Weighted Gene Co-Expression Network Analysis Reveals a New Survival Model for Prognostic Prediction in Ewing Sarcoma

Debao Li (lidebao0952@163.com)
Fifth People's Hospital of Ningxia

Lei Wang
Fifth Peoples Hospital of Ningxia

Guanghui Wang
Fifth People's Hospital of Ningxia

Yaowen Yang
Fifth People's Hospital of Ningxia

Weiyu Yang
Fifth People's Hospital of Ningxia

Hurong Wang
Fifth People's Hospital of Ningxia

Yingui Ma
Fifth People Hospital of Ningxia

Jin Zhang
Fifth People's Hospital of Ningxia

Jianing Tian
Fifth People's Hospital of Ningxia

She Jia
Fifth People's Hospital of Ningxia

Yujie Cong
Fifth People's Hospital of Ningxia

Jing Li
Fifth People's Hospital of Ningxia

Liang Xia
Zhejiang Cancer Hospital

Research

Keywords: Ewing sarcoma, Prognostic model, Gene signature, WGCNA.

Posted Date: May 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-470100/v1

Abstract

Background: Ewing sarcoma (ES) is a malignant bone or soft-tissue cancer that mainly arises in children and young adults. However, the prognosis of Ewing sarcoma remains very poor, and there is no effective prediction method. The aim of our study was to identify a prognostic model for ES patients based on prognosis-associated mRNA expression profiles.

Methods: The GSE17679 dataset was downloaded from the Gene Expression Omnibus (GEO) database. Differently expressed genes (DEGs) between ES and normal control were identified using R package "*limma*". A weighted gene co-expression network analysis (WGCNA) was used to screen gene modules associated with recurrence/metastasis and survival status based on DEGs.

Results: The prognostic model was constructed based on genes in MEbrown module, which was most associated with recurrence/metastasis and survival status, using Kaplan-Meier survival and lasso regression analysis. Sixteen genes were screened to construct the prognostic model. ES patients were grouped into high- and low-risk groups based on the median of risk score calculated for each of them. ES patients in high-risk group have worse survival than patients in low-risk group. The AUCs (Area under the ROC curve) for 1-year, 3-year, and 6-year overall survival were 0.903, 0.995, 0.953.

Conclusions: Taken together, our research constructed a prognostic model which has excellent prediction performance for overall survival of ES patients.

Background

Ewing sarcoma (ES) is an aggressive and rare tumor developed usually in bone, but sometimes in soft tissues as well, which mainly occurred in children and young adults [1]. With the development of therapeutic methods and diagnostic techniques, the prognosis of ES patients has elevated markedly [2, 3]. However, the 5-year survival rate remains less than 20% for patients with metastatic or recurrent tumors [4, 5]. Therefore, the screening of gene modules associated with recurrence/metastasis and survival status and construction of an effective prognostic model which could contribute to predict the survival rate to help the treatment of ES patients is urgently required.

In recent years, high-throughput technologies have been widely used to identify essential driver genes in the processes of tumor [6–9]. Weighted gene co-expression network analysis (WGCNA) is a valuable tool for the analysis of the gene expression patterns in multiple samples, which could efficiently identify and screen co-expressed gene modules and key biomarkers [10]. The prognosis-related gene signature and prognostic model have shown a great advantage for tumor prognosis prediction [11]. However, there are few effective prognostic models in ES for now.

In our study, human ES expression and clinical data in GSE17679 were downloaded from Gene Expression Omnibus (GEO). The R package "*Limma*" was used to get differently expressed genes (DEGs) between tumor and normal tissues; the WGCNA tool was used to screen gene modules associated with

recurrence/metastasis and survival status based on DEGs. We further performed enrichment analysis to reveal the biological functions and pathways of selected gene modules. In addition, a prognostic model was established based on prognosis-related gene signature, which was screened from the gene module associated with recurrence/metastasis and survival status, using Kaplan-Meier survival and lasso regression analysis. More importantly, we tested the prognostic model using ROC analysis, which showed a great prediction performance for overall survival of ES patients. The current study provided new possibilities for developing prognosis risk assessment models of ES patients.

Result

Identification of DEGs in ES

We obtained the gene expression profile and clinical information of ES and normal tissues from GSE17679. Differently expressed genes (DEGs) between ES and normal tissues were identified using R package "*limma*". A total of 4855 DEGs (2739 up-regulated genes and 2116 down-regulated genes) were screened based on the cut-off criteria of a adjust p-value < 0.01 and |log2FC (fold change) | > 1, in which 3779 protein-coding genes were chosen for subsequent analysis (**Figure 1A-B**).

Construction and identification of key modules

Next, 3779 protein-coding DEGs were used to perform WGCNA using R package "*WGCNA*". A total of 88 ES samples with clinical features including age, sex (female and male), survival time (follow-up time), Status (alive and dead), and RM.status (recurrence/metastasis status) (**Figure 2A**). The most appropriate soft threshold to construct a scale-free network was 5 (R^2 = 0.9) (**Figure 2B**). A total of twelve modules were identified (module size ≥ 30 and abline = 0.25) in the network (MEbrown, MEmagenta, MEgreen, MEblue, MEpurple, MEblack, MEsalmon, MEtan, MEgreenyellow, MEpink, MEturquoise, and MEgrey) (**Figure 3A**). Results indicated MEbrown module to be significantly positively associated with recurrence/metastasis status (Status) and survival status (RM.status) and negatively associated with survival time (**Figure 3B-E**). Thus, the MEbrown module was selected as a clinically pivotal module for subsequent analysis.

