Identification and validation of a three-gene signature as a candidate prognostic biomarker for lower grade glioma

Kai Xiao [1], **Qing Liu** [1], **Gang Peng** [1], **Jun Su** [1], **Chao-Ying Qin** [1], **Xiang-Yu Wang** [Corresp. 1]

[1] Department of Neurosurgery，Xiangya hospital of Central South University, Xiangya hospital, changsha, hunan, CHN

Corresponding Author: Xiang-Yu Wang
Email address: wxyMGH@gmail.com

Background Lower grade glioma (LGG) are a heterogeneous tumor that may develop into high-grade malignant glioma seriously shortens patient survival time. The clinical prognostic biomarker of lower-grade glioma is still lacking. The aim of our study is to explore novel biomarkers for LGG that contribute to distinguish potential malignancy in low-grade glioma, to guide clinical adoption of more rational and effective treatments. **Methods** The RNA-seq data for LGG was downloaded from the UCSC Xena and Chinese Glioma Genome Atlas (CGGA). By robust likelihood-based survival model, LASSO regression and multivariate Cox regression analysis, we developed a three-gene signature and established a risk score to predict the prognosis of patient with LGG. The three-gene signature was an independent survival predictor compared to other clinical parameters. Based on the signature related risk score system, stratified survival analysis was performed in patients with different age group, gender, and pathologic grade. The prognostic signature was validated in CGGA dataset. Finally, Weighted Gene Co-expression Network Analysis (WGCNA) was carried out to find the co-expression genes related to the member of the signature and enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway were conducted for those co-expression network. To prove the superiority of the model, time-dependent ROC curves of our model and other models are constructed. **Results** In this study, a three-gene signature (WEE1, CRTAC1, SEMA4G) was constructed. Based on the model, the risk score of each patient was calculated with LGG (low-risk vs. high- risk, hazard ratio[HR]=0.198, 95%CI= 0.120-0.325) and patients in the high-risk group had significantly poorer survival results than those in the low-risk group. Furthermore，the model was validated in CGGA dataset. Lastly, by WGCNA, we constructed the co-expression network of the three genes and conducted the enrichment of GO and KEGG. Our study identified a three-gene model that showed better performance in predicting the 1-, 3- and 5-year survival of LGG patients compared to other models and may be promising independent biomarker of LGG.

Identification and validation of a three-gene signature as a candidate prognostic biomarker for lower grade glioma

Kai Xiao[1], Qing Liu[1], Gang Peng[1], Jun Su[1], Chao-Ying Qin[1], Xiang-Yu Wang[1]

1. Department of Neurosurgery, Xiangya hospital of Central South University, Changsha 410008, Hunan Province, People's Republic of China

Corresponding Author:

Xiang-Yu Wang, wxyMGH@gmail.com

Xiangya road, kaifu district, Changsha, Hunan province, 410008, People's Republic of China

Abstract

Background

Lower grade glioma(LGG) are a heterogeneous tumor that may develop into high-grade malignant glioma seriously shortens patient survival time. The clinical prognostic biomarker of lower-grade glioma is still lacking. The aim of our study is to explore novel biomarkers for LGG that contribute to distinguish potential malignancy in low-grade glioma, to guide clinical adoption of more rational and effective treatments.

Methods

The RNA-seq data for LGG was downloaded from the UCSC Xena and Chinese Glioma Genome Atlas（CGGA）. By robust likelihood-based survival model、LASSO regression and multivariate Cox regression analysis, we developed a three-gene signature and established a risk score to predict the prognosis of patient with LGG. The three-gene signature was an independent survival predictor compared to other clinical parameters. Based on the signature related risk score system, stratified survival analysis was performed in patients with different age group, gender, and pathologic grade. The prognostic signature was validated in CGGA dataset. Finally, Weighted Gene Co-expression Network Analysis (WGCNA) was carried out to find the co-expression genes related to the member of the signature and enrichment analysis of Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) pathway were conducted for those co-expression network. To prove the superiority of the model, time-

dependent ROC curves of our model and other models are constructed.

Results

In this study, a three-gene signature(WEE1, CRTAC1, SEMA4G)was constructed. Based on the model, the risk score of each patient was calculated with LGG(low-risk vs. high- risk, hazard ratio[HR]=0.198, 95%CI= 0.120-0.325) and patients in the high-risk group had significantly poorer survival results than those in the low-risk group. Furthermore, the model was validated in CGGA dataset. Lastly, by WGCNA, we constructed the co-expression network of the three genes and conducted the enrichment of GO and KEGG. Our study identified a three-gene model that showed better performance in predicting the 1-, 3- and 5-year survival of LGG patients compared to other models and may be promising independent biomarker of LGG.

Keyword: low-grade glioma, prognosis, WGCNA, risk score, robust likelihood-based survival model, biomarker, signature, better performance

Introduction

With the development of sequencing and bioinformatics technologies, accumulating studies have revealed that different patients may be similar in glioma grade but differ greatly in molecular characteristics, clinical prognosis, and treatment response. So many central nervous system tumors were named according to molecular parameters and histopathologic diagnosis, especially gliomas, ependymomas, and medulloblastomas in the 2016 revision of the WHO classification(Zhang et al. 2019b). As we know, some molecular markers, such as MGMT(O6-methylguanine DNA methyltransferase)(Binabaj et al. 2018), IDH(isocitrate dehydrogenase) (Kwon et al. 2019), EGFR(epidermal growth factor receptor)(Chistiakov et al. 2017), and PTEN(phosphatase and tensin homolog)(Koshiyama et al. 2017)that have contributed to personalized therapeutic approaches and targeted anti-glioblastoma therapies have been routinely tested in glioblastoma patients clinically. However there are few specific clinical indicators and therapeutic targets for LGGs compared to glioblastoma, so it is urgent to elucidate the mechanism of glioma development and progression, which can provide potential treatment targets for LGGs.

In this study, gene expression RNAseq data and corresponding clinical information of LGG

patients were downloaded from UCSC Xena(https://xenabrowser.net/hub/) and Chinese Glioma Genome Atlas（CGGA,http://www.cgga.org.cn/）.By analyzing data from UCSC Xena using robust likelihood-based survival model and Cox regression, we developed a three-gene signature that provides effective survival risk stratification of patients with LGG and validated the signature in the CGGA dataset. These results demonstrate the potential of the three-gene signature for survival prediction of patients with LGG and provide new potential molecular treatment targets for LGGs.

Materials and Methods

Dataset of Patients with LGG

The LGGs RNA sequencing（RNAseq）data and corresponding clinical information were downloaded from The Cancer Genome Atlas (TCGA) hub by the University of California, Santa Cruz, Xena browser(https://xenabrowser.net/hub/) and CGGA data portal(http://www.cgga.org.cn/) respectively. The TCGA RNAseq data (level 3) shows the gene-level transcription estimates, as in log2(x+1) transformed RSEM normalized count. The CGGA data displays the gene expression level as fragments per kilobase transcriptome per million fragments(FPKM),which has been standardized. Expressed gene defined only if its expressed level is larger than 0 at half of samples. Only patients with a clear information of survival and detailed history of radiotherapy and chemotherapy/molecular therapy were included in the study. Finally,456 cases from TCGA dataset and 159 cases from the CGGA dataset were included in the training set and validation set respectively. Table 1 summarized the clinical characteristics and therapy information of the training set and validation set. The workflow presentation of this study is shown in Figure 1.

Identification of survival-related genes and construction of the prognostic model

By using the rbsurv package in R, a robust likelihood-based survival model was conducted to identify survival-related genes(Cho et al. 2009). The rbsurv package is a software program, which selects survival-associated genes based on the partial likelihood of the Cox model and adopts a cross-validation approach for robustness. According to the description of the rbsurv package, prior

gene selection such as univariate survival modeling can be performed if necessary and the univariate survival modeling can be performed in this software program. Compared to the survival modeling without an adjustment of risk factors, the robust likelihood-based survival model can improve the ability to discover truly survival-associated genes by modeling genes after adjusting for certain risk factor. Thus, we directly conduct a robust likelihood-based survival model to screen for the prognostic genes. The robustness test was performed on 20530 genes and 456 samples. After 10 iterations, 29 prognostic related genes were selected. With the help of glmnet and survival package in R, least absolute shrinkage and selection operator (LASSO)regression and the multivariable Cox proportional hazard regression method were used to further identify the survival-related prognostic model. The same approach was used to identify gene signatures for endometrial carcinoma (Ouyang et al. 2019). At last, three prognostic survival-related genes that were independent survival predictors and their regression coefficients were obtained at a threshold of $P < 0.05$. Based on the median expression value of each survival-related gene, we dichotomized 456 LGGs patients into low and high expression groups and compare the survival rate between the two groups by Kaplan-Meier plots and Log-rank test. According to the estimated regression coefficients, a prognostic risk score for each patients was then calculated(Wang et al. 2019). The risk score=(0.4470×expression level of WEE1)＋(-0.1530×expression level of CRTAC1)＋(-0.3723×expression level of SEMA4G). With the three-gene signature, 456 LGGs patients were divided into high-risk and low-risk groups with the median risk score as the cut-off value. Kaplan-Meier curves were performed to estimate and compare the survival for TCGA LGGs patients with a high score or a low score. The receiver operating characteristic(ROC) curve and area under the curve(AUC) were applied to evaluate the prediction accuracy of the risk score model. Furthermore stratified survival analysis was performed in patients with different age group(younger,old) , gender(male,female), and pathologic grade(G2,G3).

Univariate and multivariate Cox hazard regression analysis were conducted for the potential prognostic factors such as age group(younger vs. old), gender(male vs. female), pathologic

grade(G2 vs. G3), radiation therapy(Yes vs. No), molecular therapy(Yes vs. No) and risk score(High vs. Low).

Validation of the prognostic model in the CGGA

The prognostic model was validated in the CGGA mRNAseq_325 cohort. Only patients with a clear information of survival, detailed history of radiotherapy and chemotherapy were included in the study. Finally, 159 cases from the CGGA mRNAseq_325 cohort were included in the validation set.

Exploring co-expression genes by WGCNA

To explore the regulatory network of the three genes, WGCNA was performed in training set by the R package WGCNA(Langfelder & Horvath 2008). The top 50% variance of genes were selected for WGCNA. In other words, WGCNA based on 456 samples and 10256 genes. First, RNAseq data were filtered to reduce outliers. Using the absolute value of the correlation between the expression levels of transcripts, a co-expression similarity matrix was constructed. Then, the co-expression similarity matrix was transformed to the adjacency matrix by choosing 9 as a soft threshold. Co-expression gene module was established by the topological overlap measure. In order to identify the significance of each module, gene significance(GS) was calculated to estimate the correlation between genes and sample traits. Module significance(MS) was defined as the average GS within modules and was calculated to measure the correlation between modules and sample traits(vital status). Finally, the “vital status” related modules that contain the 3 genes as members and genes belong to such modules were identified. Genes interacted with those three genes were screened and the co-expression network was constructed by Cytoscape software(Shannon et al. 2003).

