Investigation of potential prognostic biomarkers for colorectal cancer

Keywords
Treatment, Colorectal cancer, Prognostic biomarker

Abstract
Introduction
Colorectal cancer (CRC) is the third leading cause of cancer-related death. Since CRC is largely asymptomatic until the alert features develop to an advanced stage, implementation of a screening program is important to reduce cancer morbidity and mortality. Current screening methods have significant limitations.

Material and methods
CRC-related microarray datasets were collected from the GEO database and differentially expressed genes (DEGs) were identified. Next, Venn analysis, functional enrichment analysis, protein interaction network (PPI) analysis, and survival analysis were performed.

Results
A total of 5267 and 4233 DEGs were identified in two datasets (GSE20916, GSE33133). The intersection of up-regulated genes in the two datasets was obtained by Venn Analysis as 1058 DEGs. Among the 1058 genes, 992 genes with survival and clinical information in TCGA were screened. Eleven DEGs were identified as potential prognostic markers. Model results show that the time period with the most obvious prognostic effect is 5 years, and the AUC value is the highest. ROC curve results are consistent with the model results of the survival analysis. The survival curve showed that LRRC8A, PCAT6, PLA2G15, SRD5A1, T1GD1 may be oncogenes, and DSN1, ERI1, EIT1, GLMN, MAPKAPK, NOP14 may be tumor suppressor genes.

Conclusions
This study discovers novel prognostic markers through Cox regression and survival analysis, and provides a theoretical basis for the treatment of CRC.
Investigation of potential prognostic biomarkers for colorectal cancer

ABSTRACT

Colorectal cancer (CRC) is the third leading cause of cancer-related death. More than 1 million new cases of CRC are reported worldwide and over 500,000 people die yearly. Besides, most new cases are in developed countries. Since CRC is largely asymptomatic until the alert features develop to an advanced stage, implementation of a screening program is important to reduce cancer morbidity and mortality. Current screening methods have significant limitations. Consequently, considerable scientific effort has been invested in the development of new diagnostic and prognostic biomarkers for CRC. This study aimed to discover novel prognostic markers through Cox regression and survival analysis, and provide a theoretical basis for the treatment of CRC.

Keyword: Colorectal cancer, Treatment, Prognostic biomarker

1 Introduction

Colorectal cancer (CRC) is currently the third and second most common cancer in men and women, respectively[1][2], and one of the leading causes of cancer-related deaths[3][4]. Colorectal cancer (CRC) is the third most common cancer (9%) and the second most common cancer-related mortality worldwide, according to the 2020 Global Cancer Statistics[5]. According to statistics, in 2008, there were nearly 150,000 newly diagnosed cases and 50,000 deaths[6]. By 2012, the Global Cancer Observatory estimated that 1.3 million new cases were diagnosed with CRC and nearly 700,000 patients died from it[7]. CRC is more common in Western countries, accounting for more than 65% of cases[8]. In Europe, approximately 400,000 new patients are diagnosed with CRC each year, and more than 200,000 people die yearly (Hibner, Kimsa-Furdzik and Francuz, 2018). Within four years, both morbidity and mortality had risen at an alarming rate. According to research, CRC usually occurs in the lining of the colon (large intestine) or rectum[9], and it usually grows on the lining of the colon and rectum as polyps, which are lumps that protrude in the lumen. Not all polyps are neoplastic and develop in cancer, but it is well known that most CRC develop from adenomatous polyps, the so-called adenoma-to-cancer sequence[10]. Most CRC occur in patients over the age of 50, with more than 75% of those over the age of 60. Risk factors associated with CRC incidence include older age, male gender, lifestyle, inflammatory bowel disease, and previous history of CRC[11]. Besides, the risk of CRC increases with age, is 1.5-2 times higher in men than in women, and the mortality rate in the male population has been steadily increasing over the past three decades[12]. Despite the high mortality rate and high prevalence of CRC in cancer, it is an indolent disease in its early stages and usually develops...
Symptoms when it progresses to a more advanced stage\textsuperscript{[13]}. Therefore, mortality can be reduced through early prevention and detection.

