Research Paper

Two machine learning methods identify a metastasis-related prognostic model that predicts overall survival in medulloblastoma patients

Kui Chen[1,*], Bingsong Huang[1,*], Shan Yan[2,*], Siyi Xu[1], Keqin Li[1], Kuiming Zhang[1], Qi Wang[1], Zhongwei Zhuang[1], Liang Wei[1], Yanfei Zhang[1], Min Liu[1,&], Hao Lian[1], Chunlong Zhong[1]

[1]Department of Neurosurgery, Shanghai East Hospital, Tongji University School of Medicine, Shanghai 200120, P.R. China
[2]Huamu Community Health Service Center, Shanghai 201204, P.R. China
*Equal contributions

Correspondence to: Chunlong Zhong, Hao Lian, Min Liu; **email:** drchunlongzhong@126.com, https://orcid.org/0000-0002-0605-7273; lhysybyb@sjtu.edu.cn, rodger_lm@163.com
Keywords: medulloblastoma, machine learning, prognostic model, overall survival
Received: February 26, 2020 **Accepted:** July 30, 2020 **Published:** November 5, 2020

ABSTRACT

Approximately 30% of medulloblastoma (MB) patients exhibit metastasis at initial diagnosis, which often leads to a poor prognosis. Here, by using univariate Cox regression analysis, two machine learning methods (Lasso-penalized Cox regression and random survival forest-variable hunting (RSF-VH)), and multivariate Cox regression analysis, we established two metastasis-related prognostic models, including the 47-mRNA-based model based on the Lasso method and the 21-mRNA-based model based on the RSF-VH method. In terms of the results of the receiver operating characteristic (ROC) curve analyses, we selected the 47-mRNA metastasis-associated model with the higher area under the curve (AUC). The 47-mRNA-based prognostic model could classify MB patients into two subgroups with different prognoses. The ROC analyses also suggested that the 47-mRNA metastasis-associated model may have a better predictive ability than MB subgroup. Multivariable Cox regression analysis demonstrated that the 47-mRNA-based model was independent of other clinical characteristics. In addition, a nomogram comprising the 47-mRNA-based model was built. The results of ROC analyses suggested that the nomogram had good discrimination ability. Our 47-mRNA metastasis-related prognostic model and nomogram might be an efficient and valuable tool for overall survival (OS) prediction and provide information for individualized treatment decisions in patients with MB.

INTRODUCTION

Medulloblastoma (MB) is the most frequent type of malignant pediatric brain tumor and comprises at least four distinct molecular subgroups: WNT, SHH, Group 3, and Group 4 [1, 2]. The presence of metastatic disease often results in a less favorable outcome for MB patients, and unfortunately, approximately 25-33% of MB patients present with metastases at the time of diagnosis [3, 4]. Currently, the standard protocol, including surgery followed by craniospinal radiation and chemotherapy, achieves an overall survival (OS) rate of about 85% at 5 years for standard-risk patients with MB [5–7]. However, a number of survivors suffer from serious treatment-related effects of radiotherapy and cytotoxic chemotherapy, resulting in a decline in cognition and intellect, endocrine disorders and an increased incidence of secondary cancers [8, 9]. In addition, some MB patients receive unnecessary or excessive therapies, while others may be faced with metastasis or recurrence because of a lack of appropriate treatment. The risk stratification of MB

patients is mainly based on age at diagnosis, size of the residual disease, metastatic status, histology, subgroup, and some cytogenetic biomarkers [4, 10, 11]. The current therapies and the risk stratification used since the late 1980s pose tremendous challenges [10]. The survival rate of MB patients has been stagnant for approximately 30 years despite the multipronged approach to therapy [12]. These limitations have prompted a search for more precise and comprehensive biomarkers for the discrimination of MB patients to improve precision MB treatment.

With the advances of high-throughput microarray and RNA sequencing technologies, an increasing number of metastasis-related prognostic signatures have been identified in various types of cancers [13–18]. A six-gene metastasis signature has been reported to be robust for predicting the survival of hepatocellular carcinoma patients in multicenter cohorts [15]. Another study showed that some novel tissue- and serum-based metastasis-specific microRNA biomarkers could be clinically applicable to predict prognosis in colorectal cancer [13]. In addition, a lymph-node-metastasis-related gene signature had stronger predictive power than other clinical information for the prognostic evaluation of esophageal cancer [17]. These studies suggest that metastasis-related signatures might serve as potentially accurate biomarkers for predicting the outcome of cancer patients. Therefore, searching for a metastasis-related biomarker signature may have concrete prognostic and predictive value in the management of MB. Moreover, the identification of metastasis-related molecular markers might pave the way for precisely targeted metastasis-related molecular therapies for MB.

In the present study, we focused on the mRNA expression profiles of large cohorts of patients with MB from the Gene Expression Omnibus (GEO) database. The differentially expressed genes (DEGs) were screened by analyzing the gene expression data between MB tissues with and without metastasis. Then, by employing Cox regression analysis and two machine learning algorithms, including the Lasso-penalized Cox regression model and random survival forest-variable hunting (RSF-VH) algorithm, we identified a metastasis-related prognostic signature that can accurately predict survival in MB patients. Moreover, the metastasis signature was a survival-related factor independent of well-known clinical characteristics. Finally, we built a predictive nomogram that showed good discrimination ability and was clinically useful. Overall, the metastasis signature and nomogram may be reliable and practical prognostic tools for OS evaluation and might facilitate individualized therapy for MB patients with different risks of disease.

RESULTS

Development and validation of a 47-mRNA metastasis-related prognostic model

Differential expression analysis using metastatic status as the group variable identified a total of 2,429 DEGs (adjusted $P < 0.2$). To define the association of the DEGs with the OS of MB patients, univariate Cox regression analysis was conducted, and the results revealed that 307 of the 2,429 DEGs were significantly related to OS in MB patients. Then, we employed the Lasso-penalized Cox regression and RSF-VH methods to identify the DEGs with the greatest prognostic value in which we required the selected prognostic DEGs to appear > 100 times out of 1,000 repetitions. Finally, by using multivariable Cox regression analysis, a 47-mRNA metastasis-related prognostic model based on Lasso-penalized Cox regression and a 21-mRNA metastasis-related prognostic model based on RSF-VH were established. Supplementary Tables 1 and 2 show the multivariate Cox regression coefficients of the genes in the 47-mRNA metastasis-related prognostic model and the 21-mRNA metastasis-related prognostic model, respectively. In addition, Supplementary Tables 3 and 4 show the repeat occurrence frequencies of the genes in the 47-mRNA metastasis-related prognostic model and the 21-mRNA metastasis-related prognostic model, respectively. To investigate the predictive efficiency of the afore-mentioned two metastasis-related prognostic models, the receiver operating characteristic (ROC) curve analysis was performed. The resulting area under the curve (AUC) of the 47-mRNA metastasis-associated prognostic model was 0.817 (95% CI: 0.762-0.872) (Figure 1A), while the resulting AUC of the 21-mRNA metastasis-associated prognostic model was 0.691 (95% CI: 0.625-0.757) (Figure 1B). Therefore, the 47-mRNA metastasis-associated prognostic model with the higher AUC was selected for further analysis.

Among the genes in the 47-mRNA metastasis-related prognostic model, 31 genes (AK7, ARL1, ARSG, BACH2, C9orf153, COPS7B, CPB2, EIF2B3, FABP4, GAGE1, GPR126, GUCY2C, GYG2, HIST1H2AE, ICOS, IDI2, MAGEB5, MEIS2, NTHL1, NUP210L, POLN, POU1F1, PSORS1C1, RN7SKP226, RN7SL432P, RNA5SP53, SAA3P, STXBP5L, TBCC, ZIC1, and ZPBP2) had positive coefficients, indicating an association between their higher expression levels and shorter OS, while the higher expression levels of the remaining 16 genes with negative coefficients (ARHGEF40, CAMKK1, CCDC125, FAM81A, GSDMC, IL22, KCNAB3, MDN1, PAPPA2, POLE3, RN7SL187P, RN7SL581P, RNASE9, RNU1-75P,

SLC25A11, and TBCK) may correlate with longer OS. The distributions of the 47-gene-based risk scores, OS, survival status, and 47-gene expression profiles of the MB patients in the discovery set and validation set are shown in Figure 2A, 2B. The 31 risk-related mRNAs tended to be more highly expressed in the high-risk group, whereas the 16 protective mRNAs tended to exhibit higher expression in the low-risk group. K-M survival analysis showed that in the discovery set, patients in the high-risk group (n = 110) had a significantly shorter OS than those in the low-risk group (n = 193; $P < 0.0001$; Figure 3A). Similar results were observed in the validation set ($P = 0.00034$; Figure 3C). To evaluate the predictive performance of the 47-mRNA metastasis-related prognostic model, we performed time-dependent ROC curve analysis. The AUCs of the metastasis-related prognostic model were 0.901 at 1 year, 0.806 at 3 years, and 0.782 at 5 years for the discovery set (Figure 3B), and 0.804 at 1 year, 0.759 at 3 years, and 0.69 at 5 years for the validation set (Figure 3D). All AUCs exceeded 0.6, indicating that the metastasis-related forecast model had a good performance for OS prediction in MB patients. The 47-mRNA metastasis-related prognostic model had a better predictive performance than MB subgroup in the discovery set (0.817 vs 0.586) (Figure 3E) and in the validation set (0.693 vs 0.64) (Figure 3F). When we employed the 47-mRNA metastasis-related prognostic model to predict the survival of patients in each MB subgroup, we found that there were statistically significant differences in OS between the high-risk group and the low-risk group (Figure 3G–3J). In addition, the results of multivariable Cox regression analysis revealed that the metastasis-associated prognostic model was a powerful and independent prognostic factor related to OS (Figure 4).

