Gene Clusters Based on OLIG2 and CD276 Could Distinguish Molecular Profiling in Glioblastoma

Minjie Fu
Huashan Hospital Fudan University
Jinsen Zhang
Huashan Hospital Fudan University
Weifeng Li
University of Birmingham
Shan He
University of Birmingham
Jingwen Zhang
Huashan Hospital Fudan University
Daniel Tennant
University of Birmingham
Wei Hua (hs_huawei@126.com)
Huashan Hospital Fudan University https://orcid.org/0000-0001-6409-5078
Ying Mao
Huashan Hospital Fudan University

Research Article

Keywords: Glioblastoma, Molecular subtype, OLIG2, CD276, Random forest

Posted Date: June 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-599543/v1

Abstract

Background

The molecular classification of glioblastoma (GBM) based on transcriptomic analysis could provide precise treatment and prognosis. However, current subtyping (Classic, Mesenchymal, Neural, Proneural) is a time-consuming and cost-intensive process, which hinders its clinical application. A simple and efficient method for classification was imperative.

Methods

Random forest algorithm was applied to conduct a gene cluster featured with hub genes, OLIG2 and CD276. Functional enrichment analysis and Protein-protein interaction were performed using the genes in this gene cluster. The classification efficiency of the gene cluster was validated by WGCNA and LASSO algorithms, and tested in GSE84010 and Gravandeel's GBM datasets.

Results

The gene cluster (n = 26) could distinguish mesenchymal and proneural excellently (AUC = 0.92), which could be validated by multiple algorithms (WGCNA, LASSO) and datasets (GSE84010 and Gravandeel's GBM dataset). The gene cluster could be functionally enriched in DNA elements and T cell associated pathways. Additionally, five genes in the signature could predict the prognosis well (p = 0.0051 for training cohort, p = 0.065 for test cohort).

Conclusions

This study proved the accuracy and efficiency of random forest classifier for GBM subtyping and provided a convenient and efficient method for subtyping Proneural and Mesenchymal GBM.

Introduction

Glioblastoma (GBM) is the most common malignant primary brain tumor. The past decades have witnessed considerable advances in neurosurgery and radio-chemotherapy but with limited survival benefit [1]. Many efforts were made to set several classification schemes and capture this variability by using gene expression data in an attempt to identify more homogeneous sub-categories for prognosis and drug sensitivity. Molecular signatures such as 1p/19q co-deletion, MGMT methylation, IDH mutation, TERT promoter mutation and H3F3A mutation have advanced the integrative subtype profiling of gliomas [2]. The most commonly used classification scheme was proposed by Verhaak et al., which divides GBMs into Proneural, Classical, Neural, and Mesenchymal types based on gene expression measured with mRNA microarrays from TCGA [3]. Generally, the proneural GBM patients are young and characterized with a good prognosis, while the mesenchymal has the worst prognosis. CGGA group also classified Chinese glioma patients into G1, G2, G3 groups based on gene expression, which matched with

TCGA classification system well [4]. However, the application of transcriptome subtype is still limited in clinic, for reason that RNA sequencing and bioinformatic analysis are time-consuming and expensive.

Many biomarkers associated with genomics, epigenetics and metabolism have been found in situ or in blood [5-7]. Specific biomarkers favor their expression in particular tumor types. For examples, OLIG2, one of the four critical transcriptional factors of glioma stem cells, has a high expression in proneural GBM, which associate with relatively poor prognosis and drug resistance [8, 9]. Besides, CD276, also known as B7H3, have been proven to be associated with tumor progression and poor prognosis in gliomas [10, 11]. Interestingly, CD276 was found co-expressed with stem genes in GSCs and favored its expression in midline gliomas in our previous report [12]. Our previous *in-silico* analysis also revealed it could influence clinical survival and mediate the G1/S transition in myc-drived neuroblastoma [13]. Similar efforts were also made in medulloblastoma [14], and four distinct molecular subgroups of WNT, sonic hedgehog, group 3 and group 4 could provide guidance for therapeutic strategies [15]. Many other vital oncogenes like EGFR, CDK4, MDM2, GLI, PDGFRA, MET and MYC, were also investigated [16], and gene panels were also used in the molecular profiling of GBM [17]. However, efficacy and efficiency are far from satisfactory, so that new biomarkers and methods to distinguish subtypes more efficiently are imperative.

To this end, we employed random forest, a decision tree ensemble algorithm developed by Breiman [18]. Random Forest has the following characteristics which distinguish it from other machine learning algorithms: 1) its ability to extract features, 2) the robust performance on noisy data with highly correlated variables, and 3) its ability to reduce the effect of the curse of dimensionality, i.e., high dimensional data with small sample size [19]. These characteristics make the random forest classifier an appropriate choice for gene expression datasets. And with the development of the next-generation sequencing (NGS), random forest has already been used extensively in the biomedical field, such as neurology [20], cancer classification and even protein-protein interaction sites prediction [21].

In the current study, we aim to establish a gene cluster that could distinguish different molecular subtypes and are suitable for clinical application. Gene cluster featured with hub genes of OLIG2 and CD276 from our analysis is to be conducted to distinguish molecular profiles.

Methods

Data collection and processing

All data, including clinical information and RNA-seq and array expression data, were downloaded from Gliovis (http://gliovis.bioinfo.cnio.es/). RNA seq and array expression data of the Cancer Genome Atlas (TCGA), Gravandeel, and GSE84010 datasets were retrieved. The expression data of TCGA were assigned into two cohorts randomly for training and validation. And Gravandeel's and GSE84010 datasets were used for the random forest model test. The flowchart of this study is shown in Figure 1, and the baseline of the two cohorts was tabulated in Table 1. For further machine learning, all expression data was normalized.

Data analysis was conducted using R 3.6.3 (R Core Team, 2020). A random number table generated by R 3.6.3 was used to randomly assign 70% of the patients to the training cohort (n=342) and 30% of the patients to the validation cohort (n=147).

Random forest training

Random forest was trained with R packages “randomForest”. The expression profiles were set as input and the expression subtypes were set as labels. The number of decision trees (ntree) was set as 3000 and the max features (mtry) were set as 3. After every training, the input genes were ordered according to their importance, reflected by mean decreased accuracy and Gini. Genes whose mean decreased accuracy and Gini lower than OLIG2 and CD276 are excluded from the candidate genes until all genes were not lower than OLIG2 and CD276 in the form of mean decreased accuracy and Gini.

Principal component analysis

Principal component analysis (PCA) based on the the transcription matrix was performed using R package "ggbiplot" (https://github.com/vqv/ggbiplot), and every gene is displayed in the coordinates with arrows from the origin.

WGCNA construction and key module identification

The TCGA expression data profile was used for network generation by the R package WGCNA [22]. Initially, correlation of gene adjacency was conducted using a power function. Afterwards, the modules were generated by the hierarchical average linkage clustering approach. The modules were assigned to different colors for visualization.

Functional enrichment analysis

The functional enrichment analysis is performed with R package “Clusterprofiler” [23]. We conducted the enrichment of genes in 3 modules we conducted from the PCA analysis respectively and altogether. The results of the enriched terms of the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Disease Oncology (DO) pathways and were generated after running ClusterProfiler. The adjusted P-value < 0.05 was chosen as the threshold for the identification of significant GO terms and pathways.