Therefore, the ultimate goal should be to implement and improve screening strategies\textsuperscript{[14]}\textsuperscript{[15]}. Numerous efforts have been made to establish appropriate screening methods. Screening techniques can be divided into non-invasive and invasive with varying sensitivity and specificity, including fecal occult blood test (FOBT), fecal DNA test, Double contrast barium enema (DCBE) and colonoscopy\textsuperscript{[16]}. Among them, colonoscopy is an invasive procedure with a risk of bowel perforation. These methods are either invasive, resulting in low participation rates in healthy populations\textsuperscript{[14]}, or have large limitations, limited by insufficient sensitivity or a high percentage of false positive results\textsuperscript{[17]}. Currently, new biologics for CRC screening tests The development of markers is the subject of intensive research. In recent years, many biomarkers that aid in the treatment and prognosis of colorectal cancer have received more attention. In order to discover these potential biomarkers that can be detected in blood and stool by non-invasive methods, it is important to study the genetics and pathogenesis of CRC\textsuperscript{[18]}\textsuperscript{[19]}.

A biomarker, or biomarker, is “a defined characteristic measured as an indicator of normal biological processes, pathogenic processes, or responses to exposures or interventions”\textsuperscript{[20]}. There are several subtypes of biomarkers that can be identified and quantified to aid in the diagnosis, monitoring, assessment and prediction of disease and patient response to therapy, including diagnostic biomarkers, pharmacodynamic biomarkers, predictive biomarkers, prognostic biomarkers Markers, susceptibility biomarkers\textsuperscript{[21]}. Among them, diagnostic biomarkers refer to detecting the presence of a certain disease or condition. For example, the hemoglobin A1c (HbA1c) level for the diagnosis of type 2 diabetes. Prognostic biomarkers are used to identify the likelihood of disease recurrence or progression in people who already have the disease, eg, prostate-specific antigen (PSA) levels, to assess the likelihood of cancer progression. This study aimed to identify potential novel prognostic tests and biomarkers for CRC. Two microarray datasets obtained from the Gene Expression Omnibus (GEO) database were analyzed to further identify DEGs in CRC by comparing expression levels between CRC and normal samples, and subsequently, differences shared across the datasets were identified using Venny gene DEG. Next, enrichment analysis was performed to further analyze the biological processes of CRC-related DEGs and construct a protein-protein interaction (PPI) network.

Survival analysis refers to the method of analyzing and inferring the survival time of organisms or people based on the data obtained from experiments or surveys, and studying
the relationship between survival time and outcomes and many influencing factors and their degree, also known as survival rate analysis or survival rate analysis, is often used in the evaluation of common clinical efficacy in tumors, such as exploring how long the remission period lasts for leukemia patients or the observation time of death\cite{22}\cite{23}\cite{24}. In survival analysis, what we get is not a single specific value, but a curve, called a survival curve\cite{25}, and the function corresponding to the curve is called a survival function. To validate the prognostic role of DEG in CRC, CRC samples from The Cancer Genome Atlas (TCGA) and GEO databases were collected. Survival analysis was performed based on the sample segmentation method and Cox regression model to obtain prognostic genes and their survival curves. Next, using the dataset from TCGA as the training set and the GEO dataset as the validation set, the identified key genes and risk score values were used to predict the survival rate of CRC patients. Finally, the diagnostic value of genetic biomarkers for CRC was assessed using ROC curve analysis\cite{26}\cite{27}. Complete identification of prognostic biomarkers for CRC.

2 Materials and methods

Data selection

The Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) is a publicly available repository of gene expression datasets. Its data sources include microarray and next-generation sequencing\cite{28}. The CRC dataset for this study was derived from GEO's microarray dataset. In the GEO database, according to the keyword "colorectal cancer (CRC)”, the sample size is sorted from large to small, and the sample size is uniformly distributed, and two GSE series, GSE20916 and GSE33133, are selected from the GPL570 microarray. Array data platform as a research sample for CRC. The microarray data were also obtained from The Cancer Genome Atlas TCGA (http://cancergenome.nih.gov/)\cite{29} and the GSE72968 dataset in GEO for survival analysis.

