A Glioma Stem Cell–Associated Transcriptomic Program Predicts Survival Across Adult and Pediatric High-Grade Gliomas

Qiqi Xie [1], Bokai Wang [1, 2], Jia Shen [1, 3, 4, *]

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

[4] 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 (in-house DIPG13 RNA-seq and public GSE54791 dataset). The biological relevance of this signature was confirmed through 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 that integrates the gene signature–derived risk score with key clinical factors (age, sex, race, and radiation therapy status), enabling individualized survival prediction. 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

63 # 1. Introduction

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

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

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

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 and
89 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 glioma cell (DGC) pairs. Differential expression analysis
92 identified genes consistently upregulated in GSCs across both datasets, and their intersection yielded a core set of 603 GSC-
93 associated genes (Table S1; Fig. 2A). Functional relevance of these genes was assessed through protein–protein interaction
94 (PPI) network analysis, which highlighted known stemness regulators such as SOX2 and OLIG1/2 as top hub genes (Fig.
95 2B and C), supporting the validity of the approach. Gene Ontology (GO) (Fig. 2D) and Kyoto Encyclopedia of Genes and
96 Genomes (KEGG) pathway analyses (Fig. 2E) further demonstrated that these genes are enriched in key biological processes
97 related to GSC biology, including neurogenesis, RNA splicing, and cancer signaling.

98 **2.2. Prognostic value of the GSC-associated gene signature**

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

112 To further validate its robustness, the GSC-based prognostic model was compared with multiple previously published
113 glioma signatures across all four datasets. Using both AUC and C-index as evaluation metrics, our model consistently
114 ranked among the top performers, achieving the highest predictive accuracy in the CGGA cohort and comparable or superior
115 performance in the Gravendeel and pediatric HGG cohorts (Fig. 5 and 6). Time-dependent ROC analyses across all datasets

confirmed high discrimination for 1-, 3-, and 5-year overall survival predictions (Fig. 7A–C). Univariate Cox regression analysis also demonstrated that the GSC-derived gene signature risk score (GSRScore) was a significant independent predictor of mortality in CGGA (Hazard ratio (HR) = 40.10, 95% CI: 26.67–60.27, $p < 0.001$), Gravendeel (HR = 2.65, 95% CI: 2.00–3.51, $p < 0.001$), 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, $p < 0.001$) (Fig. 7D). A pooled meta-analysis yielded a summary HR of 3.08 (95% CI: 0.48–19.89) under the random-effects model and 2.91 (95% CI: 2.50–3.40) under the fixed-effects model (Fig. 7D). Collectively, these results demonstrate that the GSC-associated gene signature provides robust and consistent prognostic value across both adult and pediatric HGG datasets.

2.3. Integrated nomogram for prognostic prediction

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

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

3. Discussion

143 By contrasting patient-derived GSCs with their differentiated counterparts from both adult and pediatric HGGs, we defined
144 a coherent set of 603 stemness-associated genes enriched in neurogenesis, RNA splicing, and oncogenic signaling pathways,
145 underscoring their direct relevance to GSC biology. Network analysis highlighted SOX2, OLIG1, and OLIG2 as central
146 regulators, consistent with their established roles in neural stem/progenitor cell identity and glioma plasticity (8, 13, 14).
147 Accordingly, the risk score generated by our StepCox[forward] + RSF model likely quantifies the activity of this stemness
148 program, with high-risk tumors reflecting a stronger GSC phenotype and poorer prognosis.

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

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

161 Nevertheless, several limitations warrant consideration. First, our analyses were retrospective, and prospective, multi-center
162 clinical trials will be essential to confirm real-world performance. Second, although the genes in our signature are
163 biologically relevant, we did not experimentally validate their functional contribution within this study. Future CRISPR-
164 based perturbation screens could uncover causal drivers and potential therapeutic targets. Finally, tissue availability remains
165 a barrier for HGG diagnostics. Future efforts may explore liquid biopsy approaches, such as circulating tumor DNA or
166 extracellular vesicles, to detect the GSC signature non-invasively, which would greatly enhance clinical feasibility.

167 In conclusion, our integration of GSC transcriptomic profiling with machine learning and multi-cohort validation produced
168 a robust prognostic model for HGGs. By distilling complex biology into a clinically usable nomogram, we offer a practical
169 tool for individualized survival prediction while reinforcing glioma stemness as a key driver of disease behavior and a
170 promising target for therapeutic intervention.

171

172 # 4. Materials and methods

173 ## 4.1. Cell culture

174 DIPG13 cells were used as a representative pHGG model (16, 17). Cells were maintained in tumor stem neurobasal-A
175 medium mixed 1:1 with DMEM/F-12, supplemented with 10 mmol/L HEPES, sodium pyruvate, nonessential amino acids,
176 GlutaMAX-I, and B27 (all from Thermo Fisher Scientific). Growth factors included human FGF (20 ng/mL), human EGF
177 (20 ng/mL), and PDGF-AA and PDGF-BB (10 ng/mL each; FUJIFILM Biosciences), along with 10 ng/mL heparin
178 (Stemcell Technologies). Under these conditions, DIPG13 cells formed neurospheres, consistent with GSC properties. For
179 differentiation assays, cells were maintained in DMEM/F-12 supplemented with 10% fetal bovine serum (FBS) for 21 days,
180 resulting in DGCs that grew adherently.

181 ## 4.2. Data acquisition and preprocessing

182 To identify genes associated with the GSC phenotype, we analyzed both in-house and publicly available datasets. The in-
183 house dataset was generated from DIPG13 cells, with total RNA extracted from GSC and DGC populations and sequenced
184 by Novogene Corporation. In parallel, a publicly available GBM dataset (GSE54791; (18)) containing matched GSC and
185 DGC pairs was included for comparative analysis.

186 For model construction and validation, one training cohort and three independent validation cohorts were assembled. The
187 CGGA RNA-seq dataset (http://www.cgga.org.cn/) served as the training cohort. Two well-established microarray datasets,
188 Gravendeel (19) and Rembrandt (20), were used as validation cohorts. In addition, a pediatric HGG validation cohort was
189 generated by integrating raw data from five high-quality studies: Hoffman (21), Paugh (22), Schwartzentruber (23), Sturm
190 (24), and Witt (25). All data were uniformly curated and processed. Patients lacking complete survival information were
191 excluded. RNA-seq data were normalized and log-transformed as log2(TPM+1) or log2(FPKM+1). Microarray data were
192 processed from raw .CEL files using the Robust Multi-array Average (RMA) algorithm for background correction,
193 normalization, and summarization. To ensure comparability across platforms, all expression values were standardized using
194 z-scores prior to downstream analyses.

195 ## 4.3. Identification of GSC–associated genes and pathway enrichment analysis

196 Differentially expressed genes (DEGs) between GSCs and DGCs were identified in both the in-house DIPG13 and
197 GSE54791 datasets using the *limma* package (26) 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 PPI network was constructed using STRING (https://string-db.org/) with a confidence score cutoff of 0.4 and visualized in Cytoscape (27). Hub modules were identified using the cytoHubba plugin (28). GO (29) and KEGG (30) enrichment analyses were carried out with *clusterProfiler* (31), with terms and pathways considered significant at adjusted $p < 0.05$.