Protein-protein Interaction

The Protein-protein Interaction (PPI) data was downloaded from String (Protein-protein Interaction). And the web plot was performed by the R package “igraph” (https://github.com/igraph).

Prognostic prediction model construction

In order to find the predictive value of the individual genes, proportional hazards (Cox) model was conducted. And the results were summarized in the form of forest plots and 2-dimension plots to show the risk factor for both PFS and OS. Calculations and graphing were conducted with the “survminer” and “ggforest” packages in R (https://cran.r-project.org/web/packages/survminer/, https://cran.r-project.org/web/packages/ggforest/).

With the aim to explore the prognosis predictive significance of the gene cluster, the least absolute shrinkage and selection operator (LASSO) was conducted with the “glmnet” package in R (https://cran.r-project.org/web/packages/glmnet/). The risk score was calculated by LASSO based on the gene expression level and regression coefficients.

Table 1: the baseline of training cohort and test cohort.There is no significant difference of the population baseline between the training cohort and test cohort.

Characteristic	Training cohort	Test cohort
Age - yr		
Median	59.2	60.1
Range	10.9-89.3	25.2-86.6
Age - no. (%)		
< 60 yr	52.9	49
≥ 60 yr	46.2	49
NA	0	2
Sex - no. (%)		
Male	59.9	61.2
Female	38.9	35.4
NA	1.2	3.4
Primary or secondary - no. (%)		
Primary	94.4	93.2
Secondary	3.5	2.0
Recurrent	1.2	2.7
NA	0.8	2.0
IDH status - no. (%)		
Wild type	68.7	70.7
Mutant	6.1	4.1
NA	25.1	25.2
Subtype - no. (%)		
Classical	27.2	25.9
Mesenchymal	28.9	32.7

Neural	16.4	19
Proneural	27.5	32.7

Results

OLIG2 and CD276 share a mutually exclusive expression in gliomas

The bioinformatic analysis showed that OLIG2 and CD276 were negatively correlated (Pearson R= -0.38, Spearman R=-0.36, Figure 2A) in TCGA GBM dataset (n=489). The expression of OLIG2 and CD276 varied in different subtypes (Figure 2B), and the mutually exclusive expression was subtype-depended (Figure 2C), as CD276 favored in mesenchymal and OLIG2 in proneural subtypes (Figure 2D). The regression analysis showed the expression panel of CD276 and OLIG2 are in the opposite tendency (Radj2=0.19, P<0.0001, Fig 2C). In view of G-CIMP status, OLIG2 is in high expression in GBM with G-CIMP, while CD276 is high in GBM without G-CIMP (Figure 2E). This contrary of CD276 and OLIG2 also could be observed in GBM with different IDH mutation status and MGMT methylation (Figure 2F, G). All these data showed that OLIG2 and CD276 could share an exclusive expression pattern in GBM.

Since CD276 is an immune costimulatory molecule, we considered the association between the immune infiltration status and OLIG2/CD276 expression. In the comparison of $OLIG2^{hi}/CD276^{lo}$ and $OLIG2^{lo}/CD276^{hi}$, in detail, immune infiltration score is differed, especially in CD4 naïve cells, cytotoxic cells, Th1 cells, central memory cells, macrophage cells, neutrophil cells, Gamma delta cells and infiltration score (Supplementary 1A). Moreover, gender, as well as MGMT status, were in no relation with OLIG2/CD276 expression while IDH mutant and G-CIMP positive mainly belonged to $OLIG2^{hi}/CD276^{lo}$ group of GBM (Figure 2H).

Gene clusters based on OLIG2 and CD276 could be generated by PCA and WCGNA algorithm

In total, 26 genes were obtained by mean decreased accuracy and Gini with random forest algorithm under supervision according to the expression subtypes (Figure 3A, B). In order to have a visual sight of the expression of the genes in four subtypes, PCA analysis was further performed. The 26 genes could distinguish the Verhaak's subtypes well and were labeled with three modules (Module-Classic, Module-Mesenchymal, Module-Proneural) according to similar gene expression pattern. In detail, genes mainly positively reflected by principle component 1 (PC2 > |PC1|) were labled as Module-Classic. And genes whose PC1 > |PC2| and PC1 < -|PC2| were labeled as Module-Proneural and Module-Mesenchymal respectively (Fig 3C). Genes in Module-Classic, Module-Mesenchymal, and Module-Proneural have higher expression level in these modules than those in other modules (Fig 3D), respectively. In detail, OLIG2 belonged to Module-Proneural and CD276 belonged to Module-Mesenchymal, which was consistent with the previous results. A column plot was displayed to show the expression of gene modules in 4 subtypes (Figure 3E).

Furthermore, ROC was displayed to evaluate the efficiency of the random forest algorithm for subtype classification. And AUC of all classes except mesenchymal versus neural is relatively high (AUC=0.855, Figure 3F). Specifically, the ability of the gene cluster to distinguish mesenchymal and proneural was excellent with an AUC value of 0.92. It is noteworthy that proneural and mesenchymal subtype of GBM can be distinguished well by the gene cluster.

In order to validate the random forest algorithm, WCGNA algorithm was performed to match the co-expression network with the three principal component modules. A soft threshold was set as 8 in consideration of the model's accuracy and the computational expense (Figure 4A). The gene clusters were grouped into six modules altogether (Figure 4B). Further, Sankey diagram was made to match the results of PCA analysis and WCGNA, and there existed correspondence between the two kinds of modules. Genes in MEblue and MEbrown belonged to the Proneural-module, which were in high expression in proneural GBM indeed and in low expression in mesenchymal GBM meanwhile. On the opposite, the vast majority of MEturquoise and the whole MEyellow belonged to the Mesenchymal-module, which were highly expressed in mesenchymal GBM and lowly expressed in proneural GBM (Figure 4C). Based on the modules, the Protein-protein Interaction network was analyzed to show the interaction among the gene clusters. As a result, only six genes in clusters (RAB34, RAB33A, OLIG2, VDR, TCF, and DNMT1) were found interactions by biochemistry experiments. RAB33A and RAB34 were Ras-related genes, the former one was in MEblue and the latter one was in MEyellow. The two proteins were exclusively expressed in mesenchymal and proneural GBM (Figure 4D). The network plot was displayed according to the correlation of the gene cluster, which matched the WGCNA results and PPI network very well (Figure 4E).

Gene clusters based on OLIG2 and CD276 could be validated in independent datasets

GSE84010 and Gravandeel's GBM datasets were used to test classifying efficacy. 7 of 26 genes are included in GSE84010 dataset and 21 of 26 genes in Gravandeel's dataset (Figure 5A, D). ROC analysis showed the results of validation on GSE84010 and Gravandeel's datasets (Figure 5C, D). The AUC of random forest algorithm is still ideal (0.816 in GSE84010 dataset, 0.820 in Gravandeel's dataset). In detail, AUCs for mesenchymal-proneural classification in these two datasets both exceeded 0.9, indicating the excellent efficacy of the gene cluster to distinguish mesenchymal and proneural subtypes. This feature was also displayed in the PCA plot, in which orange circle (referring to proneural subtypes) and red circle (referring to mesenchymal subtypes) were well distinguished (Figure 5B, E). In general, the gene cluster showed good efficacy of the four expression subtypes generally.

