JBUON 2021; 26(4): 1252-1259
ISSN: 1107-0625, online ISSN: 2241-6293 • www.jbuon.com
Email: editorial_office@jbuon.com

ORIGINAL ARTICLE

Identification of sixteen metabolic genes as potential biomarkers for colon adenocarcinoma

Fuqiang Zhao[1,2], Yanlong Liu[1], Xinyue Gu[1], Bomiao Zhang[1], Chengxin Song[1], Binbin Cui[1]

[1]Department of Colorectal Surgery, Affiliated Tumor Hospital of Harbin Medical University, Harbin 150040, Heilongjiang Province, China. [2]Department of Surgical Oncology, The Second Affiliated Hospital of Qiqihar Medical University, Qiqihar 161006, Heilongjiang Province, China.

Summary

Purpose: *To identify some key prognosis-related metabolic genes (PRMG) and establish a clinical prognosis model for colon adenocarcinoma (COAD) patients.*

Methods: *We used The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) to obtain gene expression profiles of COAD, and then identified differentially expressed prognostic-related metabolic genes through R language and Perl software, Through univariate Cox analysis and least absolute shrinkage and selection operator (LASSO) Cox analysis to obtain target genes, established metabolic genes prognostic models and risk scores. Through Cox regression analysis, independent risk factors affecting the prognosis of COAD were analyzed, and receiver operating characteristics (ROC) curve analysis of independent prognostic factors was performed and a nomogram for predicting overall survival was constructed. We performed the consistency index (C-index) test and decision curve analysis (DCA) on the nomogram, and used gene set enrichment analysis (GSEA) to identify the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of model genes. We selected PRMG based on the expression of metabolic genes, and used LASSO Cox regression to construct 16 metabolic gene models (SEPHS1, P4HA1, ENPP2, PTGDS, GPX3, CP, ASPA, POLR3A, PKM, POLR2D , XDH, EPHX2, ADH1B, HMGCL, GPD1L and MAOA).*

Results: *The risk score generated from our model can well predict the survival prognosis of COAD. A nomogram based on the clinicopathological characteristics and risk scores of COAD can personally predict the overall survival rate of COAD patients.*

Conclusions: *The risk score based on the expression of 16 metabolic genes can effectively predict the prognosis of patients with COAD.*

Key words: *colon cancer, GEO, metabolism, prognostic, TCGA*

Introduction

Although significant progress has been made in surgery, radiation therapy, chemotherapy and targeted therapy, colorectal cancer (CRC) is still one of the main types of cancer in terms of global morbidity and cancer-related deaths. All the time, the TNM staging system has been used as three prognostic indicators of the risk of recurrence in CRC patients [1]. But the TNM staging system only considers the anatomical characteristics of the tumor, not its biological ones. The metabolic recoding of tumor cells helps them adapt to the tumor microenvironment. The tumor microenvironment can provide the energy needed to maintain the growth of their malignant tumor cells, including accelerating proliferation, anti-apoptosis, evading immune attack and maintaining cancer stem cell status [2]. Certain genetic drivers of CRC, such as p53 [3] and KRas [4], are well-known regulators of cancer metabolism, and metabolic gene variants promote colorectal cancer [5]. It is currently known that a

Corresponding author: Binbin Cui, MD. Department of Colorectal Surgery, The Affiliated Tumor Hospital of Harbin Medical University, Harbin 150040, Heilongjiang Province, China.
Tel: +86-13351112888; Email: Cuibb201609@163.com
Received: 14/12/2020; Accepted: 21/01/2021

single gene or molecular marker cannot provide a good diagnosis or predict the progression of the disease. A single biomolecular marker is usually unable to predict the survival of patients with COAD, and more and more research institutions are using multi-gene combination to build predictive models for disease diagnosis. TCGA and GEO provide a lot of tumor-related information, such as gene expression, methylation, mutations and clinical parameters [6], which are of clinical significance and cancer biology has created unprecedented opportunities. In this study, we first screened the PRMG through univariate Cox regression based on the expression of metabolic genes, and then used the LASSO to construct an important gene prognostic model. In addition, ROC curve analysis of independent prognostic factors was performed and a nomogram for predicting overall survival was constructed. GSEA shows the way of KEGG enrichment

Methods

TCGA and GEO data download

We downloaded COAD mRNA expression data and clinical data from TCGA (http://portal.gdc.cancer.gov/). A total of 398 COAD samples and 39 normal colon samples were obtained from TCGA for gene expression and prognosis analysis. We organized and annotated the RNA sequencing matrix files of different samples to the genome. The mRNA expression was obtained from the RNA sequencing data matrix file. We downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/) a data set containing 1048 colon carcinoma samples (GSE40976). The data of the TCGA-COAD and GSE40976 samples were collated, extracted, annotated and standardized by "Strawberry Perl 5.32.0" and R language (version 4.0.2) (https://www.r-project.org/).