4.4. 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 RSF, CoxBoost, and Supervised Principal Components (SuperPC). Model performance was assessed using the 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.

4.5. 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 ROC curves and the corresponding 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.

4.6. 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. HRs with 95% CIs were pooled using a random-effects model based on the DerSimonian and Laird method (32), yielding a summary HR. This approach provided a robust estimate of prognostic value while accounting for heterogeneity among cohorts.

4.7. 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 (33) 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 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.

4.8. 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 ROC analyses were carried out with the *timeROC* package (34), and meta-analyses were implemented using the *meta* package (35). A two-sided p value < 0.05 was considered statistically significant.

4.9. Data availability

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

Author contributions

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

Funding support

This study was supported by the ACS-IRG Grant Mechanism (22-147-37; J. Shen), the Schwarz Family and Friends Cancer Research Fund (J. Shen), and Indiana University Seed Grant (J. Shen).

Acknowledgments

The authors gratefully acknowledge Dr. Q. Richard Lu (Cincinnati Children's Hospital Medical Center) for providing the DIPG13 cells, originally developed by Dr. Michelle Monje at Stanford University.

308 **References**

309 1. Poorva, P., Mast, J., Cao, B., Shah, M.V., Pollok, K.E., and Shen, J. 2025. Killing the killers: Natural killer cell
310 therapy targeting glioma stem cells in high-grade glioma. *Mol Ther* 33:2462-2478.
311 2. Sturm, D., Bender, S., Jones, D.T., Lichter, P., Grill, J., Becher, O., Hawkins, C., Majewski, J., Jones, C., Costello,
312 J.F., et al. 2014. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer* 14:92-
313 107.
314 3. Vredenburgh, J.J., Desjardins, A., Herndon, J.E., 2nd, Dowell, J.M., Reardon, D.A., Quinn, J.A., Rich, J.N.,
315 Sathornsumetee, S., Gururangan, S., Wagner, M., et al. 2007. Phase II trial of bevacizumab and irinotecan in
316 recurrent malignant glioma. *Clin Cancer Res* 13:1253-1259.
317 4. Monje, M., Cooney, T., Glod, J., Huang, J., Peer, C.J., Faury, D., Baxter, P., Kramer, K., Lenzen, A., Robison, N.J.,
318 et al. 2023. Phase I trial of panobinostat in children with diffuse intrinsic pontine glioma: A report from the Pediatric
319 Brain Tumor Consortium (PBTC-047). *Neuro Oncol* 25:2262-2272.
320 5. Prager, B.C., Bhargava, S., Mahadev, V., Hubert, C.G., and Rich, J.N. 2020. Glioblastoma Stem Cells: Driving
321 Resilience through Chaos. *Trends Cancer* 6:223-235.
322 6. Suva, M.L., and Tirosh, I. 2020. The Glioma Stem Cell Model in the Era of Single-Cell Genomics. *Cancer Cell*
323 37:630-636.
324 7. Bao, S., Wu, Q., McLendon, R.E., Hao, Y., Shi, Q., Hjelmeland, A.B., Dewhirst, M.W., Bigner, D.D., and Rich,
325 J.N. 2006. Glioma stem cells promote radioresistance by preferential activation of the DNA damage response.
326 *Nature* 444:756-760.
327 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.
328 2025. SUV39H1 maintains cancer stem cell chromatin state and properties in glioblastoma. *JCI Insight* 10.
329 9. Surowiec, R.K., Ferris, S.F., Apfelbaum, A., Espinoza, C., Mehta, R.K., Monchamp, K., Sirihorachai, V.R., Bedi,
330 K., Ljungman, M., and Galban, S. 2021. Transcriptomic Analysis of Diffuse Intrinsic Pontine Glioma (DIPG)
331 Identifies a Targetable ALDH-Positive Subset of Highly Tumorigenic Cancer Stem-like Cells. *Mol Cancer Res*
332 19:223-239.
333 10. Niu, W., Yu, H., Fan, X., Li, S., Sun, S., Gong, M., Zhang, S., Bi, W., Chen, X., and Fang, Z. 2024. Development
334 of stemness-related signature to optimize prognosis prediction and identify XMD8-85 as a novel therapeutic
335 compound for glioma. *Cell Signal* 120:111231.

336 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
337 Subtypes Based on the Stemness Index Predict Prognosis in Glioma Patients. *Front Genet* 12:616507.
338 12. Wang, Z., Wang, Y., Yang, T., Xing, H., Gao, L., Guo, X., Xing, B., and Ma, W. 2021. Machine learning revealed
339 stemness features and a novel stemness-based classification with appealing implications in discriminating the
340 prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Brief Bioinform* 22.
341 13. 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,
342 D.H. 2004. The oligodendroglial lineage marker OLIG2 is universally expressed in diffuse gliomas. *J Neuropathol*
343 *Exp Neurol* 63:499-509.
344 14. Balasubramaniyan, V., Timmer, N., Kust, B., Boddeke, E., and Copray, S. 2004. Transient expression of Olig1
345 initiates the differentiation of neural stem cells into oligodendrocyte progenitor cells. *Stem Cells* 22:878-882.
346 15. Li, Y., Wang, T., Wan, Q., Wang, Q., Chen, Z., Gao, Y., Ye, Y., Lin, J., Zhao, B., Wang, H., et al. 2022. TRAF4
347 Maintains Deubiquitination of Caveolin-1 to Drive Glioblastoma Stemness and Temozolomide Resistance. *Cancer*
348 *Res* 82:3573-3587.
349 16. Nagaraja, S., Vitanza, N.A., Woo, P.J., Taylor, K.R., Liu, F., Zhang, L., Li, M., Meng, W., Ponnuswami, A., Sun,
350 W., et al. 2017. Transcriptional Dependencies in Diffuse Intrinsic Pontine Glioma. *Cancer Cell* 31:635-652 e636.
351 17. Liao, Y., Luo, Z., Deng, Y., Zhang, F., Rao, R., Wang, J., Xu, L., Kumar, S.S., Sengupta, S., DeWire-Schottmiller,
352 M., et al. 2021. OLIG2 maintenance is not essential for diffuse intrinsic pontine glioma cell line growth but regulates
353 tumor phenotypes. *Neuro Oncol* 23:1183-1196.
354 18. Suva, M.L., Rheinbay, E., Gillespie, S.M., Patel, A.P., Wakimoto, H., Rabkin, S.D., Riggi, N., Chi, A.S., Cahill,
355 D.P., Nahed, B.V., et al. 2014. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma
356 stem-like cells. *Cell* 157:580-594.
357 19. Gravendeel, L.A., Kouwenhoven, M.C., Gevaert, O., de Rooi, J.J., Stubbs, A.P., Duijm, J.E., Daemen, A., Bleeker,
358 F.E., Bralten, L.B., Kloosterhof, N.K., et al. 2009. Intrinsic gene expression profiles of gliomas are a better predictor
359 of survival than histology. *Cancer Res* 69:9065-9072.
360 20. Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.C., Fine, H., and Madhavan, S. 2018. The REMBRANDT
361 study, a large collection of genomic data from brain cancer patients. *Sci Data* 5:180158.
362 21. Hoffman, L.M., Veldhuijzen van Zanten, S.E.M., Colditz, N., Baugh, J., Chaney, B., Hoffmann, M., Lane, A.,
363 Fuller, C., Miles, L., Hawkins, C., et al. 2018. Clinical, Radiologic, Pathologic, and Molecular Characteristics of