Functional analysis of MEbrown module

To further explore the potential function of genes in MEbrown module, we conducted the functional analysis using R package "*clusterprofiler*". The functional enrichment results revealed that these genes were mainly associated with chromosome segregation and nuclear division in BP (biological process) category. For CC (cellular component) category, the genes in MEbrown module were mainly distributed in chromosomal region and spindle region. In MF (molecular function) category, the genes in MEbrown module mainly enriched in single-stranded DNA binding and catalytic activity, acting on DNA (**Figure 4A**). In addition, KEGG pathway analysis indicated the enrichment of MEbrown module mainly in Cell cycle pathway (**Figure 4B**).

Construction of the PPI network

To further explore the relationship between genes in MEbrown module, The STRING website was used to construct the PPI network. We identified three significant sub-modules using cluster analysis of PPI network in Cytoscape MCODE plug-in. These three sub-modules are respectively composed of 82, 24 and 11 genes (**Figure 5A-C**).

Construction of the Prognostic Model for ES

We further performed Kaplan-Meier survival analysis and screened 49 genes with $p < 0.00001$ (**Supplementary Table 1**). The prognostic model was constructed based on the 49 genes using lasso regression (**Figure 6A-B**). A total of 16 genes, including NOP2, DDB2, PTPN22, ASPM, GLB1L2, EMILIN1, TTK, CDCA3, NEK2, P4HTM, CENPW, PSMG3, RAD51AP1, SQLE, YWHAG, and SLC16A4 were selected by lasso regression (**sup-Figure 1A-P**). The prognosis index for each patient was calculated by the established formula: The Prognosis Index (PI) = (0.42 * expression of NOP2) + (-0.84 * expression of DDB2) + (-0.21 * expression of PTPN22) + (1.02 * expression of ASPM) + (0.77 * expression of GLB1L2) + (-0.07 * expression of EMILIN1) + (-0.04 * expression of TTK) + (-0.50 * expression of CDCA3) + (-0.15* expression of NEK2) + (-0.11 * expression of P4HTM) + (-0.07 * expression of CENPW) + (0.08 * expression of PSMG3) + (-0.43 * expression of RAD51AP1) + (-0.10 * expression of SQLE) + (1.10 * expression of YWHAG) + (-0.17 * expression of SLC16A4). All patients were divided into two groups (high- and low-risk groups) according to the median risk score. Kaplan-Meier survival analysis showed the prognostic model to be a good predictor of ES patient prognosis. The overall survival of ES patients was significantly poorer in the high-risk group (**Figure 6C-D**). Furthermore, ROC analysis was used to determine the accuracy of this signature in predicting the survival. As shown in Figure 6E, this prognostic model showed high sensitivity and specificity with the AUCs for 1-year, 3-year, and 6-year overall survival to be 0.903, 0.995, 0.953. Taken together, the prognostic model could serve as a good predictor of overall survival of ES patients.

Discussion

Understanding the pathogenesis of ES is essential for identifying new biomarkers to develop new therapeutic strategies [12, 13]. However, metastasis and recurrence are the main causes of poor prognosis of ES [14, 15]. Thus, identification of novel biomarkers related to metastasis and recurrence to predict ES patients' survival might help to customize more personalized therapy and would be able to improve their prognosis. Researchers have established a predictive model of OS of ES patients based on T2-Weighted Images and Contrast-Enhanced T1-Weighted Images [16]. However, there are still few studies on ES prognosis model based on gene expression profiles.

In this study, we first identified 4855 DEGs (2739 up-regulated genes and 2116 down-regulated genes) in ES tissues compared to normal tissues from GSE17679. Next, a total of 88 ES samples with clinical features including age, sex (female and male), survival time (follow-up time), status (alive and dead), and RM.status (recurrence/metastasis status) were included for WGCNA. The results indicated MEbrown

module to be significantly positively associated with recurrence/metastasis status (Status) and survival status (RM.status) and negatively associated with survival time. In addition, KEGG pathway analysis indicated the enrichment of genes in MEbrown module mainly in Cell cycle pathway.

We further performed Kaplan-Meier survival analysis using genes in MEbrown module. A total of 49 genes with $p < 0.00001$ were selected to construct prognostic model using lasso regression analysis. We identified a gene signature including NOP2, DDB2, PTPN22, ASPM, GLB1L2, EMILIN1, TTK, CDCA3, NEK2, P4HTM, CENPW, PSMG3, RAD51AP1, SQLE, YWHAG, and SLC16A4 and established a prognostic model to predict the risk of ES patients. The prognostic model showed high sensitivity and specificity with the AUCs for 1-year, 3-year, and 6-year overall survival to be 0.903, 0.995, 0.953.

In conclusion, we established an efficient prognostic model with great efficacy to predict overall survival, which may help improve early diagnosis and treatment, enhancing the clinical outcomes of ES patients.

Materials And Methods

Data collection

Series matrix files with expression profile and clinical data of GSE17679 were downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database. The platform of GSE17679 was GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array). The datasets of GSE17679 comprised 88 ES patient samples and 18 normal samples.

Identification of DEGs

R package "*Limma*" was used to screen DEGs between the ES samples and normal control samples. The adjust $P < 0.01$ and the absolute value of log2 fold change $|log2FC| > 1$ was the screening threshold. A total of 3779 protein coding DEGs were selected for subsequent analysis.