Functional enrichment analysis

Using Enrichment analysis of Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) pathway were conducted via the clusterProfiler package in R language(Yu et al. 2012) for those genes that belong to the “vital status” related modules associated with the three genes. Benjamini-Hochberg (BH)-adjusted p-value <0.05 were considered significant.

Results

Three prognostic genes were identified in TCGA dataset and validated in CGGA dataset

456 patients and 20530 genes were included in the TCGA-LGG to train the prognostic model. The robust likelihood-based survival model found 29 survival-related genes, 13 genes were obtained through LASSO Cox method(Fig.2). We further reduced the dimensionality of these high-dimensional data by multivariate Cox proportional hazard regression model. Finally, three genes that were independent survival predictors were identified as survival prediction signature. Those three genes included in the model were WEE1, SEMA4G, CRTAC1. It has been reported that WEE1 is closely related to the growth, invasion and migration of glioma(Wu et al. 2019). Currently, there is no study revealing the role of SEMA4G and CRTAC1 in gliomas. After calculating the risk score, patients were divided into a high- and low-risk group based on the median cut-off point of the risk score. The three-gene signature risk score distribution is shown in Fig 3A. Besides, the relationship between risk score and the status of the LGGs was calculated(Fig.3B). As shown in the heat map of the Figure 3C, a remarkable high expression was noted for WEE1 in the high-risk group, while a lower expression was observed for the other genes in the high-risk group.(Fig.3C). Patients in the high-risk group were significantly worse off the overall survival time compared to the low-risk group($P<0.0001$)(Fig.4A). The area under ROC curve of the signature for 1-,3- and 5-year overall survival was 0.904,0.878 and 0.805, respectively, in training set. (Fig.4B). A similar result can be noted in the validation dataset(Fig.4C). The area under ROC curve of the signature for 1-,3- and 5-year overall survival was 0.783,0.813 and 0.813, respectively, in validation set. (Fig.4D).Moreover, the predicting power of the risk score model was not decreased in subgroup analysis for age group(younger,P=0.00012;old,$P<0.0001$) , gender(male,$P<0.0001$;female,$P<0.0001$), and pathologic grade(G2,P=0.00013;G3,$P<0.0001$) in the training set(Fig.5A-5F). The same trend can be observed in the validation dataset(Fig.6A-6F).For the WEE1, the member of high expression group had significantly shorter survival than those in low expression group($P<0.0001$)(Fig.7A). For the SEMA4G and CRTAC1, the member of high expression group had

significantly longer survival than those in low expression group(P＜0.0001)(Fig.7B-7C). The expression level of WEE1 was significantly higher in grade III compared to grade II(P＜0.0001), while the other are opposite(Fig.8A). These results can also be verified in the validation dataset(Fig.7D-7F、8B).

Multivariate Cox proportional hazard regression demonstrated that age group (HR=0.274, P=2.21E-09), pathologic grade (HR=2.49, P=0.00011) and risk score(HR=0.198, P＜0.000000000168) were independent prognostic factors in the training set, while pathologic grade(HR=3.799, P=0.00000151), 1p19q status(HR=4.566, P=0.0000388), radiation therapy (HR=0.524, P=0.046), and risk score (HR=0.415, P=0.000653)were independent prognostic factors in validation dataset(Table 2) .

Calculation of module-trait correlation in LGGs and module visualization of the network connections

Using the R package WGCNA, gene modules were identified based on the top 50% variance of genes. To analyze the relationship between gene modules and sample clinical information, we used the module eigengene(ME) as the overall gene expression level of the corresponding modules and calculated correlations with clinical phenotypes, for example, vital status. we obtained 16 gene modules (SupplementFigure1.A-D) with size ranging from 31 to 1501 genes. We assigned each co-expression module an arbitrary color for reference: black, blue, brown, cyan, green, greenyellow, lightcyan, magenta, midnightblue, pink, purple, red, salmon, tan, turquoise, and yellow. These modules contained 449, 1352, 850, 46, 519, 91, 31, 201, 43, 336, 135, 462, 51, 90, 1501 and 845 genes, respectively. As a single group, the non-co-expressed group designated as ‘grey’ based on the WGCNA developer’s rationale. Vital status related modules, such as yellow, green, black modules that contain the 3 genes as members and genes belong to such modules were screened(SupplementFigure1.D). Finally, 32 genes were discovered to be co-expressed with CRTAC1, 181 genes were co-expressed with WEE1, 6 genes with SEMA4G. We exported the screened genes and three prognostic survival-related genes into Cytoscape and constructed the co-expression network(Fig.9).

GO and KEGG analysis of screened genes interacted with three-gene signature

For the "biological processes"(BP), chromosome segregation, nuclear division, mitotic nuclear division, organelle fission, mitotic sister chromatid segregation were the commonly enriched categories(Fig.10A). For the "cellular component"(CC), the enriched categories were correlated with condensed chromosome, chromosome/centromeric region, chromosomal region, kinetochore, condensed chromosome/centromeric region(Fig.10B). For the "molecular function" (MF), these screened genes mainly enriched in microtubule binding, tubulin binding, histone kinase activity, DNA-dependent ATPase activity, protein serine/threonine kinase activity(Fig.10D).KEGG pathway enrichment analysis suggested that cell cycle was the most important pathway for these selected genes. The following pathway also involved many screened genes, including, oocyte meiosis, progesterone-mediated oocyte maturation, fanconi anemia pathway, homologous recombination(Fig.10C).Additionally, For the Gene Ontology analysis, these 3 co-expression gene modules (yellow, green, black) enriched results can been seen in supplemental files(SupplementalFigure2、3).

A comparison between our and other models

Recently, Chen X.P et al reported a model containing 3 genes(EMP3、GSX2、EMILIN3) based on integrative analysis of DNA methylation and gene expression in TCGA dataset(Zeng et al. 2018).Chuang Zhang et al also reported a 4-gene(EMP3、GNG12、KIF2C、IFI44) prognostic signature based on genes encodes by chr1p/19q(Zhang et al. 2019a). To compared the prognostic values of our prognostic signature and their model, we performed time-dependent ROC curve analysis in our model and other models based on the risk score calculated by the regression coefficients which obtained by themselves and the expression level of members in their signature showed in the TCGA dataset, which has a larger number of samples compared to CGGA dataset, that might be able to ensure the credibility of the comparison results. The results exhibited that our model displayed a better predictive value in predicting 1-, 3- and 5-year overall survival compared to other models, especially in 1-, 3-year overall survival(Fig.11A-11C). In other words, our 3-gene model had a better efficiency in predicting both short- and long-term

prognosis.

Discussion

From the perspective of traditional pathology, the diagnosis of low-grade glioma depends on pathological type and pathological grade. With the development of sequencing technology, molecular biomarkers for the diagnosis of LGG have attracted widespread attention(Cancer Genome Atlas Research et al. 2015). Prognostic factors for the low-grade glioma that are well known include IDH mutations(Batsios et al. 2019), 1p/19q co-deficiency(Zhang et al. 2019a), ATRX mutation(Ren et al. 2019), TERT promoter mutations(Chan et al. 2015), CIC loss(Sahm et al. 2012), FUBP1 loss(Sahm et al. 2012) and PTEN loss(Sabha et al. 2014) and the above prognostic marker contribute to clinicians to understand the mechanism of low-grade gliomas. The complex pathogenesis of LGG encourages us to explore more prognostic markers for further understand it and develop an efficient treatment.

In this study, we identified three genes that were closely correlated with LGG prognosis. Considering the differentially expressed genes(DEGs) between tumor and normal tissue might not be associated with survival(Liu et al. 2019) and the univariate survival modeling can be performed in rbsurv package, the robust likelihood-based survival model was performed using the rbsurv package in R as the first step instead of screening for DEGs and conducting the univariate Cox regression. LASSO and Cox proportional hazard regression model are widely used to generate prognostic genes in the context of high dimensional data, thus were adopted in subsequent analysis. Compared to a single predictive biomarker, integrating multiple biomarkers into a signature is believed to be more predictive. The risk score calculated by the risk model was considered to have good predictive capabilities and was demonstrated to be an independent prognostic factor after adjusting the effects of age, sex, tumor grade, molecular therapy and radiation therapy. The risk score was confirmed to be effective in different age groups、gender and pathologic grade. Regardless of the training set or the validation set, the AUC value of 1-, 3- and 5-year was greater than 0.75. The pathologic grade and the risk level were confirmed to be independent prognostic factors both in training set and validation set.

In order to construct a co-expression network of the three genes, WGCNA was used in the training set. We found the survival related modules to which these three genes belong, and extracted the genes of these three modules to construct a co-expression network. Finally, 32 genes were discovered to be co-expressed with CRTAC1, 181 genes were co-expressed with WEE1, 6 genes with SEMA4G. The co-expression network of the three genes is visualized by Cytoscape in Figure 9.

Based on the result of GO and KEGG enrichment analysis of these co-expression genes, "condensed chromosome" was the most significant enrichment in CC. Coincidentally, Rebecca C et al. found that interference with chromatin condensation results in failure to fully activate DNA damage response(Burgess et al. 2014) and the DNA damage response triggers multiple cellular events including activation of DNA repair pathway, arrest of the cell cycle to allow time for repair, and, in certain cases, initiation of senescence or apoptosis programs(Ciccia & Elledge 2010). For the BP category, chromosome segregation was the most enrichment and research has proven that chromosome instability contributes to the development of genetic heterogeneity in tumors and allows the outgrowth of tumorigenic cells with advantageous karyotypes(Conde et al. 2017). Regarding the MF category, microtubule binding was the most influential and the drug targeted microtubule was proven effective in glioma. For example, the drug EM011 functions by disrupting microtubule dynamics and modules several oncogenic mediators causing a decrease in cell viability, proliferation and migration/invasion in the astrocytoma cell lines(Ajeawung et al. 2013). For KEGG pathway enrichment analysis, cell cycle was the most significant pathway. Stephen D has explained that signaling pathway converge on the cell cycle machinery to regulate developmental genes and execute cell fate decisions(Dalton 2015).