Gene clusters could be functionally enriched in DNA elements and T cell associated pathways

To further explore the function of the 26-gene clusters, the GO, KEGG and DO database functional enrichment was performed for genes in different modules[24-26]. GO pathway enrichment revealed both classic and mesenchymal modules are enriched in the promoter-specific chromatin binding pathway (Figure 6A). Most enriched pathways from GO pathway analysis correlated with the DNA elements which regulate the expression of genes such as promoter-specific chromatin binding, methyl-CpG binding, E-box binding and catalytic activity, acting on DNA. Notably, both GO biological process terms of lymphocyte differentiation and T cell activation associated with four genes in the cluster, indicating the immune activity occupied an essential position in GBM as shown in the chord plot (Figure 6B). KEGG pathways revealed that signaling regulating pluripotency of stem cells and microRNAs in cancer pathways were enriched and indicated that genes in cluster associate with functions of pluripotency regulation (Figure 6C). In DO database, genes in the cluster were found to have a close connection with cancers in other systems like non-small cell lung carcinoma, bladder carcinoma and integumentary system disease (Figure 6D). Furthermore, we found genes in classic modules are also related to other common epithelial diseases like skin disease, dermatitis, and genes in Module-mesenchymal are associated with tumors originated from mesenchymal tissues, such as non-small cell lung carcinoma.

Five genes in the gene clusters could construct a survival prediction model

Regarding risk factors of the overall survival and progression free survival, multivariate Cox regression analysis revealed that VDR (HR for OS, 1.71; 95% CI, 1.15–2.54; p = 0.008; HR for PFS, 1.49; 95% CI, 1.04–2.13; p = 0.029), LMNB2 (HR for OS, 1.68; 95% CI, 1.16–2.43; p = 0.006; HR for PFS, 1.70; 95% CI, 1.23–2.36; p = 0.001), TCF3 (HR for OS, 0.60; 95% CI, 0.38–0.94; p = 0.027; HR for PFS, 0.47; 95% CI, 0.31–0.71; p < 0.001) and TNFAIP8 (HR for OS, 0.61; 95% CI, 0.38–1.00; p = 0.050; HR for PFS, 0.54; 95% CI, 0.35–0.84; p = 0.006) as the independent factors for both of OS and PFS (Figure 7A).

Furthermore, a 2-dimension plot was used to visualize the predictive value of genes in cluster for both of OS and PFS. The closer the plot to the top right, the better prognostic the genes indicated (Figure 7B). Through the LASSO regression algorithm, a signature of 5 genes was finally obtained (Figure 7C, D). Among the five genes in the signature, only TCF3 expression is in negative correlation with risk score (Figure 7E). As shown in Figure 7F, patients in the low risk-score group (Blue) benefit from longer survival time than those in the high risk-score group (Red). The prognosis of the low risk-score group is better than that of the high risk-score group in the training cohort and test cohort (Figure 7G).

Discussion

In this study, we found that random forest algorithm performs efficiently in the subtype classification based on CD276 and OLIG2. Random forest algorithm-based classification was also studied in GBM before. Crisman TJ. et. al. simplified Verhaak's 840 total genes into 48 genes [27]. Though the gene panel

consists of as many as 48 genes, efficacy for classification of Neural subtype over others is not satisfactory.

We concentrated on two genes (OLIG2 and CD276), both of which favored expression in proneural and mesenchymal expression subtypes. Therefore, we suppose that the two biomarkers, OLIG2 and CD276, represents two tendency or subtypes of GBM, which could result in the different expression profiles and prognosis status. And previous studies found proneural GBM expressing OLIG2 is more prevalent in younger patients with a better prognosis [26]. Thus, through the identification of the two biomarkers, we can make a prediction for prognosis. Since these genes are potential biomarkers for some special subtypes, target drugs can be administrated accordingly.

In the process of WGCNA analysis, we found some contradiction with the results of PCA analysis. In detail, OLIG2 belonged to Proneural-module but also belonged to MEturquoise, which is in high expression in mesenchymal but in low expression in proneural. The bias of the model is not elusive to explain this mismatch. The mesenchymal-proneural transition might result in this phenomenon, based on the evidence that increased OLIG2 expression is considered as a biomarker of mesenchymal-to-proneural transition (PMT). We also hyposise that OLIG2 not only represented the expression subtype but also might be a driver gene of the biological transition process. Loss of OLIG2 function in GSCs was found to result in mesenchymal transformation, which indicate OLIG2 play a vital role in PMT[28]. It is known that Proneural subtype has a correlation with IDH mutated GBM. So, OLIG2 drived gliomas might have a better prognosis, which is also demonstrated in this study.

In the meantime, CD276 could be down-regulated in IDH mutated gliomas, which mainly caused by autophagy induced by 2-HG accumulation [29]. CD276 also seems to favor its expression in H3 mutated gliomas [30], which is exclusive to IDH mutation. The malignancy represented by CD276 seems correlated with TGF-beta pathway [31]. All these interesting findings indicate that different driving gene cluster might dominate in different subtype gliomas, which cause different biological behaviour and clinical features and prognosis.

There are many attempts to sort gliomas, and the subtypes could be linked with some biomarkers, such as PDGFRA with proneural and NF1 with mesenchymal, ATRX mutation with astrocytomas and pTERT mutation with oligdendrocytomas [32].Wang et al. classified GBM into two types based on lineage markers, as type 2 shared oligodendrocyte differentiation and better prognosis [33].

Also, there are some limitations in this study. For one thing, the sample size of training cohorts in this study is small. As a result, our random forest algorithm generated smaller number of decision trees to avoid overfitting, which might miss some important genes. Another drawback is that, although glioblastoma is charaterised by its inter-tumor and intra-tumor heterogeneity, which we didn't take into consideration when applying the random forest algorithm. Moreover, in order to obtain a simplified gene panel, small number of genes were obtained. The weak association of small number of genes leads to the low significance of the functional enrichment analysis. Important pathways might be missed.

In conclusion, the random forest algorithm is proved efficient in the multi-classification of GBM expression subtypes, which would pave the way for Precision Medicine. With the development of NGS, the combination of machine learning and NEG is likely to play an essential role in the diagnosis and prognosis prediction of GBM.

Abbreviations

AUC area under curve CGGA Chinese Glioma Genome Atlas CI confidence interval COX proportional hazards model DO Disease Ontology GBM glioblastoma GO Gene Ontology KEGG Kyoto Encyclopedia of Genes and Genomes LASSO least absolute shrinkage and selection operator NGS next Generation Sequencing PCA principal component analysis PMT mesenchymal-to-proneural transition PPI protein-protein Interaction ROC receiver operating characteristic curve TCGA the Cancer Genome Atlas WGCNA weighted correlation network analysis

Declarations

1. Tan AC, Ashley DM, Lopez GY, Malinzak M, Friedman HS, Khasraw M: **Management of glioblastoma: State of the art and future directions.** *CA Cancer J Clin* 2020, **70:**299-312.

2. Yan W, Zhang W, You G, Zhang J, Han L, Bao Z, Wang Y, Liu Y, Jiang C, Kang C, et al: **Molecular classification of gliomas based on whole genome gene expression: a systematic report of 225 samples from the Chinese Glioma Cooperative Group.** *Neuro Oncol* 2012, **14:**1432-1440.

3. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer Cell* 2010, **17:**98-110.

4. Chai R, Li G, Liu Y, Zhang K, Zhao Z, Wu F, Chang Y, Pang B, Li J, Li Y, et al: **Predictive value of MGMT promoter methylation on the survival of TMZ treated IDH-mutant glioblastoma.** *Cancer Biol Med* 2021, **18:**272-282.

5. Nicolaidis S: **Biomarkers of glioblastoma multiforme.** *Metabolism* 2015, **64:**S22-27.

6. Linhares P, Carvalho B, Vaz R, Costa BM: **Glioblastoma: Is There Any Blood Biomarker with True Clinical Relevance?** *Int J Mol Sci* 2020, **21**.

7. Fu M, Hussain A, Dong Y, Fei Y: **A retrospective analysis of GSE84010: Cell adhesion molecules might contribute to bevacizumab resistance in glioblastoma.** *Journal of Clinical Neuroscience* 2021, **86:**110-115.

8. Bouchart C, Trepant AL, Hein M, Van Gestel D, Demetter P: **Prognostic impact of glioblastoma stem cell markers OLIG2 and CCND2.** *Cancer Med* 2020, **9:**1069-1078.

9. Liu H, Weng W, Guo R, Zhou J, Xue J, Zhong S, Cheng J, Zhu MX, Pan SJ, Li Y: **Olig2 SUMOylation protects against genotoxic damage response by antagonizing p53 gene targeting.** *Cell Death Differ* 2020, **27:**3146-3161.

10. Zhang T, Jin Y, Jiang X, Li L, Qi X, Mao Y, Hua D: **Clinical and Prognostic Relevance of B7-H3 and Indicators of Glucose Metabolism in Colorectal Cancer.** *Front Oncol* 2020, **10:**546110.

11. Lemke D, Pfenning PN, Sahm F, Klein AC, Kempf T, Warnken U, Schnolzer M, Tudoran R, Weller M, Platten M, Wick W: **Costimulatory protein 4IgB7H3 drives the malignant phenotype of glioblastoma by mediating immune escape and invasiveness.** *Clin Cancer Res* 2012, **18:**105-117.

12. Johnston MJ, Nikolic A, Ninkovic N, Guilhamon P, Cavalli FMG, Seaman S, Zemp FJ, Lee J, Abdelkareem A, Ellestad K, et al: **High-resolution structural genomics reveals new therapeutic vulnerabilities in glioblastoma.** *Genome Res* 2019, **29:**1211-1222.

13. Zhang H, Zhang J, Li C, Xu H, Dong R, Chen CC, Hua W: **Survival Association and Cell Cycle Effects of B7H3 in Neuroblastoma.** *J Korean Neurosurg Soc* 2020, **63:**707-716.

14. Kijima N, Kanemura Y: **Molecular Classification of Medulloblastoma.** *Neurol Med Chir (Tokyo)* 2016, **56:**687-697.

15. Archer TC, Mahoney EL, Pomeroy SL: **Medulloblastoma: Molecular Classification-Based Personal Therapeutics.** *Neurotherapeutics* 2017, **14:**265-273.

16. Hui AB, Lo KW, Yin XL, Poon WS, Ng HK: **Detection of multiple gene amplifications in glioblastoma multiforme using array-based comparative genomic hybridization.** *Lab Invest* 2001, **81:**717-723.

17. Guo XX, Su J, He XF: **A 4-gene panel predicting the survival of patients with glioblastoma.** *J Cell Biochem* 2019, **120:**16037-16043.

18. Breiman L: **Random Forests.** *Machine Learning* 2001, **45**.

19. Sarica A, Cerasa A, Quattrone A: **Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review.** *Front Aging Neurosci* 2017, **9:**329.

20. Castaneda-Vega S, Katiyar P, Russo F, Patzwaldt K, Schnabel L, Mathes S, Hempel J-M, Kohlhofer U, Gonzalez-Menendez I, Quintanilla-Martinez L, et al: **Machine learning identifies stroke features between species.** *Theranostics* 2021, **11:**3017-3034.

21. Wang X, Yu B, Ma A, Chen C, Liu B, Ma Q: **Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique.** *Bioinformatics* 2019, **35:**2395-2402.

22. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9:**559.

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

24. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28:**27-30.

25. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Research* 2004, **32:**D258-D261.

26. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu LJ, Danila MI, Feng G, Chisholm RL: **Annotating the human genome with Disease Ontology.** *BMC Genomics* 2009, **10 Suppl 1:**S6.

27. Crisman TJ, Zelaya I, Laks DR, Zhao Y, Kawaguchi R, Gao F, Kornblum HI, Coppola G: **Identification of an Efficient Gene Expression Panel for Glioblastoma Classification.** *PLoS One* 2016, **11:**e0164649.

28. Kupp R, Shtayer L, Tien AC, Szeto E, Sanai N, Rowitch DH, Mehta S: **Lineage-Restricted OLIG2-RTK Signaling Governs the Molecular Subtype of Glioma Stem-like Cells.** *Cell Rep* 2016, **16:**2838-2845.

29. Zhang M, Zhang H, Fu M, Zhang J, Zhang C, Lv Y, Fan F, Zhang J, Xu H, Ye D, et al: **The Inhibition of B7H3 by 2-HG Accumulation Is Associated With Downregulation of VEGFA in IDH Mutated Gliomas.** *Frontiers in Cell and Developmental Biology* 2021, **9**.

30. Zhou Z, Luther N, Ibrahim GM, Hawkins C, Vibhakar R, Handler MH, Souweidane MM: **B7-H3, a potential therapeutic target, is expressed in diffuse intrinsic pontine glioma.** *J Neurooncol* 2013, **111:**257-264.

31. Zhang J, Wang J, Marzese DM, Wang X, Yang Z, Li C, Zhang H, Zhang J, Chen CC, Kelly DF, et al: **B7H3 regulates differentiation and serves as a potential biomarker and theranostic target for human glioblastoma.** *Lab Invest* 2019, **99:**1117-1129.

32. Karsy M, Guan J, Cohen AL, Jensen RL, Colman H: **New Molecular Considerations for Glioma: IDH, ATRX, BRAF, TERT, H3 K27M.** *Curr Neurol Neurosci Rep* 2017, **17:**19.

33. Wang Z, Sun D, Chen YJ, Xie X, Shi Y, Tabar V, Brennan CW, Bale TA, Jayewickreme CD, Laks DR, et al: **Cell Lineage-Based Stratification for Glioblastoma.** *Cancer Cell* 2020, **38:**366-379 e368.