Weighted gene co-expression network analysis

A total of 88 ES samples with clinical data (including age, sex, survival time, survival status and recurrence/metastasis status) were included in the WGCNA using R package "WGCNA" [10]. The WGCNA was performed as previously described [17,18]. Briefly, the pearson correlation coefficient was used to construct the correlation matrix. The co-expression similarity matrix was constructed by the absolute value of correlation between transcriptional data. We used the formula $Axy = |Cxy| \beta$ (Axy: the adjacency between gene x and gene y; Cxy: Pearson correlation between gene x and gene y) to establish the weighted adjacency matrix. The topological overlap matrix (TOM) adds the adjacent genes generated by related networks to calculate the corresponding differences. Based on the "hclust" algorithm, the TOM diversity measure was clustered by gene hierarchy, and the genes with highly collaborative changes were divided into a module. The minimum gene number of the gene tree was 40. The module was constructed by dynamic branch cutting method.

Enrichment analysis

Gene Ontology (GO) analysis was performed using EnrichGO function of R package "*clusterProfiler*" [19], with the parameters: ont = "all," pvalue-Cutoff = 0.05, qvalue-Cutoff = 0.05. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was performed by EnrichKEGG function of R package "clusterProfiler", with the parameters: keyType = "kegg" pvalue-Cutoff = 0.05 and qvalue-Cutoff = 0.05.

PPI network construction and module analysis

The protein-protein interaction network (PPI) analysis was performed by STRING database (https://string-db.org/) using modules associated with survival status in WGCNA. The Network was visualized by cytoscape software (version:3.7.2) (https://cytoscape.org/). The significant modules in PPI were identified by the Molecular Complex Detection (MCODE) plug-in of cytoscape software.

Construction of the prognostic model

Kaplan-Meier survival and lasso regression analyses were employed to investigate the OS related genes in modules associated with survival status in WGCNA. A total of 49 genes that meet the standard of P<0.00001 in Kaplan-Meier survival were selected for Lasso-penalized Cox regression. Lasso statistical algorithm was conducted using "glmnet" package in the R software (version 3.5.3, https://www.r-project.org/). Based on the expression level of each gene, Lasso identified the eligible mRNAs for the risk system and generated the corresponding coefficients for each of them. A sixteen-gene signature was established finally. The Prognosis Index (PI) = (0.42 * expression of NOP2) + (-0.84 * expression of DDB2) + (-0.21 * expression of PTPN22) + (1.02 * expression of ASPM) + (0.77 * expression of GLB1L2) + (-0.07 * expression of EMILIN1) + (-0.04 * expression of TTK) + (-0.50 * expression of CDCA3) + (-0.15* expression of NEK2) + (-0.11 * expression of P4HTM) + (-0.07 * expression of CENPW) + (0.08 * expression of PSMG3) + (-0.43 * expression of RAD51AP1) + (-0.10 * expression of SQLE) + (1.10 * expression of YWHAG) + (-0.17 * expression of SLC16A4). The ES patients of GSE17679 with survival data were grouped into high- and low-risk groups based on the median risk score. The Kaplan-Meier survival curves for the patients in high- and low-risk groups were performed. To estimate the predictive efficiency of the gene-signature model, time-dependent receiver operating characteristic (ROC) curve analyses were conducted.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

D.L., L.W., G.W. and Y.Y. designed the study. D.L., W.Y., H.W., Y.M. and J.Z. performed data analysis. J.T., S.J. and L.X. wrote the manuscript and helped with validation.

References

1. Granowetter L, West DC. The Ewing's sarcoma family of tumors: Ewing's sarcoma and peripheral primitive neuroectodermal tumor of bone and soft tissue. Cancer Treat Res. 1997;92:253–308.
2. Ozaki T. Diagnosis and treatment of Ewing sarcoma of the bone: a review article. J Orthop Sci. 2015;20:250–63.
3. Grünewald TGP, Cidre-Aranaz F, Surdez D, Tomazou EM, de Álava E, Kovar H, Sorensen PH, Delattre O. Dirksen. Ewing sarcoma. Nat Rev Dis Primers. 2018;4:5.
4. Bacci G, Mercuri M, Longhi A, Bertoni F, Barbieri E, Donati D, et al. Neoadjuvant chemotherapy for Ewing's tumour of bone: recent experience at the Rizzoli Orthopaedic Institute. Eur J Cancer. 2002;38:2243–51.
5. Burdach S, Jürgens H. High-dose chemoradiotherapy (HDC) in the Ewing family of tumors (EFT). Crit Rev Oncol Hematol. 2002;41:169–89.
6. Moriarity BS, Otto GM, Rahrmann EP, Rathe SK, Wolf NK, Weg MT, et al. A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis. Nat Genet. 2015;47:615–24.
7. Smida J, Xu H, Zhang Y, Baumhoer D, Ribi S, Kovac M, et al. Genome-wide analysis of somatic copy number alterations and chromosomal breakages in osteosarcoma. Int J Cancer. 2017;141:816–28.
8. Kuijjer ML, Rydbeck H, Kresse SH, Buddingh EP, Lid AB, Roelofs H, et al. Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. Genes Chromosomes Cancer. 2012;51:696–706.
9. Zhang H, Guo L, Zhang Z, Sun Y, Kang H, Song C, et al. Co-expression network analysis identified gene signatures in osteosarcoma as a predictive tool for lung metastasis and survival. J Cancer. 2019;10:3706–16.
10. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
11. Liu GM, Xie WX, Zhang CY, Xu JW. Identification of a four-gene metabolic signature predicting overall survival for hepatocellular carcinoma. J Cell Physiol. 2020;235:1624–36.
12. Kondo T. Current status and perspectives of patient-derived models for Ewing's sarcoma. Cancers (Basel). 2020;12:10.