Screening for differentially expressed metabolic genes

We downloaded the KEGG genes containing metabolic genes from the GSEA website (https://www.gsea-msigdb.org/gsea/index.jsp) and set c2.cp.kegg.v7.0.symbols.gmt, using Strawberry Perl 5.32.0 to extract a total of 921 metabolism-related genes from TCGA-COAD samples. Using R language "Limma" software package and "sva" software package [7], a total of 645 crossover genes were extracted from the TCGA-COAD and GSE40976 data sets and reduced the heterogeneity between the two studies. We used the Wilcoxon test of the R language "Limma" software package to screen for differentially expressed genes (DEG). P value <0.05 and |log2-fold change|> 0.5 were defined as DEG. A total of 225 DEG were screened out. The heat map and volcano map are created by the "pheatmap" software package of the R language (https: //cran.r-project.org/web/packages/pheatmap/).

Construction and verification of prognostic models

TCGA-COAD was used in the trial cohort, and GES40976 was used in the verification cohort. In the mRNA expression data in TCGA-COAD, we used univariate Cox regression analysis to obtain a total of 22 DEG related to patient survival (p value <0.05). In the LASSO Cox regression analysis, p value <0.05 was filtered as a statistically different prognostic gene. After re-sampling 1000 times using the R language "glmnet" software package [8], a predictive model that affected the prognosis was established. The median risk score value was selected as the cutoff value of the COAD cohort, and divided into high-risk group and low-risk group. The "survival" software package and "survminer" software package of R language were used to perform Kaplan-Meier analysis and draw survival curves to verify the correlation between the prognostic model and overall survival. According to the different risk scores of the patients, the risk curve diagram, survival status diagram and heat map were drawn.

Construction and verification of nomogram prediction model

Univariate and multivariate Cox regression analysis were performed to verify whether the model can be used as an independent prognostic factor in TCGA-COAD and GSE40976. The "survival ROC" software package of the R language was used to analyze the independent prognosis in the TCGA-COAD and GSE40976 data sets. Time-dependent ROC analysis was performed on influencing factors, and the sensitivity and specificity of survival prediction were analyzed by genetic marker risk score. The area under the curve (AUC) can be used as an indicator of prognostic accuracy. If not specifically stated, p value <0.05 was considered statistically different for survival analysis. According to the results of multivariate Cox regression analysis, the R language "rms", "Hmisc", "lattice", "survival", "Formula", "ggplot2", "SparseM" software packages were used to calculate and visualize the nomogram. We carried out the calibration curve and the C-index analysis to verify the predictive ability of the nomogram. DCA was used to evaluate the net rate of return of the nomogram in clinical practice.

Gene set enrichment analysis

GSEA is implemented using the developed Java software, and uses default parameters by comparing high-risk (above the median) and low-risk groups (below the median). We used GSEA_4.0.1 to identify the enriched KEGG pathway of model genes in TCGA-COAD and GSE40976. The representative enrichment pathway was plotted using the "ggplot2" software package of the R language.

Results

Screening of differentially expressed metabolic genes related to survival

We obtained a total of 398 COAD samples and 39 normal colon samples from TCGA for gene expression and prognostic analysis, and downloaded a dataset containing 1048 colon carcinoma samples (GSE40976) from the GEO database. We downloaded the KEGG gene set c2.cp.kegg.v7.0.symbols.

gmt containing metabolic genes from the GSEA website, and extracted a total of 921 metabolism-related genes. Then, 645 cross-expressed genes were screened in TCGA-COAD and GSE40976. P value <0.05 and |log2-fold change|> 0.5 were defined as DEG. A heat map analysis was performed to show cluster analysis of gene characteristics (Figure 1A), and a volcano map was constructed to reveal 225 significantly DEG (Figure 1B). A total of 377 cases were extracted from TCGA-COAD, and 556 cases were extracted from GSE40976 for prognostic analysis. We deleted missing data and