Identification of differentially expressed genes(DEGs)

The number of GSE20916 samples was 145, including 111 cancer tissues and 34 normal controls. The number of GSE33133 samples was 96, including 90 cancer tissues and 6 normal controls. Differential genes were defined with adj.P.Val<0.05 for differential analysis. In the TCGA colorectal cancer data, the number of samples was 624, and 35 samples with expression but no clinical survival information or with clinical survival information but no expression were excluded, and 589 samples were finally included.

Venn analysis
In order to obtain the intersection of up-regulated genes of the two datasets, the Venn diagram network tool (http://bioinformatics.psb.ugent.be/webtools/Venn)\(^1\) was used to analyze the gene expression profiles in the two groups, and the identification of genes with CRC Prognosis related genes.

**Functional and pathway enrichment analysis**

GO is a database established by the Gene Ontology Consortium, which aims to keep the functional description of gene products in each database consistent, so that the properties of gene products can be queried at different levels\(^2\), and GO analysis has become a popular tool for large-scale transcriptome data. A common method for functional studies and genomic data analysis\(^3\). The Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/pathway) is a reference database for systematic analysis of gene function, linking genomic information with metabolism, membrane transport, signal transduction linked to higher-order functional information such as the cell cycle\(^4\). In this study, GO functional and KEGG pathway enrichment analysis was performed on the identified DEGs using R 4.0.3. P value <0.005.

**Protein interaction network (PPI) analysis**

The protein interaction network (PPI network) is composed of individual protein interactions and participates in all aspects of life processes such as biological signal transmission, gene expression regulation, energy and material metabolism, and cell cycle regulation\(^5\). The Search Tool for Retrieving Interacting Genes (STRING; http://string-db.org) database is an online software, and this study used the STRING database to build a PPI network of differential genes related to survival to present interactions\(^6\).

**Survival analysis of hub genes**

CRC datasets from TCGA and GEO were used for survival analysis based on sample segmentation method and Cox regression model. The identified key genes can be used to characterize CRC survival and serve as potential therapeutic targets and prognostic biomarkers. The above studies have screened 1058 up-regulated genes, and due to sequencing factors, 922 genes were sequenced in the clinical data of 589 CRC samples in the TCGA database. In order to study the prognostic impact of the identified hub genes on CRC patients, univariate Cox analysis is usually performed to screen out the associated variables, and then a multivariate model is constructed to further confirm whether the association between variables and survival is independent\(^7\). However, this approach does not take into account the effect of multicollinearity among variables, leading to the possibility of contradictory results of hazard ratios obtained by univariate and multivariate Cox regression, traditional
variable selection methods such as stepwise regression, forward regression, and backward regression in Cox regression are no longer applicable\cite{37}.

Compared with Cox analysis method, Lasso (Least absolute shrinkage and selection operator) realizes variable selection while estimating model parameters, and performs better in solving the multicollinearity problem in regression analysis, and can explain the results well\cite{38}. Therefore, when there is multicollinearity between variables or the number of variables is larger than the sample size, the method that Cox regression combined with Lasso is recommended to initially screen variables, and then further establish a Lasso regression model to analyze its impact on prognosis. In a word, the Lasso-Cox survival analysis mode was used, and the differential genes related to prognosis were analyzed by using R4.03, glmnet package and survival package in this paper.

3 results

Screening for DEGs

In this study, CRC-related microarray datasets were collected from the GEO database, and a total of 5267 and 4233 DEGs were identified from the GSE20916 and GSE33133 datasets, respectively. P value<0.05. Among them, GSE20916 has 2847 up-regulated genes and 2420 down-regulated genes; GSE33133 has 2104 up-regulated genes and 2129 down-regulated genes. Create volcano plots of DEG distributions for these two datasets (Figure 1A). Volcano plots (Figure 1B) show significant and non-significant DEGs for the dataset, where blue dots represent down-regulated genes, red dots represent up-regulated genes, and grey dots represent genes that were not significantly different.

Venn analysis

Venn analysis of gene expression in tumor and normal tissues, taking the intersection of all up-regulated genes in GSE20916 and GSE33133 expression data, identified 1058 up-regulated genes (Figure 2).

