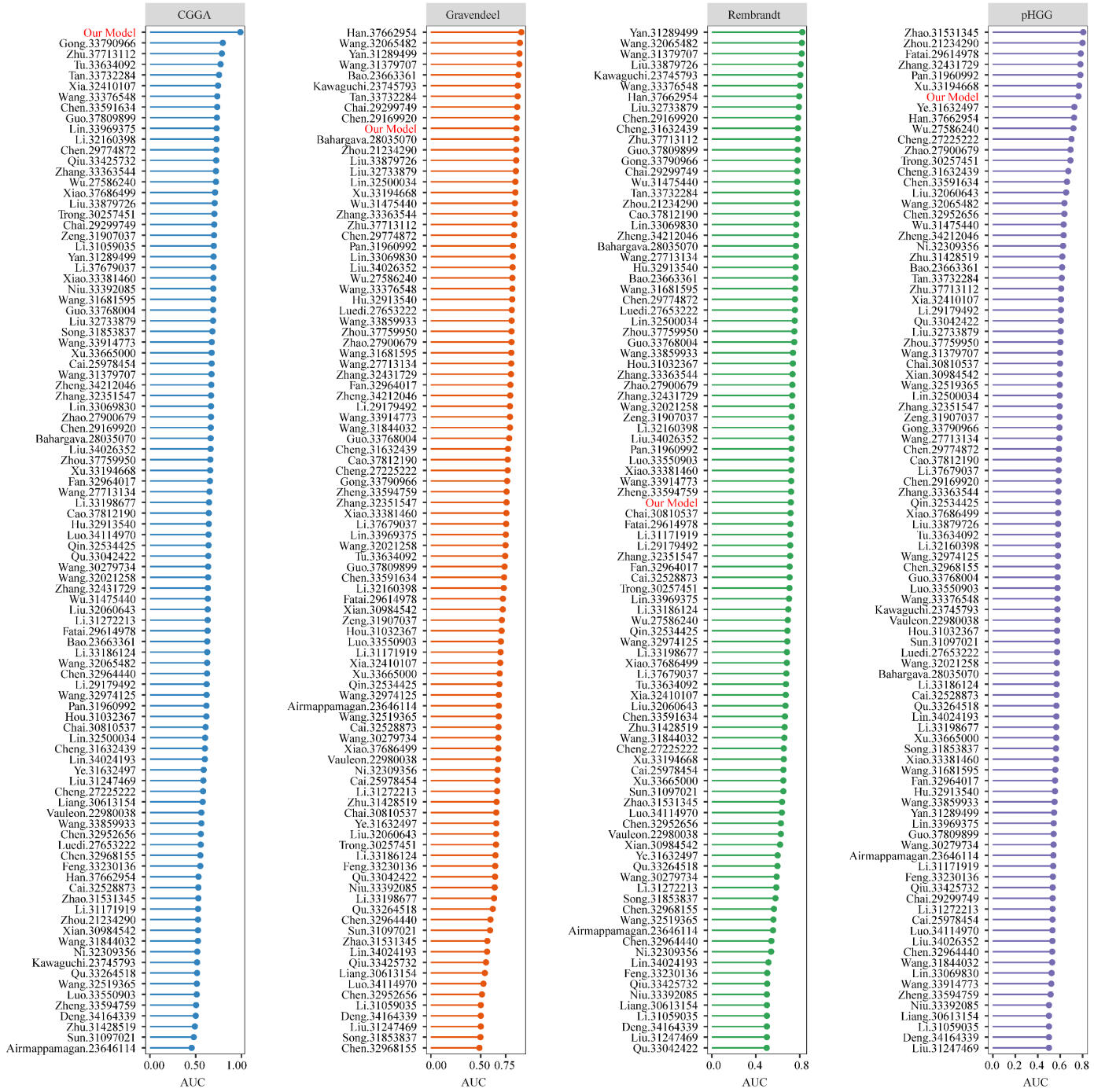


Fig. 5. Benchmarking of the proposed prognostic model against published signatures using AUC. Dot plots comparing the AUC of the newly developed prognostic model (highlighted in red) with a comprehensive set of previously published prognostic gene signatures. Analyses were performed independently across four HGG cohorts: CGGA, Gravendeel, Rembrandt, and pHGG. Each dot represents the AUC performance of a signature within a cohort.