Weighted gene co-expression network analysis and gene ontology enrichment analysis for identifying the pathways significantly associated with the 47-mRNA-based risk score model

All genes from entire MB dataset were applied to construct a gene co-expression network using weighted gene co-expression network analysis (WGCNA). The original 55 modules were obtained with Dynamic Tree Cut method (Figure 5A). The module dissection threshold was set at 0.3 to merge correlated modules and 41 modules were finally generated (Figure 5B). The correlations between co-expression modules and clinical phenotypes were calculated and visualized through a heatmap (Figure 6). The scatterplot of gene significance (GS) for the 47-mRNA-based risk score model vs. module membership (MM) was plotted in the co-expression magenta module (Figure 7A). Gene

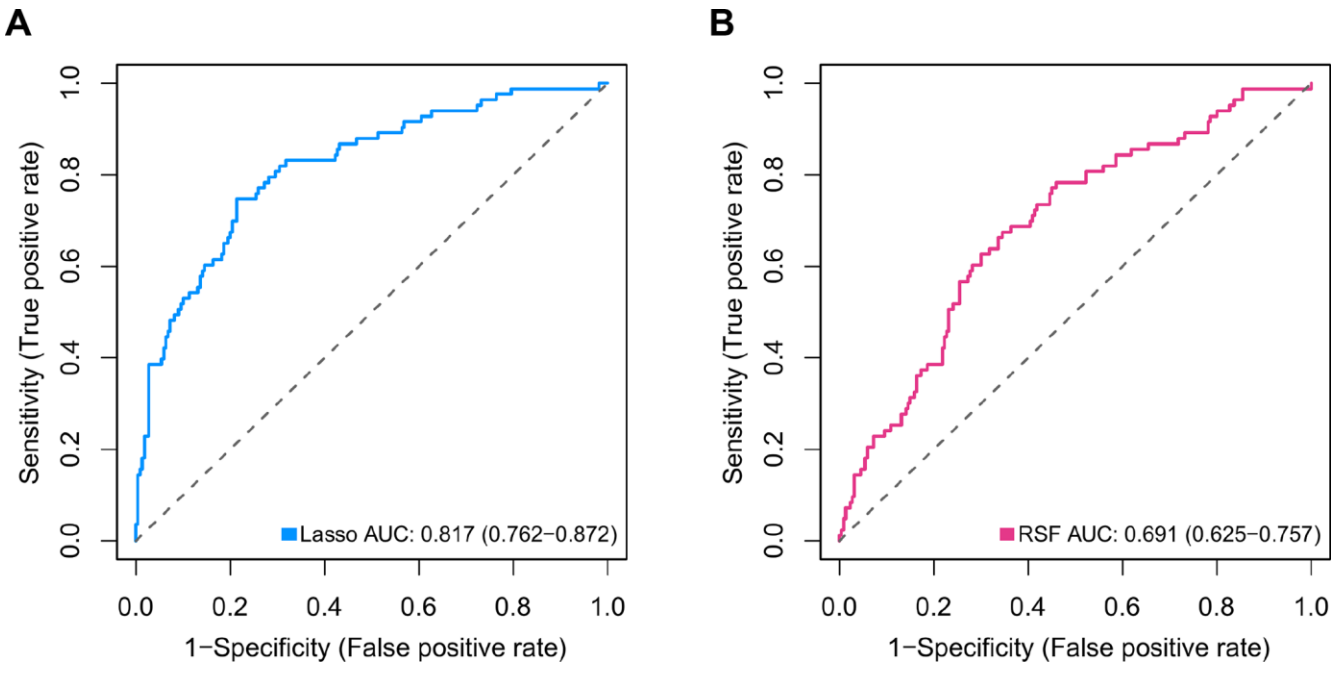

Figure 1. Comparison of the predictive power of the 47-mRNA metastasis-related model based on Lasso-penalized Cox regression and the 21-mRNA metastasis-related model based on random survival forest-variable hunting (RSF-VH). The receiver operating characteristic (ROC) curves of the 47-mRNA-based model based on Lasso-penalized Cox regression (**A**) and the 21-mRNA-based model based on RSF-VH (**B**).

ontology (GO) enrichment analysis of hub genes revealed a significantly relationship between sensory perception of smell, cellular process involved in reproduction, and regulation of STAT cascade and the 47-mRNA-based risk score model (Figure 7B).

Construction and validation of a metastasis-related nomogram

To provide clinicians with a quantitative method that could predict the probability of 1-, 3-, and 5-year OS in patients with MB, a metastasis-associated nomogram was generated by integrating the 47-mRNA metastasis-related prognostic model and five clinicopathological factors (Figure 8). Calibration plots showed that the metastasis-related nomogram performed well compared with the ideal curve (the 45-degree line) (Figure 9A–9C). Decision curve analysis (DCA) indicated that if the threshold probability of patients or doctors is more than 10%, then utilizing the metastasis-related nomogram to predict the probability of 1-, 3-, and 5-year OS adds more net benefit than the treat-none scheme or the treat-all-patients scheme (Figure 9D–9F). The AUCs of the metastasis-related nomogram were 0.887 at 1 year, 0.834 at 3 years, and 0.805 at 5 years for the discovery set (Figure 10A), and 0.84 at 1 year, 0.775 at 3 years, and 0.73 at 5 years for the validation set (Figure 10B).

DISCUSSION

Although MB consists of four primary molecular subgroups with disparate clinical outcomes, molecular markers that could precisely predict survival in MB patients are still lacking. Given that nonmetastatic and metastatic patients with MB often have distinct outcomes, metastasis-associated mRNAs may be accurate predictors of outcome in MB patients. In our study, the differentially expressed genes between metastatic and nonmetastatic MB tissues were screened, and univariate Cox analysis, two different machine learning algorithms including Lasso-penalized Cox regression and RSF-VH, and multivariate Cox analysis were performed to construct two metastasis-related prognostic models (the 47-

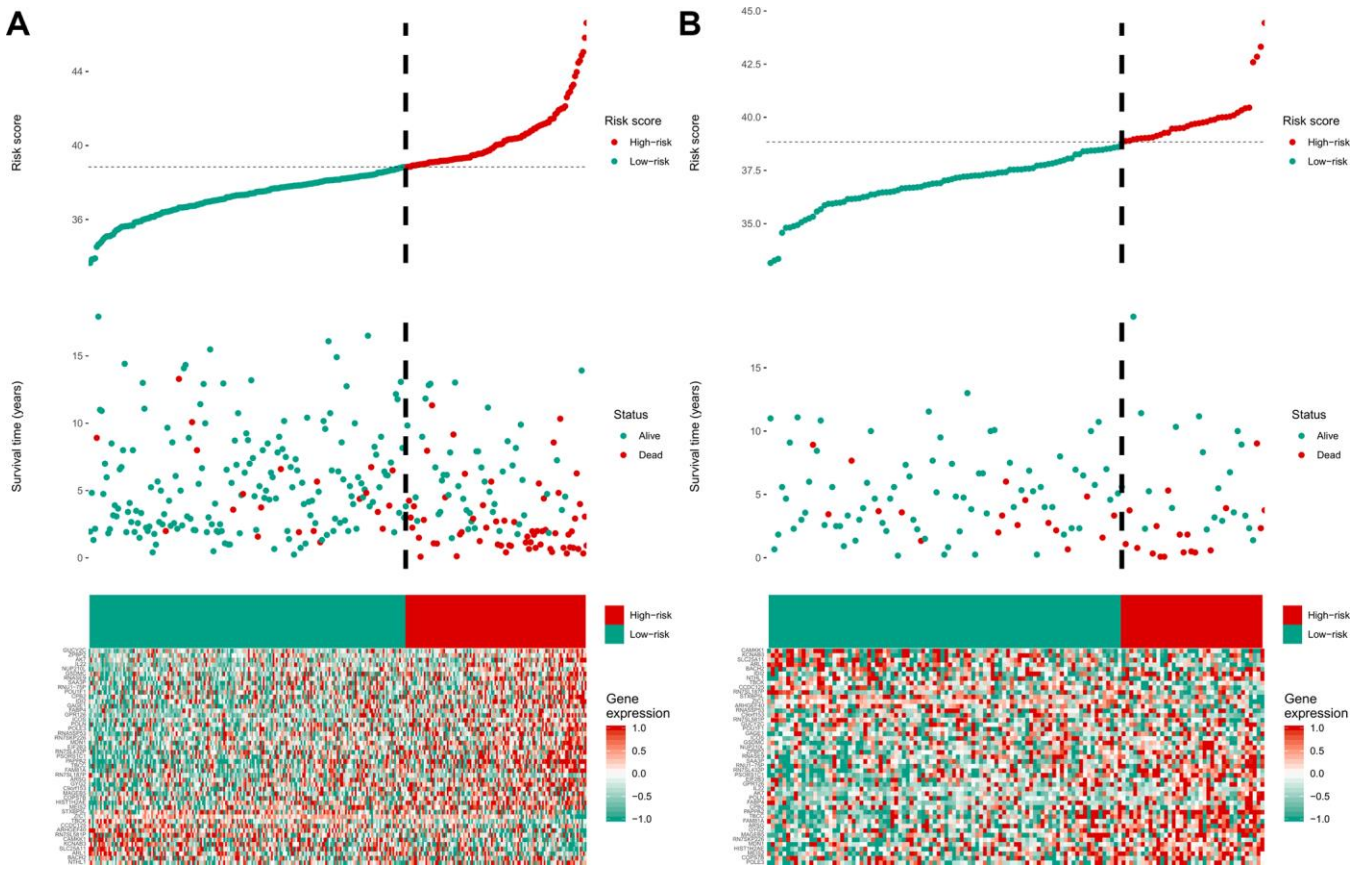

Figure 2. The distribution of the risk score, overall survival (OS), OS status, and heatmap of the 47-mRNA metastasis-related model in the discovery set (**A**) and validation set (**B**).

