Identification Sixteen Metabolic Genes as Potential Biomarkers for Colon Adenocarcinoma

Fu-qiang Zhao[1], Yan-long Liu[1], Xin-yue Gu[1], Bomiao Zhang[1], Chengxin Song[1], and Bin-bin Cui[1]

[1]Tumor Hospital of Harbin Medical University

November 18, 2020

Abstract

Purpose Colon adenocarcinoma is the most common primary malignant tumor of the digestive tract. It is still important to find important markers that affect the prognosis of COAD. This research aims to identify some key prognosis-related metabolic genes (PRMG) and establish a clinical prognosis model for COAD patients. Method 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 Characteristic (ROC) curve analysis of independent prognostic factors was performed and a nomogram for predicting overall survival was constructed. Perform the consistency index (C-index) test and decision curve analysis (DCA) on the nomogram, and use Gene Set Enrichment Analysis (GSEA) to identify the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of model genes. Result We selected PRMG based on the expression of metabolic genes, and used LASSO Cox regression to construct 16 metabolic gene (SEPHS1, P4HA1, ENPP2, PTGDS, GPX3, CP, ASPA, POLR3A, PKM, POLR2D , XDH, EPHX2, ADH1B, HMGCL, GPD1L and MAOA) models. 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. Conclusion We comprehensively identified metabolic genes related to the prognosis of COAD. The risk score based on the expression of 16 metabolic genes can effectively predict the prognosis of patients with COAD.

Identification Sixteen Metabolic Genes as Potential Biomarkers for Colon Adenocarcinoma

Abstract Purpose Colon adenocarcinoma (COAD) is the most common primary malignant tumor of the digestive tract. It is still important to find important markers that affect the prognosis of COAD. This research aims to identify some key prognosis-related metabolic genes (PRMG) and establish a clinical prognosis model for COAD patients. **Method**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 Characteristic (ROC) curve analysis of independent prognostic factors was performed and a nomogram for predicting overall survival was constructed. Perform the consistency index (C-index) test and decision curve analysis (DCA) on the nomogram, and use Gene Set Enrichment Analysis (GSEA) to identify the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of model genes. **Result** We selected PRMG based on the expression of metabolic genes, and used LASSO Cox regression to construct 16 metabolic gene (SEPHS1, P4HA1, ENPP2, PTGDS,

GPX3, CP, ASPA, POLR3A, PKM, POLR2D , XDH, EPHX2, ADH1B, HMGCL, GPD1L and MAOA) models. 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. **Conclusion** We comprehensively identified metabolic genes related to the prognosis of COAD. The risk score based on the expression of 16 metabolic genes can effectively predict the prognosis of patients with COAD.

Keywords colon cancer; metabolism; prognostic; TCGA; GEO

Background

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 (tumor-nodes-matastasis) staging system have been used as three prognostic indicators of the risk of recurrence in CRC patients. But the TNM staging system only considers the anatomical characteristics of the tumor, not the biological characteristics of the tumor. 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[1]. Certain genetic drivers of CRC, such as p53 [2]and KRas[3], are well-known regulators of cancer metabolism. And metabolic gene variants promote colorectal cancer[4]. It is currently known that a 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[5], which are of clinical significance 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.

Method

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 analysis and prognosis analysis. Organize and annotate the RNA sequencing matrix files of different samples to the genome. The mRNA expression is obtained from the RNA sequencing data matrix file. Download from GEO (https://www.ncbi.nlm.nih.gov/geo/) to a data set containing 1048 COAD samples (GSE40976). The data of the TCGA-COAD and GSE40976 samples are 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) 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 [6], 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 Wilcox test of the R language "Limma" software package to screen for differentially expressed genes (DEG). P Value <0.05 and $|\log2\text{-fold change}| > 0.5$ 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[7], a predictive model that affects 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 Figure 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 are drawn.

Construction and verification of nomogram prediction model

Univariate Cox regression 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 "survivalROC" 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 is performed on influencing factors, and the sensitivity and specificity of survival prediction are analyzed by genetic marker risk score. Area Under 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 are used to calculate and visualize the nomogram. Carry out the calibration curve and the C-index analysis to verify the predictive ability of the nomogram. DCA is used to evaluate the net rate of return of the nomogram in clinical practice.

Gene set enrichment analysis

GSEA is implemented using 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 is plotted using the "ggplot2" software package of the R language.

Result

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 analysis and prognostic analysis, and downloaded a dataset containing 1048 COAD samples (GSE40976) from the GEO database. 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. Delete missing data and cases with survival time less than 30 days from the cases. Univariate Cox regression analysis revealed 22 PRMG in TCGA-COAD (Fig. 1C, p value <0.05).

Construction and verification of risk scoring prognostic model

After 1000 resampling, 22 metabolic genes were subjected to LASSO Cox regression analysis to construct a prognostic model, which containing 16 metabolic genes. It shows the 16 genes and their coefficients in the risk scoring model (Table 1). Calculate 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 (Figure 2A, B). The risk score distribution showed that the mortality rate of the high-risk group was higher than that of the low-risk group (Figure 2C, D). A heat map was developed to show the high-risk and low-risk TCGA-COAD and GSE40976 gene expression profiles (Fig. 2E, F). The heat map shows the expression of 16 gene markers. SEPHS1, P4HA1, ENPP2, PTGDS, GPX3, CP, ASPA, POLR3A, PKM and POLR2D are 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 time. P Value$<$0.05 is considered statistically different.