The three-gene signature provided a wealth of potential biological and therapeutic information about LGG.WEE1(WEE1 G2 checkpoint kinase), located on the short arm of human chromosome 11(11p15.4), encodes a nuclear protein, which is a tyrosine kinase belonging to the Ser/Thr family of protein kinases. The protein catalyzes the inhibitory tyrosine phosphorylation

of CDC2/cyclin B kinase, and appears to coordinate the transition between DNA replication and mitosis by protecting the nucleus from cytoplasmically activated CDC2 kinase. WEE1 has been confirmed that its protein expression increases with malignancy grade(Music et al. 2016). Moreover, patients with high WEE1 expression had poor survival than did patients with low WEE1 expression in grade III gliomas(Music et al. 2016).CRTAC1, cartilage acidic protein 1, a novel human marker which allowed discrimination of human chondrocytes from osteoblasts and mesenchymal stem cells in culture can be divided into CRTAC1-A and CRTAC1-B two subtypes according to the last exon. Previous study found that inhibition of CRTAC1 reduces ultraviolet B irradiation induced-apoptosis through P38 mitogen-activated protein kinase and jun Amino-Terminal kinase pathway(Ji et al. 2016). It means that the relationship between the expression of CRTAC1 and apoptosis is positively correlated after ultraviolet B irradiation. To some extent, this is consistent with our finding that CRTAC1 high expression prolongs survival time in LGG patients. However, its detailed mechanism in LGG remains to be further explored. Semaphorins are a large family of conserved secreted and membrane associated proteins which possess a semaphoring(Sema) domain and a PSI domain in the N-terminal extracellular portion. Based on sequence and structural similarities, semaphorins are put into eight classes: invertebrates contain classes 1 and 2, viruses have class 8, and vertebrates contain class 3-7. Semaphorins serves as axon guidance ligands via multimeric receptor complexs, some containing plexin proteins. Semaphorins and Plexins are cognate ligand-receptor families that regulate important steps during nervous system development(Maier et al. 2011). A low-expression of SEMA4G was detected in colorectal cancer tissues compared with normal tissues. It means that SEMA4G might be a tumor suppressor gene related to colorectal cancer(Wang et al. 2008). However, little work has been done to elucidate the role of SEMA4G in glioma. Our study demonstrated that SEMA4G was significantly down-regulated in grade III patients compared to grade II and the high-expression of SEMA4G was associated with a good prognosis in LGG patients. Further work is needed to explore its functions in LGG. To sum up, the three-gene signature could predict LGG survival based on a risk score model. We firmly believed that

these genes are potential prognostic markers or therapeutic targets for LGG patients. Nevertheless, the molecular mechanisms how the three-gene signature affected the prognosis of LGG patients should be further elucidated by a series of experiments.

Conclusion

In conclusion, Our study identified a 3-gene model that showed better performance in predicting short- and long-term survival of LGG patients compared to other models. Moreover, our finding provided new insights into the pathogenesis and prognosis of LGG.

Acknowledgments

We thank editors for constructive comments on a previous version of this manuscript.

Additional information and declarations

Funding statement

This work was funded by grants from the National Nature Science Foundation of China(NO 81801908)and the National Key Technology Research and Development Program of the Ministry of Science and Technology of China(NO 2014BAI04B01, NO 2014BAI04B02 and NO 2015BAI12B04). There was no additional external funding received for this study.

Competing Interests

The authors report no conflicts of interest in this work.

Author Contributions

All authors contributed to data analysis, drafting and revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Availability of data and material

The following information was supplied regarding data availability:

All relevant data have been provided in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found at Supplemental Files.

References

Ajeawung NF, Joshi HC, and Kamnasaran D. 2013. The microtubule binding drug EM011 inhibits the growth of paediatric low grade gliomas. *Cancer Lett* 335:109-118. 10.1016/j.canlet.2013.02.004

Batsios G, Viswanath P, Subramani E, Najac C, Gillespie AM, Santos RD, Molloy AR, Pieper RO, and Ronen SM.

2019. PI3K/mTOR inhibition of IDH1 mutant glioma leads to reduced 2HG production that is associated with increased survival. *Sci Rep* 9:10521. 10.1038/s41598-019-47021-x

Binabaj MM, Bahrami A, ShahidSales S, Joodi M, Joudi Mashhad M, Hassanian SM, Anvari K, and Avan A. 2018. The prognostic value of MGMT promoter methylation in glioblastoma: A meta-analysis of clinical trials. *J Cell Physiol* 233:378-386. 10.1002/jcp.25896

Burgess RC, Burman B, Kruhlak MJ, and Misteli T. 2014. Activation of DNA damage response signaling by condensed chromatin. *Cell Rep* 9:1703-1717. 10.1016/j.celrep.2014.10.060

Cancer Genome Atlas Research N, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, Cooper LA, Rheinbay E, Miller CR, Vitucci M, Morozova O, Robertson AG, Noushmehr H, Laird PW, Cherniack AD, Akbani R, Huse JT, Ciriello G, Poisson LM, Barnholtz-Sloan JS, Berger MS, Brennan C, Colen RR, Colman H, Flanders AE, Giannini C, Grifford M, Iavarone A, Jain R, Joseph I, Kim J, Kasaian K, Mikkelsen T, Murray BA, O'Neill BP, Pachter L, Parsons DW, Sougnez C, Sulman EP, Vandenberg SR, Van Meir EG, von Deimling A, Zhang H, Crain D, Lau K, Mallery D, Morris S, Paulauskis J, Penny R, Shelton T, Sherman M, Yena P, Black A, Bowen J, Dicostanzo K, Gastier-Foster J, Leraas KM, Lichtenberg TM, Pierson CR, Ramirez NC, Taylor C, Weaver S, Wise L, Zmuda E, Davidsen T, Demchok JA, Eley G, Ferguson ML, Hutter CM, Mills Shaw KR, Ozenberger BA, Sheth M, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Ayala B, Baboud J, Chudamani S, Jensen MA, Liu J, Pihl T, Raman R, Wan Y, Wu Y, Ally A, Auman JT, Balasundaram M, Balu S, Baylin SB, Beroukhim R, Bootwalla MS, Bowlby R, Bristow CA, Brooks D, Butterfield Y, Carlsen R, Carter S, Chin L, Chu A, Chuah E, Cibulskis K, Clarke A, Coetzee SG, Dhalla N, Fennell T, Fisher S, Gabriel S, Getz G, Gibbs R, Guin R, Hadjipanayis A, Hayes DN, Hinoue T, Hoadley K, Holt RA, Hoyle AP, Jefferys SR, Jones S, Jones CD, Kucherlapati R, Lai PH, Lander E, Lee S, Lichtenstein L, Ma Y, Maglinte DT, Mahadeshwar HS, Marra MA, Mayo M, Meng S, Meyerson ML, Mieczkowski PA, Moore RA, Mose LE, Mungall AJ, Pantazi A, Parfenov M, Park PJ, Parker JS, Perou CM, Protopopov A, Ren X, Roach J, Sabedot TS, Schein J, Schumacher SE, Seidman JG, Seth S, Shen H, Simons JV, Sipahimalani P, Soloway MG, Song X, Sun H, Tabak B, Tam A, Tan D, Tang J, Thiessen N, Triche T, Jr., Van Den Berg DJ, Veluvolu U, Waring S, Weisenberger DJ, Wilkerson MD, Wong T, Wu J, Xi L, Xu AW, Yang L, Zack TI, Zhang J, Aksoy BA, Arachchi H, Benz C, Bernard B, Carlin D, Cho J, DiCara D, Frazer S, Fuller GN, Gao J, Gehlenborg N, Haussler D, Heiman DI, Iype L, Jacobsen A, Ju Z, Katzman S, Kim H, Knijnenburg T, Kreisberg RB, Lawrence MS, Lee W, Leinonen K, Lin P, Ling S, Liu W, Liu Y, Liu Y, Lu Y, Mills G, Ng S, Noble MS, Paull E, Rao A, Reynolds S, Saksena G, Sanborn Z, Sander C, Schultz N, Senbabaoglu Y, Shen R, Shmulevich I, Sinha R, Stuart J, Sumer SO, Sun Y, Tasman N, Taylor BS, Voet D, Weinhold N, Weinstein JN, Yang D, Yoshihara K, Zheng S, Zhang W, Zou L, Abel T, Sadeghi S, Cohen ML, Eschbacher J, Hattab EM, Raghunathan A, Schniederjan MJ, Aziz D, Barnett G, Barrett W, Bigner DD, Boice L, Brewer C, Calatozzolo C, Campos B, Carlotti CG, Jr., Chan TA, Cuppini L, Curley E, Cuzzubbo S, Devine K, DiMeco F, Duell R, Elder JB, Fehrenbach A, Finocchiaro G, Friedman W, Fulop J, Gardner J, Hermes B, Herold-Mende C, Jungk C, Kendler A, Lehman NL, Lipp E, Liu O, Mandt R, McGraw M, McLendon R, McPherson C, Neder L, Nguyen P, Noss A, Nunziata R, Ostrom QT, Palmer C, Perin A, Pollo B, Potapov A, Potapova O, Rathmell WK, Rotin D, Scarpace L, Schilero C, Senecal K, Shimmel K, Shurkhay V, Sifri S, Singh R, Sloan AE, Smolenski K, Staugaitis SM, Steele R, Thorne L, Tirapelli DP, Unterberg A, Vallurupalli M, Wang Y, Warnick R, Williams F, Wolinsky Y, Bell S, Rosenberg M, Stewart C, Huang F, Grimsby JL, Radenbaugh AJ, and

Zhang J. 2015. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* 372:2481-2498. 10.1056/NEJMoa1402121

Chan AK, Yao Y, Zhang Z, Chung NY, Liu JS, Li KK, Shi Z, Chan DT, Poon WS, Zhou L, and Ng HK. 2015. TERT promoter mutations contribute to subset prognostication of lower-grade gliomas. *Mod Pathol* 28:177-186. 10.1038/modpathol.2014.94

Chistiakov DA, Chekhonin IV, and Chekhonin VP. 2017. The EGFR variant III mutant as a target for immunotherapy of glioblastoma multiforme. *Eur J Pharmacol* 810:70-82. 10.1016/j.ejphar.2017.05.064

Cho H, Yu A, Kim S, Kang J, and al e. 2009. Robust likelihood-based survival modeling with microarray data. *Journal of Statistical Software* 29.