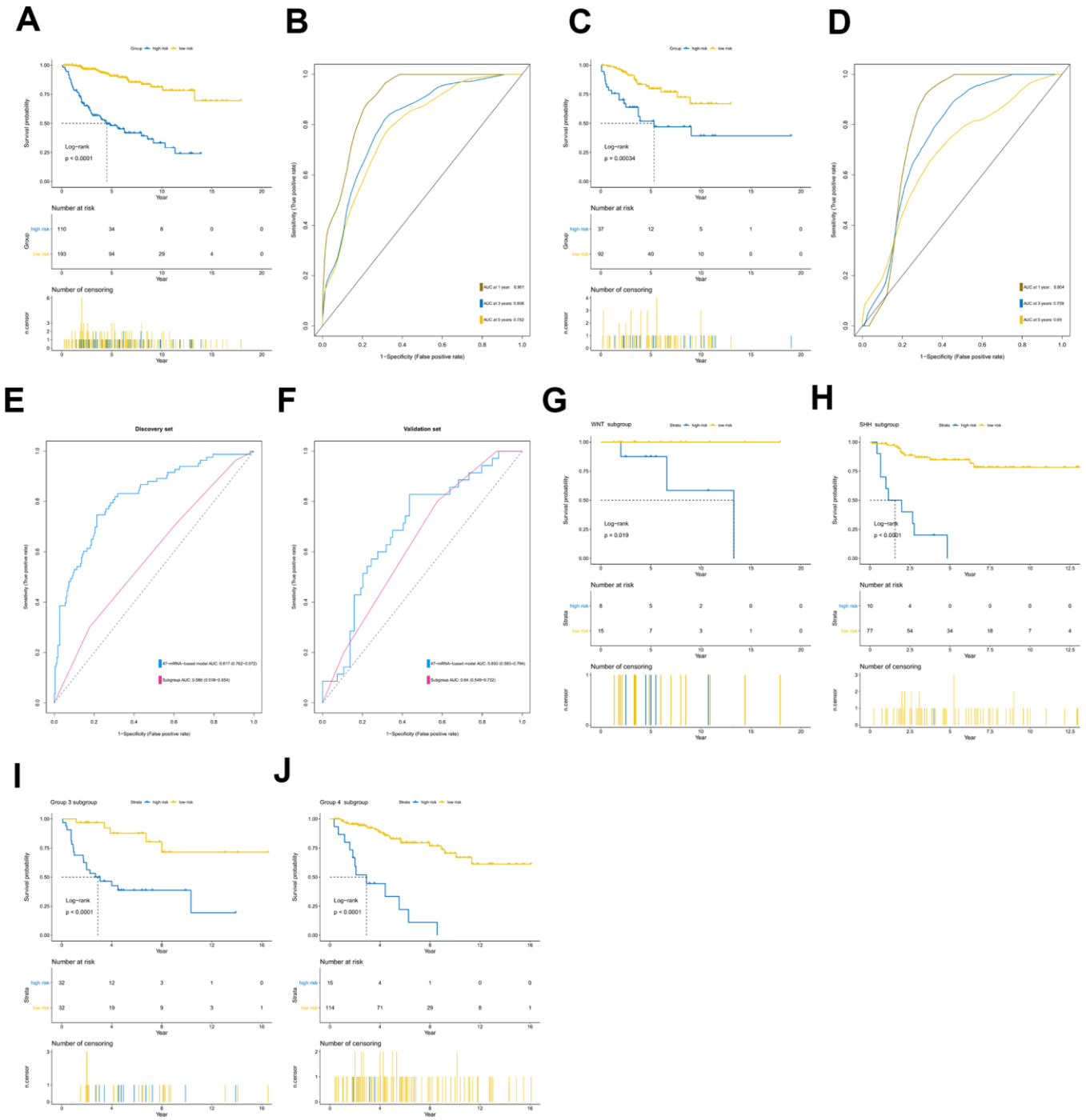

Figure 3. Prognostic value of the 47-mRNA metastasis-related model. The Kaplan-Meier (K-M) curves show the OS of the high- and low-risk patients with MB classified by the optimal cutoff value. (**A**) K-M curves for the discovery set. (**B**) ROC curves for the 47-mRNA-based model in the discovery set. (**C**) K-M curves for the validation set. (**D**) ROC curves for the 47-mRNA-based model in the validation set. (**E**) The comparison of the area under the ROC of the 47-mRNA-based model versus that of subgroup in the discovery set. (**F**) The comparison of the area under the ROC of the 47-mRNA-based model versus that of subgroup in the validation set. (**G**) K-M curves showing the OS for the high- and low-risk patients with WNT MB using the 47-mRNA-based model in the discovery set. (**H**) K-M curves showing the OS for the high- and low-risk patients with SHH MB using the 47-mRNA-based model in the discovery set. (**I**) K-M curves showing the OS for the high- and low-risk patients with group 3 MB using the 47-mRNA-based model in the discovery set. (**J**) K-M curves showing the OS for the high- and low-risk patients with group 4 MB using the 47-mRNA-based model in the discovery set.

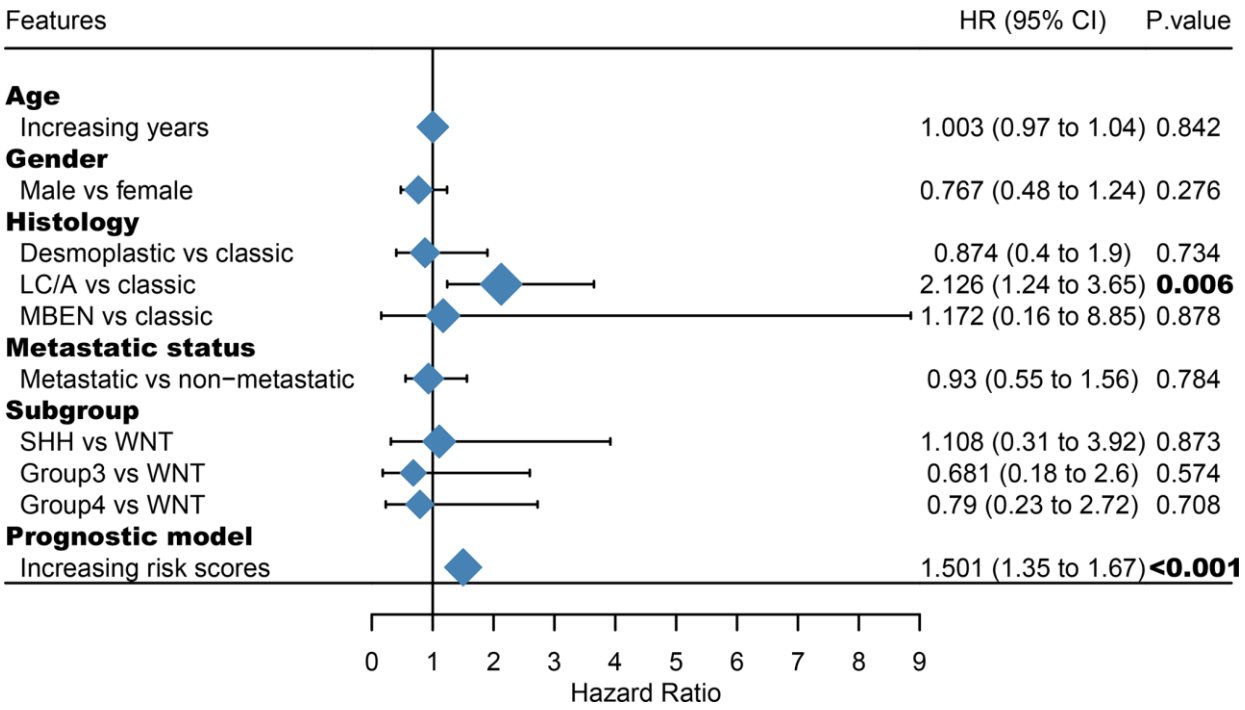

Figure 4. Multivariate Cox regression analysis incorporating the 47-mRNA metastasis-related model and known prognostic clinical characteristics. LC/A, large cell/anaplastic; MBEN, medulloblastoma with extensive nodularity.

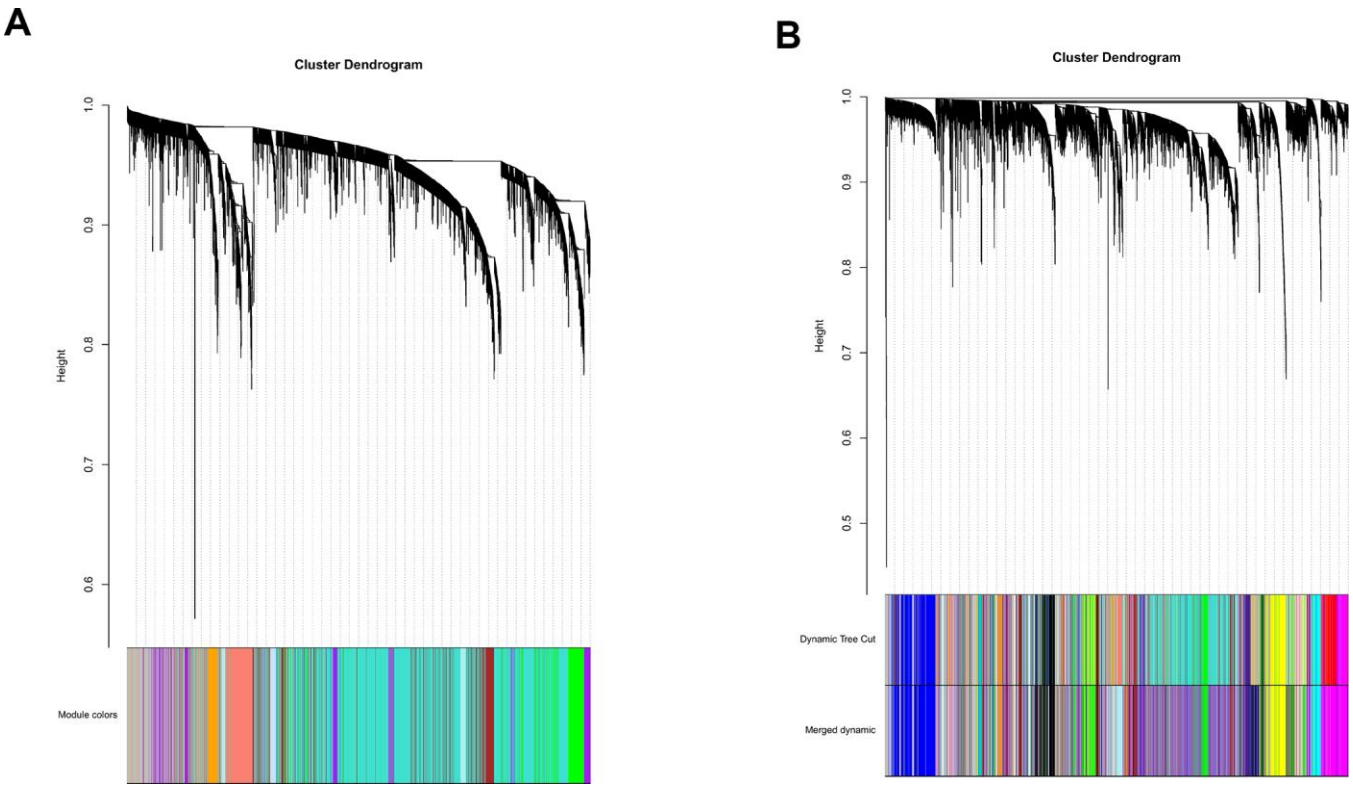

Figure 5. Establishment of co-expression modules of MB. The colored bars below the clustering dendrogram represent the original modules (**A**) and merged modules (**B**). Fifty-five modules were generated by the Dynamic Tree Cut method. Forty-one modules were identified after merging according to the module dissection threshold.