Construction and verification of nomogram prediction model

We used univariate Cox regression analysis and multivariate Cox regression 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 show that the risk score is a valuable prognostic indicator (Fig. 3A, B). The AUC curve of one-year survival includes age, gender, T-stage, N-stage, M-stage, and risk score. The AUC curve are 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 shows a better forecast value (Fig. 3C). Multivariate analysis in the GSE40976 data set showed that age, gender, T-stage, N-stage, and M-stage are independent prognostic factors that affect overall survival. In the TCGA-COAD data set, the risk score is an independent factor affecting overall survival.

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 , Assign a score to each factor, and use a nomogram to predict the 1-year, 2-year, and 3-year overall survival rates of COAD patients (Fig. 3D). Use 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 OS indicated that the nomogram-predicted survival closely corresponded with actual survival outcomes (Fig. 3E). The nomogram C-index of the GEO data set is 0.732, 95% CI (0.692-0.772). DCA shows that the clinical net rate of return represented by the nomogram is higher than the TNM staging system (Fig. 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. Implement GSEA in the high-risk and low-risk groups to investigate 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 are concentrated in metabolic pathways, such as arginine and proline metabolism, fructose and mannose metabolism, galactose metabolism, and nicotinate and nicotinamide metabolism. The results also include some well-known cancer-related pathways, such as antigen processing and presentation, basic transcription factors, endometrial cancer, glycolytic gluconeogenesis, erbb signaling pathway, and glycosylphosphatidylinositol gpi anchor biosynthesis. It shows some representative pathways in figure 4. The relevant parameters of the channel are listed in Table 2. The results further illustrate the role of metabolic mechanisms in COAD.

Discussion

Recent studies have shown that metabolic pathways play an important role in regulating tumor progression [8,9]. The survival and proliferation of cancer cells depends on metabolic reprogramming [10]. 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 is the biological process of tumor growth [11]. Abdel-Wahab and other reports pointed out that controlling glucose metabolism may be a new way to inhibit cancer progression [12,13]. 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 [14]. 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 analysis 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 shows 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 shows 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. Several studies have explored prognostic biomarkers and found that gene expression profiles play a crucial role in the prognosis of cancer [15]. 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 are rich in a variety of cancer-related pathways. Based on these results, we concluded that the risk score can accurately predict 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 adenocarcinoma at 1, 2, and 3 years survival rate. The calibration curve for predicting OS 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.

Reference

[1] Lee N, Kim D. Cancer Metabolism: Fueling More than Just Growth. Mol Cells. 2016 Dec;39(12):847-854.

[2] Labuschagne CF, Zani F, Vousden KH. Control of metabolism by p53 - Cancer and beyond. Biochim Biophys Acta Rev Cancer. 2018 Aug;1870(1):32-42.

[3] Kawada K, Toda K, Sakai Y. Targeting metabolic reprogramming in KRAS-driven cancers. Int J Clin Oncol. 2017 Aug;22(4):651-659.

[4] Hlavata I, Vrana D, Smerhovsky Z, Pardini B, Naccarati A, Vodicka P, Novotny J, Mohelnikova-Duchonova B, Soucek P. 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 Nov;24(5):1347-53.

[5] Bezzecchi E, Ronzio M, Dolfini D, Mantovani R. NF-YA Overexpression in Lung Cancer: LUSC. Genes

(Basel). 2019 Nov 17;10(11):937.

[6] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

[7] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1-22.

[8] 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 Nov;40(11):866-882.

[9] 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 Mar;33:2-22.

[10] Hoxhaj G, Manning BD. The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. Nat Rev Cancer. 2020 Feb;20(2):74-88.

[11] Vander Heiden MG, DeBerardinis RJ. Understanding the Intersections between Metabolism and Cancer Biology. Cell. 2017 Feb 9;168(4):657-669.

[12] Agrawal B. New therapeutic targets for cancer: the interplay between immune and metabolic checkpoints and gut microbiota. Clin Transl Med. 2019 Aug 27;8(1):23.

[13] Abdel-Wahab AF, Mahmoud W, Al-Harizy RM. Targeting glucose metabolism to suppress cancer progression: prospective of anti-glycolytic cancer therapy. Pharmacol Res. 2019 Dec;150:104511.

[14] Wong SH, Yu J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. Nat Rev Gastroenterol Hepatol. 2019 Nov;16(11):690-704.

[15] Shen S, Kong J, Qiu Y, Yang X, Wang W, Yan L. Identification of core genes and outcomes in hepatocellular carcinoma by bioinformatics analysis. J Cell Biochem. 2019 Jun;120(6):10069-10081.

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

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

	SIZE	ES	NES	P	FDR-c
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

Hosted file

figure.pdf available at https://authorea.com/users/336139/articles/493642-identification-sixteen-metabolic-genes-as-potential-biomarkers-for-colon-adenocarcinoma