Figures

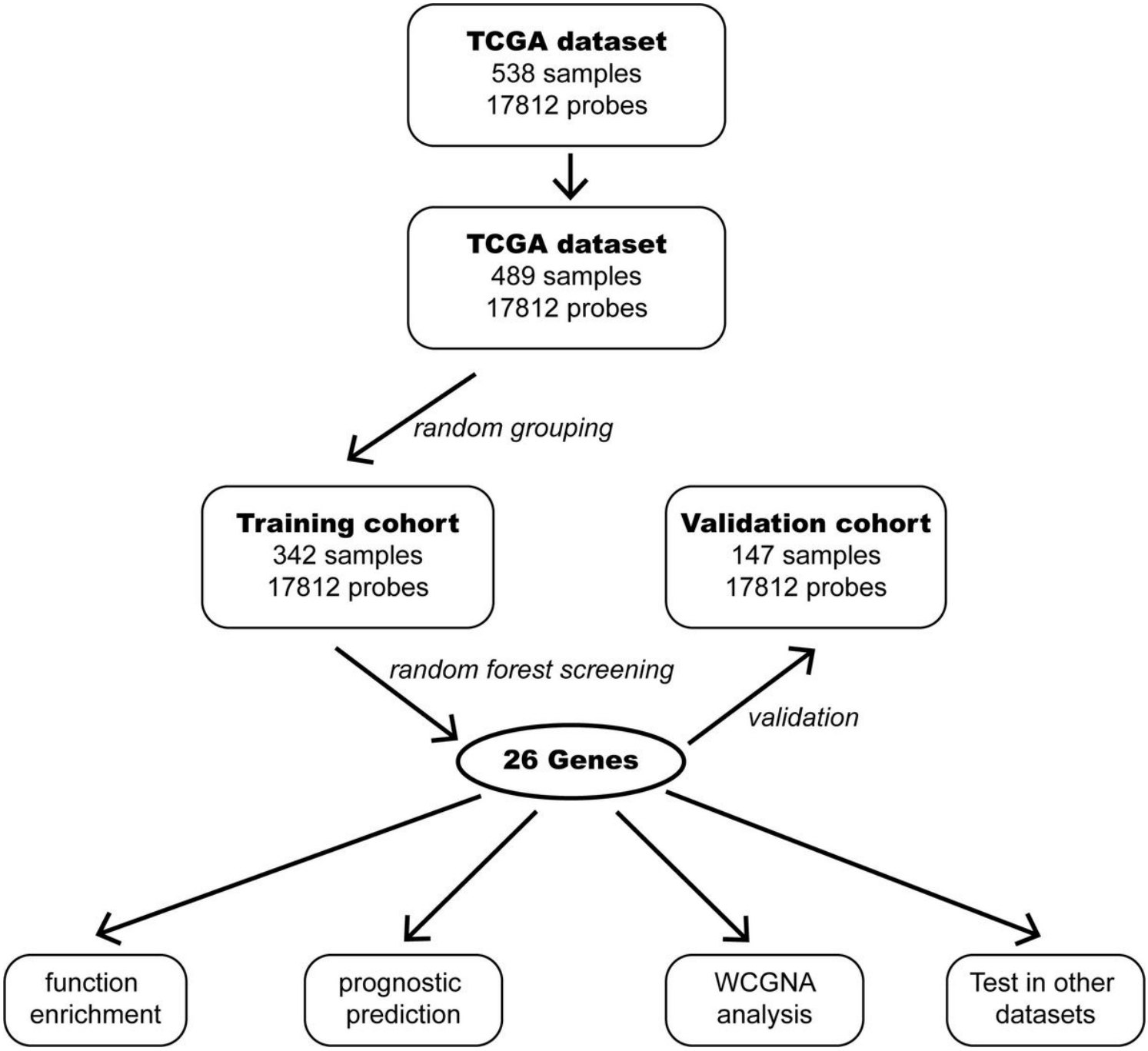

Figure 1

The flowchart of study design.

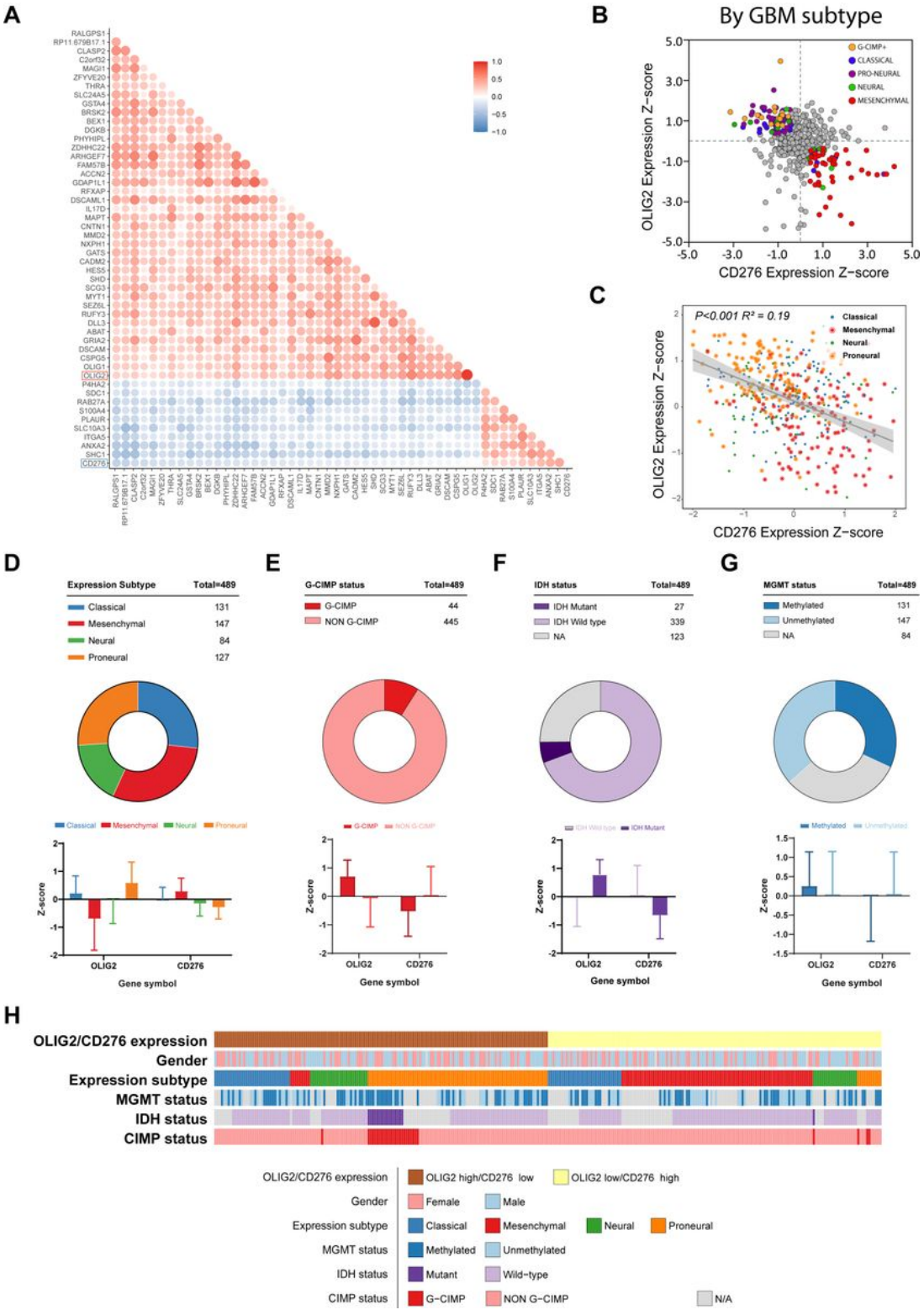

Figure 2

The exclusive expression correlation of OLIG2 and CD276 in different GBM subtypes. The expression of OLIG2 and CD276 is negatively correlated in TCGA GBM dataset (A). OLIG2 expression is high in proneural subtypes, while CD276 in mesenchymal (B,C,D). In GBM with G-CIMP status, IDH mutation status and MGMT methylation status, OLIG2 is highly expressed and CD276 shared exclusive expression

pattern (E,F,G). The full view of the correlation of OLIG2/CD276 expression and other phenotypes is shown (G).