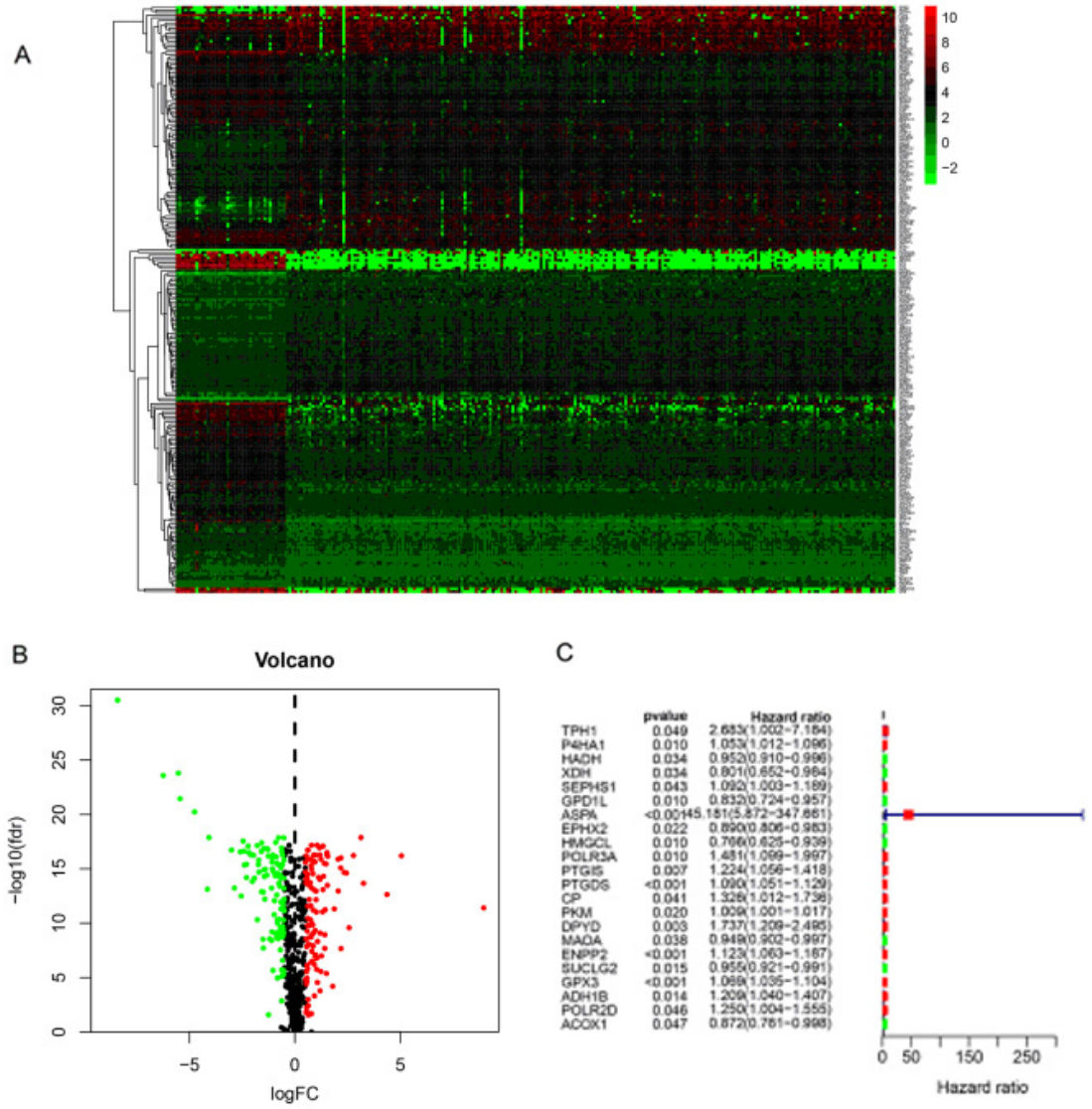

Figure 1. Differentially expressed metabolic genes related to survival. **A:** Heat map showing hierarchical cluster analysis of DEG in TCGA-COAD. **B:** Volcano plot of gene expression data. Green dots are down-regulated RNAs. Red dots are up-regulated RNA; black dots are RNA that is not differentially expressed. P value <0.05 and |log2-fold change|> 0.5 are considered to be statistically significant. **C:** Univariate Cox regression analysis forest plot shows the differential genes related to survival in TCGA-COAD. DEG, differentially expressed genes; TCGA-COAD, colon adenocarcinoma gene expression profile.

Table 1. Genes and coefficients of prognostic models

Gene	*Coef*	*Gene*	*Coef*	*Gene*	*Coef*	*Gene*	*Coef*
P4HA1	0.025	XDH	-0.122	ASPA	2.651	PKM	0.007
POLR3A	0.258	GPD1L	-0.034	CP	0.030	ENPP2	0.095
GPX3	0.041	HMGCL	-0.097	POLR2D	0.094	MAOA	-0.012
PTGDS	0.044	EPHX2	-0.042	SEPHS1	0.039	ADH1B	-0.187