Ciccia A, and Elledge SJ. 2010. The DNA damage response: making it safe to play with knives. *Mol Cell* 40:179-204. 10.1016/j.molcel.2010.09.019

Conde M, Michen S, Wiedemuth R, Klink B, Schrock E, Schackert G, and Temme A. 2017. Chromosomal instability induced by increased BIRC5/Survivin levels affects tumorigenicity of glioma cells. *BMC Cancer* 17:889. 10.1186/s12885-017-3932-y

Dalton S. 2015. Linking the Cell Cycle to Cell Fate Decisions. *Trends Cell Biol* 25:592-600. 10.1016/j.tcb.2015.07.007

Ji Y, Rong X, Li D, Cai L, Rao J, and Lu Y. 2016. Inhibition of Cartilage Acidic Protein 1 Reduces Ultraviolet B Irradiation Induced-Apoptosis through P38 Mitogen-Activated Protein Kinase and Jun Amino-Terminal Kinase Pathways. *Cell Physiol Biochem* 39:2275-2286. 10.1159/000447920

Koshiyama DB, Trevisan P, Graziadio C, Rosa RFM, Cunegatto B, Scholl J, Provenzi VO, de Sa AP, Soares FP, Velho MC, de APFN, Oliveira CA, and Zen PRG. 2017. Frequency and clinical significance of chromosome 7 and 10 aneuploidies, amplification of the EGFR gene, deletion of PTEN and TP53 genes, and 1p/19q deficiency in a sample of adult patients diagnosed with glioblastoma from Southern Brazil. *J Neurooncol* 135:465-472. 10.1007/s11060-017-2606-6

Kwon MJ, Kang SY, Cho H, Lee JI, Kim ST, and Suh YL. 2019. Clinical relevance of molecular subgrouping of gliomatosis cerebri per 2016 WHO classification: a clinicopathological study of 89 cases. *Brain Pathol*. 10.1111/bpa.12782

Langfelder P, and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. 10.1186/1471-2105-9-559

Liu B, Liu J, Liu K, Huang H, Li Y, Hu X, Wang K, Cao H, and Cheng Q. 2019. A prognostic signature of five pseudogenes for predicting lower-grade gliomas. *Biomed Pharmacother* 117:109116. 10.1016/j.biopha.2019.109116

Maier V, Jolicoeur C, Rayburn H, Takegahara N, Kumanogoh A, Kikutani H, Tessier-Lavigne M, Wurst W, and Friedel RH. 2011. Semaphorin 4C and 4G are ligands of Plexin-B2 required in cerebellar development. *Mol Cell Neurosci* 46:419-431. 10.1016/j.mcn.2010.11.005

Music D, Dahlrot RH, Hermansen SK, Hjelmborg J, de Stricker K, Hansen S, and Kristensen BW. 2016. Expression and prognostic value of the WEE1 kinase in gliomas. *J Neurooncol* 127:381-389. 10.1007/s11060-015-2050-4

Ouyang D, Li R, Li Y, and Zhu X. 2019. A 7-lncRNA signature predict prognosis of Uterine corpus endometrial carcinoma. *J Cell Biochem* 120:18465-18477. 10.1002/jcb.29164

Ren Y, Zhang X, Rui W, Pang H, Qiu T, Wang J, Xie Q, Jin T, Zhang H, Chen H, Zhang Y, Lu H, Yao Z, Zhang J,

and Feng X. 2019. Noninvasive Prediction of IDH1 Mutation and ATRX Expression Loss in Low-Grade Gliomas Using Multiparametric MR Radiomic Features. *J Magn Reson Imaging* 49:808-817. 10.1002/jmri.26240

Sabha N, Knobbe CB, Maganti M, Al Omar S, Bernstein M, Cairns R, Cako B, von Deimling A, Capper D, Mak TW, Kiehl TR, Carvalho P, Garrett E, Perry A, Zadeh G, Guha A, and Sidney C. 2014. Analysis of IDH mutation, 1p/19q deletion, and PTEN loss delineates prognosis in clinical low-grade diffuse gliomas. *Neuro Oncol* 16:914-923. 10.1093/neuonc/not299

Sahm F, Koelsche C, Meyer J, Pusch S, Lindenberg K, Mueller W, Herold-Mende C, von Deimling A, and Hartmann C. 2012. CIC and FUBP1 mutations in oligodendrogliomas, oligoastrocytomas and astrocytomas. *Acta Neuropathol* 123:853-860. 10.1007/s00401-012-0993-5

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498-2504. 10.1101/gr.1239303

Wang X, Zbou C, Qiu G, Fan J, Tang H, and Peng Z. 2008. Screening of new tumor suppressor genes in sporadic colorectal cancer patients. *Hepatogastroenterology* 55.

Wang Z, Wang Z, Niu X, Liu J, Wang Z, Chen L, and Qin B. 2019. Identification of seven-gene signature for prediction of lung squamous cell carcinoma. *Onco Targets Ther* 12:5979-5988. 10.2147/OTT.S198998

Wu M, Li X, Liu Q, Xie Y, Yuan J, and Wanggou S. 2019. miR-526b-3p serves as a prognostic factor and regulates the proliferation, invasion, and migration of glioma through targeting WEE1. *Cancer Manag Res* 11:3099-3110. 10.2147/CMAR.S192361

Yu G, Wang LG, Han Y, and He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16:284-287. 10.1089/omi.2011.0118

Zeng WJ, Yang YL, Liu ZZ, Wen ZP, Chen YH, Hu XL, Cheng Q, Xiao J, Zhao J, and Chen XP. 2018. Integrative Analysis of DNA Methylation and Gene Expression Identify a Three-Gene Signature for Predicting Prognosis in Lower-Grade Gliomas. *Cell Physiol Biochem* 47:428-439. 10.1159/000489954

Zhang C, Yu R, Li Z, Song H, Zang D, Deng M, Fan Y, Liu Y, Zhang Y, and Qu X. 2019a. Comprehensive analysis of genes based on chr1p/19q co-deletion reveals a robust 4-gene prognostic signature for lower grade glioma. *Cancer Manag Res* 11:4971-4984. 10.2147/CMAR.S199396

Zhang GH, Zhong QY, Gou XX, Fan EX, Shuai Y, Wu MN, and Yue GJ. 2019b. Seven genes for the prognostic prediction in patients with glioma. *Clin Transl Oncol*. 10.1007/s12094-019-02057-3

Figure 1

Study outline.

The outline indicates the exploration process.

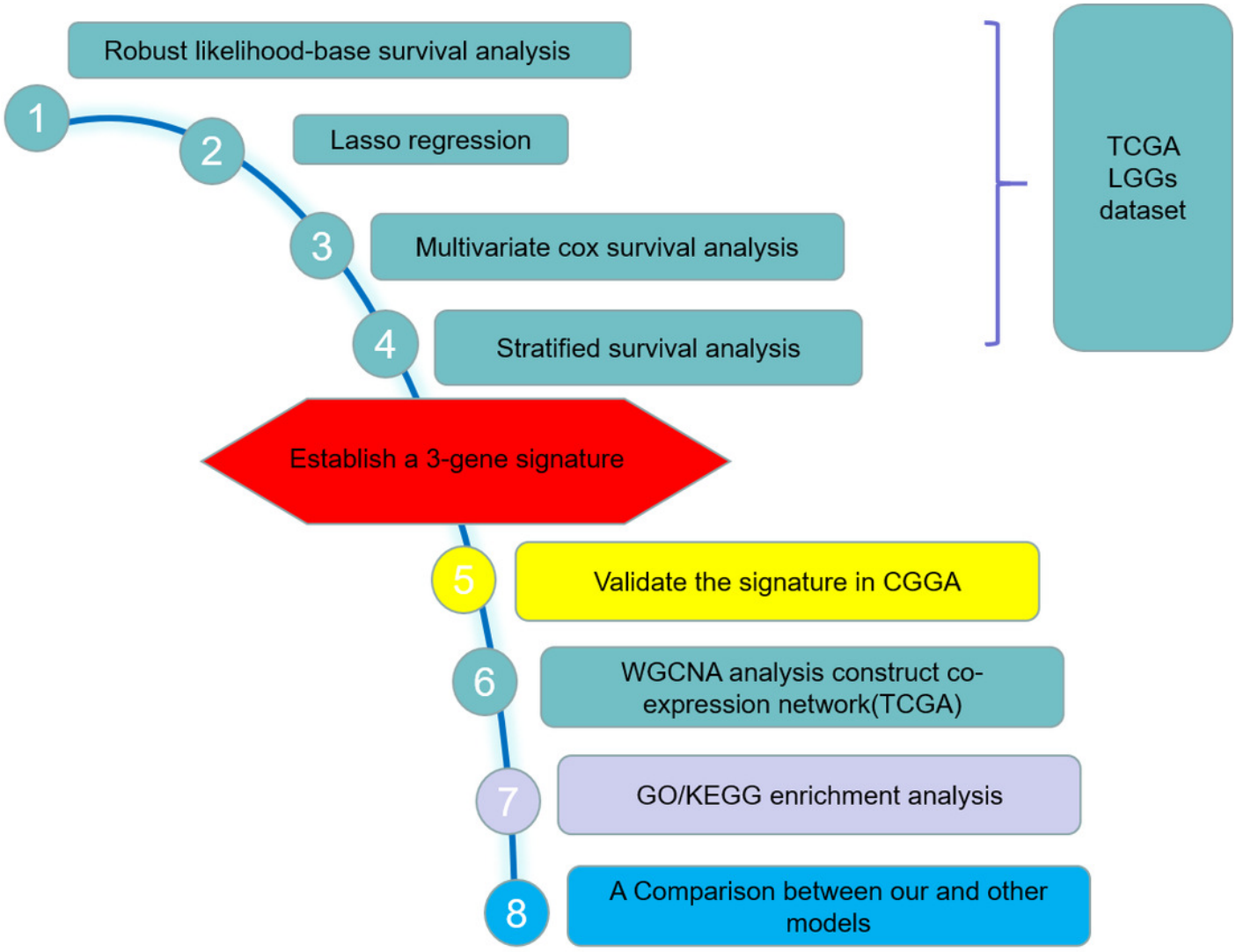


Figure 2

The LASSO regression used to reduce the dimensionality of survival related genes.

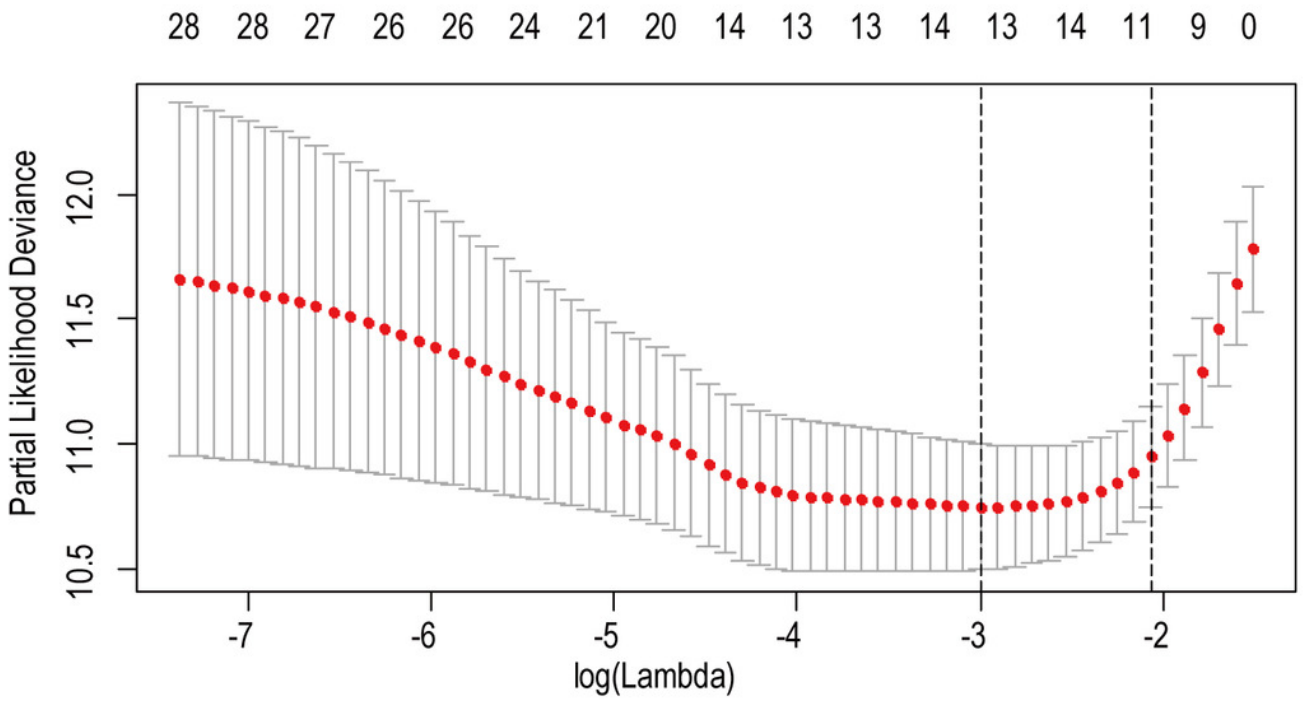


Figure 3

Risk score analysis, survival status and survival time between two risk group and expression distribution of the three-gene signature in TCGA dataset.