364 Long-Term Survivors of Diffuse Intrinsic Pontine Glioma (DIPG): A Collaborative Report From the International
365 and European Society for Pediatric Oncology DIPG Registries. *J Clin Oncol* 36:1963-1972.
366 22. Paugh, B.S., Broniscer, A., Qu, C., Miller, C.P., Zhang, J., Tatevossian, R.G., Olson, J.M., Geyer, J.R., Chi, S.N.,
367 da Silva, N.S., et al. 2011. Genome-wide analyses identify recurrent amplifications of receptor tyrosine kinases and
368 cell-cycle regulatory genes in diffuse intrinsic pontine glioma. *J Clin Oncol* 29:3999-4006.
369 23. Schwartzentruber, J., Korshunov, A., Liu, X.Y., Jones, D.T., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A.M.,
370 Quang, D.A., Tonjes, M., et al. 2012. Driver mutations in histone H3.3 and chromatin remodelling genes in
371 paediatric glioblastoma. *Nature* 482:226-231.
372 24. Sturm, D., Capper, D., Andreiuolo, F., Gessi, M., Kolsche, C., Reinhardt, A., Sievers, P., Wefers, A.K., Ebrahimi,
373 A., Suwala, A.K., et al. 2023. Multiomic neuropathology improves diagnostic accuracy in pediatric neuro-oncology.
374 *Nat Med* 29:917-926.
375 25. Mueller, S., Fangusaro, J., Thomas, A.O., Jacques, T.S., Bandopadhayay, P., de Blank, P., Packer, R.J., Fouladi,
376 M., van Meeteren, A.S., Jones, D., et al. 2024. Consensus framework for conducting phase I/II clinical trials for
377 children, adolescents, and young adults with pediatric low-grade glioma: Guidelines established by the International
378 Pediatric Low-Grade Glioma Coalition Clinical Trial Working Group. *Neuro Oncol* 26:407-416.
379 26. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. 2015. limma powers differential
380 expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47.
381 27. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker,
382 T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome*
383 *Res* 13:2498-2504.
384 28. Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T., and Lin, C.Y. 2014. cytoHubba: identifying hub objects
385 and sub-networks from complex interactome. *BMC Syst Biol* 8 Suppl 4:S11.
386 29. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight,
387 S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
388 *Nat Genet* 25:25-29.
389 30. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of
390 Genes and Genomes. *Nucleic Acids Res* 27:29-34.

391 31. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. 2021. clusterProfiler
392 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2:100141.
393 32. DerSimonian, R., and Laird, N. 1986. Meta-analysis in clinical trials. *Control Clin Trials* 7:177-188.
394 33. Balachandran, V.P., Gonen, M., Smith, J.J., and DeMatteo, R.P. 2015. Nomograms in oncology: more than meets
395 the eye. *Lancet Oncol* 16:e173-180.
396 34. Blanche, P., Dartigues, J.F., and Jacqmin-Gadda, H. 2013. Estimating and comparing time-dependent areas under
397 receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 32:5381-5397.
398 35. Balduzzi, S., Rucker, G., and Schwarzer, G. 2019. How to perform a meta-analysis with R: a practical tutorial. *Evid
399 Based Ment Health* 22:153-160.
400
401

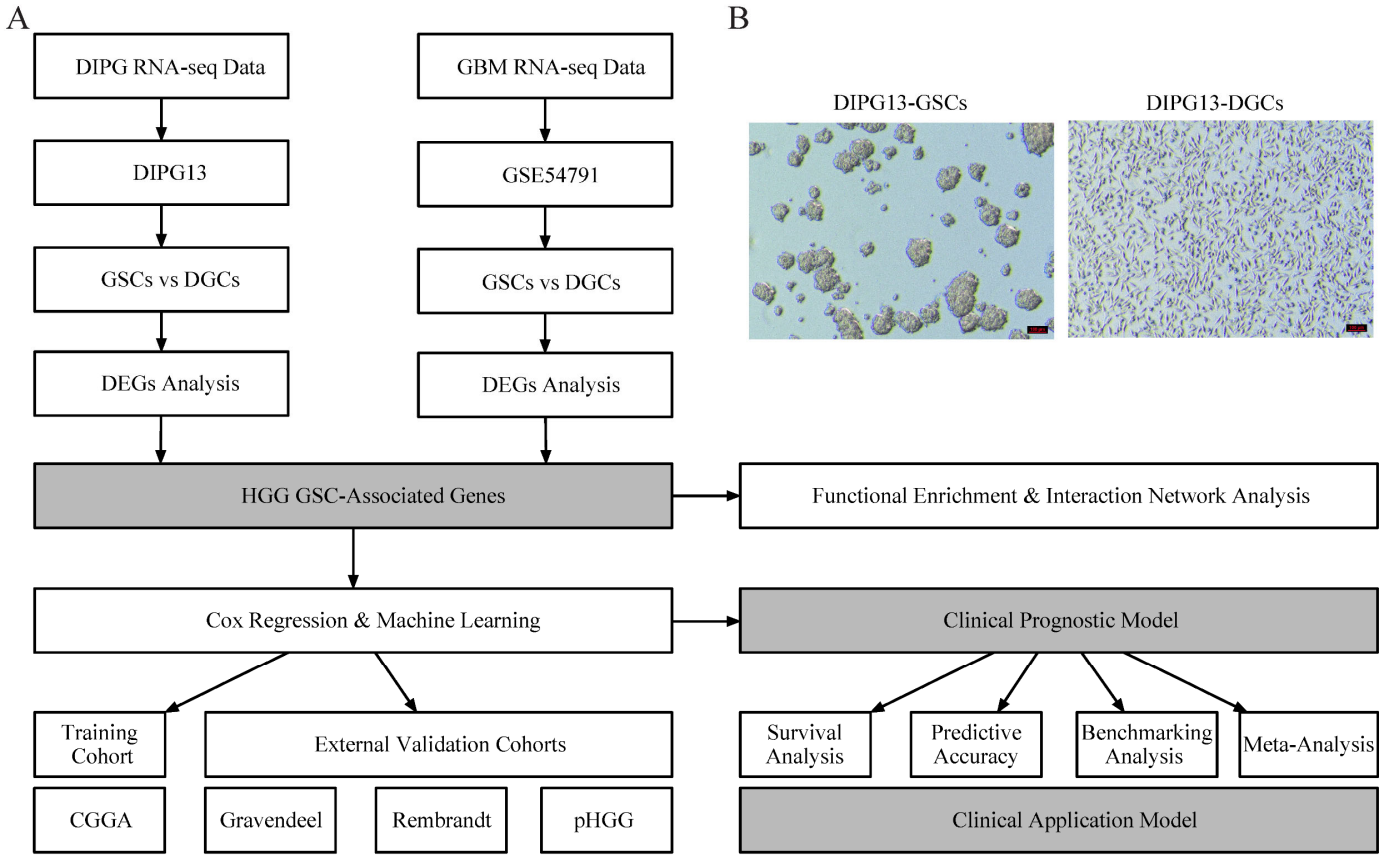


Fig. 1. Schematic overview of the study design. (A) RNA-seq datasets from DIPG13 (in-house, pediatric HGG) and GSE54791 (adult GBM, HGG) were analyzed to identify differentially expressed genes (DEGs) between glioma stem cells (GSCs) and differentiated glioma cells (DGCs). The intersecting DEGs defined a core set of HGG GSC-associated genes, which were subjected to functional enrichment, network analysis, Cox regression, and multiple 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 tool. (B) Representative images of DIPG13 cells under stem cell conditions (left) and following serum-induced differentiation (10% FBS for 21 days, right). DIPG13 cells displayed neurosphere morphology in stem cell medium and adopted an adherent morphology upon differentiation. Scale bars, 100 μm.