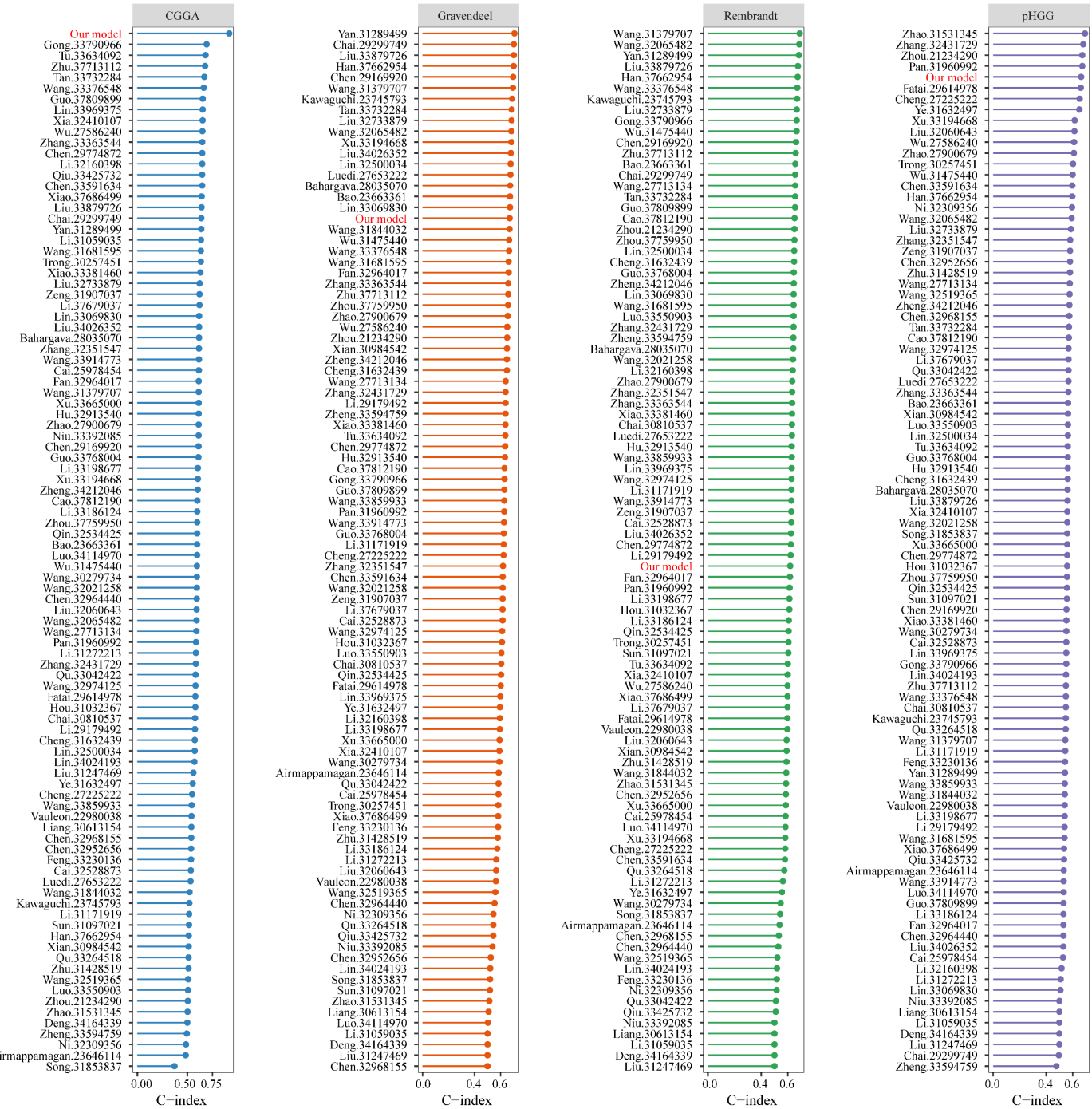


Fig. 6. Benchmarking of the proposed prognostic model against published signatures using C-index. Dot plots illustrating the C-index of the new prognostic model (highlighted in red) compared with a wide range of previously published prognostic gene signatures. Analyses were conducted across the CGGA, Gravendeel, Rembrandt, and pHGG cohorts. Each dot indicates the C-index for an individual signature within a given cohort.