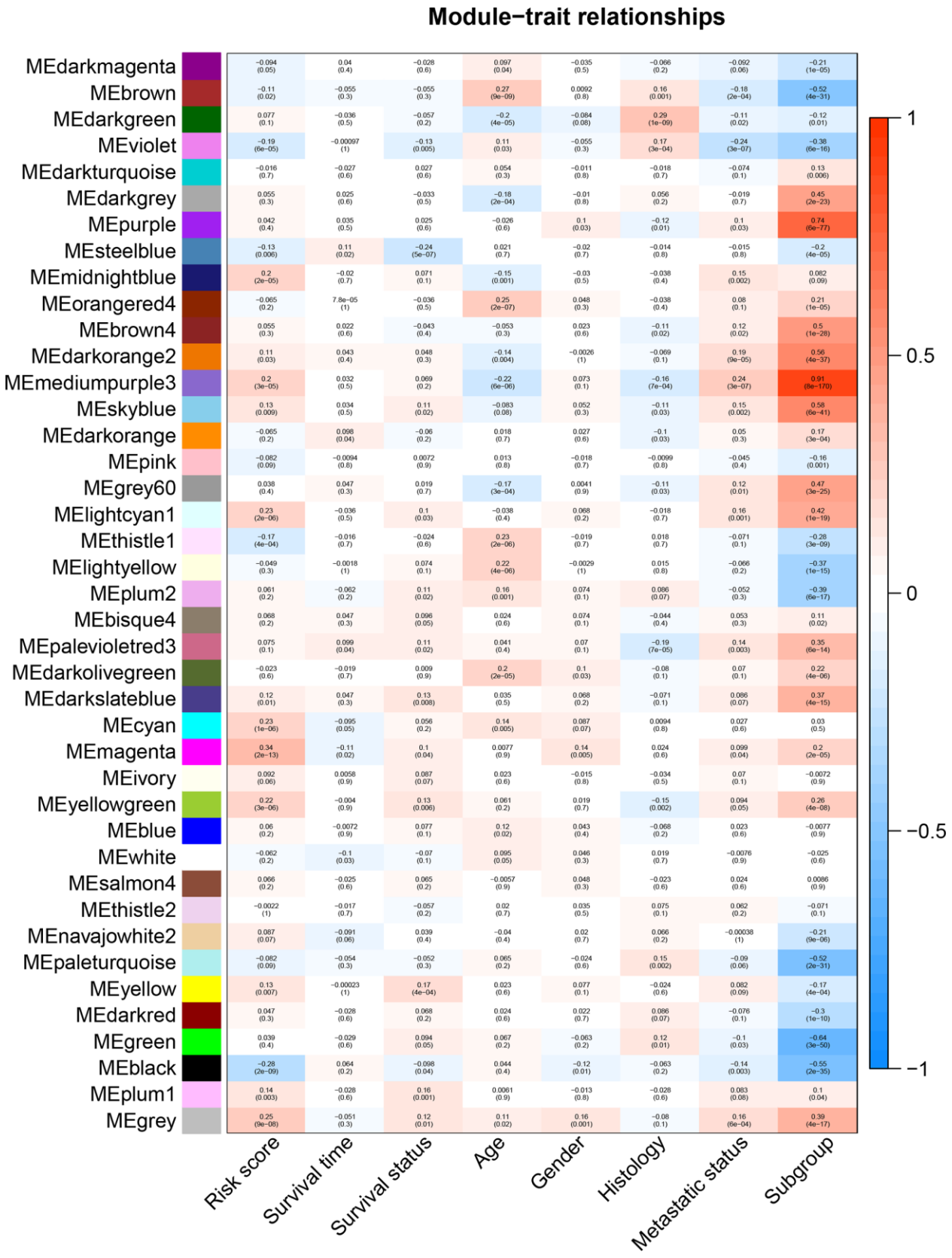

Figure 6. Heatmap of the correlation between gene modules and clinical traits of MB. Each row in the heatmap corresponds to a module, and each column in the heatmap corresponds to a specific clinical characteristic. Each cell contains the corresponding correlation coefficient and *p*-value.

mRNA metastasis-related model based on Lasso Cox regression and the 21-mRNA metastasis-related model based on RSF-VH). According to the predictive performance, we chose the 47-mRNA metastasis-associated model based on Lasso Cox regression with the higher AUC. More recently, machine learning approaches have been successfully utilized for identifying novel diagnostic molecular markers, tracking cancer development, predicting cancer prognosis and monitoring treatment responses to allow the accurate classification of cancer [19]. Lasso and RSF are two common machine learning methods used for building cancer-related prognostic models. Several recent studies showed that in the case of low data dimensions, linear models such as Lasso regression can separate samples more ideally, whereas more complex machine learning models such as random forest are more prone to overfitting, leading to a less precise prediction [20–22]. Thus, considering that the sample size used in the present study is relatively small, Lasso Cox regression may be an appropriate method to establish MB-related prognostic models.

The 47-mRNA metastasis-related model developed in this study categorized MB patients into low- and high-risk groups with significantly different OS outcomes in the discovery and validation cohorts. The clinicians could design the MB patients' treatment plans based on the predicted outcome of the metastasis-associated model to achieve individualized treatment of patients with MB. Strategies should be developed to prevent metastasis or detect MB metastases as early as possible in high-risk MB populations.

Among the 47 genes of the metastasis-related model based on Lasso Cox regression, except for ZIC1 [23, 24], the other 46 genes were either poorly investigated or have not been reported in MB. In addition, five genes involved in the 47-mRNA metastasis-related model, including FABP4 [25, 26], GUCY2C [27–29], MEIS2 [30, 31], POU1F1 [32–34], and SLC25A11 [35] have been reported to be related to the metastasis of other human cancers. Although the roles of these five genes in MB are presently unclear, our results suggest that they deserve further biological and mechanistic investigation. Interestingly, we found that

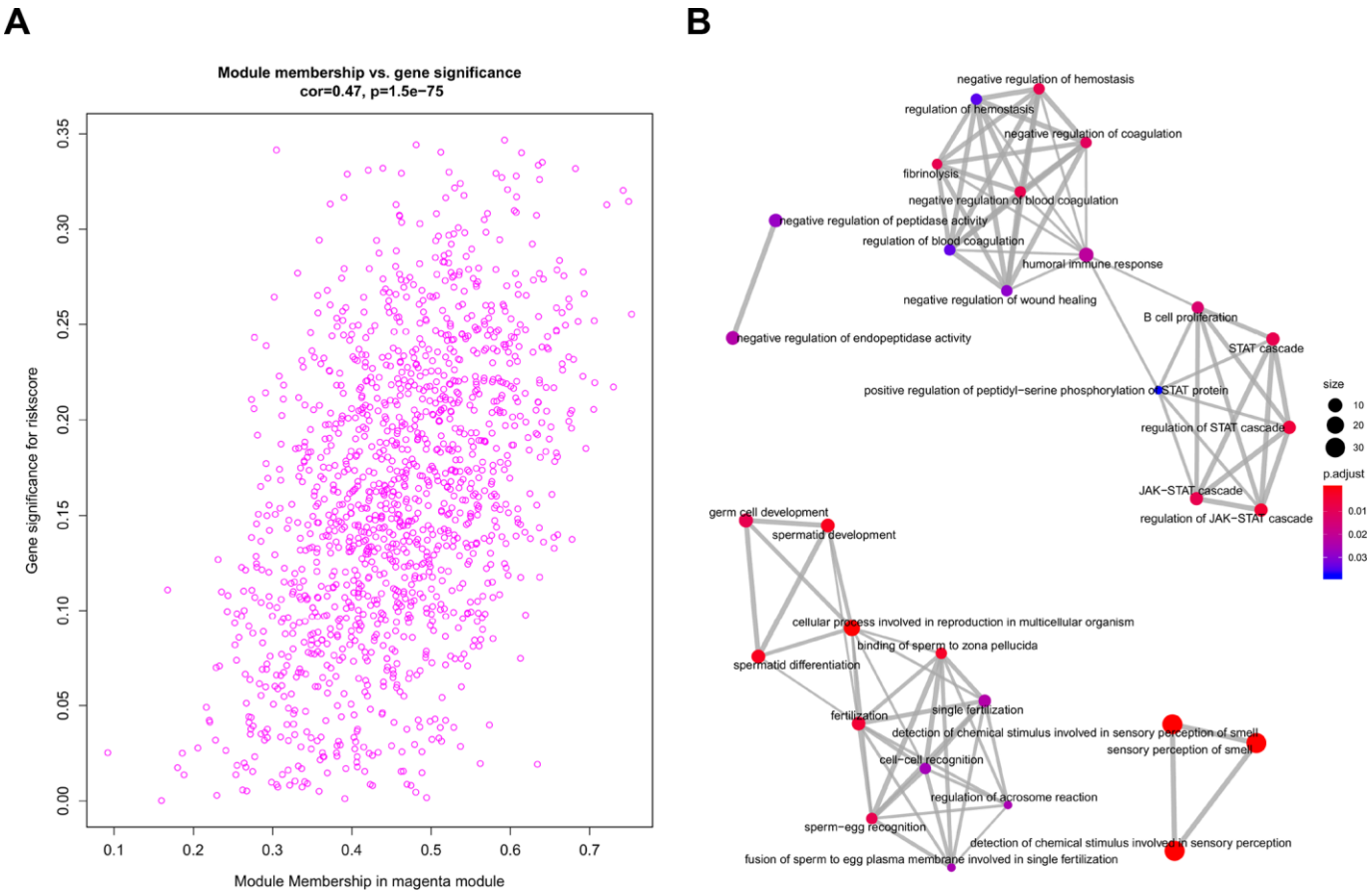

Figure 7. Functional annotation for magenta module. (**A**) Scatter plot of module eigengenes associated with risk score in the magenta module. (**B**) GO analysis involved in the co-expression magenta module.