13. Lin SH, Sampson JN, Grünewald T, Surdez D, Reynaud S, Mirabeau O, et al. Low-frequency variation near common germline susceptibility loci are associated with risk of Ewing sarcoma. PLoS One. 2020;15:e237792.
14. Sciandra M, De Feo A, Parra A, Landuzzi L, Lollini PL, Manara MC, et al. Circulating miR34a levels as a potential biomarker in the follow-up of Ewing sarcoma. J Cell Commun Signal. 2020;14:335–47.
15. Zhou Z, Yang Y, Wang F, Kleinerman ES. Neuronal repressor REST controls ewing sarcoma growth and metastasis by affecting vascular pericyte coverage and vessel perfusion. Cancers (Basel). 2020;12:1405.
16. Dai Y, Yin P, Mao N, Sun C, Wu J, Cheng G, et al. Differentiation of pelvic osteosarcoma and ewing sarcoma using radiomic analysis based on T2-weighted images and contrast-enhanced t1-weighted images. Biomed Res Int. 2020; 9078603.
17. Liu Z, Li M, Hua Q, Li Y, Wang G. Identification of an eight-lncRNA prognostic model for breast cancer using WGCNA network analysis and a Cox–proportional hazards model based on L1-penalized estimation. Int J Mol Med. 2019;44:1333–43.
18. Ding M, Li F, Wang B, Chi G, Liu H. A comprehensive analysis of WGCNA and serum metabolomics manifests the lung cancer-associated disordered glucose metabolism. J Cell Biochem. 2019;120:10855–63.
19. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16, 284–287.

Figures

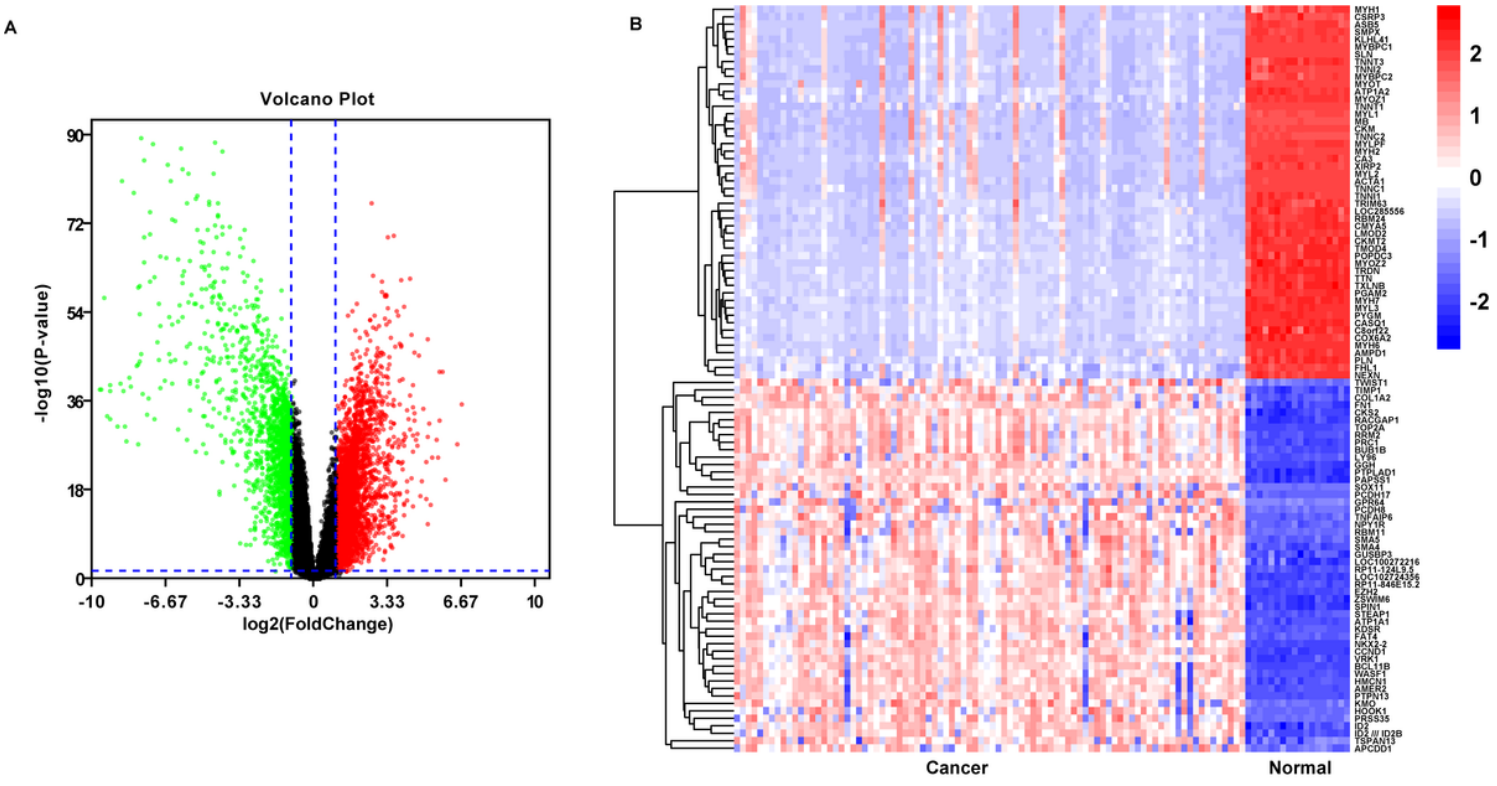

Figure 1

Identification of DEGs in ES. (A), Volcano plot of significance of gene expression difference between ES tissues and normal tissues in GSE17679. (B), heatmap of top 50 up-regulated genes and top 50 down-regulated genes of cancer vs. normal group.