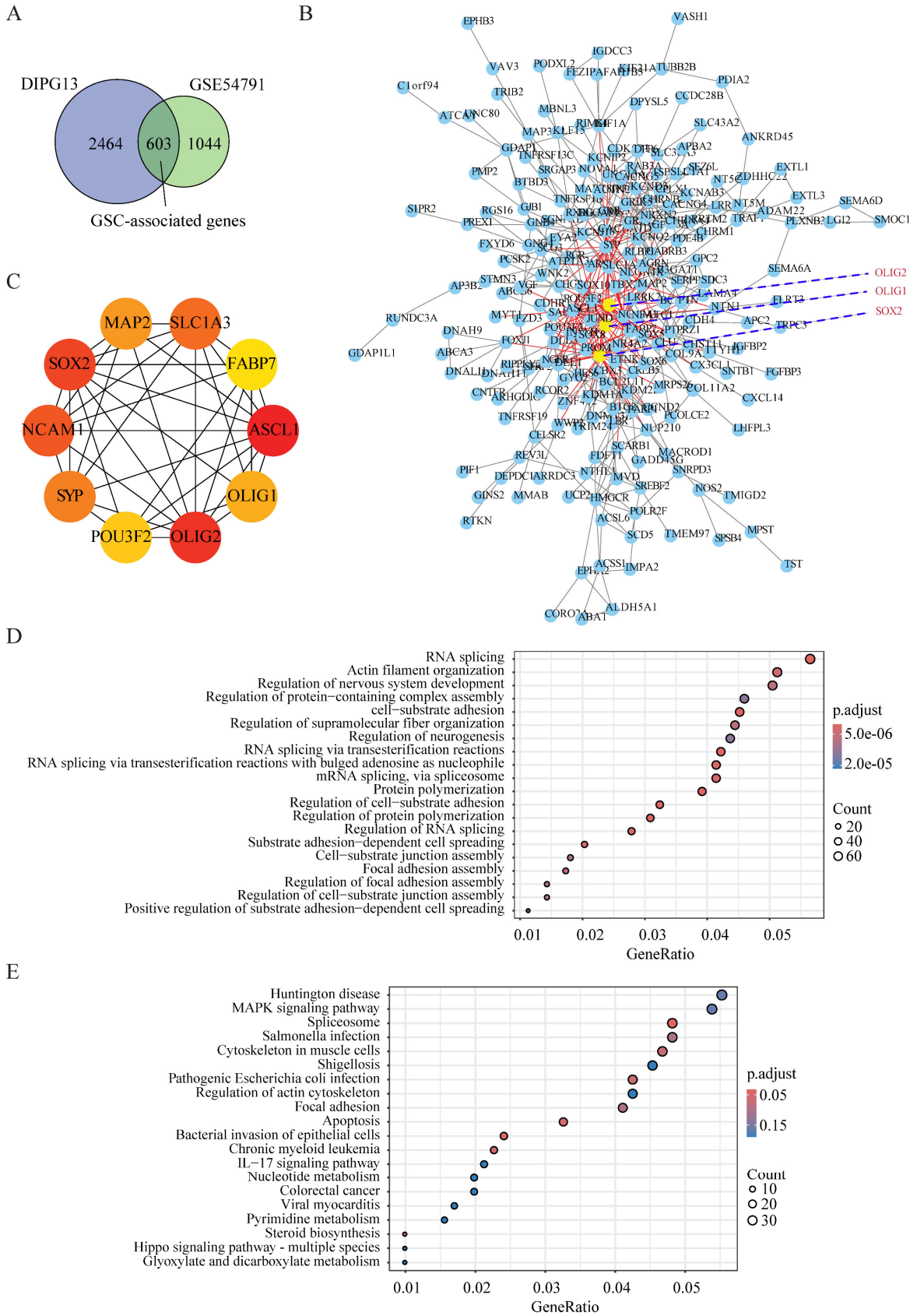


Fig. 2. Functional characterization of HGG GSC-associated genes. (A) Venn diagram showing overlap of DEGs between GSCs and DGCs in DIPG13 and GSE54791 datasets. The 603 shared DEGs were defined as the core HGG GSC-associated genes. (B) PPI network of these 603 genes generated using STRING, with key regulators 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.