seven pseudogenes, including RN7SKP226, RN7SL187P, RN7SL432P, RN7SL581P, RNA5SP53, RNU1-75P, and SAA3P, were included in the 47-mRNA metastasis- related model, indicating that expression analysis of these pseudogenes might become a new paradigm for investigating MB mechanisms and discovering prognostic biomarkers in MB. Therefore, the 47-mRNA metastasis-related model

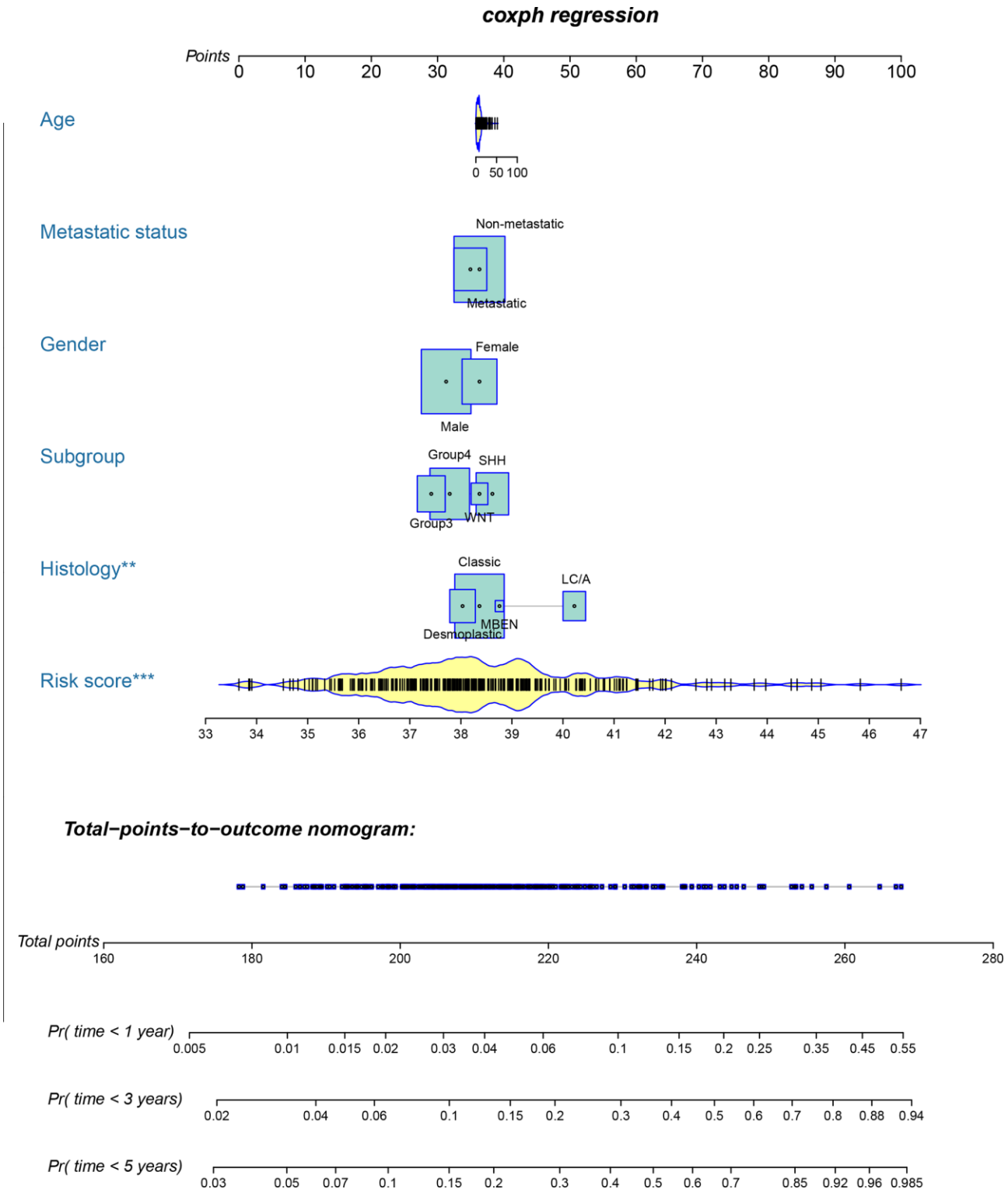

Figure 8. Nomogram for predicting 1-, 3-, and 5-year OS in patients with MB.

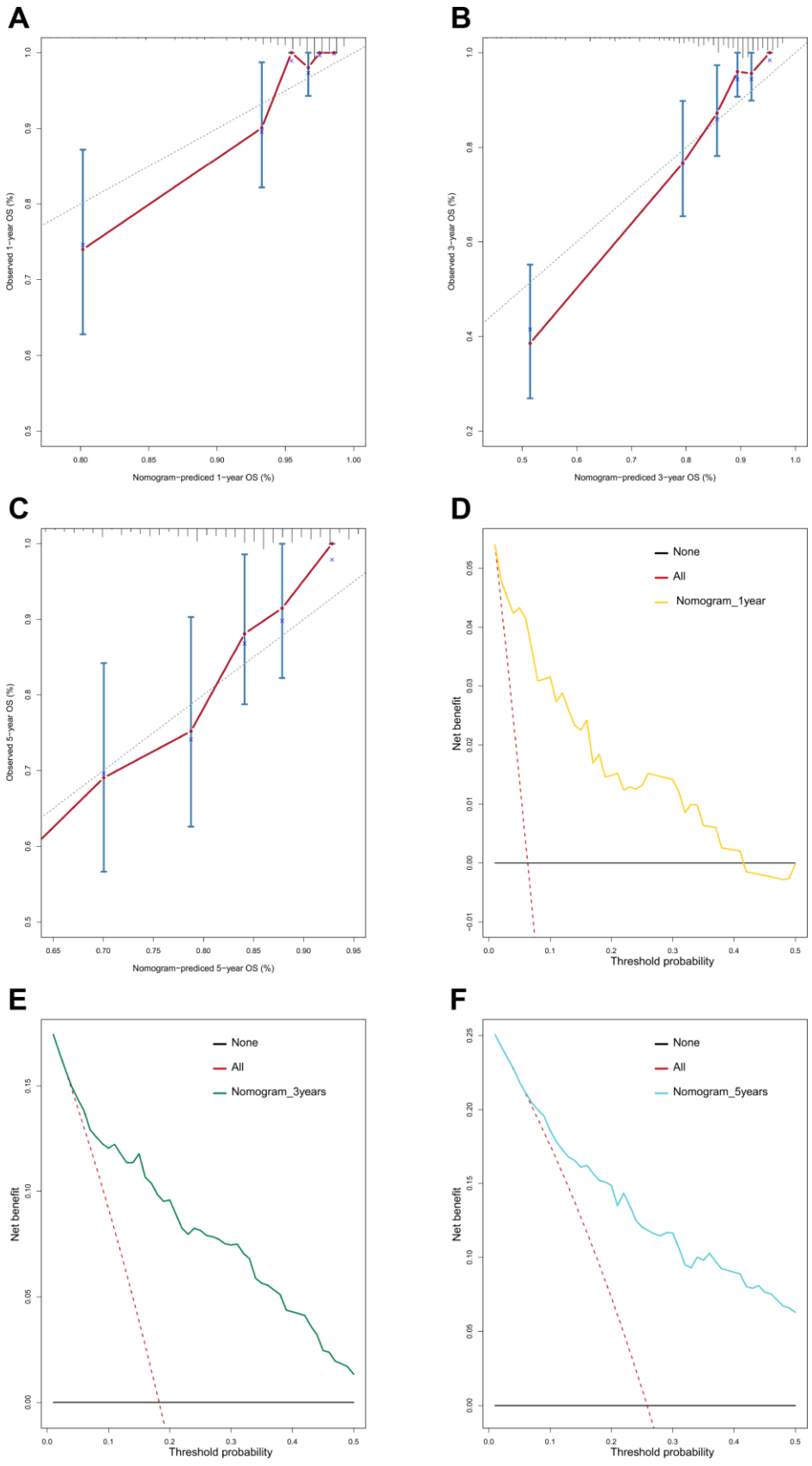

Figure 9. Calibration curves and decision curve analysis (DCA) of the nomogram. Calibration curves of the nomogram for predicting OS at 1 year (**A**), 3 years (**B**), and 5 years (**C**). DCA of the nomogram for predicting OS at 1 year (**D**), 3 years (**E**), and 5 years (**F**).

may offer potential therapeutic targets for MB treatment.

Furthermore, we demonstrated that the 47-mRNA metastasis-related model remained an independent prognostic factor after adjusting for other clinical characteristics. This finding suggests that a comprehensive model incorporating the 47-mRNA metastasis-related model and other clinicopathological factors may achieve a more reliable and favorable OS prediction efficacy for MB patients. Therefore, we built a nomogram that combined the 47-mRNA prognostic model and other clinical features (age, sex, histology, metastatic status, and subgroup). The calibration curves showed that the actual OS corresponded closely with the predicted OS, indicating that the predictive performance of the metastasis-related nomogram was good. DCA demonstrated that the metastasis-related nomogram was clinically useful. According to the results of ROC analyses, this nomogram showed good discrimination ability. Thus, our nomogram could be a promising tool for facilitating patient counselling, treatment decision-making, and follow-up scheduling.