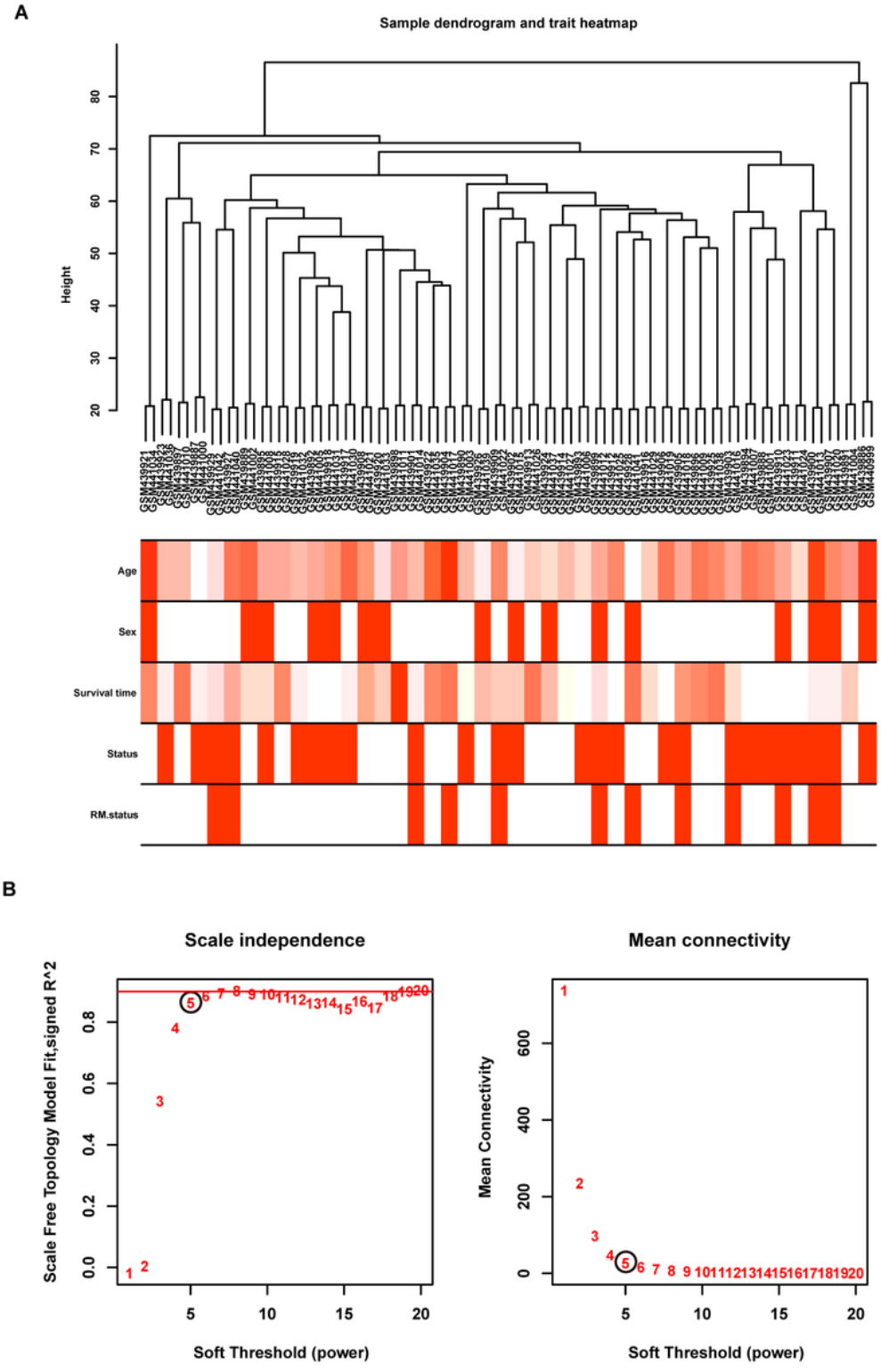

Figure 2

Clustering dendrogram of ES samples. (A), Clustering dendrogram of 88 ES samples with clinical features including age, sex (female and male), survival time (follow-up time), Status (alive and dead), and RM.status (recurrence/metastasis status). (B), The scale-free index for various soft-threshold powers and the mean connectivity for various soft-threshold powers.