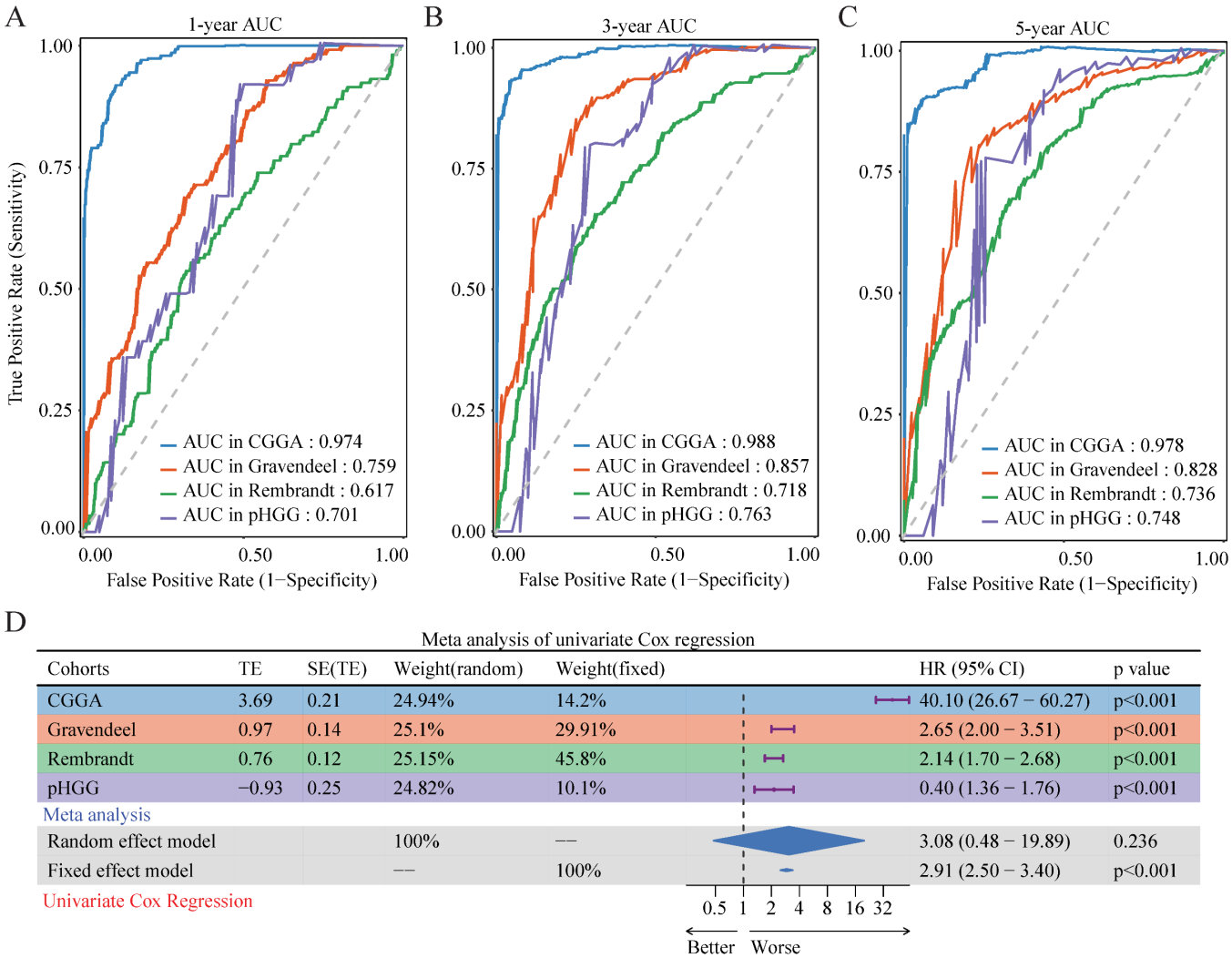


Fig. 7. Validation of the final prognostic model. (A–C) Time-dependent ROC curves showing the predictive accuracy of the final StepCox[forward] + RSF model for overall survival at (A) 1 year, (B) 3 years, and (C) 5 years in the CGGA, Gravendeel, Rembrandt, and pHGG cohorts. The AUC for each cohort is indicated on the plots. (D) Forest plot summarizing a meta-analysis of the model's prognostic value using univariate Cox regression. The analysis displays HRs and 95% CIs for each cohort, along with pooled HR estimates derived from both random- and fixed-effects models.

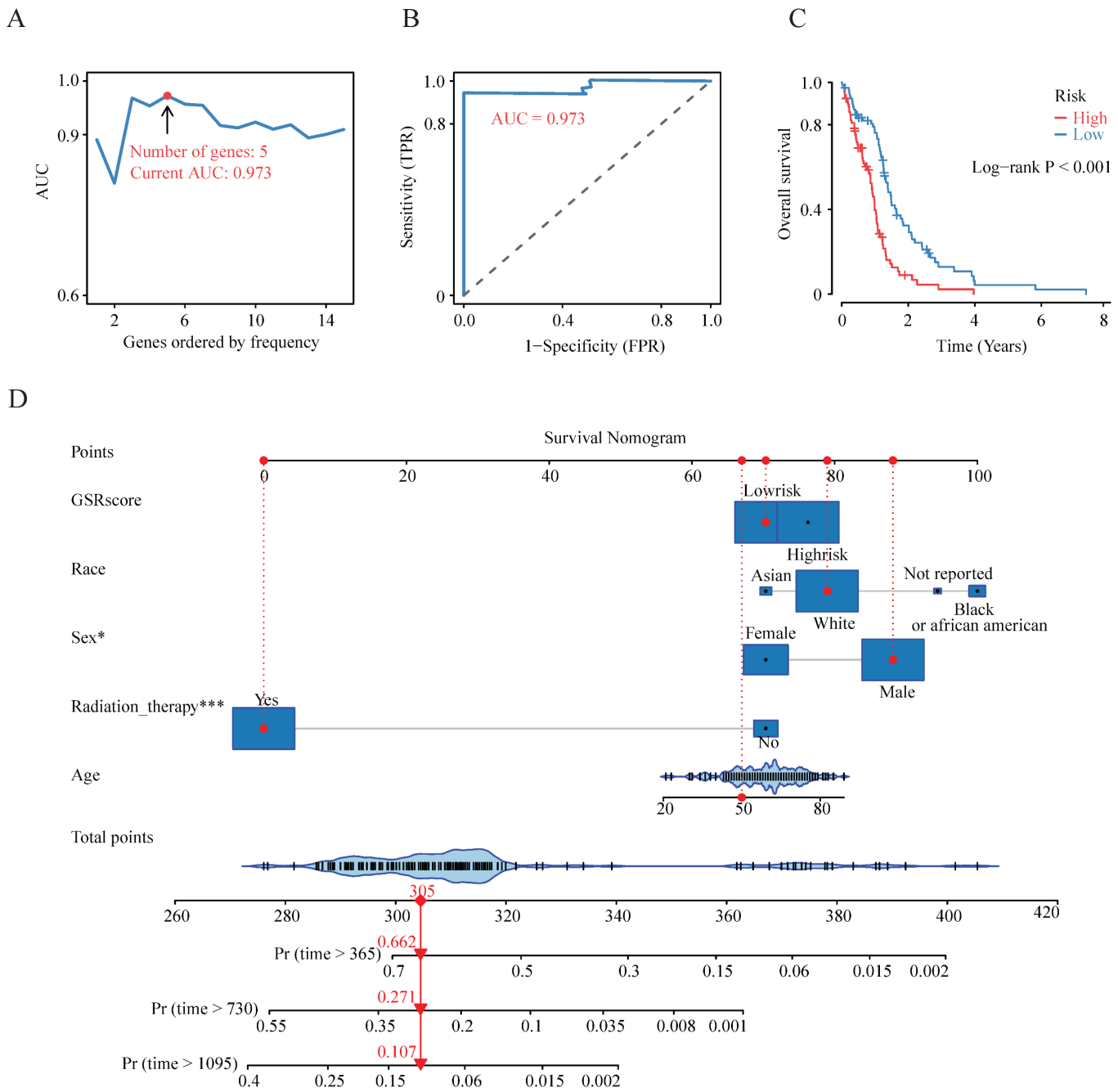


Fig. 8. Development of a prognostic nomogram for clinical application. (A) Determination of the optimal number of genes for the