Coef: coefficient

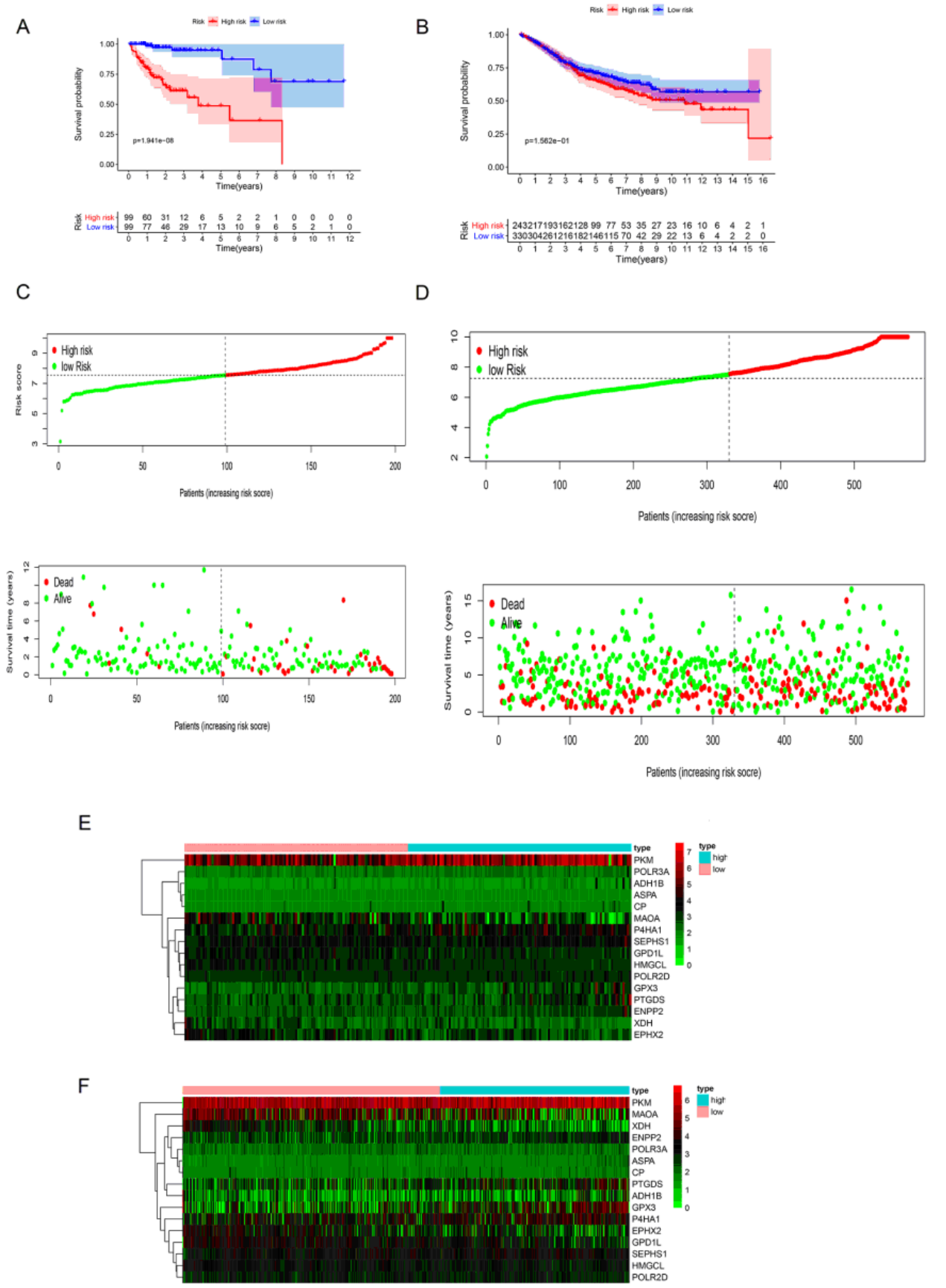

Figure 2. Risk score for predicting survival. **A,B:** Kaplan-Meier analysis of the TCGA-COAD and GSE40976 prognostic models has a longer overall survival (OS) time for the high-risk group than the low-risk group. **C,D:** Distribution of risk scores of 16 genes in TCGA-COAD and GSE40976 (each point represents a sample). **E,F:** Cluster enrichment analysis of 16 genes in TCGA-COAD and GSE40976. GEO, Gene Expression Omnibus.

cases with survival time less than 30 days from the cases. Univariate Cox regression analysis revealed 22 PRMG in TCGA-COAD (Figure 1C, $p<0.05$).