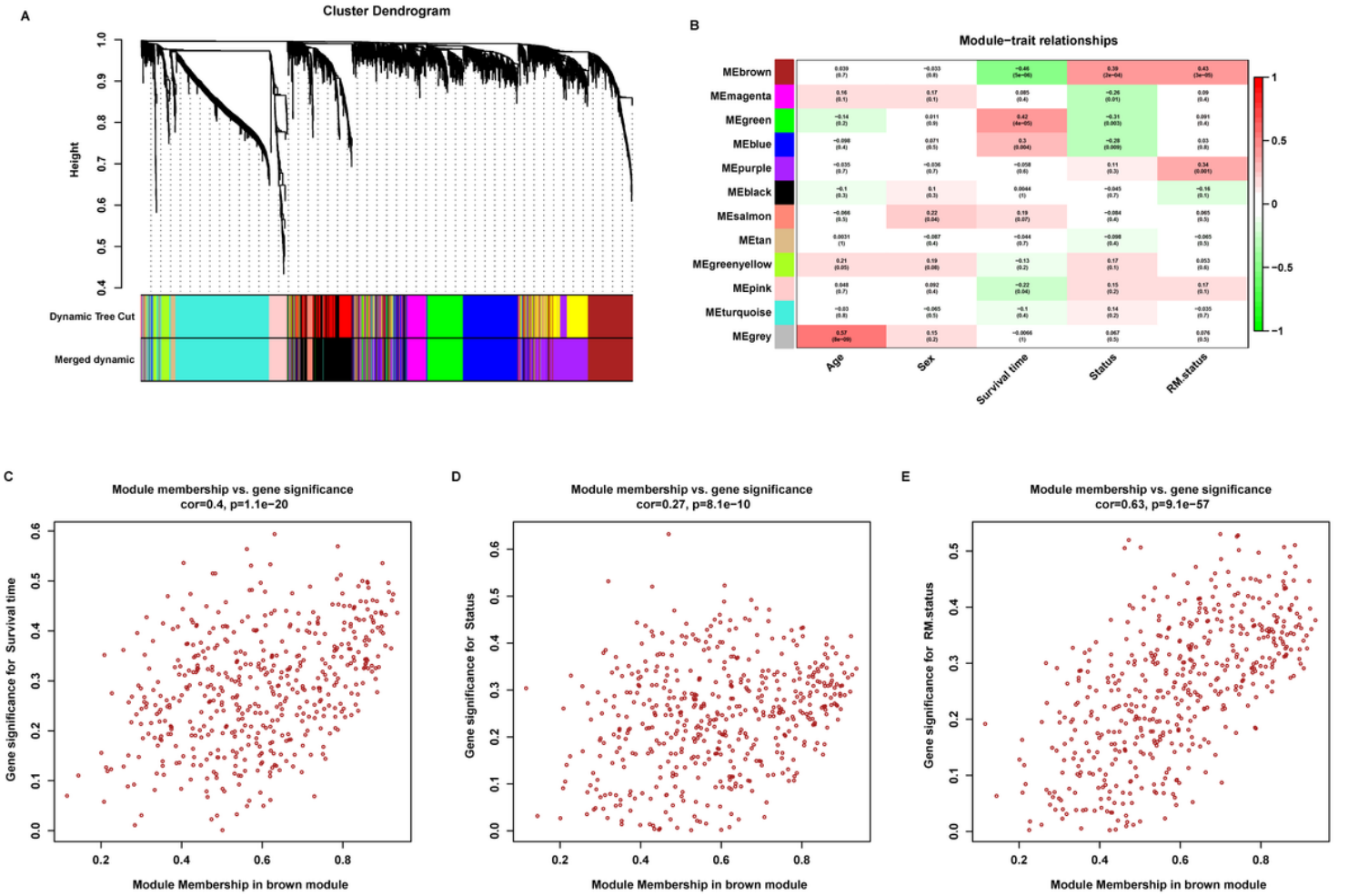

Figure 3

Identification of key modules. (A), The TOM cluster dendrogram: A branch of the tree corresponds to a set of highly related genes. Dynamic tree cut represents the original module and merged dynamic represents the final module. (B), The correlation heatmap represents the relationship between module eigengenes and clinical traits of ES patients. (C-E), Scatterplots represent the MEbrown module to have strong correlation with survival time (C), survival status (D), and recurrence/metastasis status (E).