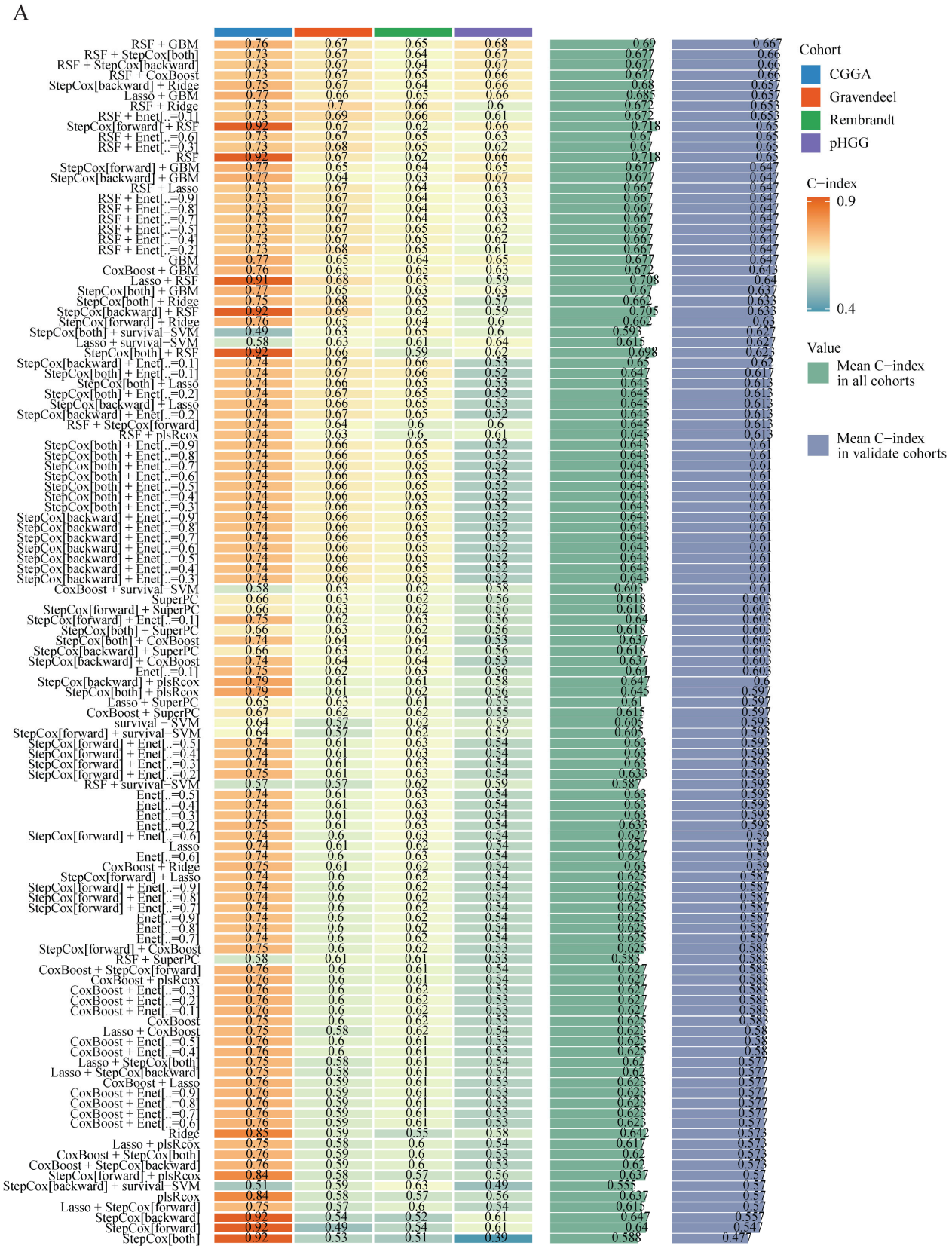
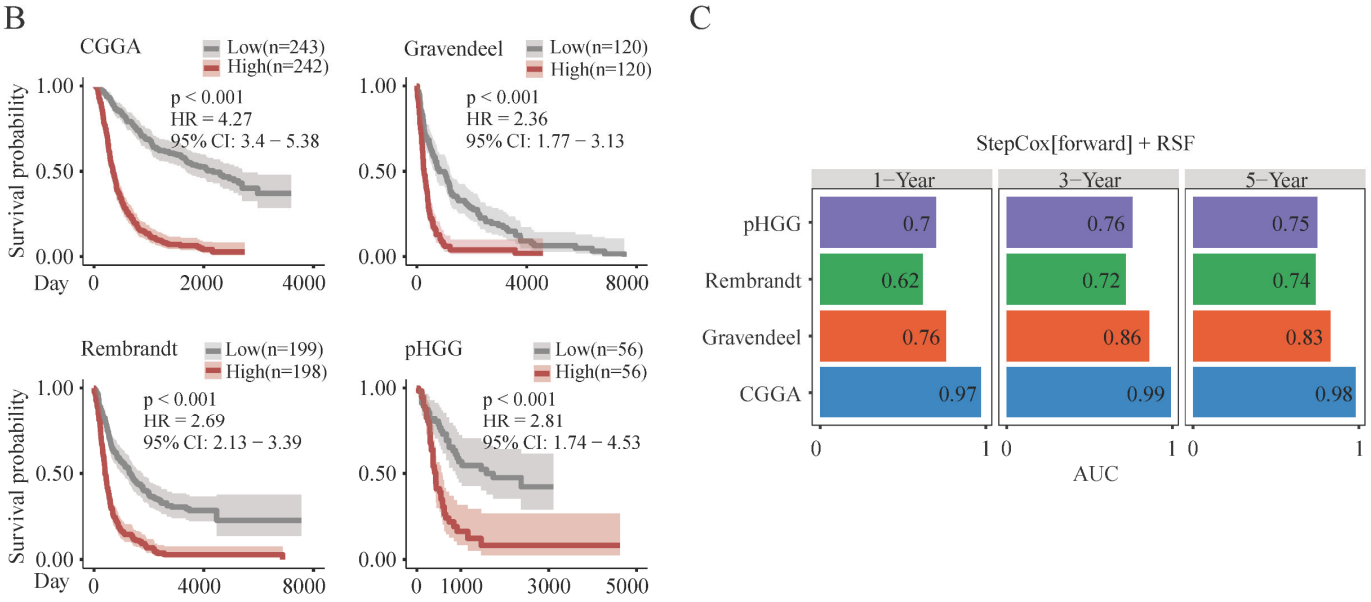


Fig. 3. Development of the optimal machine learning–based prognostic model. (A) Heatmap summarizing the performance of 135 machine learning algorithm combinations across four HGG cohorts (CGGA, Gravendeel, Rembrandt, and pHGG), measured by C-index. Bars on the right indicate mean C-index across all cohorts 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.