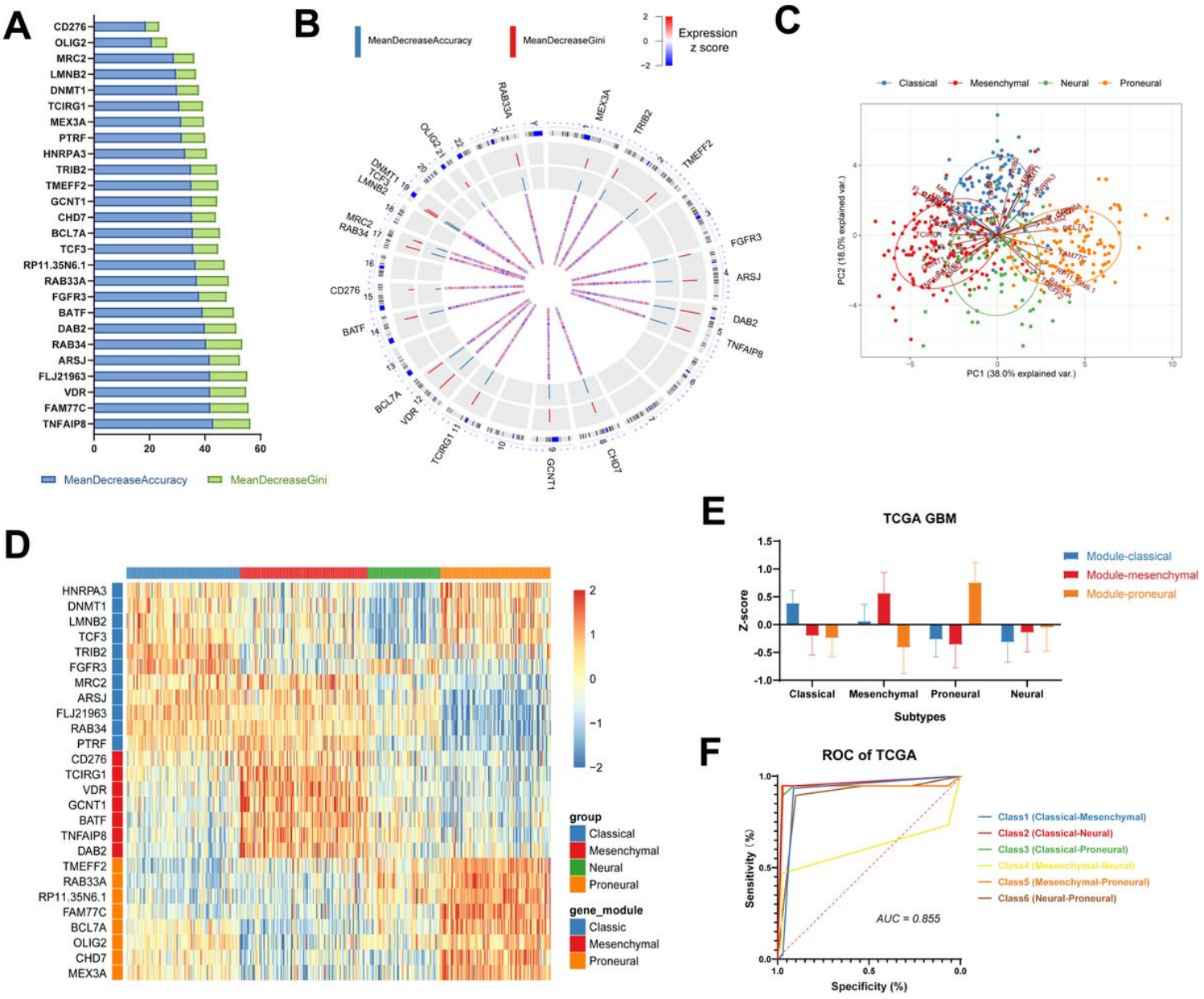

Figure 3

Gene clusters based on OLIG2 and CD276 generated by PCA analysis. 26 genes are obtained by random forest algorithm according to the expression subtypes (A). The full view of the locations of genes on chromatin is shown (B). PCA analysis revealed GBM subtypes can be identified clearly (C). Gene expression of three modules (Module-Classic, Module-Mesenchymal, and Module-Proneural) generated by PCA are shown in heatmap (D). Three gene modules expressed differently in four subtypes (E). ROC curve of the random forest algorithm for subtype classification is shown and the AUC reaches 0.855 (F).