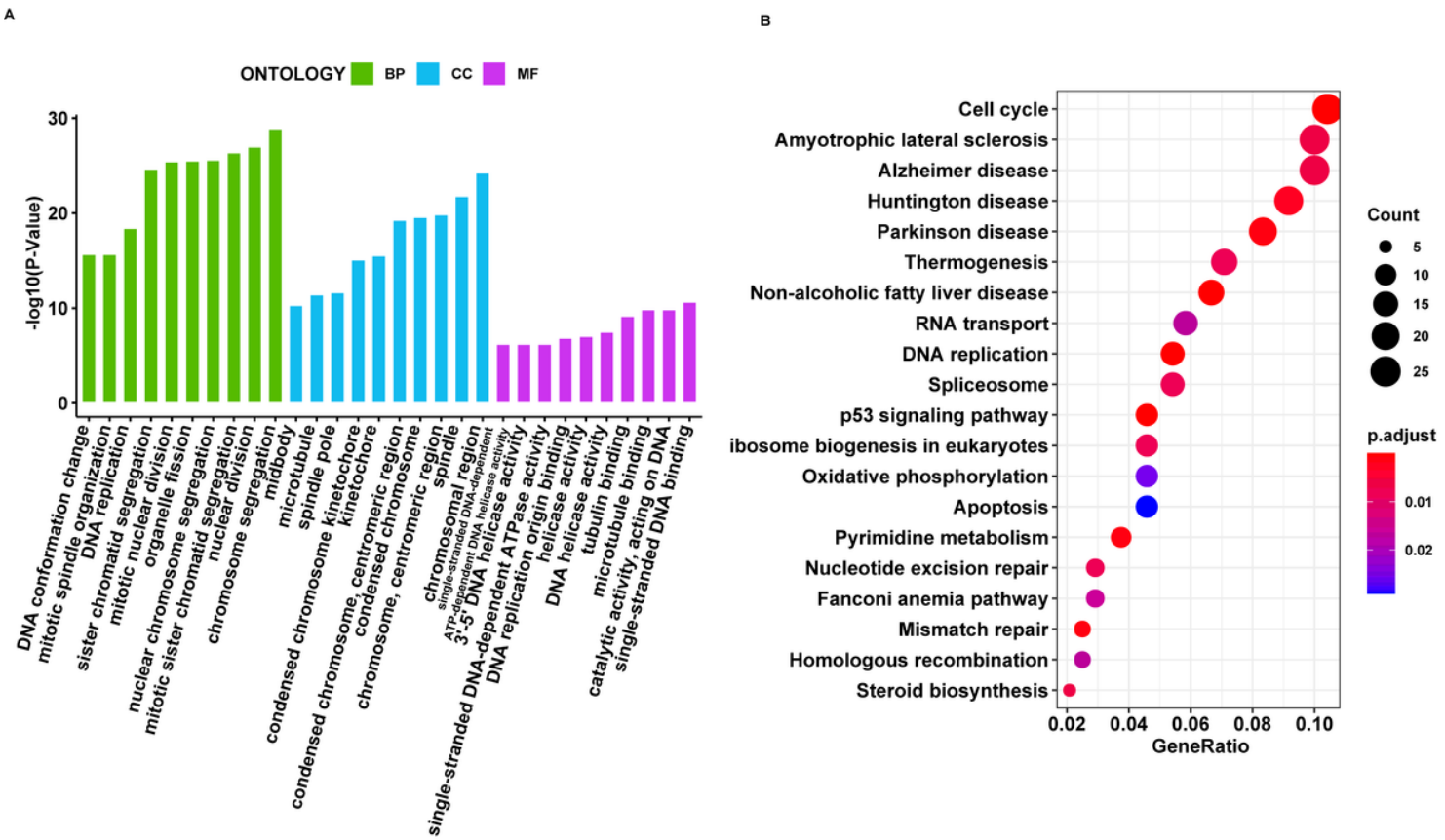

Figure 4

Functional analysis of MEbrown module. (A), Gene Ontology (GO) enrichment analysis of genes in MEbrown module. Top 10 terms in BP, CC, and MF were showed. (B), Significant KEGG pathways of genes in MEbrown module.