406

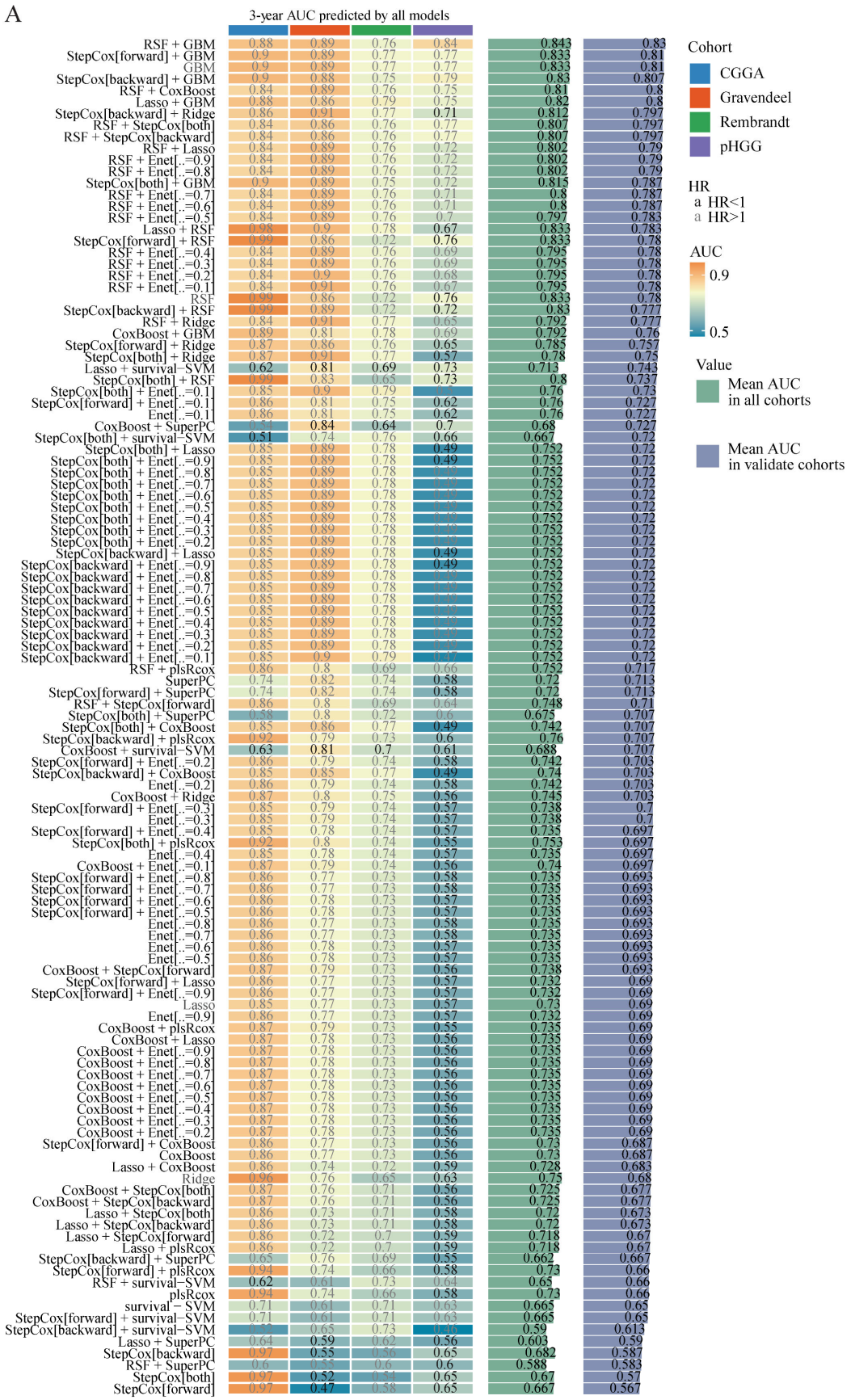
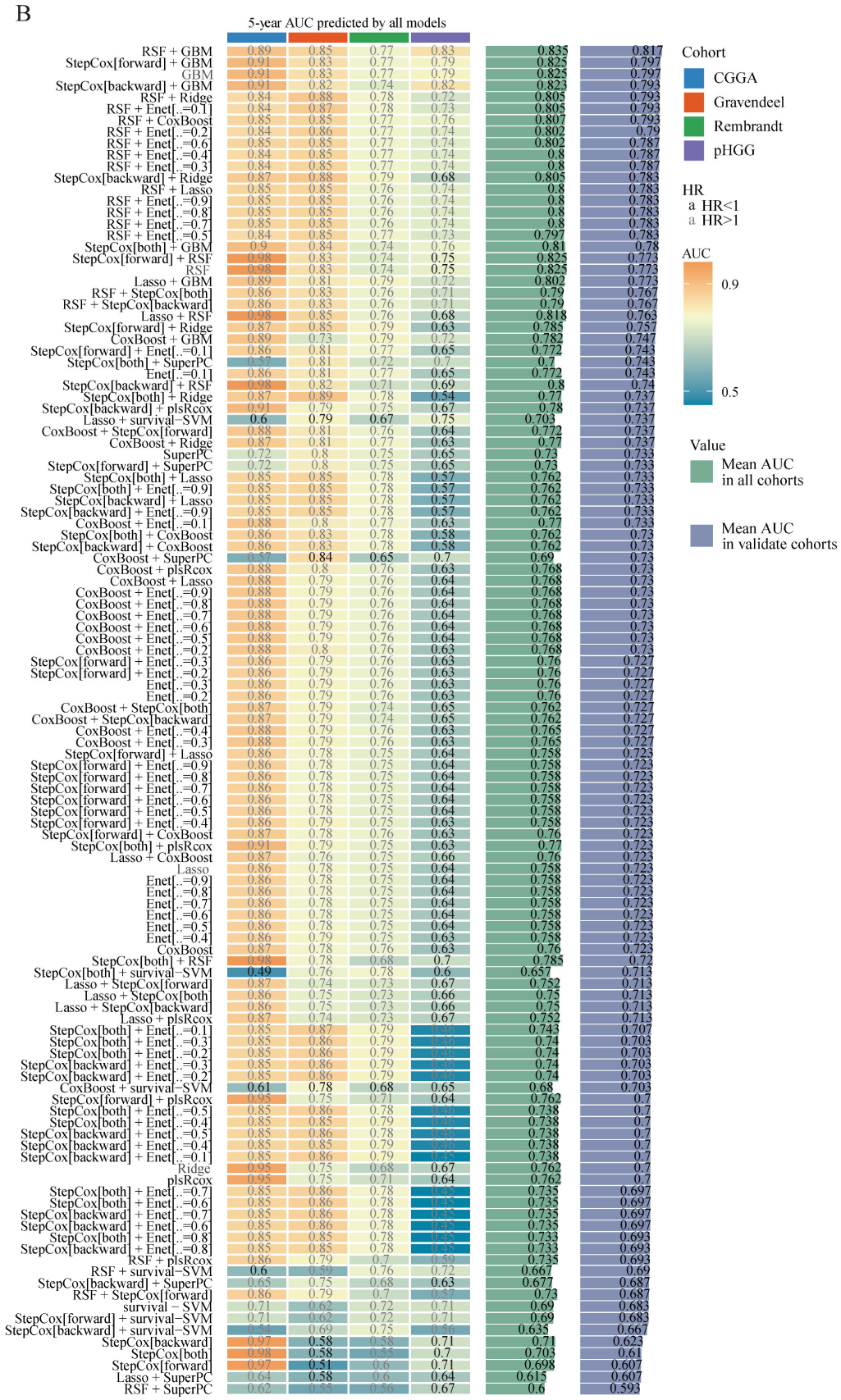


Fig. 4. Comparative evaluation of predictive accuracy across candidate models. (A–B) Heatmaps comparing the predictive performance of all 135 candidate models across four HGG cohorts (CGGA, Gravendeel, Rembrandt, and pHGG). Performance was evaluated using time-dependent AUC at (A) 3 years and (B) 5 years. 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.