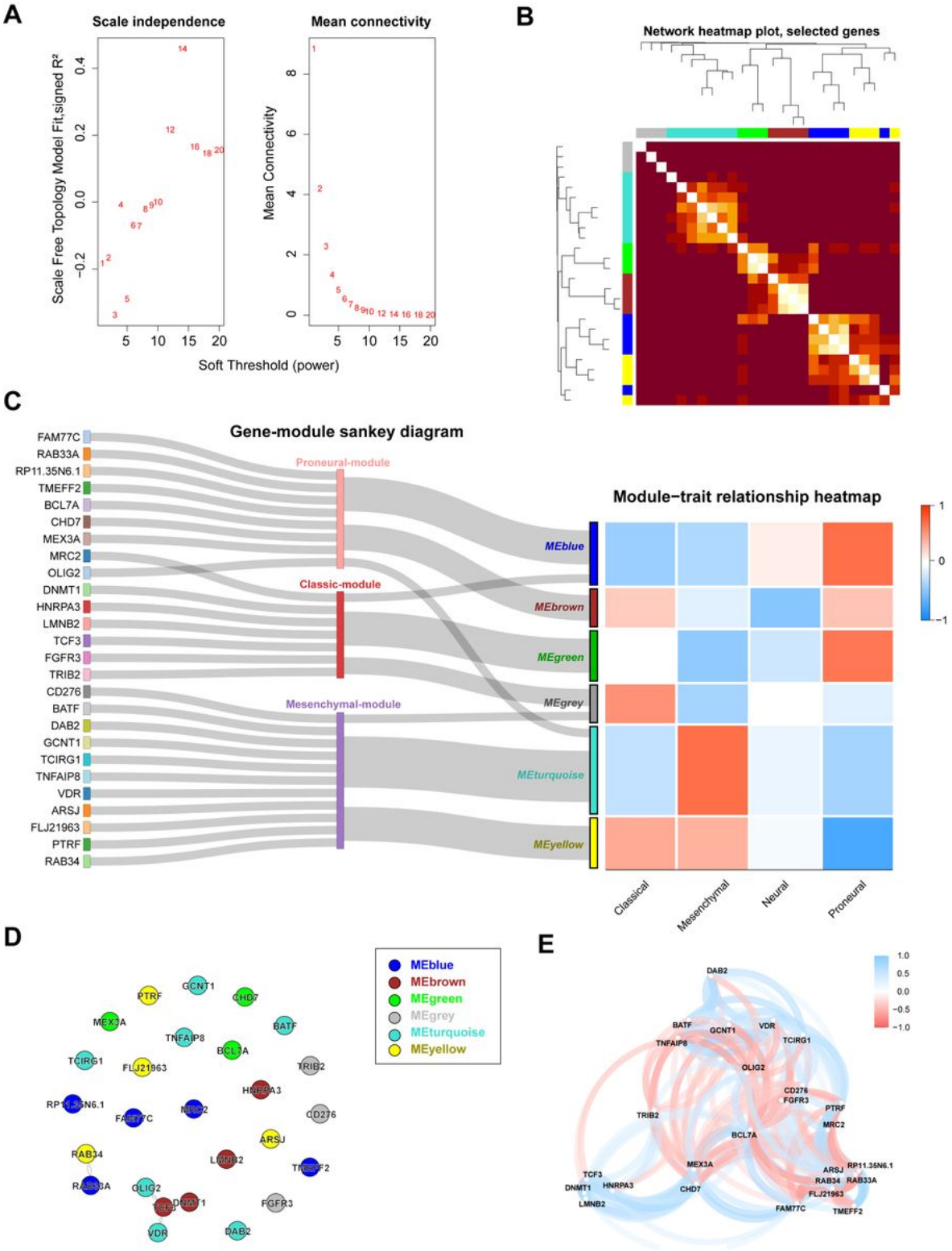

Figure 4

Gene clusters based on OLIG2 and CD276 generated by WCGNA algorithm. The soft threshold with corresponding scale free topology model fit and mean connection is set as 8 (A). TOM heatmap shows good cohesion of six modules generated by WCGNA algorithm (B). The Sankey diagram reveals existed correspondence between the two kinds of modules generated by PCA and WCGNA algorithm (C). RAB33A

and RAB34 were exclusively expressed in mesenchymal and proneural GBM (D). Protein-protein network shows interaction among the gene clusters.

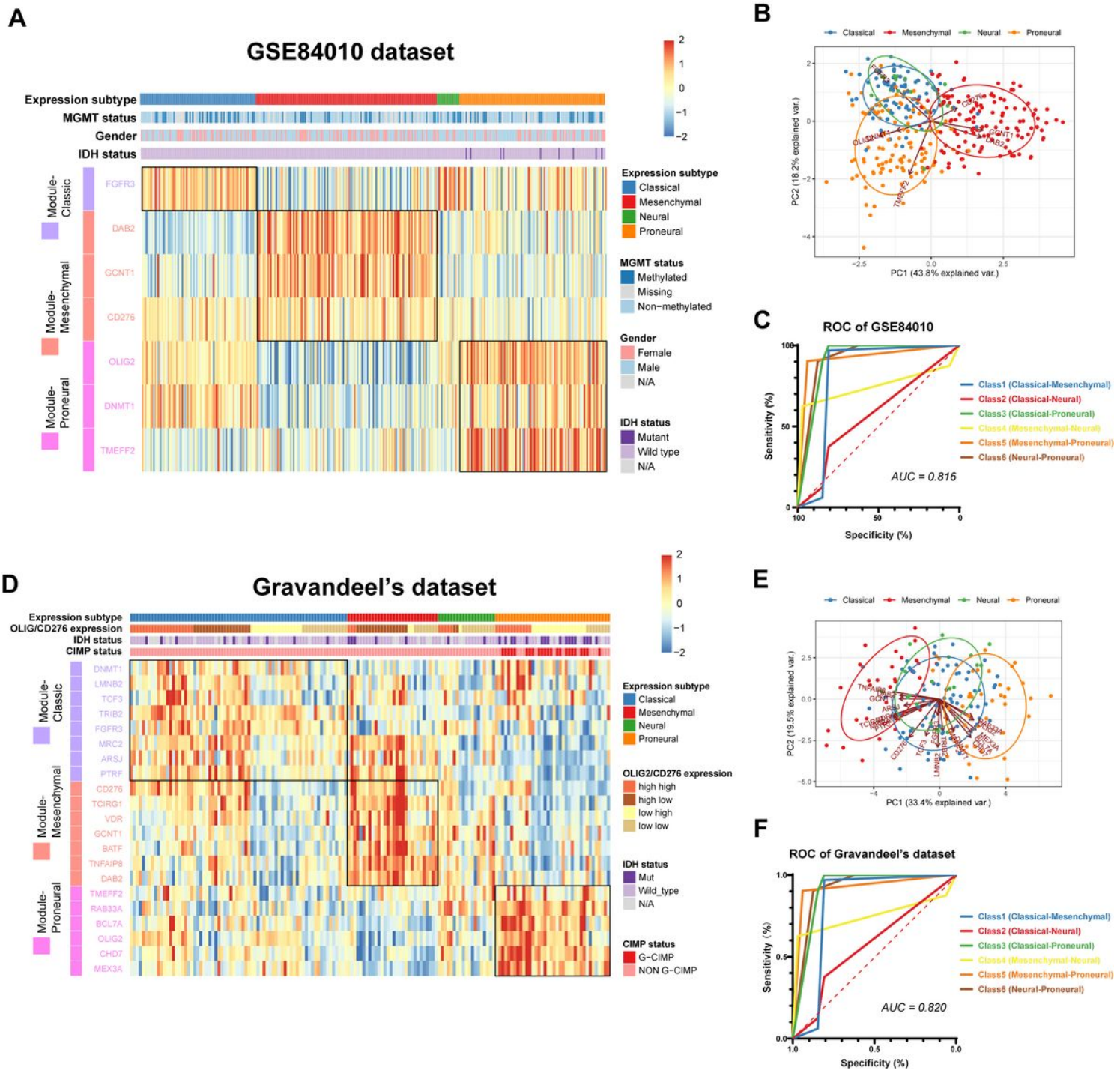

Figure 5

Validation of gene clusters in GSE84010 and Gravandeel's GBM datasets. Heatmap shows good classifying ability of gene clusters in two independent datasets (A, E). PCA analysis reveals gene clusters could distinguish mesenchymal and proneural subtypes (B, E). ROC of random forest classification model reveals good efficacy (0.816 in GSE84010 dataset, 0.820 in Gravandeel's dataset).