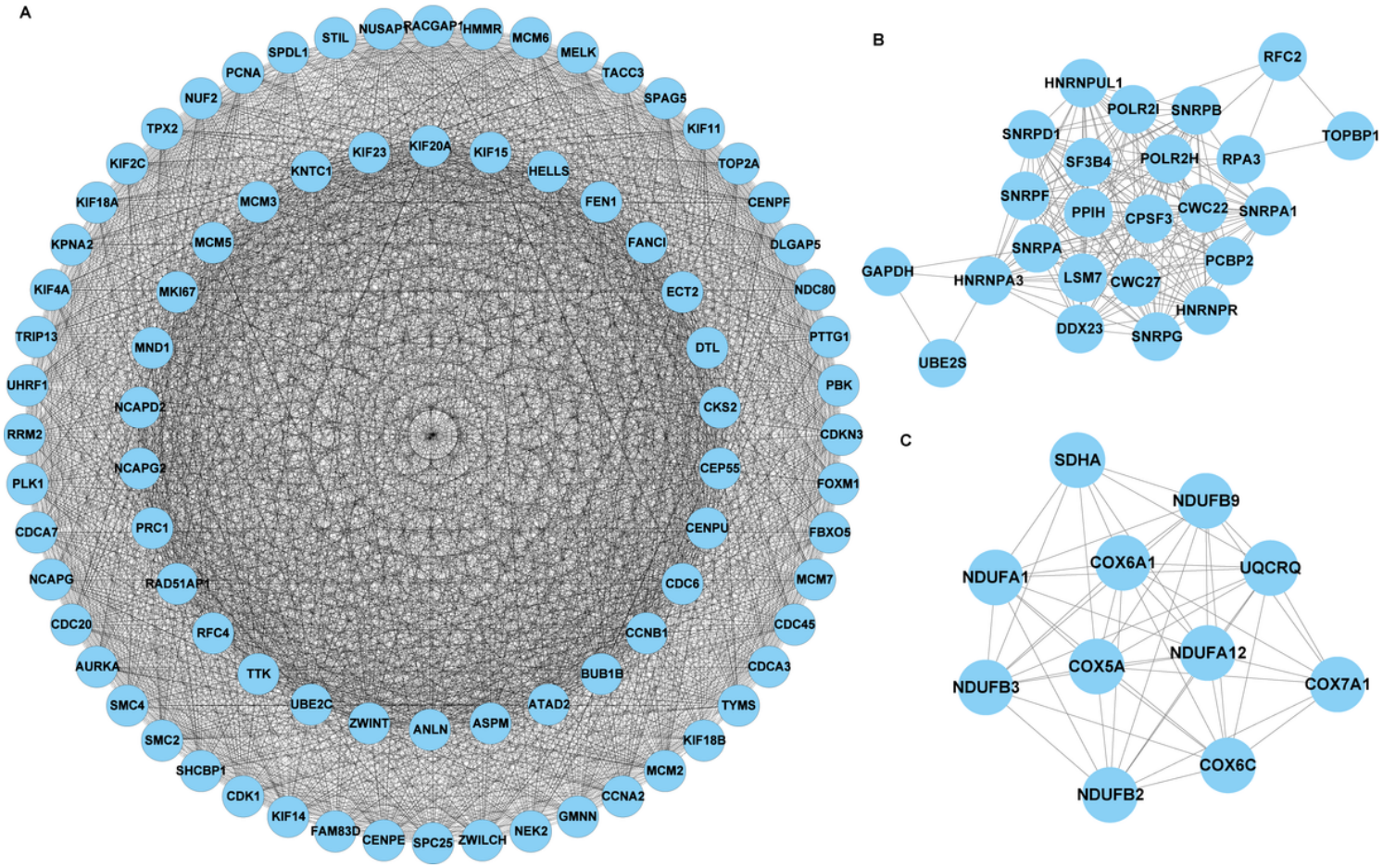

Figure 5

Protein-protein interactions network of genes in MEbrown module. (A-C), Top 3 sub-modules of protein-protein interactions network were showed.

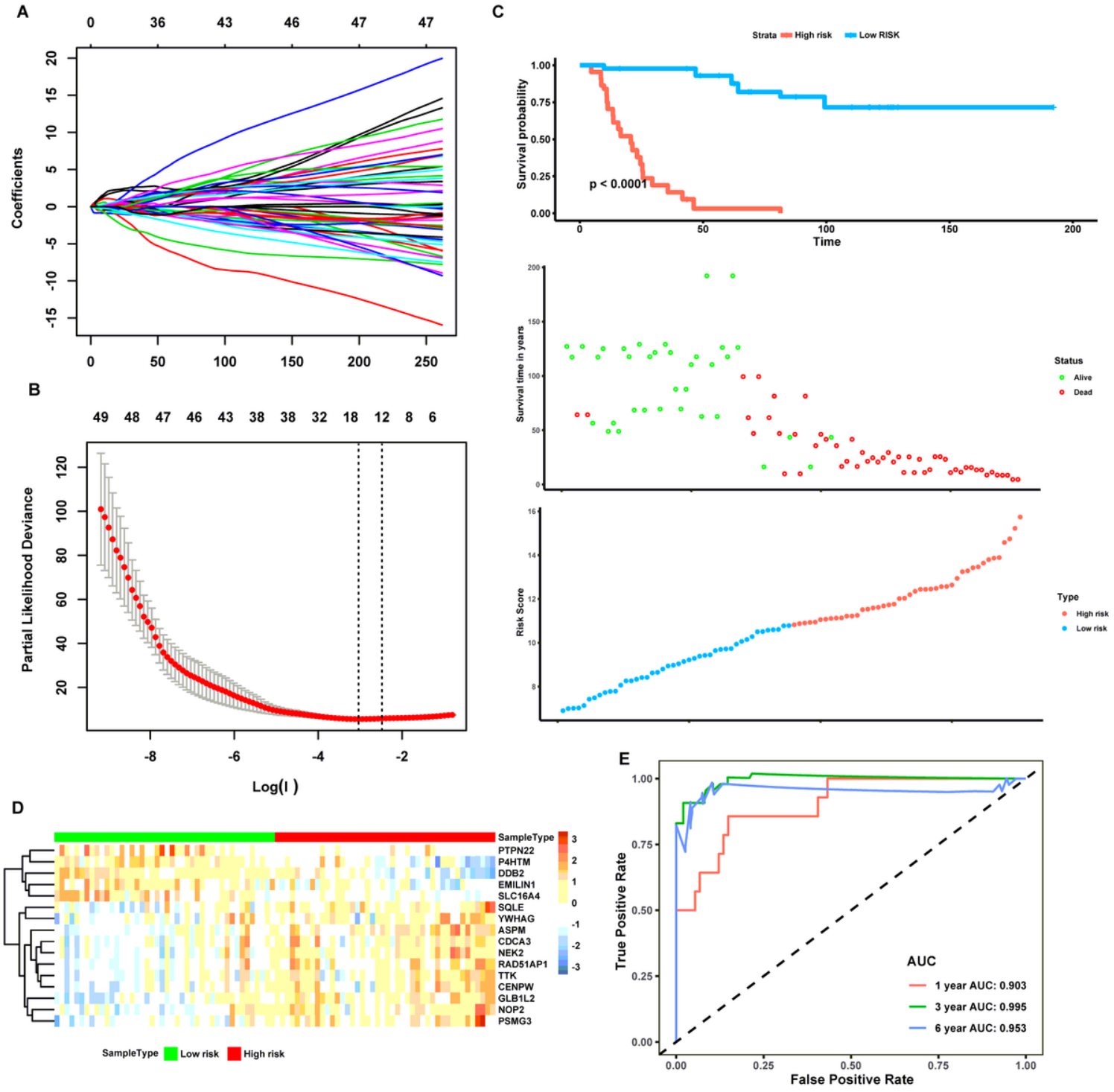

Figure 6

The construction of prognostic model. (A), Log (Lambda) value of the 49 genes in LASSO model. (B), The most appropriate log (Lambda) value in the LASSO model. (C), Kaplan–Meier analysis, survival status distribution, and risk score distribution in the GSE17679. (D), Heatmap of the expression of 16 genes in the GSE17679. (E), ROC curves of the sensitivity and specificity of the 1-, 3-, and 6-year survival using the prognostic model.

Supplementary Files

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

- SupplementaryTable1KaplanMeiersurvivalanalysisresults.csv
- supFigure1.tif
- supplementaryfigurelegends.docx