prognostic risk signature based on the AUC. The model containing 8 genes achieved the highest predictive accuracy with an AUC of 0.99. (B) ROC curve analysis further validated the robust predictive performance of the 8-gene signature, yielding an AUC of 0.993. (C) Kaplan–Meier survival analysis demonstrating that the 8-gene signature provides significant prognostic stratification of HGG patients (Log-rank p < 0.001). (D) Construction of a prognostic nomogram integrating the gene signature risk score (GSRSscore) with independent clinical risk factors, including patient age, sex, race, and radiation therapy status. The nomogram was designed to predict 1-, 2-, and 3-year overall survival probabilities.

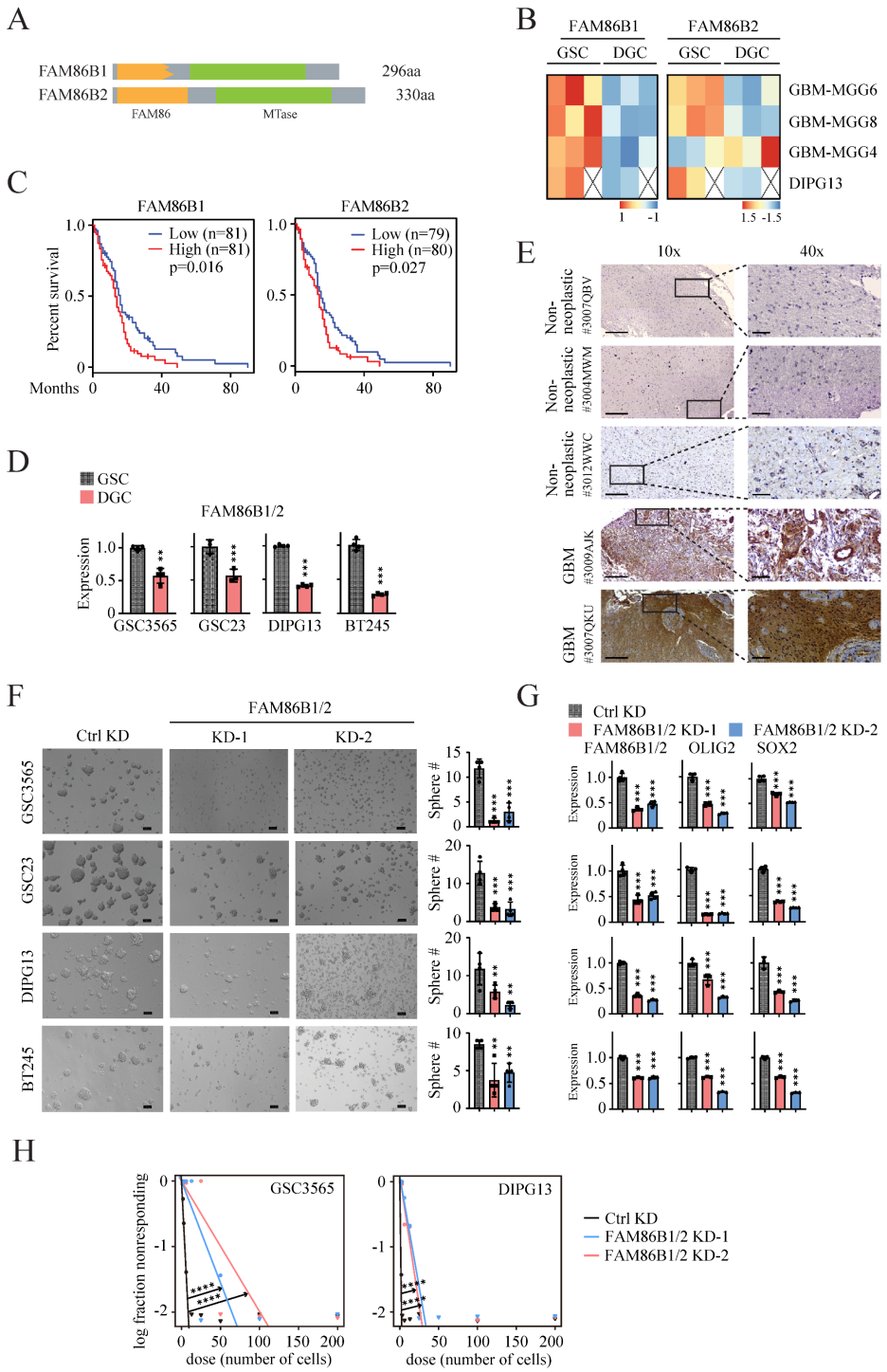