Construction and verification of risk scoring prognostic model

After 1000 resamplings, 22 metabolic genes were subjected to LASSO Cox regression analysis to construct a prognostic model, which contained 16 metabolic genes (Table 1). We calculated the risk score of each patient based on the mRNA expression level and risk coefficient of each gene.We divided the TCGA-COAD and GSE40976 samples into high-risk groups and low-risk groups based on the median risk score. Kaplan–Meier analysis was performed to prove that the overall survival of the high-risk group was poor (Figures 2A,B). The risk score distribution showed that the mortality rate of the high-risk group was higher compared with the low-risk group (Figures 2C,D). A heat map was developed to show the high-risk and low-risk TCGA-COAD and GSE40976 gene expression profiles

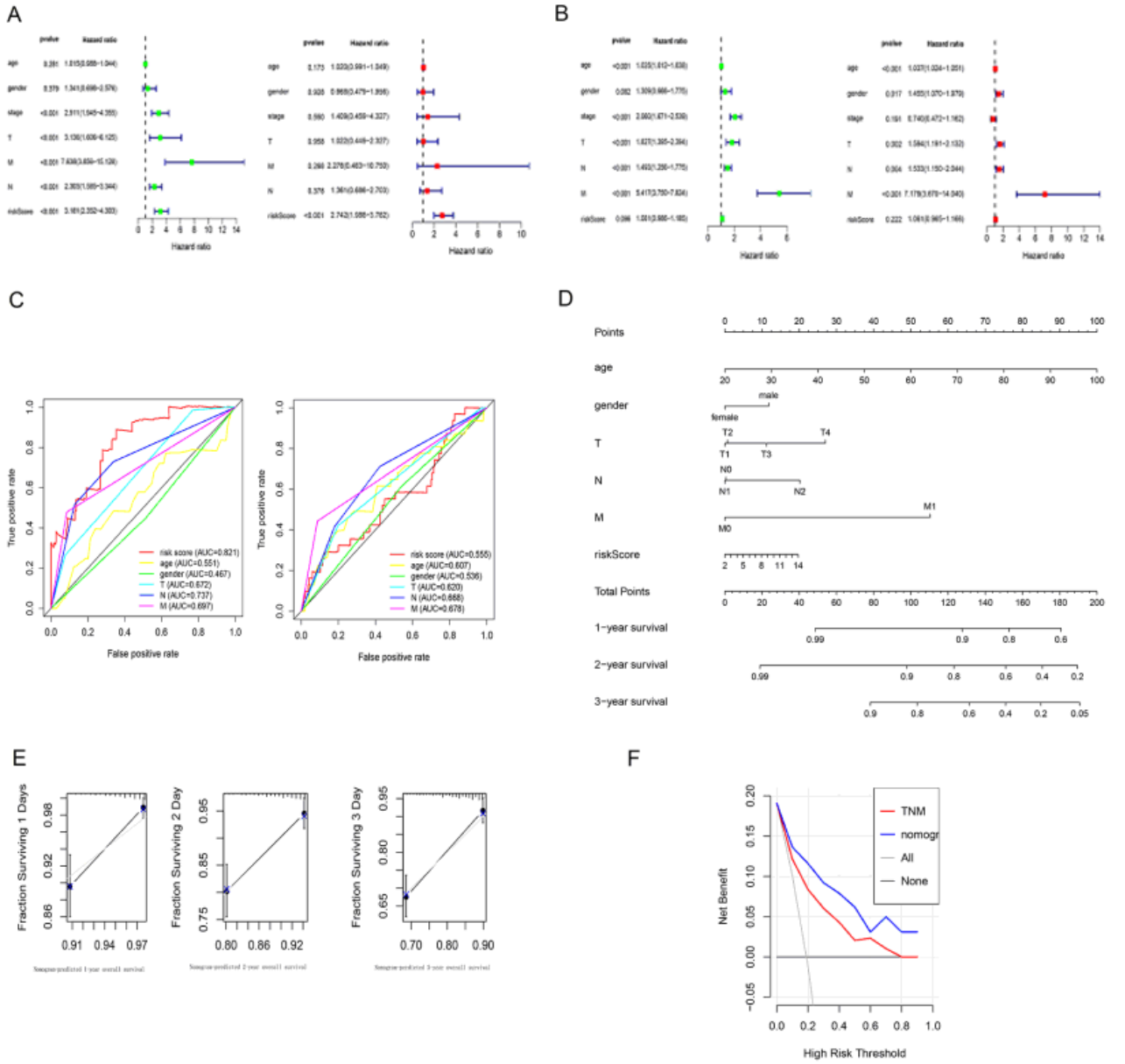

Figure 3. Establishment and verification of the nomogram prediction model. **A,B:** univariate and multivariate analysis forest plots in TCGA-COAD and GSE40976. **C:** Time-dependent ROC analysis in TCGA-COAD and GSE40976 showed the most AUC curve area. **D:** Establish a nomogram on GEO. **E:** Calibration curve for nomogram 1-year, 2-year and 3-year overall survival. **F:** The clinical decision curve shows that the clinical benefit rate of the nomogram is higher than that of the TNM staging system. Black horizontal line: all patients died; black oblique line: no patient died; blue solid line: nomogram prediction model; Red solid line: TNM staging system model. AUC: area under the ROC curve; ROC: receiver operating characteristics.