However, our study had some limitations. First, our 47-mRNA metastasis-related model and nomogram need further validation in multicenter, large-scale, prospective studies. Second, functional and mechanistic studies on the 47 genes alone and in combination should be performed to support their clinical application. Third, information on radiotherapy protocols, chemotherapeutic regimens, patient neurological/clinical status, and cytogenetic aberrations is not available in the MB cohort included in the present study. Finally, we constructed the 47-mRNA metastasis-related model based on the gene expression data without considering the DNA methylation, mutation, or other genetic events of genes that likely have an effect on the metastasis of MB.

In summary, our findings indicate that the 47-mRNA metastasis-related prognostic model derived from Lasso-penalized Cox regression might be a reliable and useful tool for predicting OS in MB patients. A nomogram comprising the metastasis-related prognostic model may assist clinicians in selecting personalized therapeutic regimens for MB patients.

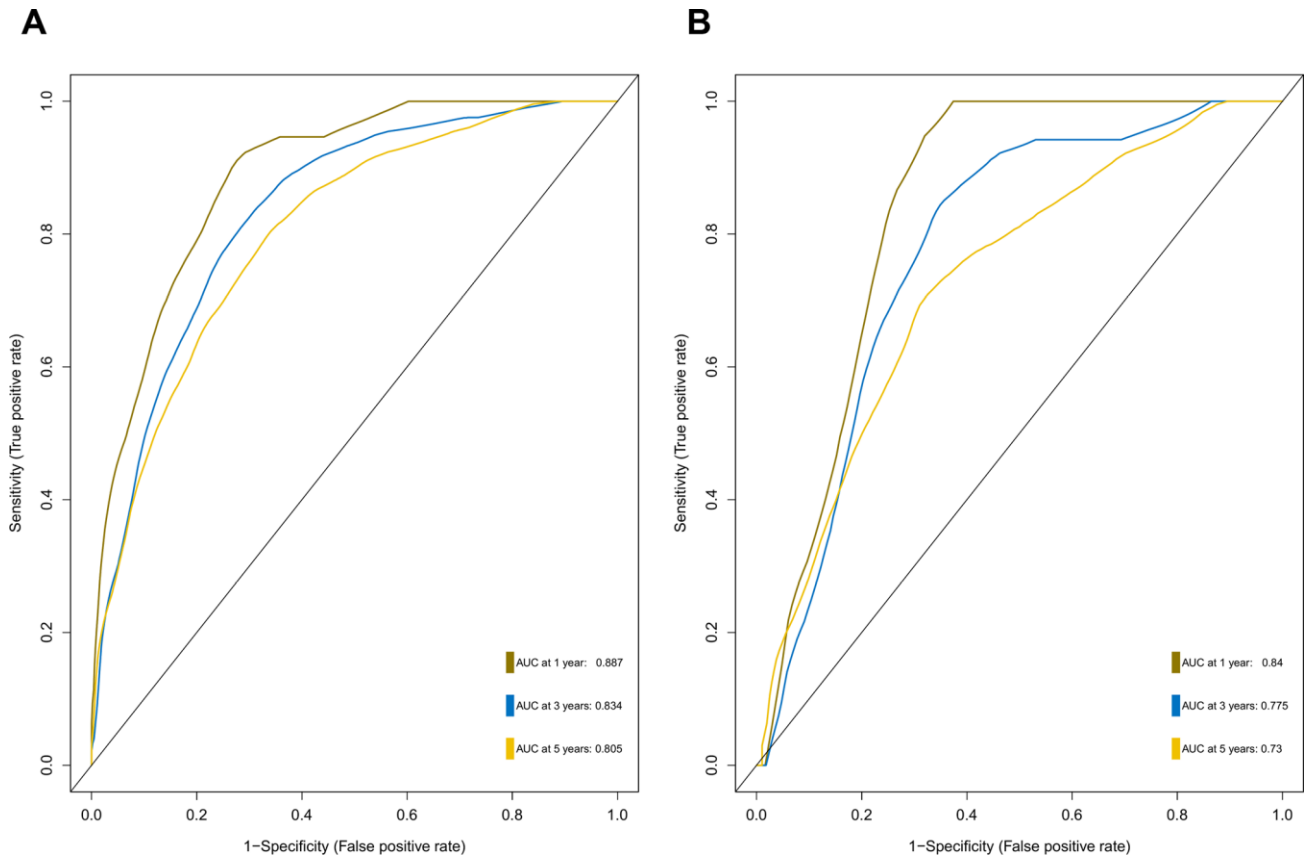

Figure 10. ROC curves for the nomogram in the discovery set (**A**) and validation set (**B**).

MATERIALS AND METHODS

Data source

MB gene expression data were directly downloaded from the GEO GSE85218 dataset (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85218) [36]. The corresponding clinical information was obtained from the supplementary data in the relative literature [36]. Then, MB patients with no information on at least one of the following clinical characteristics were excluded from further analysis: age, sex, metastatic status, histology, survival time, and survival state. Finally, a total of 432 MB patients were included in this study; the median age was 8.00 years (range, 0.24 to 52.00 years) and the median OS was 4.06 years (range, 0.08 to 19.03 years). These 432 MB patients were randomly divided into a discovery set (70%) and validation set (30%) by utilizing the createDataPartition function of the caret package for R software. The distribution of the baseline clinical characteristics of the two groups was balanced (all *P* values > 0.05). The clinical information of these MB patients is summarized in Supplementary Table 5.

Construction and validation of a metastasis-associated gene signature

In the discovery set, DEGs between MB tissues from patients with and without metastatic disease were calculated using the limma package for R software. The DEGs with an adjusted *P* value of < 0.2 were considered for downstream analysis. Next, univariate Cox proportional hazards regression analyses were performed to investigate the association between the OS of MB patients in the discovery set and the expression level of each DEG. In the univariate Cox regression analyses, the genes whose parameter *P* values were < 0.05 were selected for subsequent analysis. To further select primary predictive features, we applied two well-established machine learning algorithms (Lasso-penalized Cox regression and RSF-VH) on the discovery set. Within these two analyses, we subsampled the discovery set at a ratio of 7:3 with 1,000 replacements and selected the prognostic DEGs with repeat occurrence frequencies of more than 100. Then, two metastasis-related risk score staging models (one derived from Lasso-penalized Cox regression and another derived from RSF-VH) were constructed based on a linear combination of the regression coefficient obtained from the multivariate Cox regression analysis (β_i) multiplied by its expression level ($expr_i$). The formula for computing the risk scores of these two prognostic models is described as follows:

$$\text{Risk score} = \sum_{i=1}^{n} (\beta_i * expr_i)$$

The area under the curve (AUC) of the ROC curve was calculated to assess the prediction efficiency of the two metastasis-related prognostic models by using the pROC package for R software. The metastasis-associated risk score model with the higher AUC was kept for subsequent analysis. MB patients in the discovery set were classified into high and low risk score groups according to the optimal risk score cutoff point yielded by utilizing the surv_cutpoint function of the survminer package for R software. We also performed ROC analyses to compare the specificity and sensitivity of OS prediction based on the metastasis-related prognostic model and MB subgroup. Given that the validation set size is small (WNT subgroup, n=12; SHH subgroup, n=35; Group 3, n=17; Group 4, n=65), we only employed the metastasis-related prognostic model to predict survival of patients for each MB subgroup in the discovery set. To test whether the metastasis-related prognostic model was independent of other clinical features (including age, sex, histology, metastatic status, and molecular subgroup), multivariate Cox regression analyses were performed. In the validation set, we used the same risk score formula and cutoff value and divided the MB patients into high- and low-risk groups to test the robustness of the metastasis-related risk score model. The survival difference between the low- and high-risk groups in each set was evaluated by the Kaplan-Meier (K-M) method and compared with the log-rank test.

WGCNA and GO enrichment analysis for discovering the pathways significantly correlated with the 47-mRNA-based risk score model

The expression data of all genes and clinical data (risk score, survival time, survival status, age, gender, histology, metastatic status, and subgroup) in entire MB cohort were included in the WGCNA and analyzed by using the WGCNA package for R software [37]. The all genes were classified into some co-expression modules utilizing an appropriate soft-thresholding parameter β which was calculated by using the pickSoftThreshold function. The Eigengenes for each co-expression module were calculated and correlated modules were merged according to the module dissection threshold. By calculating the correlation between co-expression modules and clinical features by the module-trait relationship of WGCNA, we could screen the module most associated with the clinical trait we were interested in. In this study, the 47-mRNA-based risk score model was selected as the interested clinical trait for subsequent analysis. After the interesting module was chosen, we defined the cor.geneTraitSignificance > 0.2 (the correlation between the gene expression profile and the module eigengene) and the cor.geneModule

Membership > 0.4 (the correlation between a certain clinical phenotype and the gene) as the threshold for screening hub genes in a module. GO enrichment analysis was performed by using the clusterProfiler package for R software.

Establishment and validation of a predictive nomogram

A metastasis-related nomogram was constructed to predict 1-, 3-, and 5-year OS for MB patients by combining the metastasis-related risk score model with clinical variables (age, sex, histology, metastatic status, and subgroup) by using the regplot package for R software. Subsequently, validation, comprising the discrimination ability and predictive accuracy of the nomogram, was performed. Time-dependent ROC curve analyses, which were conducted with the R package "survivalROC", were performed to assess the discrimination ability of the nomogram. The predictive accuracy of the nomogram was determined using the calibration plots, which were generated with the R package "rms". Additionally, decision curve analysis was conducted to assess the clinical usefulness of the nomogram by quantifying the net benefits for a range of threshold probabilities using the "stdca.R" package [38].

Statistical analysis

All statistical analyses were executed by R 3.5.2. Lasso Cox regression analysis and RSF-VH were performed with the R package "glmnet" and "randomForestSRC", respectively. For survival analyses including the K-M method and Cox regression, a two-sided P value < 0.05 was considered statistically significant. The adjusted P values for multiple testing were calculated by using the Benjamini-Hochberg method.