(A) The three-gene signature risk score distribution. **(B)** Scatterplot of patient survival status ordered by risk score. **(C)** The heat-map of the three-gene expression profiles after standardized and centralized.

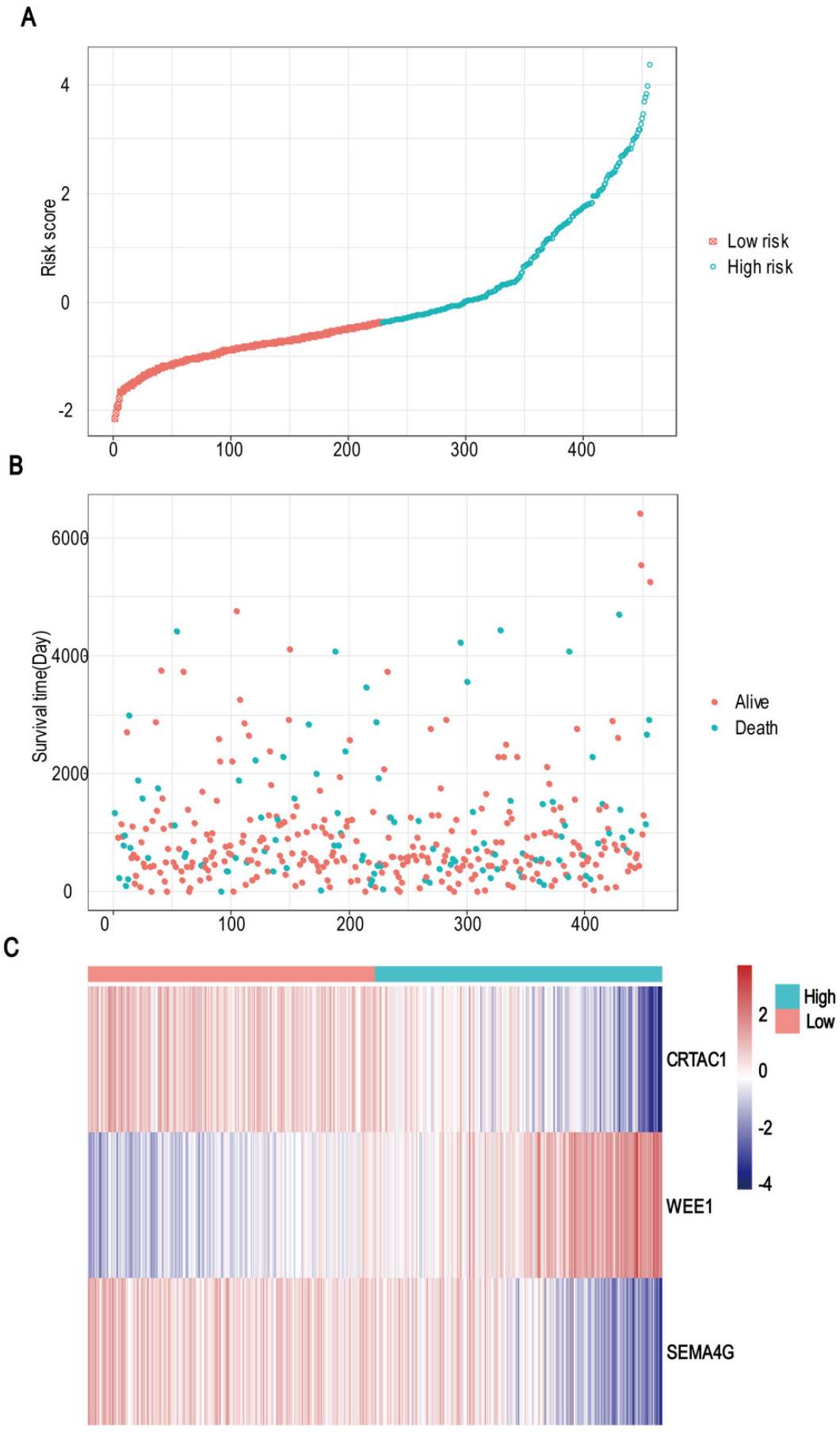

A
Risk score
4
2
0
-2
0
100
200
300
400
Low risk
High risk
B
Survival time(Day)
6000
4000
2000
0
0
100
200
300
400
Alive
Death
C
CRTAC1
WEE1
SEMA4G
High
Low
2
0
-2
-4

Figure 4

Establishment and verification of the risk model in the training set and validation set.

(A) Patient in high-risk group displayed significantly shorter survival time compared to those in low-risk group in training set(P < 0.0001). **(B)**The ROC for predicting the 1-, 3- and 5-year survival and AUC for the risk score model showed good accuracy in training set. **(C, D)**The same result can be observed in the validation set.

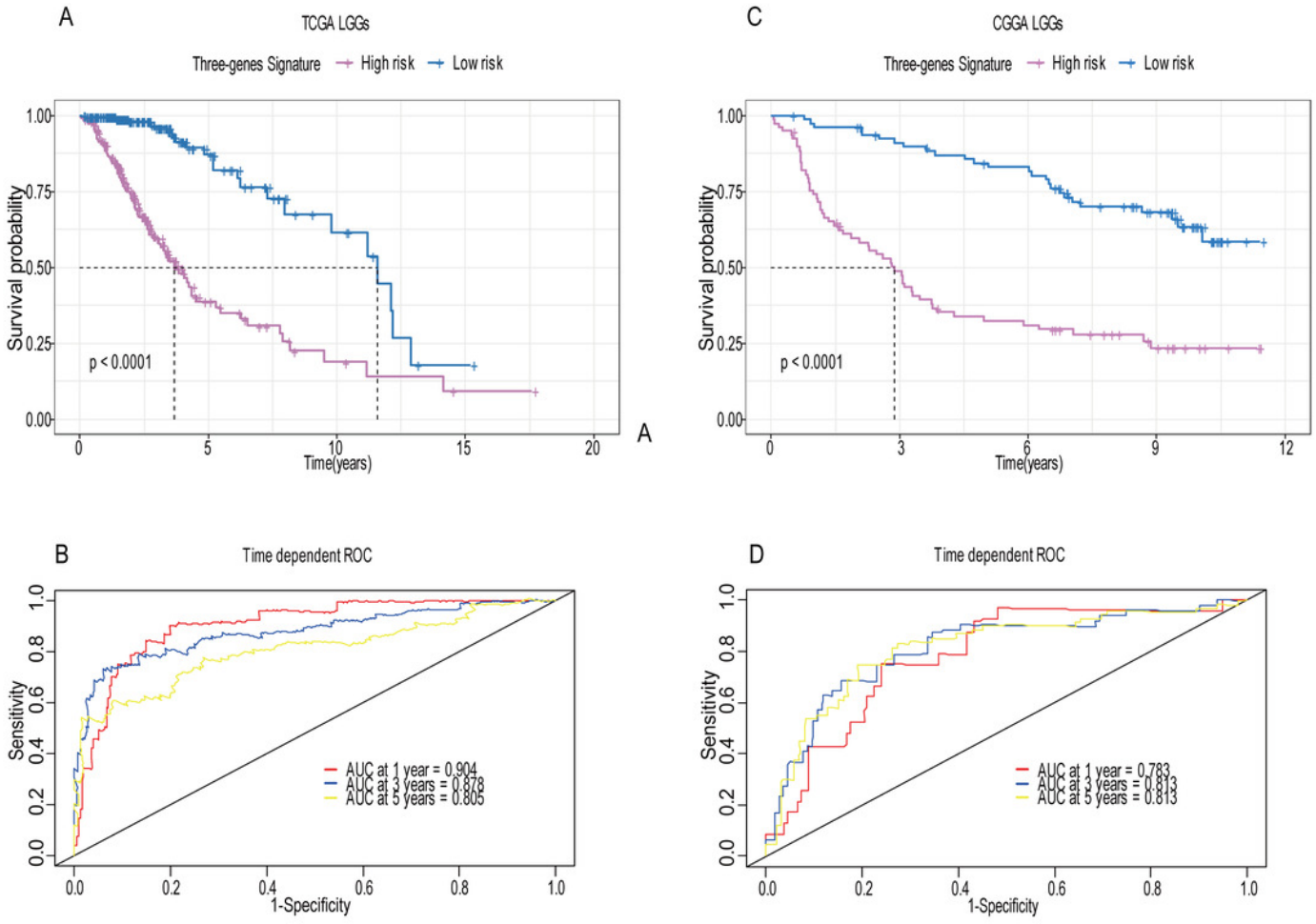


Figure 5

Stratified survival analysis based on the risk model in the training set.

Based on the risk score model, stratified survival analysis performed in patients with different age group(A□D)□gender(B□E)□and pathologic grade(C□F) in the training set.