(Figures 2E,F). The heat map showed the expression of 16 gene markers. SEPHS1, P4HA1, ENPP2, PTGDS, GPX3, CP, ASPA, POLR3A, PKM and POLR2D were positively correlated with high-risk groups, indicating that high expression of these genes is associated with a shorter overall survival time. XDH, EPHX2, ADH1B, HMGCL, GPD1L and MAOA revealed opposite effects, indicating that high expression of these genes is associated with longer overall survival. P value<0.05 was considered statistically different.

Construction and verification of nomogram prediction model

We used univariate and multivariate Cox regression analyses to analyze the significance of age, gender, T-stage, N-stage, M-stage, and risk score in predicting clinical outcomes in the TCGA-COAD and GSE40976 data sets. The results showed that the risk score is a valuable prognostic indicator (Figures 3A, B). The AUC curve of one-year survival includes age, gender, T-stage, N-stage, M-stage, and risk score. The AUC curve was 0.821 and 0.555 in the ATCGA-COD and GSE40976 data sets, respectively. Compared with other parameters such as age and gender, the risk score of metabolism-related genes showed a better forecast value (Figure 3C). Multivariate analysis in the GSE40976 data set showed that age, gender, T-stage, N-stage, and M-stage were independent prognostic factors that affect overall survival. In the TCGA-COAD data set, the risk score was an independent factor affecting overall survival.

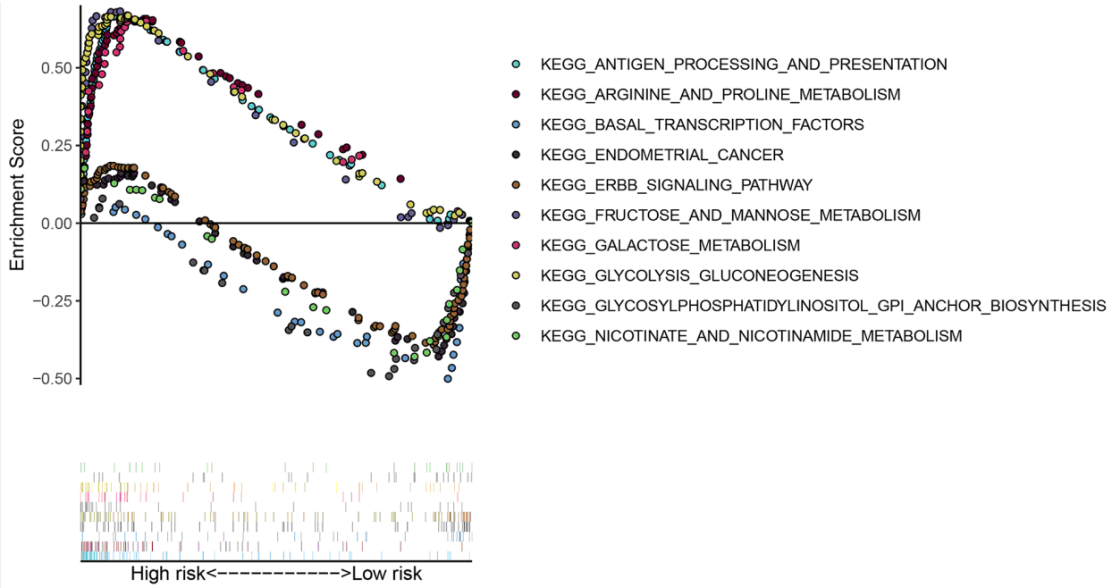

Figure 4. Ten representative KEGG enrichment pathways analyzed by GSEA. Each group contains 5 KEGG pathways. GESA: gene set enrichment analysis; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Table 2. Related parameters of Figure 4 Kyoto Encyclopedia of Genes and Genomes representative pathways