Data accessibility

The data that support the findings of the current study are available from the corresponding authors on reasonable request.

Abbreviations

MB: Medulloblastoma; RSF-VH: Random survival forest-variable hunting; ROC: Receiver operating characteristic; AUC: Area under the curve; WGCNA: Weighted gene co-expression network analysis; GO: Gene ontology; OS: Overall survival; GEO: Gene Expression Omnibus; DEGs: Differentially expressed genes; GS: Gene significance; MM: Module membership; MRI: Magnetic resonance imaging; CSF: Cerebrospinal fluid.

AUTHOR CONTRIBUTIONS

Conceptualization, C.Z., H.L. and M.L; Methodology, K.C., B.H. and S.Y.; Formal analysis: S.X., K.L., K.Z., Q.W. and Z.Z.; Writing—Original Draft Preparation, K.C., B.H. and S.Y. and H.L.; Writing—Review and Editing, C.Z. Y.Z. and L.W.; Supervision, C.Z.; Project Administration, C.Z.; Funding Acquisition, C.Z.

ACKNOWLEDGMENTS

We thank Yikeshu and Xiaoya for the technical assistance.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This work was supported by the Outstanding Leaders Training Program of Pudong Health Bureau of Shanghai (PWR12018-07) and Key Discipline Construction Project of Pudong Health Bureau of Shanghai (PWZxk2017-23, PWYgf2018-05).

Editorial Note

[&]This corresponding author has a verified history of publications using the personal email address for correspondence.

REFERENCES

1. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW. The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. 2016; 131:803–20. https://doi.org/10.1007/s00401-016-1545-1 PMID:27157931

2. Ramaswamy V, Remke M, Bouffet E, Bailey S, Clifford SC, Doz F, Kool M, Dufour C, Vassal G, Milde T, Witt O, von Hoff K, Pietsch T, et al. Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. Acta Neuropathol. 2016; 131:821–31. https://doi.org/10.1007/s00401-016-1569-6 PMID:27040285

3. Aref D, Croul S. Medulloblastoma: recurrence and metastasis. CNS Oncol. 2013; 2:377–85. https://doi.org/10.2217/cns.13.30 PMID:25054581

4. Zeltzer PM, Boyett JM, Finlay JL, Albright AL, Rorke LB, Milstein JM, Allen JC, Stevens KR, Stanley P, Li H, Wisoff JH, Geyer JR, McGuire-Cullen P, et al. Metastasis stage, adjuvant treatment, and residual tumor are prognostic factors for medulloblastoma in children: conclusions from the children's cancer group 921 randomized phase III study. J Clin Oncol. 1999; 17:832–45. https://doi.org/10.1200/JCO.1999.17.3.832 PMID:10071274

5. Packer RJ, Rood BR, MacDonald TJ. Medulloblastoma: present concepts of stratification into risk groups. Pediatr Neurosurg. 2003; 39:60–67. https://doi.org/10.1159/000071316 PMID:12845195

6. Packer RJ, Gajjar A, Vezina G, Rorke-Adams L, Burger PC, Robertson PL, Bayer L, LaFond D, Donahue BR, Marymont MH, Muraszko K, Langston J, Sposto R. Phase III study of craniospinal radiation therapy followed by adjuvant chemotherapy for newly diagnosed average-risk medulloblastoma. J Clin Oncol. 2006; 24:4202–08. https://doi.org/10.1200/JCO.2006.06.4980 PMID:16943538

7. Lannering B, Rutkowski S, Doz F, Pizer B, Gustafsson G, Navajas A, Massimino M, Reddingius R, Benesch M, Carrie C, Taylor R, Gandola L, Björk-Eriksson T, et al. Hyperfractionated versus conventional radiotherapy followed by chemotherapy in standard-risk medulloblastoma: results from the randomized multicenter HIT-SIOP PNET 4 trial. J Clin Oncol. 2012; 30:3187–93. https://doi.org/10.1200/JCO.2011.39.8719 PMID:22851561

8. Fouladi M, Gilger E, Kocak M, Wallace D, Buchanan G, Reeves C, Robbins N, Merchant T, Kun LE, Khan R, Gajjar A, Mulhern R. Intellectual and functional outcome of children 3 years old or younger who have CNS Malignancies. J Clin Oncol. 2005; 23:7152–60. https://doi.org/10.1200/JCO.2005.01.214 PMID:16192599

9. Edelstein K, Spiegler BJ, Fung S, Panzarella T, Mabbott DJ, Jewitt N, D'Agostino NM, Mason WP, Bouffet E, Tabori U, Laperriere N, Hodgson DC. Early aging in adult survivors of childhood medulloblastoma: long-term neurocognitive, functional, and physical outcomes. Neuro Oncol. 2011; 13:536–45. https://doi.org/10.1093/neuonc/nor015 PMID:21367970

10. Ramaswamy V, Taylor MD. Medulloblastoma: from myth to molecular. J Clin Oncol. 2017; 35:2355–63. https://doi.org/10.1200/JCO.2017.72.7842 PMID:28640708

11. Shih DJ, Northcott PA, Remke M, Korshunov A, Ramaswamy V, Kool M, Luu B, Yao Y, Wang X, Dubuc AM, Garzia L, Peacock J, Mack SC, et al. Cytogenetic prognostication within medulloblastoma subgroups. J Clin Oncol. 2014; 32:886–96. https://doi.org/10.1200/JCO.2013.50.9539 PMID:24493713

12. Johnston DL, Keene D, Kostova M, Lafay-Cousin L, Fryer C, Scheinemann K, Carret AS, Fleming A, Percy V, Afzal S, Wilson B, Bowes L, Zelcer S, et al. Survival of children with medulloblastoma in Canada diagnosed between 1990 and 2009 inclusive. J Neurooncol. 2015; 124:247–53. https://doi.org/10.1007/s11060-015-1831-0 PMID:26024655

13. Hur K, Toiyama Y, Schetter AJ, Okugawa Y, Harris CC, Boland CR, Goel A. Identification of a metastasis-specific MicroRNA signature in human colorectal cancer. J Natl Cancer Inst. 2015; 107:dju492. https://doi.org/10.1093/jnci/dju492 PMID:25663689

14. Qu A, Yang Y, Zhang X, Wang W, Liu Y, Zheng G, Du L, Wang C. Development of a preoperative prediction nomogram for lymph node metastasis in colorectal cancer based on a novel serum miRNA signature and CT scans. EBioMedicine. 2018; 37:125–33. https://doi.org/10.1016/j.ebiom.2018.09.052 PMID:30314890

15. Yuan S, Wang J, Yang Y, Zhang J, Liu H, Xiao J, Xu Q, Huang X, Xiang B, Zhu S, Li L, Liu J, Liu L, Zhou W. The prediction of clinical outcome in hepatocellular carcinoma based on a six-gene metastasis signature. Clin Cancer Res. 2017; 23:289–97. https://doi.org/10.1158/1078-0432.CCR-16-0395 PMID:27449498

16. Xie X, Wang J, Shi D, Zou Y, Xiong Z, Li X, Zhou J, Tang H, Xie X. Identification of a 4-mRNA metastasis-related prognostic signature for patients with breast cancer. J Cell Mol Med. 2019; 23:1439–47. https://doi.org/10.1111/jcmm.14049 PMID:30484951

17. Cai W, Li Y, Huang B, Hu C. Esophageal cancer lymph node metastasis-associated gene signature optimizes overall survival prediction of esophageal cancer. J Cell Biochem. 2019; 120:592–600. https://doi.org/10.1002/jcb.27416 PMID:30242875

18. Tang XR, Li YQ, Liang SB, Jiang W, Liu F, Ge WX, Tang LL, Mao YP, He QM, Yang XJ, Zhang Y, Wen X, Zhang J, et al. Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: a retrospective, multicentre, cohort study. Lancet Oncol. 2018; 19:382–93. https://doi.org/10.1016/S1470-2045(18)30080-9 PMID:29428165

19. Londhe VY, Bhasin B. Artificial intelligence and its potential in oncology. Drug Discov Today. 2019; 24:228–32. https://doi.org/10.1016/j.drudis.2018.10.005 PMID:30342246

20. Rohm M, Tresp V, Müller M, Kern C, Manakov I, Weiss M, Sim DA, Priglinger S, Keane PA, Kortuem K. Predicting Visual Acuity by Using Machine Learning in Patients Treated for Neovascular Age-Related Macular Degeneration. Ophthalmology. 2018; 125:1028–36. https://doi.org/10.1016/j.ophtha.2017.12.034 PMID:29454659

21. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, Zhu S, Ye Z. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019; 17:119. https://doi.org/10.1186/s12967-019-1860-0 PMID:30971285

22. Sánchez Fernández I, Sansevere AJ, Gaínza-Lein M, Kapur K, Loddenkemper T. Machine learning for outcome prediction in electroencephalograph (EEG)-monitored children in the intensive care unit. J Child Neurol. 2018; 33:546–53. https://doi.org/10.1177/0883073818773230 PMID:29756499

23. Michiels EM, Oussoren E, Van Groenigen M, Pauws E, Bossuyt PM, Voûte PA, Baas F. Genes differentially expressed in medulloblastoma and fetal brain. Physiol Genomics. 1999; 1:83–91. https://doi.org/10.1152/physiolgenomics.1999.1.2.83 PMID:11015565

24. Lacroix J, Schlund F, Leuchs B, Adolph K, Sturm D, Bender S, Hielscher T, Pfister SM, Witt O, Rommelaere J, Schlehofer JR, Witt H. Oncolytic effects of parvovirus H-1 in medulloblastoma are associated with repression of master regulators of early neurogenesis. Int J Cancer. 2014; 134:703–16. https://doi.org/10.1002/ijc.28386 PMID:23852775

25. Li HY, Lv BB, Bi YH. FABP4 accelerates glioblastoma cell growth and metastasis through Wnt10b signalling. Eur Rev Med Pharmacol Sci. 2018; 22:7807–18. https://doi.org/10.26355/eurrev_201811_16405 PMID:30536325