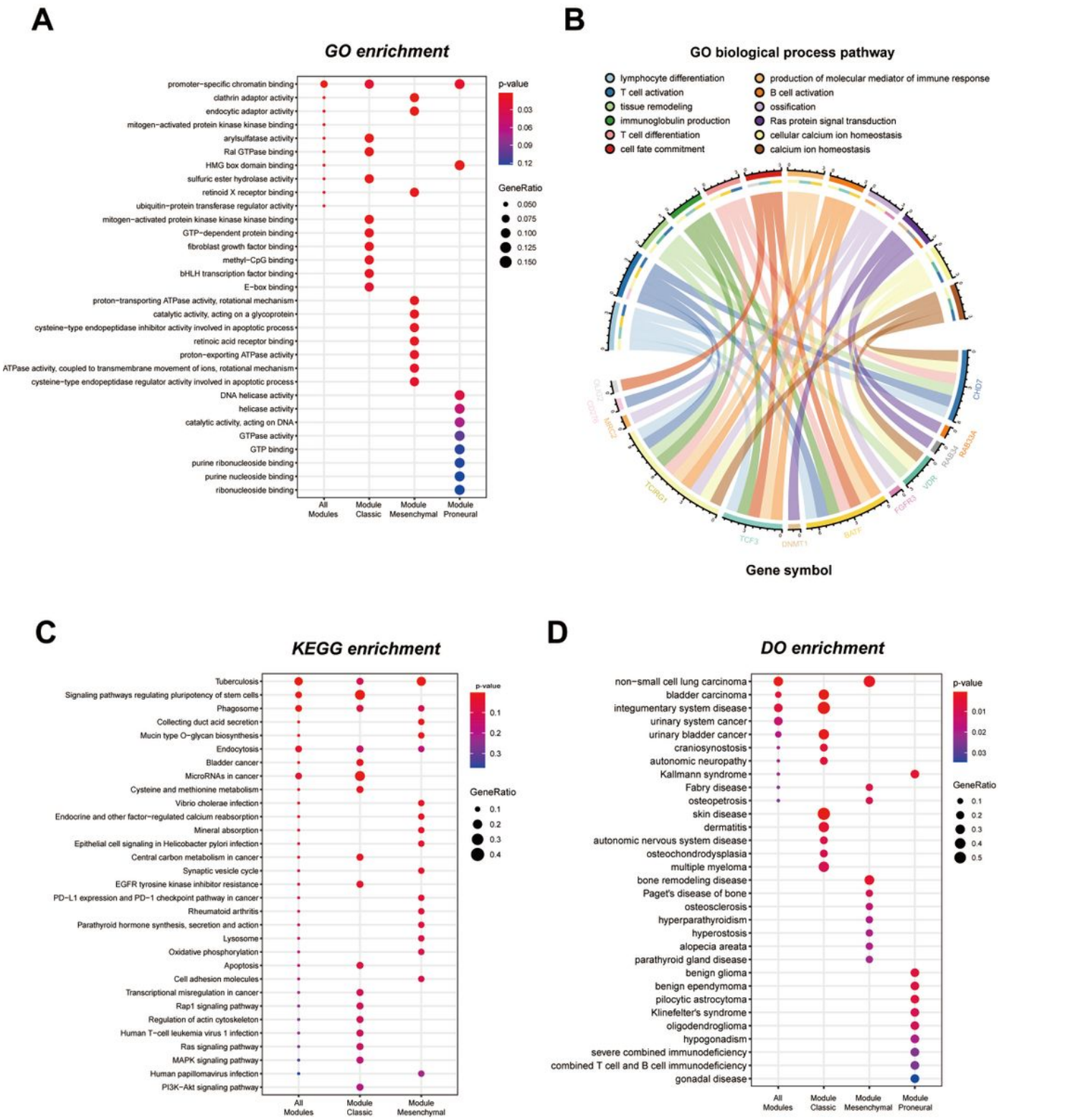

Figure 6

Functional enrichments of gene clusters. GO pathway enrichment revealed both classic and mesenchymal modules are enriched in DNA elements related pathway (A). GO biological process shows lymphocyte differentiation and T cell activation associated with four genes in the cluster (B). KEGG pathways reveals that signaling regulating pluripotency of stem cells is enriched (C). In DO database, genes in the cluster have a close connection with cancers (D).

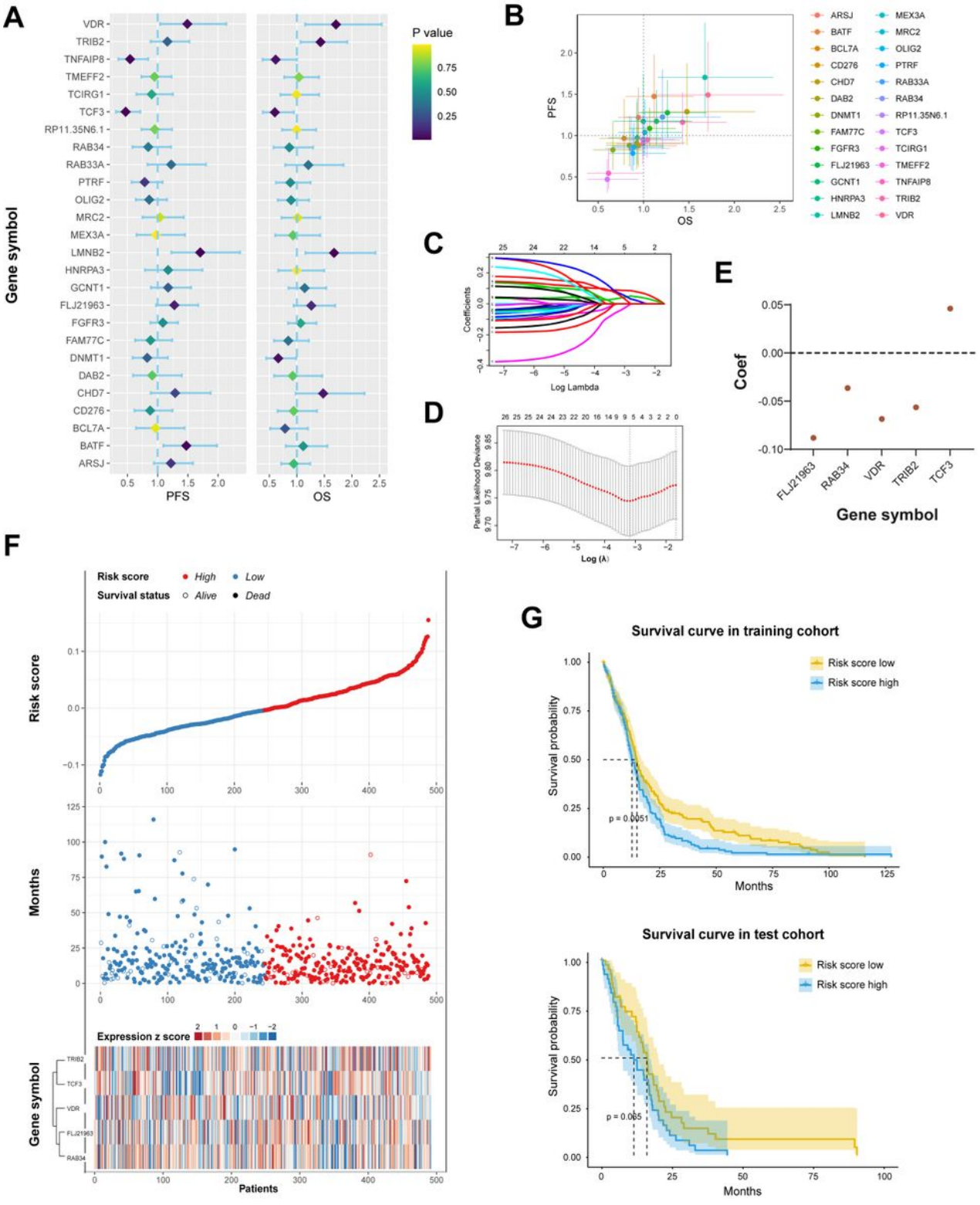

Figure 7

Survival prediction model based on genes in clusters. Multivariate Cox regression analysis revealed the association of 26 genes with OS and PFS (A). The predictive value of genes in cluster for OS and PFS is shown in 2-dimension plot (B). A signature of five genes is obtained by LASSO regression algorithm (C, D). The coefficients of five genes in the signature is shown (E). The full view of the risk score and the

survival status based on five genes signature(F). The survival prediction model is tested in training cohort and test cohort (G).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.docx