Fig. 9. FAM86B1/2 axis supports glioma stemness. (A) Schematic comparison of human FAM86B1 and FAM86B2 protein domain architecture. FAM86B1 (296 amino acids (aa)) is shorter than FAM86B2 (330 aa) and encodes a truncated FAM86 domain. (B) Heatmaps showing FAM86B1 and FAM86B2 expression in GSCs and DGCs from the public GBM dataset GSE54791 and the in-house DIPG13 RNA-seq dataset. (C) Kaplan-Meier survival curves showing overall survival of patients with high versus low FAM86B1 or FAM86B2 expression in the TCGA GBM cohort generated using GEPIA2. Statistical significance was determined by log-rank test. (D) qPCR analysis of FAM86B1/2 expression in matched GSC and DGC models from GBM and DIPG. (E) Immunohistochemical analysis of FAM86B1/2 expression in human GBM and non-neoplastic brain tissues. Unpaired, two-tailed t test. Scale bars, 100 µm. (F) Representative images (left) and quantification (right) of tumorsphere formation in GSCs 72 hours after control or FAM86B1/2 knockdown. One-way ANOVA and Dunnett's multiple-comparison test. Scale bars, 100 µm. (G) qPCR analysis of FAM86B1/2 as well as the stemness markers OLIG2 and SOX2 in control and FAM86B1/2 knockdown GSCs shown in panel F. One-way ANOVA and Dunnett's multiple-comparison test. (H) Extreme limiting dilution analysis (ELDA) of GSC self-renewal following FAM86B1/2 knockdown. Data points represent the log fraction of wells without spheres plotted against the number of cells plated per well after 10 days. Statistical significance was determined by pairwise comparison. Data represent mean ± SD. **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.