26. Gharpure KM, Pradeep S, Sans M, Rupaimoole R, Ivan C, Wu SY, Bayraktar E, Nagaraja AS, Mangala LS, Zhang X, Haemmerle M, Hu W, Rodriguez-Aguayo C, et al. FABP4 as a key determinant of metastatic potential of ovarian cancer. Nat Commun. 2018; 9:2923. https://doi.org/10.1038/s41467-018-04987-y PMID:30050129

27. Carrithers SL, Barber MT, Biswas S, Parkinson SJ, Park PK, Goldstein SD, Waldman SA. Guanylyl cyclase C is a selective marker for metastatic colorectal tumors in human extraintestinal tissues. Proc Natl Acad Sci USA. 1996; 93:14827–32. https://doi.org/10.1073/pnas.93.25.14827 PMID:8962140

28. Xiang B, Baybutt TR, Berman-Booty L, Magee MS, Waldman SA, Alexeev VY, Snook AE. Prime-boost immunization eliminates metastatic colorectal cancer by producing high-avidity effector $CD8^+$ T cells. J Immunol. 2017; 198:3507–14. https://doi.org/10.4049/jimmunol.1502672 PMID:28341670

29. Magee MS, Abraham TS, Baybutt TR, Flickinger JC Jr, Ridge NA, Marszalowicz GP, Prajapati P, Hersperger AR, Waldman SA, Snook AE. Human GUCY2C-targeted chimeric antigen receptor (CAR)-expressing T cells eliminate colorectal cancer metastases. Cancer Immunol Res. 2018; 6:509–16. https://doi.org/10.1158/2326-6066.CIR-16-0362 PMID:29615399

30. Bhanvadia RR, VanOpstall C, Brechka H, Barashi NS, Gillard M, McAuley EM, Vasquez JM, Paner G, Chan WC, Andrade J, De Marzo AM, Han M, Szmulewitz RZ, Vander Griend DJ. MEIS1 and MEIS2 expression and prostate cancer progression: a role for HOXB13 binding partners in metastatic disease. Clin Cancer Res. 2018; 24:3668–80. https://doi.org/10.1158/1078-0432.CCR-17-3673 PMID:29716922

31. Xie R, Chen X, Chen Z, Huang M, Dong W, Gu P, Zhang J, Zhou Q, Dong W, Han J, Wang X, Li H, Huang J, Lin T. Polypyrimidine tract binding protein 1 promotes lymphatic metastasis and proliferation of bladder cancer via alternative splicing of MEIS2 and PKM. Cancer Lett. 2019; 449:31–44. https://doi.org/10.1016/j.canlet.2019.01.041 PMID:30742945

32. Seoane S, Martinez-Ordoñez A, Eiro N, Cabezas-Sainz P, Garcia-Caballero L, Gonzalez LO, Macia M, Sanchez L, Vizoso F, Perez-Fernandez R. POU1F1 transcription factor promotes breast cancer metastasis via recruitment and polarization of macrophages. J Pathol. 2019; 249:381–94. https://doi.org/10.1002/path.5324 PMID:31292963

33. Ben-Batalla I, Seoane S, Garcia-Caballero T, Gallego R, Macia M, Gonzalez LO, Vizoso F, Perez-Fernandez R. Deregulation of the pit-1 transcription factor in human breast cancer cells promotes tumor growth and metastasis. J Clin Invest. 2010; 120:4289–302. https://doi.org/10.1172/JCI42015 PMID:21060149

34. Martinez-Ordoñez A, Seoane S, Cabezas P, Eiro N, Sendon-Lago J, Macia M, Garcia-Caballero T, Gonzalez LO, Sanchez L, Vizoso F, Perez-Fernandez R. Breast cancer metastasis to liver and lung is facilitated by pit-1-CXCL12-CXCR4 axis. Oncogene. 2018; 37:1430–44. https://doi.org/10.1038/s41388-017-0036-8 PMID:29321662

35. Buffet A, Morin A, Castro-Vega LJ, Habarou F, Lussey-Lepoutre C, Letouzé E, Lefebvre H, Guilhem I, Haissaguerre M, Raingeard I, Padilla-Girola M, Tran T, Tchara L, et al. Germline mutations in the mitochondrial 2-Oxoglutarate/malate carrier SLC25A11 gene confer a predisposition to metastatic paragangliomas. Cancer Res. 2018; 78:1914–22. https://doi.org/10.1158/0008-5472.CAN-17-2463 PMID:29431636

36. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, Garzia L, Torchia J, Nor C, Morrissy AS, Agnihotri S, Thompson YY, Kuzan-Fischer CM, et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell. 2017; 31:737–754.e6. https://doi.org/10.1016/j.ccell.2017.05.005 PMID:28609654

37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. https://doi.org/10.1186/1471-2105-9-559 PMID:19114008

38. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008; 8:53. https://doi.org/10.1186/1472-6947-8-53 PMID:19036144

SUPPLEMENTARY MATERIALS

Supplementary Tables

Supplementary Table 1. The multivariate Cox regression coefficients of the genes in the 47-mRNA metastasis-related prognostic model.

Gene name	Coefficient
AK7	0.179608
ARHGEF40	-0.62077
ARL1	0.670076
ARSG	0.31238
BACH2	0.387613
C9orf153	0.759155
CAMKK1	-0.42176
CCDC125	-0.50066
COPS7B	0.417233
CPB2	0.490751
EIF2B3	0.244405
FABP4	0.084767
FAM81A	-0.09479
GAGE1	0.085492
GPR126	0.176501
GSDMC	-0.03405
GUCY2C	0.926616
GYG2	0.545351
HIST1H2AE	0.757313
ICOS	0.48971
IDI2	0.382761
IL22	-0.27895
KCNAB3	-0.17314
MAGEB5	0.560332
MDN1	-0.4732
MEIS2	0.233501
NTHL1	1.20003
NUP210L	0.776257
PAPPA2	-0.05873
POLE3	-0.64073
POLN	0.177848
POU1F1	0.06912
PSORS1C1	1.19728

RN7SKP226	0.206219
RN7SL187P	-0.62164
RN7SL432P	0.854117
RN7SL581P	-1.14738
RNA5SP53	0.854826
RNASE9	-0.09789
RNU1-75P	-0.51289
SAA3P	0.462747
SLC25A11	-0.39191
STXBP5L	0.008379
TBCC	0.709829
TBCK	-0.46125
ZIC1	0.077023
ZPBP2	1.18368

Supplementary Table 2. The multivariate Cox regression coefficients of the genes in the 21-mRNA metastasis-related prognostic model.

Gene name	Coefficient
AK7	0.003188
ARHGEF40	-0.81575
FABP4	0.220511
FAM81A	0.262222
GPR126	0.0707
ICT1	-0.0832
IDI2	0.887522
LRRC45	0.913915
LUZP4	0.160284
MEIS2	0.274243
OR2W1	0.888631
PPIL1	-0.62113
PSMG3	0.348928
RN7SKP226	0.395319
RNA5SP318	-0.16243
RNU4-13P	0.072429
RNU5A-2P	0.495492
SCN8A	-0.06987
SLC8B1	0.088436
STXBP5L	-0.20713
TBCC	0.456662

Supplementary Table 3. The repeat occurrence frequencies of the genes in the 47-mRNA metastasis-related prognostic model based on Lasso-penalized Cox regression.

Gene name	Repeat occurrence frequency
NTHL1	616
SLC25A11	596
IDI2	586
ARHGEF40	509
MEIS2	440
ZPBP2	432
CCDC125	407
RN7SKP226	391
RNA5SP53	387
KCNAB3	372
HIST1H2AE	370
RN7SL581P	366
GPR126	355
FABP4	337
FAM81A	334
PSORS1C1	333
BACH2	303
GYG2	289
NUP210L	283
RN7SL187P	251
RNU1-75P	248
POLE3	247
ZIC1	235
AK7	211
SAA3P	208
ARSG	206
CAMKK1	205
EIF2B3	201
POLN	183
TBCC	179
MAGEB5	171
PAPPA2	165
CPB2	155
COPS7B	154
C9orf153	153
STXBP5L	152
GAGE1	142
MDN1	137

POU1F1	137
ARL1	121
ICOS	118
RN7SL432P	115
RNASE9	114
IL22	111
TBCK	109
GSDMC	108
GUCY2C	108

Supplementary Table 4. The repeat occurrence frequencies of the genes in the 21-mRNA metastasis-related prognostic model based on RSF-VH.

Gene name	Repeat occurrence frequency
ICT1	229
ARHGEF40	201
FABP4	164
RNA5SP318	151
IDI2	146
FAM81A	141
STXBP5L	133
LUZP4	127
SCN8A	120
TBCC	120
OR2W1	116
PSMG3	115
LRRC45	113
RNU5A-2P	109
AK7	108
RNU4-13P	107
SLC8B1	106
RN7SKP226	105
GPR126	104
PPIL1	104
MEIS2	102

Supplementary Table 5. Comparison of distribution of clinical characteristics between the discovery and validation set.

Parameter	Discovery set (n= 303)	Validation set (n=129)	*p*-value
Age (mean (SD))	8.78 (7.74)	10.30 (8.21)	0.066
Gender (n (%))			
Female	100 (33.0)	54 (41.9)	0.099
Male	203 (67.0)	75 (58.1)	
Histology (n (%))			
Classic	201 (66.3)	91 (70.5)	0.052
Desmoplastic	54 (17.8)	15 (11.6)	
LC/A	42 (13.9)	15 (11.6)	
MBEN	6 (2.0)	8 (6.2)	
Metastatic status (n (%))			
Non-metastatic	214 (70.6)	89 (69.0)	0.822
Metastatic	89 (29.4)	40 (31.0)	
Subgroup (n (%))			
Group 3	64 (21.1)	17 (13.2)	0.197
Group 4	129 (42.6)	65 (50.4)	
SHH	87 (28.7)	35 (27.1)	
WNT	23 (7.6)	12 (9.3)	

Abbreviations: SD, standard deviation; LC/A, large cell/anaplastic; MBEN, medulloblastoma with extensive nodularity.