	Size	*ES*	*NES*	*p*	*FDR-q*
High Risk					
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	62	0.67	2.1	0	0.063
KEGG_GALACTOSE_METABOLISM	26	0.67	2.09	0	0.035
KEGG_ARGININE_AND_PROLINE_METABOLISM	54	0.66	2.06	0	0.027
KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	33	0.68	2.06	0	0.022
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	81	0.67	2.01	0.012	0.029
Low Risk					
KEGG_BASAL_TRANSCRIPTION_FACTORS	35	-0.54	-1.69	0.023	0.989
KEGG_ENDOMETRIAL_CANCER	52	-0.45	-1.58	0.03	0.973
KEGG_ERBB_SIGNALING_PATHWAY	87	-0.4	-1.53	0.042	0.636
KEGG_GLYCOSYLPHOSPHATIDYLINOSITOL_GPI_ANCHOR_BIOSYNTHESIS	25	-0.53	-1.55	0.083	0.763
KEGG_NICOTINATE_AND_NICOTINAMIDE_METABOLISM	24	-0.47	-1.52	0.044	0.534

According to the multivariate Cox regression model, by combining age, gender, T-stage, N-stage, M-stage and risk score, we established a nomogram model for predicting prognosis in the GEO, based on the contribution to survival risk, assigned a score to each factor, and used a nomogram to predict the 1-year, 2-year, and 3-year overall survival rates of colon carcinoma patients (Figure 3D) and we used the above clinical information to draw nomograms to facilitate the application of risk scores. The calibration curve for predicting 1-year, 2-year and 3-year overall survival indicated that the nomogram-predicted survival closely corresponding with actual survival outcomes (Figure 3E). The nomogram C-index of the GEO data set were 0.732, with 95%CI 0.692-0.772. DCA showed that the clinical net rate of return represented by the nomogram was higher than the TNM staging system (Figure 3F). The above results indicate the importance and independence of risk score as a prognostic indicator of COAD.

Gene enrichment analysis

In order to find out why the risk score can predict the survival of patients with COAD, the samples were grouped according to the median risk score, namely high-risk group and low-risk group. The implemented GSEA in the high-risk and low-risk groups investigated the ways of change. We identified the KEGG enrichment pathway of model genes in TCGA-COAD. GSEA analysis showed that the altered genes were observed to be enriched in several common pathways. Among the 178 genomes of the high-risk phenotype group, 130 genomes were up-regulated, and 62 genomes were significant at FDR <25%. Most enrichment pathways were concentrated in metabolic pathways, such as arginine and proline metabolism, fructose and mannose metabolism, galactose metabolism, and nicotinate and nicotinamide metabolism. The results also included some well-known cancer-related pathways, such as antigen processing and presentation, basic transcription factors, endometrial cancer, glycolytic gluconeogenesis. Some representative pathways are shown in Figure 4. The relevant parameters of the channel are listed in Table 2. The results further illustrated the role of metabolic mechanisms in COAD.

Discussion

Recent studies have shown that metabolic pathways play an important role in regulating tumor progression [9-11]. The survival and proliferation of cancer cells depends on metabolic reprogramming [12]. Many studies have reported the possibility of metabolic pathways as tumor-targeted treatments. Specific metabolic activities can directly affect the transformation process or proliferation process, which are the biological processes of tumor growth [13]. Abdel-Wahab et al reports pointed out that controlling glucose metabolism may be a new way to inhibit cancer progression [14,15]. Recent studies have shown that microbial metabolites, such as secondary bile acids, can promote cancer. The metabolism of intestinal microbes related to cancer and diets rich in fat and meat, and extracellular metabolism can promote cancer progression [16]. However, the basic mechanism of metabolism in COAD has not been fully elucidated, which hinders the targeted therapy of metabolism. Therefore, the discovery of new molecular markers related to the prognosis of COAD is very important. In this study, based on LASSO Cox regression analysis, we identified 16 PRMG in the TCGA-COAD and GSE40976 data sets to construct a prognostic model for COAD patients and determine the risk score. The prognostic model is accurate and accurate Kaplan-Meier analysis proved that the risk score model can predict the overall survival rate of COAD. Univariate and multivariate regression analyses confirmed that risk score is an independent prognostic factor for COAD. The AUC curve of the gene confirms that the risk score has a good prognostic value in predicting overall survival. The C-index of the nomogram was 0.732. DCA showed that the nomogram prediction model has a higher clinical benefit rate than the TNM staging system. Many enrichment analysis pathways are concentrated in metabolic pathways. In addition to metabolic pathways, the high-risk group showed some cancer-related pathways, such as antigen processing and presentation, basal transcription factors, endometrial cancer, glycolysis gluconeogenesis, erbb signal pathway, and glycosylphosphatidylinositol gpi anchor biosynthesis . These results show that these genes are closely related to metabolic pathways and reveal the potential role of metabolic pathways in COAD.