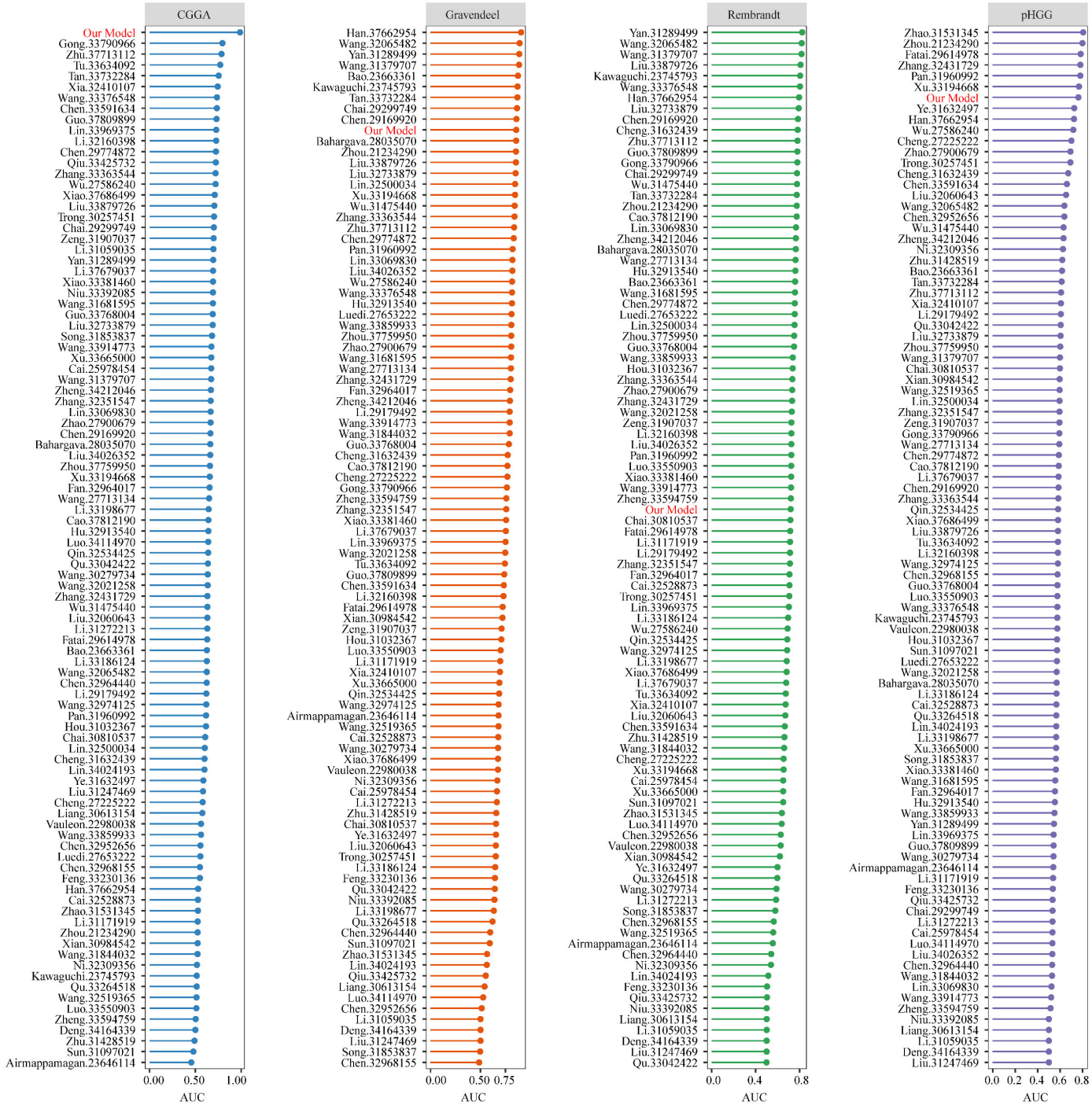


Fig. 5. Benchmarking of the proposed prognostic model against published signatures using AUC. Dot plots comparing the AUC of the newly developed prognostic model (highlighted in red) with a comprehensive set of previously published prognostic gene signatures. Analyses were performed independently across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

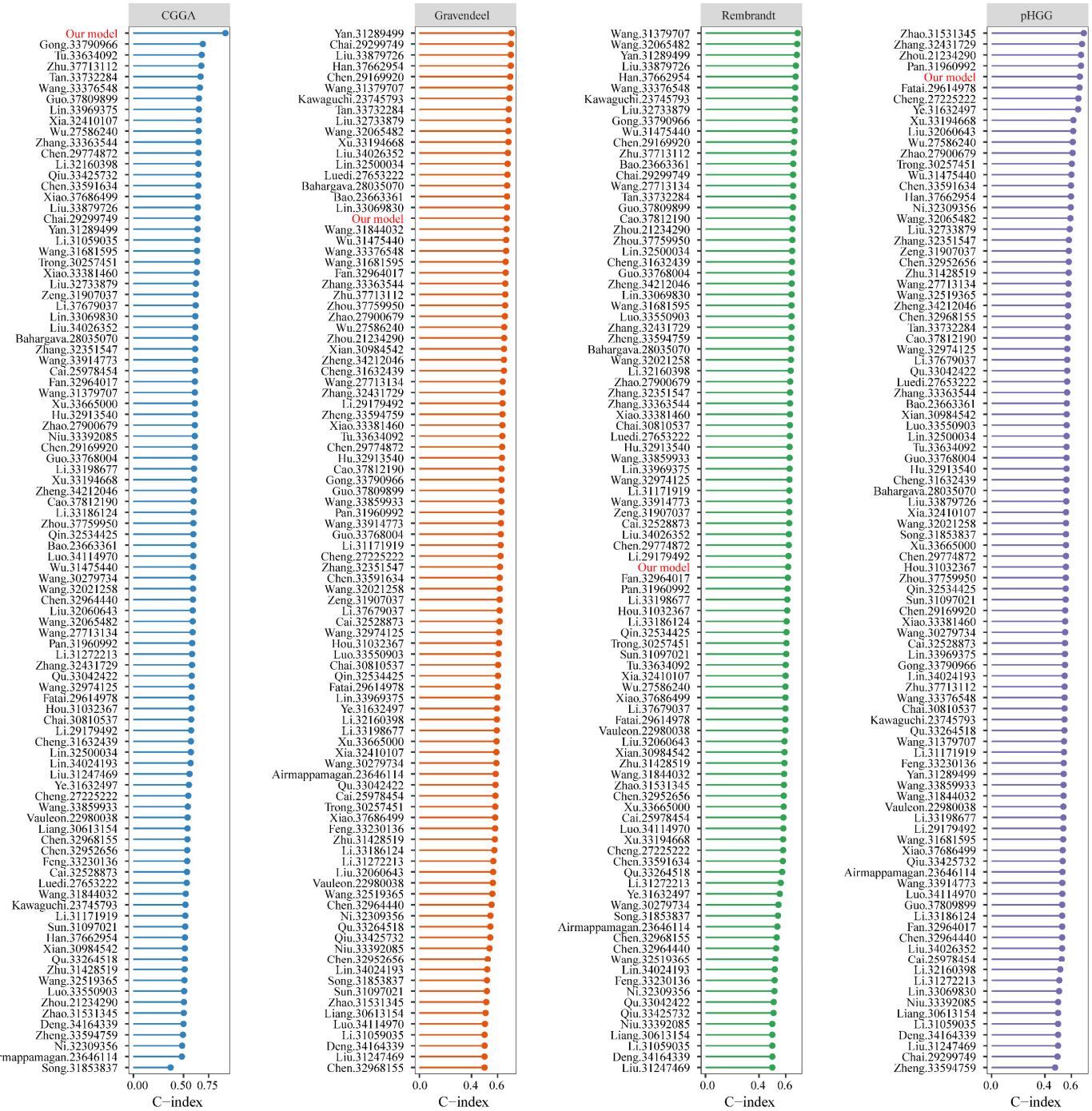


Fig. 6. Benchmarking of the proposed prognostic model against published signatures using C-index. Dot plots illustrating the C-index of the new prognostic model (highlighted in red) compared with a wide range of previously published prognostic gene signatures. Analyses were conducted across the CGGA,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