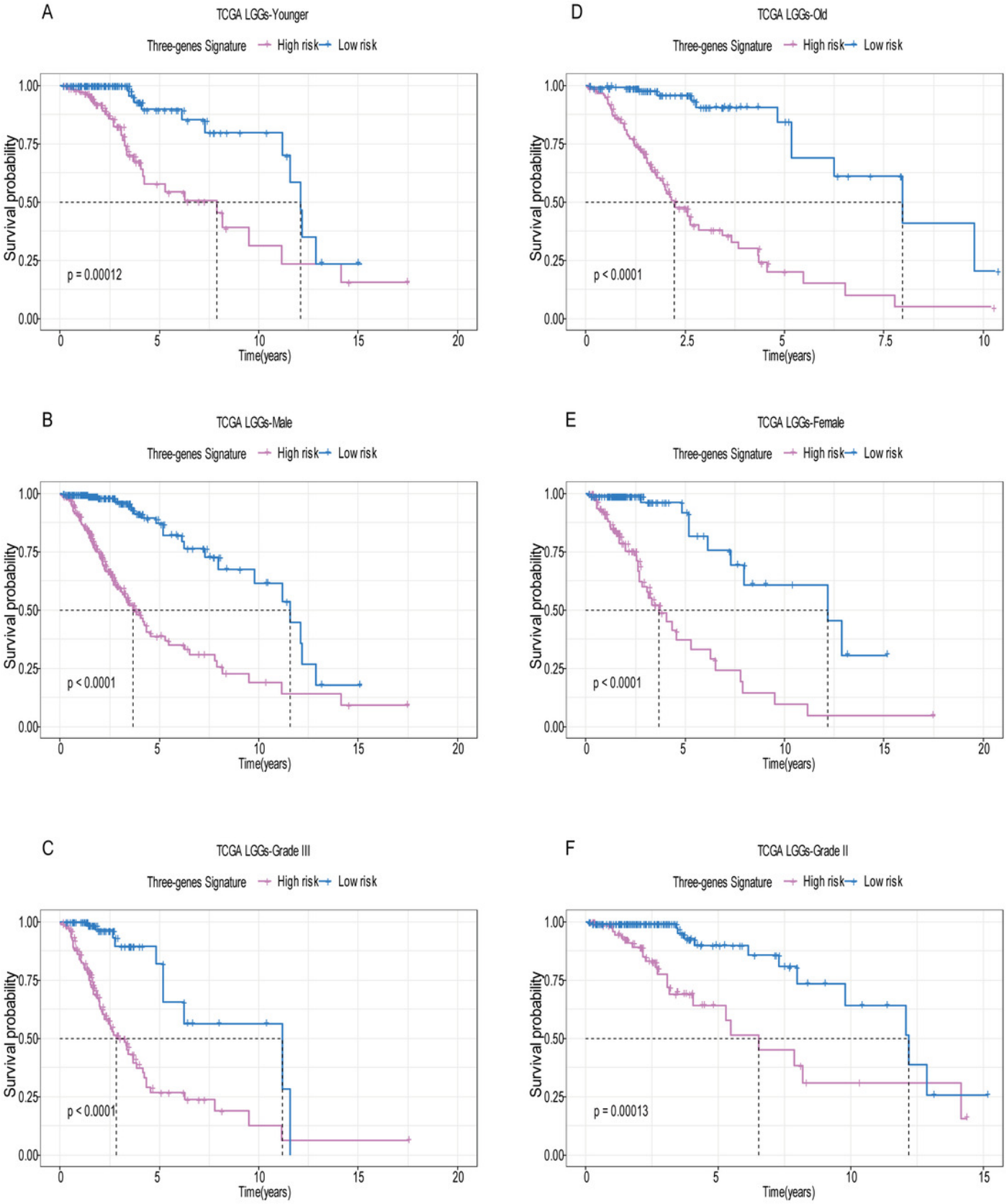

A
TCGA LGGs-Younger
Three-genes Signature
High risk
Low risk
Survival probability
p = 0.00012
Time(years)
D
TCGA LGGs-Old
Three-genes Signature
High risk
Low risk
Survival probability
p < 0.0001
Time(years)
B
TCGA LGGs-Male
Three-genes Signature
High risk
Low risk
Survival probability
p < 0.0001
Time(years)
E
TCGA LGGs-Female
Three-genes Signature
High risk
Low risk
Survival probability
p < 0.0001
Time(years)
C
TCGA LGGs-Grade III
Three-genes Signature
High risk
Low risk
Survival probability
p < 0.0001
Time(years)
F
TCGA LGGs-Grade II
Three-genes Signature
High risk
Low risk
Survival probability
p = 0.00013
Time(years)

Figure 6

Stratified survival analysis based on the risk model in the validation set

Based on the risk score model, stratified survival analysis performed in patients with different age group(A□D)□gender(B□E)□and pathologic grade(C□F) in the validation set.

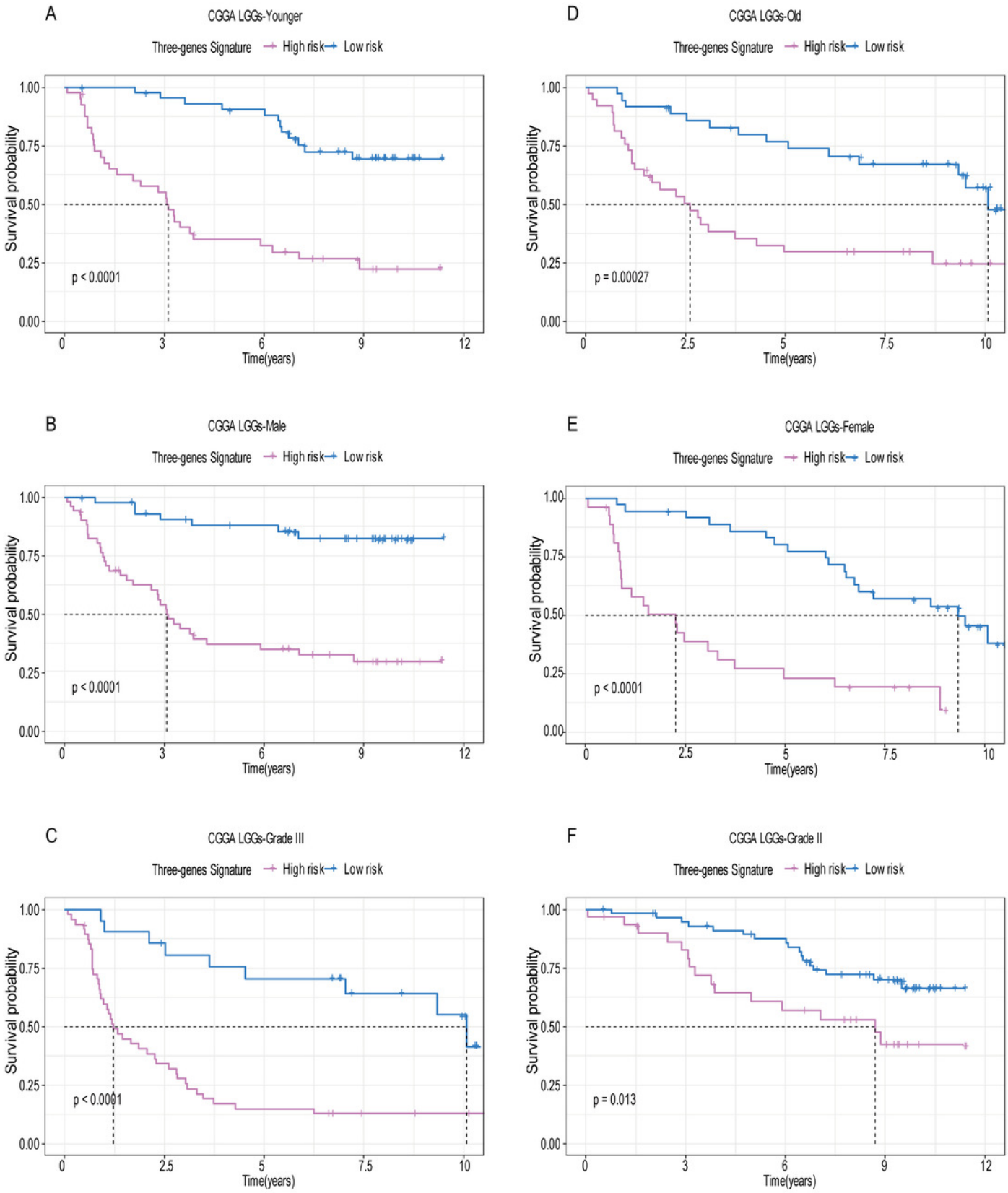

A
CGGA LGGs-Younger
Three-genes Signature High risk Low risk
Survival probability
p < 0.0001
Time(years)
D
CGGA LGGs-Old
Three-genes Signature High risk Low risk
Survival probability
p = 0.00027
Time(years)
B
CGGA LGGs-Male
Three-genes Signature High risk Low risk
Survival probability
p < 0.0001
Time(years)
E
CGGA LGGs-Female
Three-genes Signature High risk Low risk
Survival probability
p < 0.0001
Time(years)
C
CGGA LGGs-Grade III
Three-genes Signature High risk Low risk
Survival probability
p < 0.0001
Time(years)
F
CGGA LGGs-Grade II
Three-genes Signature High risk Low risk
Survival probability
p = 0.013
Time(years)

Figure 7

The expression level of the three genes can divided the patients into different prognostic group in both training set and validation set.

For the WEE1，the member of high expression group had significantly shorter survival than those in low expression group (P＜0.0001) (A，D). For the CRTAC1，the member of high expression group had significantly longer survival than those in low expression group(P ＜0.0001)(B，E). For the SEMA4G，the member of high expression group had significantly longer survival than those in low expression group(P＜0.0001)(C，F).