Target genes are important members of metabolic pathways and can serve as therapeutic targets for cancer. Prognosis prediction is very important for selecting clinical treatment options for cancer patients [17]. Several studies have explored prognostic biomarkers and found that gene expression profiles play a crucial role in the prognosis of cancer [18]. Although our screening of these genes related to cancer prognosis is rarely reported, these genes can reflect the status of cancer driver genes related to their upstream and downstream to a certain extent. The genes we screened were rich in a variety of cancer-related pathways. Based on these results, we concluded that the risk score can accurately pre-

dict the survival of patients with COAD, perhaps because the score can reflect the multi-level status of COAD. We constructed a nomogram to predict individualized clinical outcomes. The nomogram generates a graphical statistical prediction model that assigns scores to each factor, including age, gender, and clinical stage, covering important factors that affect clinical outcomes. In addition to traditional clinicopathological characteristics (such as age, gender, TNM staging), risk scores based on genetic markers can also be incorporated into the predictive nomogram model to predict clinical outcomes. The nomogram is a stable and reliable quantification of personal risk by combining clinical characteristics and risk scores. Our nomogram includes risk scores and clinicopathological characteristics, which can well predict patients with colon carcinoma at 1, 2, and 3 years survival rate. The calibration curve for predicting overall survival indicated that the nomogram-predicted survival closely corresponded with actual survival outcomes. We constructed 16 metabolic gene models based on TCGA and GEO to predict the prognosis of COAD patients. The risk score based on 16 genes may be a promising independent prognostic biomarker. However, these are not yet clear. How genes play their roles in the mechanism, therefore, more research is needed to explore the impact of metabolic enzymes on survival.

The study has limitations. First of all, this is a retrospective study. Therefore, information including recurrence time, treatment records and detailed pathological staging cannot be obtained. Second, although the model has been validated in all cohorts, it still needs more samples for further confirmation before clinical application.

Conflict of interests

The authors declare no conflict of interests.

References

1. Sakin A, Sahin S, Sakin A et al. Mean platelet volume and platelet distribution width correlates with prognosis of early colon cancer. JBUON 2020;25:227-39.
2. Lee N, Kim D. Cancer Metabolism: Fueling More than Just Growth. Mol Cells 2016; 39:847-54.
3. Labuschagne CF, Zani F, Vousden KH. Control of metabolism by p53-Cancer and beyond. Biochim Biophys Acta Rev Cancer 2018; 1870:32-42.
4. Kawada K, Toda K, Sakai Y. Targeting metabolic reprogramming in KRAS-driven cancers. Int J Clin Oncol 2017; 22:651-9.
5. Hlavata I, Vrana D, Smerhovsky Z et al. Association between exposure-relevant polymorphisms in CYP1B1, EPHX1, NQO1, GSTM1, GSTP1 and GSTT1 and risk of colorectal cancer in a Czech population. Oncol Rep 2010; 24:1347-53.
6. Bezzecchi E, Ronzio M, Dolfini D, Mantovani R. NF-YA Overexpression in Lung Cancer: LUSC. Genes (Basel) 2019; 10:937.
7. Ritchie ME, Phipson B, Wu D et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43:e47.
8. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33:1-22.
9. Phan TK, Bindra GK, Williams SA, Poon IKH, Hulett MD. Combating Human Pathogens and Cancer by Targeting Phosphoinositides and Their Metabolism. Trends Pharmacol Sci 2019; 40:866-82.
10. Koliarakis I, Psaroulaki A, Nikolouzakis TK et al. Intestinal microbiota and colorectal cancer: a new aspect of research. J BUON 2018;23:1216-34.
11. Lacroix M, Riscal R, Arena G, Linares LK, Le Cam L. Metabolic functions of the tumor suppressor p53: Implications in normal physiology, metabolic disorders, and cancer. Mol Metab 2020; 33:2-22.
12. Hoxhaj G, Manning BD. The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. Nat Rev Cancer 2020; 20:74-88.
13. Vander Heiden MG, DeBerardinis RJ. Understanding the Intersections between Metabolism and Cancer Biology. Cell 2017; 168:657-69.
14. Agrawal B. New therapeutic targets for cancer: the interplay between immune and metabolic checkpoints and gut microbiota. Clin Transl Med 2019; 8:23.
15. Abdel-Wahab AF, Mahmoud W, Al-Harizy RM. Targeting glucose metabolism to suppress cancer progression: prospective of anti-glycolytic cancer therapy. Pharmacol Res 2019; 150:104511.
16. Wong SH, Yu J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. Nat Rev Gastroenterol Hepatol 2019; 16:690-704.
17. Zenger S, Gurbuz B, Can U, Balik E, Bugra D. Clinicopathologic features and prognosis of histologic subtypes in the right-sided colon cancer. J BUON 2020;25:2154-9.
18. Shen S, Kong J, Qiu Y, Yang X et al. Identification of core genes and outcomes in hepatocellular carcinoma by bioinformatics analysis. J Cell Biochem 2019; 120:10069-81.