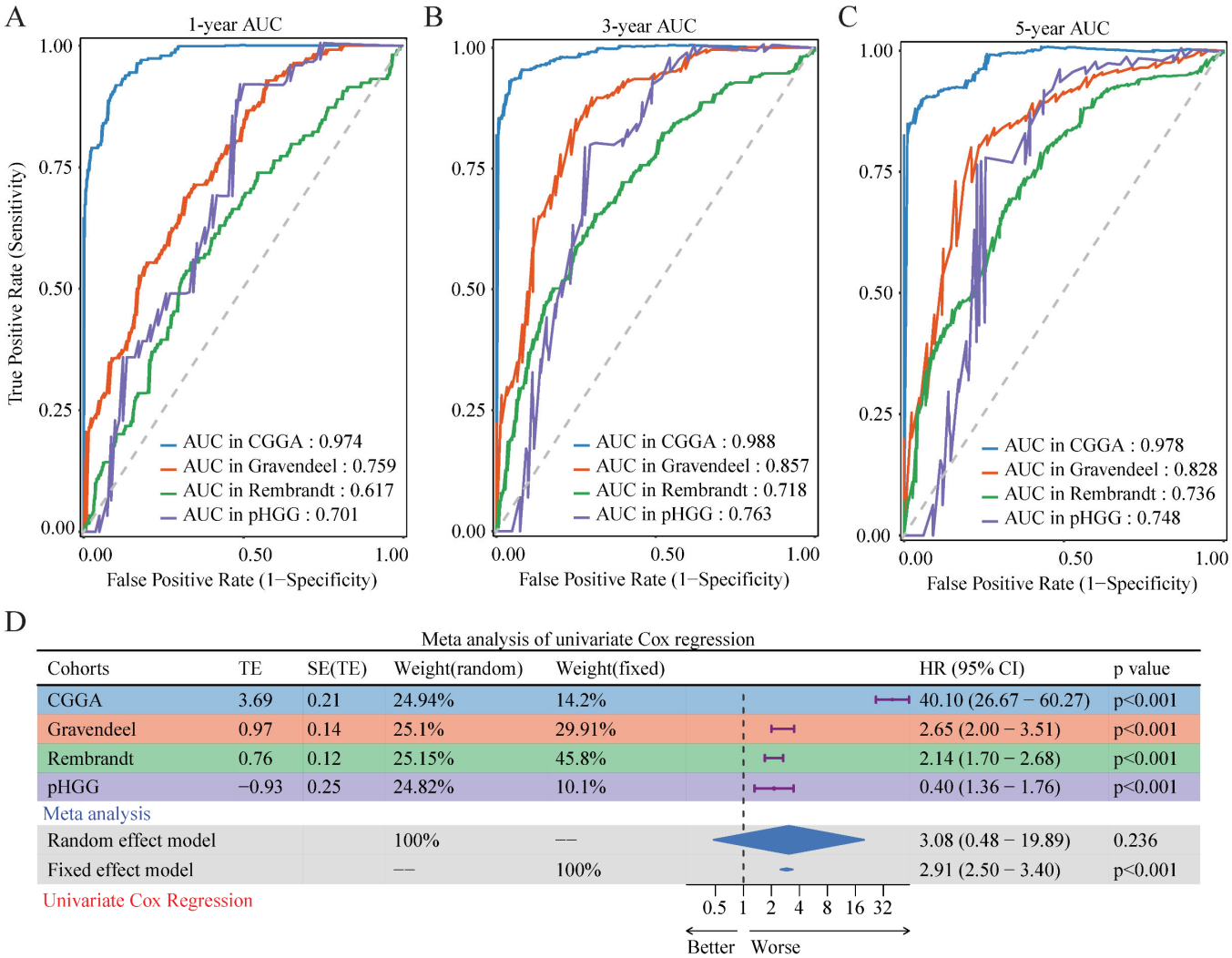


Fig. 7. Validation of the final prognostic model. (A–C) Time-dependent ROC curves showing the predictive accuracy of the final StepCox[forward] + RSF model for overall survival at (A) 1 year, (B) 3 years, and (C) 5 years in the CGGA, Gravendeel, Rembrandt, and pHGG cohorts. The AUC for each cohort is 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.

411

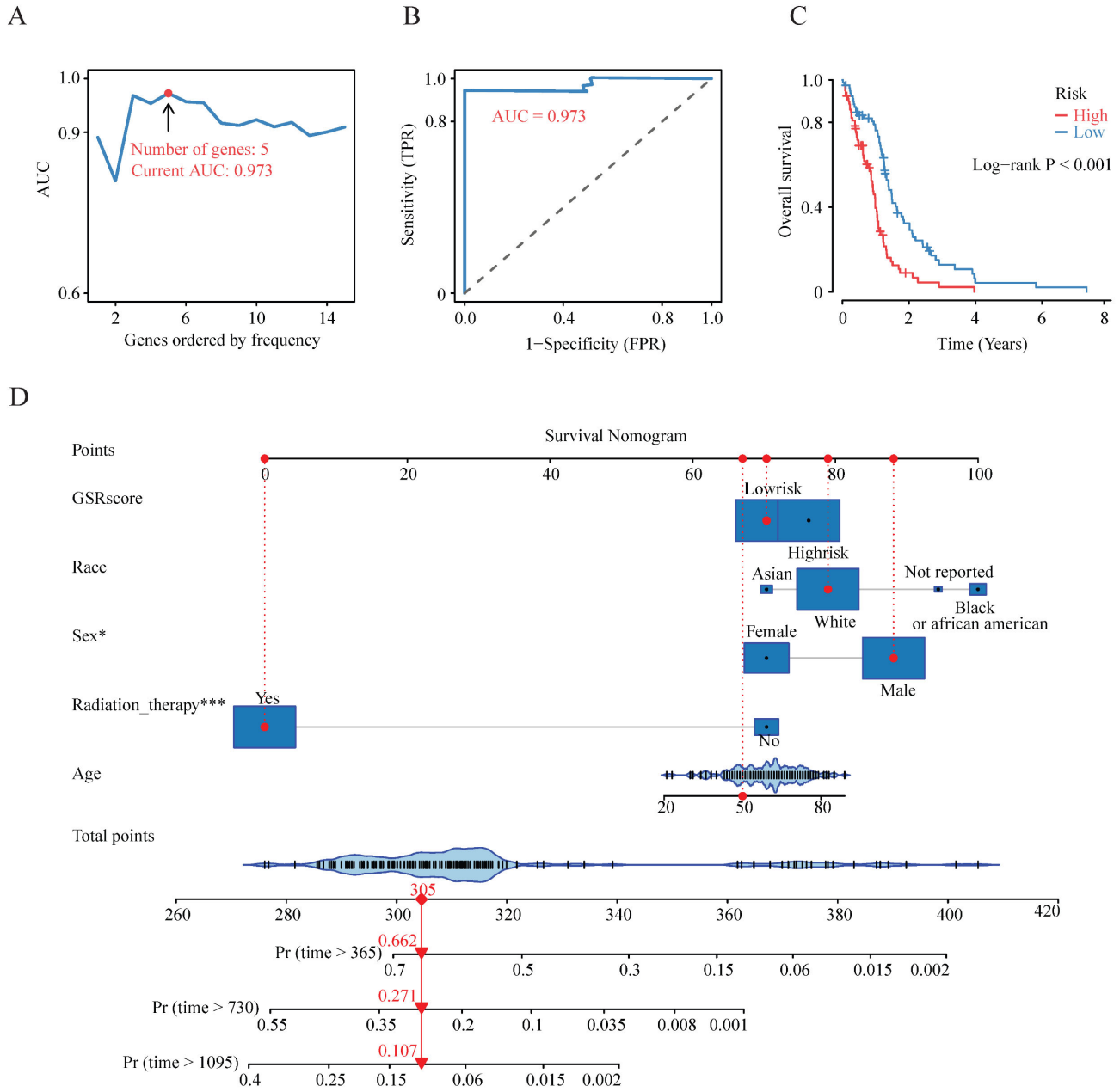


Fig. 8. Development of a prognostic nomogram for clinical application. (A) Determination of the optimal number of genes for the prognostic risk signature based on the AUC. The model containing 5 genes achieved the highest predictive accuracy with an AUC of 0.973. (B) ROC curve analysis further validated the robust predictive performance of the 5-gene signature, yielding an AUC of 0.973. (C) Kaplan–Meier survival analysis demonstrating that the 5-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.

412