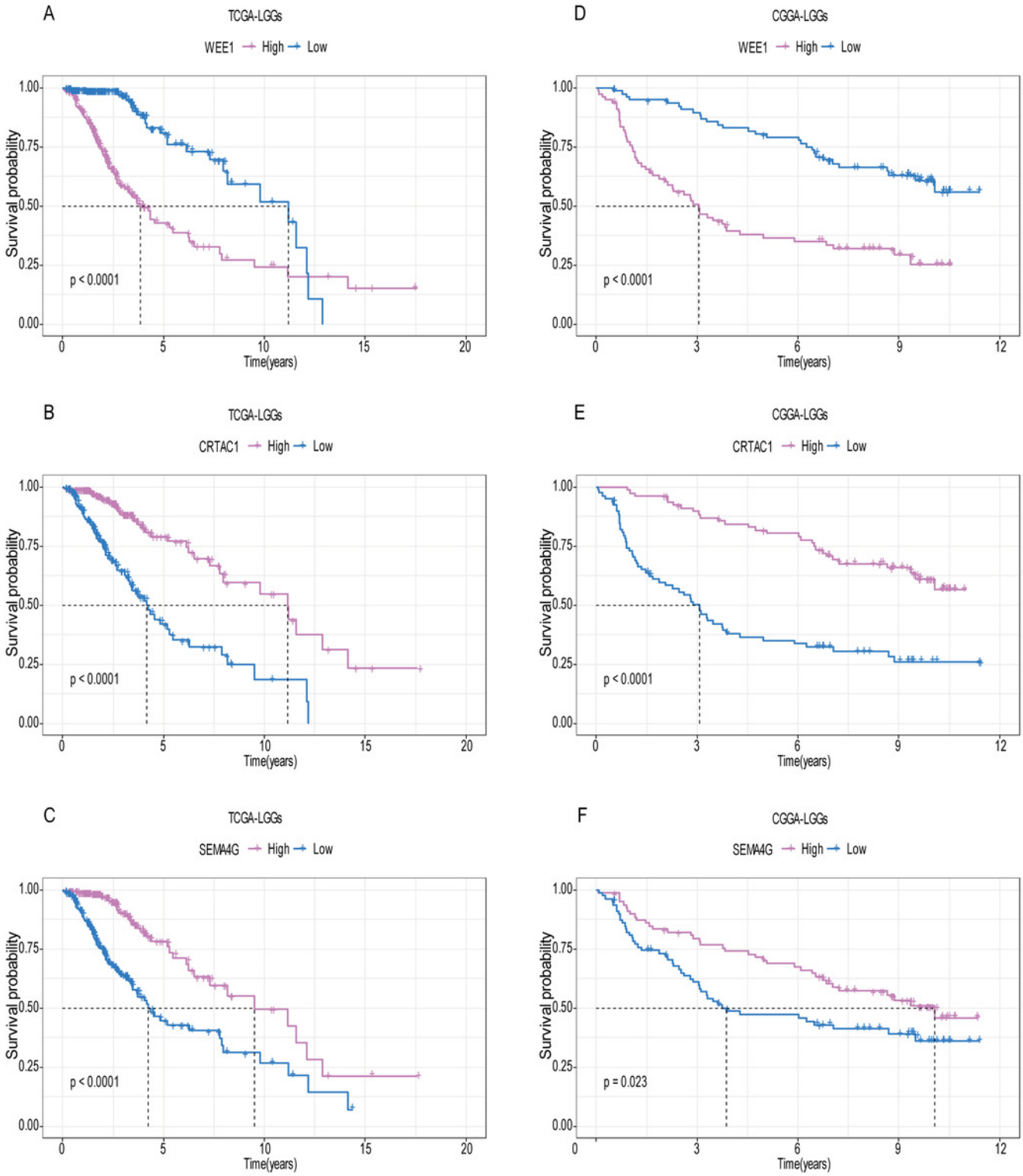

A
TCGA-LGGs
WEE1 High Low
Survival probability
p < 0.0001
Time(years)
D
CGGA-LGGs
WEE1 High Low
Survival probability
p < 0.0001
Time(years)
B
TCGA-LGGs
CRTAC1 High Low
Survival probability
p < 0.0001
Time(years)
E
CGGA-LGGs
CRTAC1 High Low
Survival probability
p < 0.0001
Time(years)
C
TCGA-LGGs
SEMA4G High Low
Survival probability
p < 0.0001
Time(years)
F
CGGA-LGGs
SEMA4G High Low
Survival probability
p = 0.023
Time(years)

Figure 8

Expression of the three genes between grade II tumor and grade III tumor in training set and validation set.

In the training set,the expression level of WEE1 was significantly higher in grade III compared to grade II(P□0.0001)□while the other are opposite(A). These results can also be verified in the validation dataset(B).

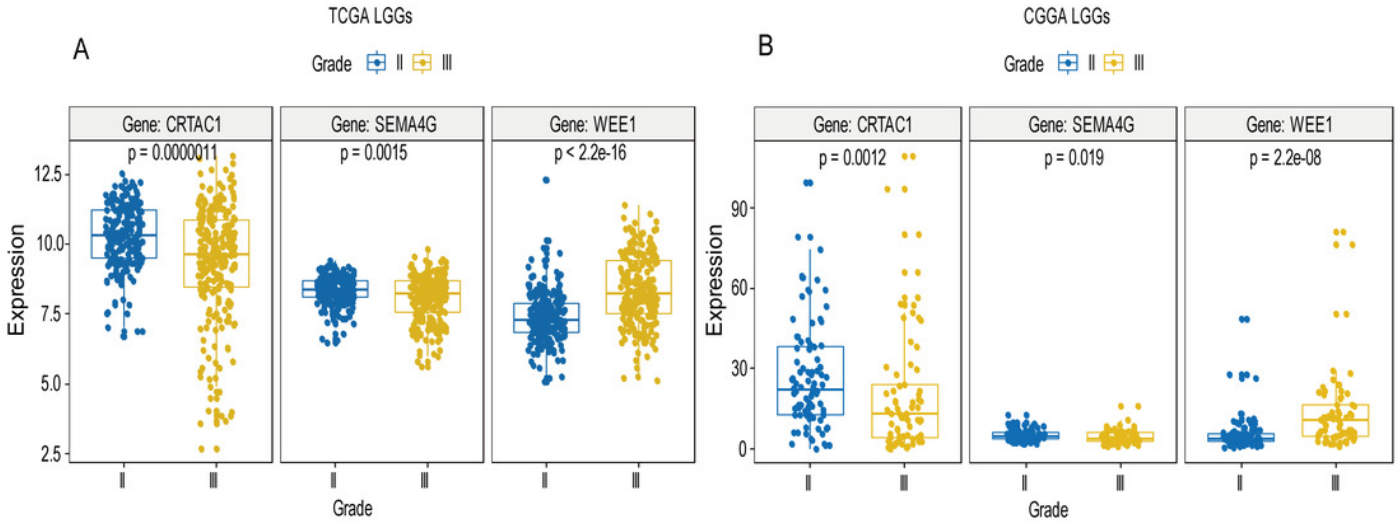


Figure 9

Co-expression network of the three-gene signature.

The co-expression networks of WEE1(A), CRTAC1(B) and SEMA4G(C) were showed. Yellow nodes showed key genes and blue nodes are genes which co-expressed with the key genes.

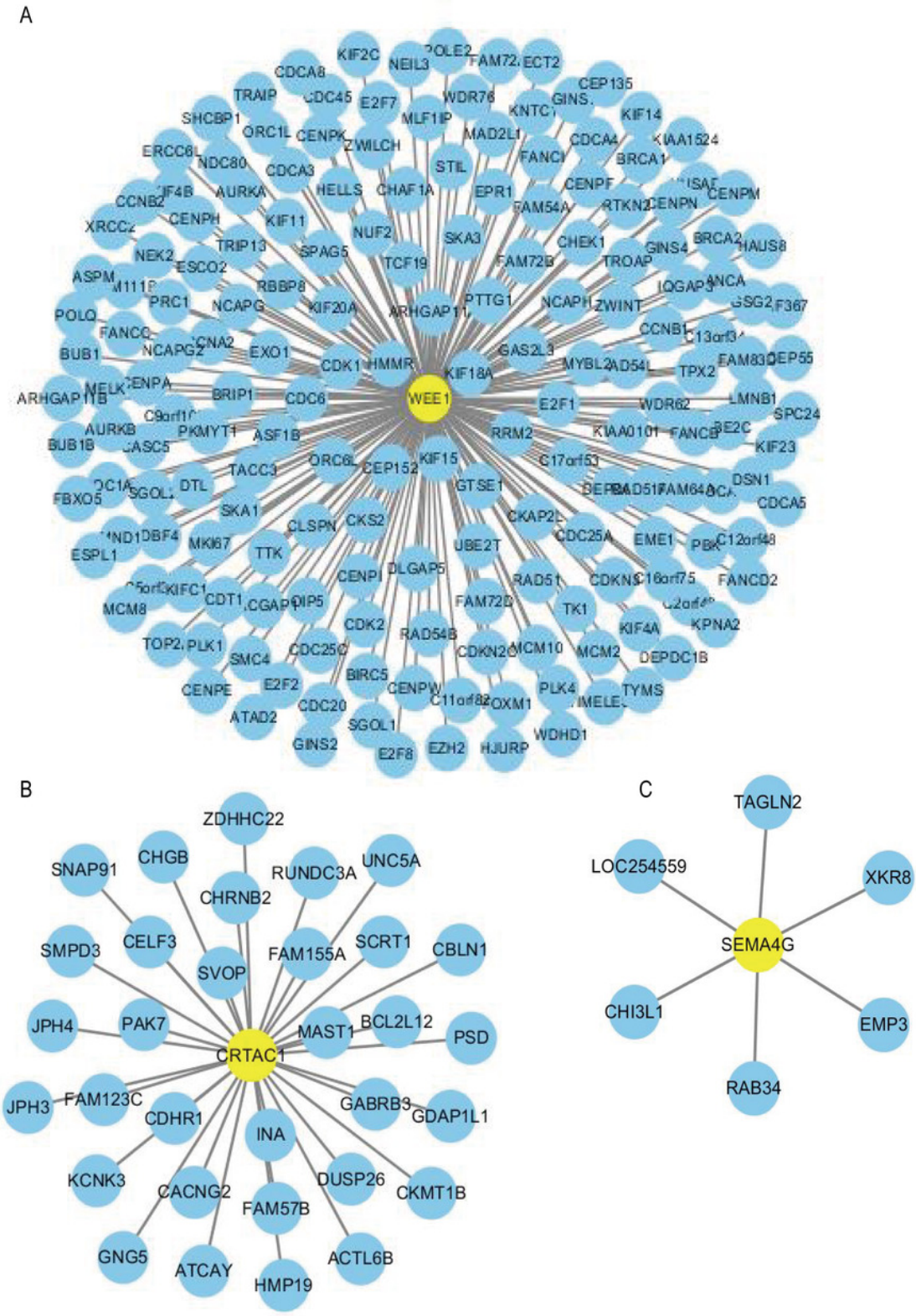

A
WEE1
B
CRTAC1
ZDHHC22
CHGB
SNAP91
RUNDC3A
UNC5A
CHRNB2
SMPD3
CELF3
FAM155A
SCRT1
CBLN1
SVOP
JPH4
PAK7
MAST1
BCL2L12
PSD
JPH3
FAM123C
CDHR1
GABRB3
GDAP1L1
INA
KCNK3
CACNG2
DUSP26
CKMT1B
FAM57B
GNG5
ATCAY
ACTL6B
HMP19
C
SEMA4G
TAGLN2
LOC254559
XKR8
CHI3L1
EMP3
RAB34

Figure 10

The most significantly enriched GO annotations and KEGG pathways of co-expressed genes. The length of the bars and the size of the dots represents the numbers of genes, and the color of the bars/dots corresponds to the P-value according to legend.

(A) Top 5 significantly enriched biological process. **(B)** Top 5 significantly enriched cellular component. **(C)** Top 5 significantly enriched KEGG pathways. **(D)** Top 5 significantly enriched molecular function.

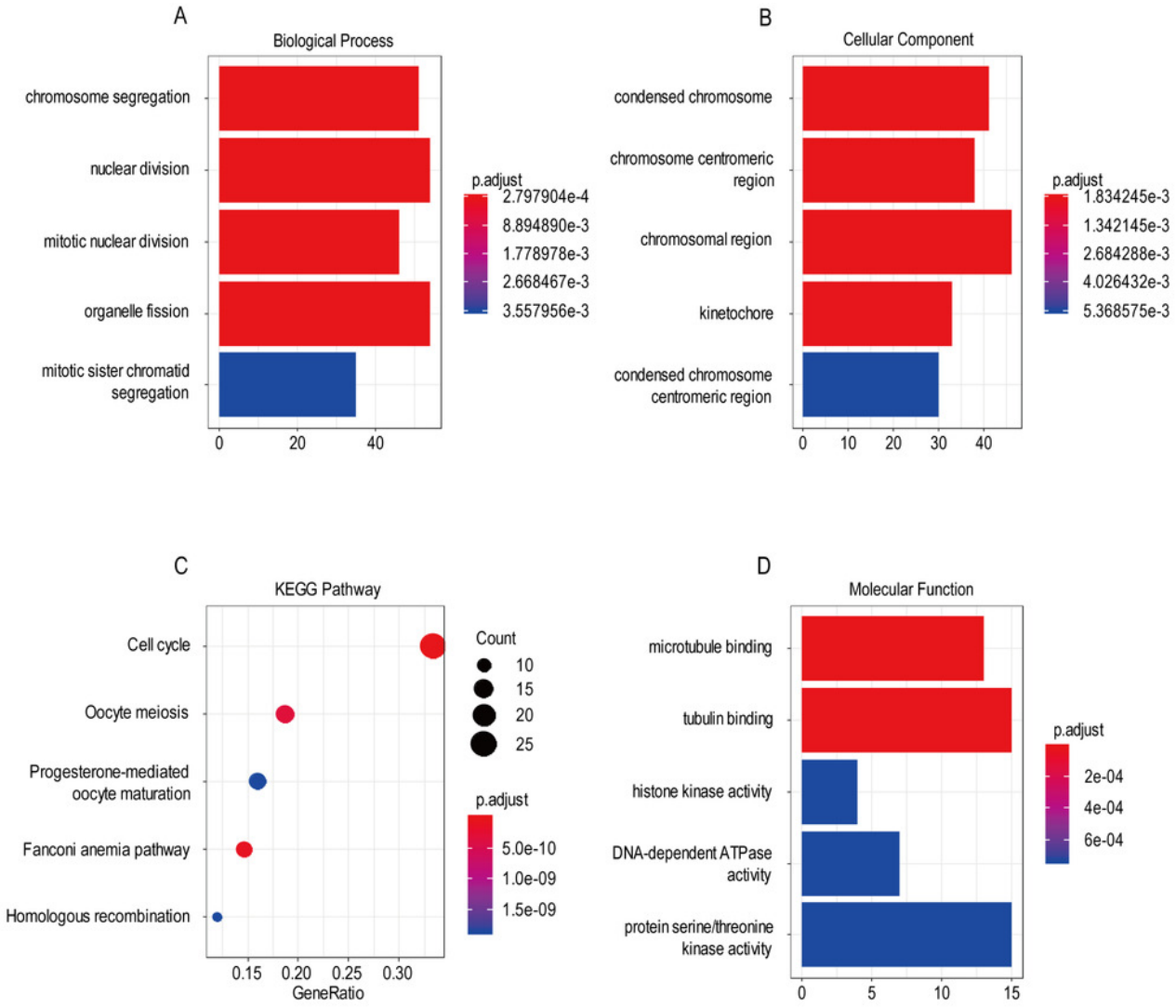
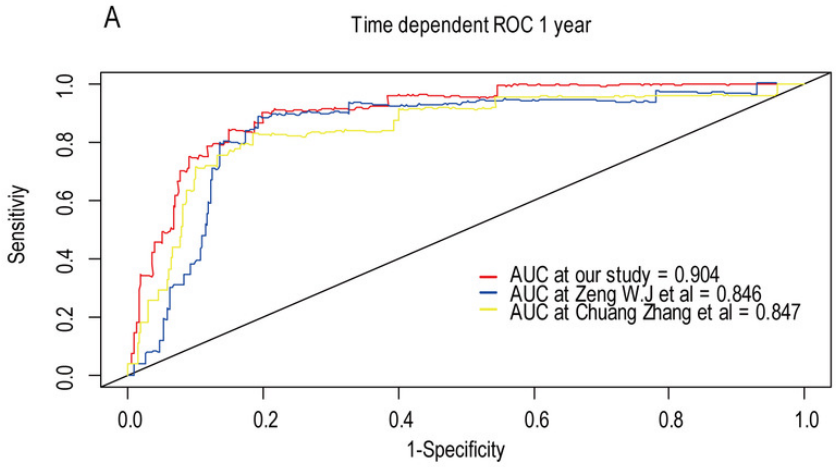
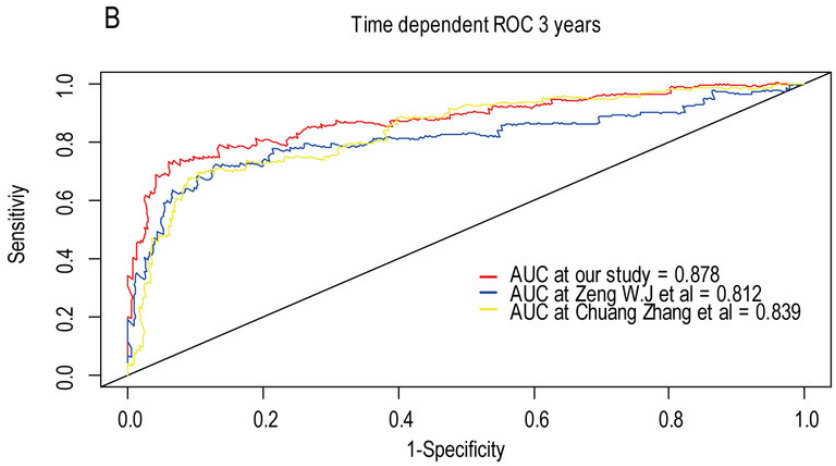


Figure 11

Comparison of our 3-gene model and other literature models.

The-dependent ROC analysis was performed to compare the three models in predicting 1-year**(A)**, 3-year**(B)** and 5-year**(C)** overall survival.


A
Time dependent ROC 1 year
Sensitivity
1-Specificity
AUC at our study = 0.904
AUC at Zeng W.J et al = 0.846
AUC at Chuang Zhang et al = 0.847


B
Time dependent ROC 3 years
Sensitivity
1-Specificity
AUC at our study = 0.878
AUC at Zeng W.J et al = 0.812
AUC at Chuang Zhang et al = 0.839

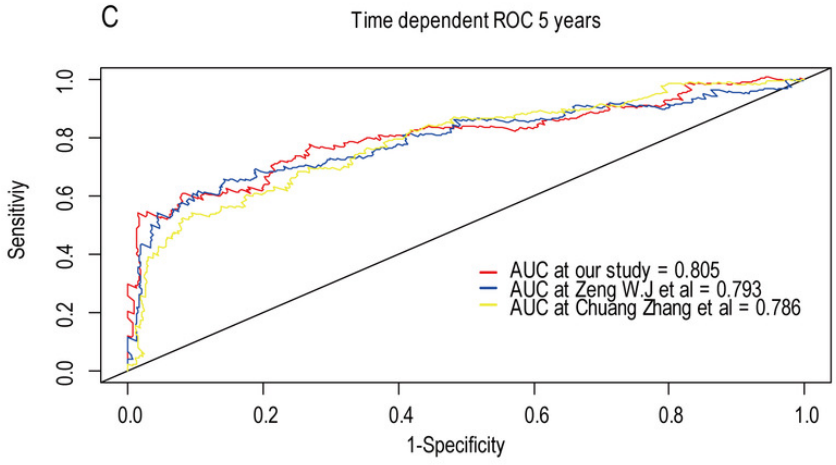

C
Time dependent ROC 5 years
Sensitivity
1-Specificity
AUC at our study = 0.805
AUC at Zeng W.J et al = 0.793
AUC at Chuang Zhang et al = 0.786

Table 1(on next page)

Clinical parameters of patients in the training set and validation set.

Table1:

Clinical parameters of patients in the training set and validation set.

Variables	Training set(n=456)	Validation set(n=159)	*P*value
Age group(Median)			0.5448
younger	232	86	
old	224	73	
Sex			0.1894
Female	210	63	
Male	246	96	
Grade			0.09385
G2	221	90	
G3	235	69	
Molecular therapy			
Yes	263	/	
NO	193	/	
Chemoterapy			
Yes	/	80	
NO	/	79	
Risk level			1
High	228	79	
Low	228	80	
IDH_status			
Wildtype	/	43	
Mutant	/	116	
1p19q_status			
Non_codel	/	52	
Codel	/	107	
Radiation therapy			3.52E-10
Yes	280	141	
NO	176	18	
Age(years)			
Mean±SD	43.4±13.3	40.7±10.9	0.989
Median	41	40	
Vital status			
Alive	341	77	
Dead	115	82	
Survival time(days)			
Mean	998.6±953.8	2024.7±1334.3	5.27E-

24

Median	714.5	2340
Histologic type		
Astrocytoma	162	34
Oligodendroglioma	171	21
Oligoastrocytoma	123	35
Anaplastic astrocytomas	/	26
Anaplastic oligoastrocytomas	/	32
Anaplastic oligodendrogliomas	/	11

4

Table 2(on next page)

Univariate and multivariate Cox regression analyses of the risk score and other clinicopathological factors in training set and validation set.

Variables		Training set(n=456)						Validation set(n=159)					
		Univariate			Multivariate			Univariate			Multivariate		
		HR	95%CI	p-value	HR	95%CI	p-value	HR	95%CI	p-value	HR	95%CI	p-value
Age group(Median)	younger vs old	0.278	0.184-0.420	**1.3E-09**	0.274	0.179-0.418	**2.21E-09**	0.817	0.530-1.261	0.363	1.051	0.654-1.691	0.837
Sex	Male vs Female	1.043	0.721-1.509	0.823	1.081	0.743-1.572	0.686	0.641	0.416-0.989	**0.044**	0.651	0.412-1.028	0.066
Grade	G3 vs G2	3.301	2.196-4.963	**9.3E-09**	2.49	1.568-3.953	**0.00011**	3.590	2.292-5.625	**2.4E-08**	3.799	2.205-6.545	**0.00000151**
Molecular therapy	Yes vs No	1.366	0.924-2.018	0.117	0.893	0.578-1.379	0.608	/	/	/	/	/	/
Chemoterapy	Yes vs No	/	/	/	/	/	/	2.216	1.409-3.485	**0.00057**	1.041	0.614-1.765	0.881
Risk level	Low vs High	0.188	0.118-0.299	**2.1E-12**	0.198	0.120-0.325	**1.68E-10**	0.246	0.153-0.394	**5.7E-09**	0.415	0.251-0.688	**0.000653**
IDH_status	Wildtype vs Mutant	/	/	/	/	/	/	2.496	1.582-3.937	**8.4E-05**	0.995	0.600-1.650	0.983
1p19q_status	Non_codel vs Codel	/	/	/	/	/	/	6.554	3.358-12.790	**3.6E-08**	4.566	2.215-9.414	**0.0000388**
Radiation therapy	Yes vs No	1.996	1.278-3.118	**0.00236**	0.814	0.488-1.358	0.43	0.475	0.262-0.861	**0.0141**	0.524	0.277-0.990	**0.046**

1