Functional enrichment analysis of DEGs

To fully understand the biological roles of these DEGs in CRC, we continue to analyze the GO functions and pathways in which they are involved. The results of GO enrichment analysis included biological process (BP), cellular component (CC), and molecular function (MF).

GO analysis
Figure 3 (sorted by p-value from small to large) shows the results of the GO enrichment analysis. Up-regulation of DEGs mainly involves chromosomal region, spindle, condensed chromosome, chromosome, centromeric region, spindle pole, kinetochore, preribosome, condensed chromosome, centromeric region, replication fork spindle microtubule cell components, in ribonucleoprotein complex biogenesis, ncRNA metabolic process, ncRNA processing, ribosome biogenesis, chromosome segregation, mitotic nuclear division, rRNA processing, rRNA metabolic process, sister chromatid segregation, mitotic sister chromatid segregation biological process significantly enriched. Molecular functions such as catalytic activity acting on RNA, catalytic activity, acting on DNA, helicase activity, single-stranded DNA binding, ribonucleoprotein complex binding, damaged DNA binding, RNA methyltransferase activity, snoRNA binding, DNA replication origin binding, and DNA polymerase binding are active.

**KEGG pathway analysis**

The 10 key pathways involved in DEG are sorted by the value of p.adjust from small to large, as shown in Figure 4. As the results show, the most significantly enriched signaling pathways for up-regulated DEGs are: Cell cycle, Ribosome biogenesis in eukaryotes, DNA replication, Nucleocytoplasmic transport, Proteasome, Nucleotide excision repair, Base excision repair, Spliceosome, Mismatch repair and IL-17 signaling pathway.

**Survival analysis of DEGs**

Variables were first screened by univariate Cox, followed by Lasso regression analysis to determine final variables and build models to analyze prognostic impact. Univariate Cox regression analysis was performed on 922 DEGs obtained in this study, and 88 prognostic variables were identified. Differential genes, followed by lasso regression analysis, yielded 11 DEGs (Figure 5), which can be used as potential prognostic indicators for CRC.

**Survival Analysis and modeling**

Subsequently, survival analysis was performed on the 11 DEGs, and the Kaplan-Meier curves were obtained as shown in Figure 6. Kaplan-Meier OS curves of these genes showed that high LRRC8A, PCAT6, PLA2G15, SRD5A1, T1GD1 expression was significantly associated with low OS in colorectal cancer patients, and high PLA2G15 expression was associated with the lowest survival after five years; high DSN1, ERI1, EIT1, GLMN, MAPKAPK, NOP14 expression was significantly associated with high OS in patients with colorectal cancer, and high expression of NOP14 was associated with the highest survival rate after five years. The former may be oncogenes, while the latter may be potential tumor suppressor genes.
**PPI analysis results of DEGs**

The PPI network of 88 survival-related DEGs was constructed using the String online tool (Figure 7), and Figure 8 was obtained after removing unrelated genes.

**Survival Analysis and modeling**

Then, based on the 11 DEGs and coefficients that can be used as prognostic indicators obtained by LASSO-Cox analysis (see Table 1), these potential prognostic indicators were further studied. A prognostic risk model was constructed: Risk score = DSN1*(-0.0006)+ERI1*(-0.0035)+ETF1*(-0.0005)+GLMN*(-0.0124)+LRRC8A*0.0015+MAPKAPK3*(-0.0029)+NOP14*(-0.0003)+PCAT6*0.0020+PLA2G15*0.0049+SRD5A1*0.0015+TGFD1*0.0231, multiply the gene expression by the coefficient to get the risk score, and then predict the patient's survival according to the score, and get the AUC value of the model result; that is, the value of each sample. Risk Score. A higher AUC value indicates a better classification performance of the CRC prognostic model for the gene. The model prognostic ability was assessed using the TCGA dataset, and the model was validated using GSE72968.

The performance of the model on the training set (TCGA): The AUCs for predicting 1-year, 3-year and 5-year survival are: 0.577, 0.569 and 0.633, respectively. Model performance on the validation set (GSE72968): The AUCs for predicting 1-year, 3-year and 5-year survival were: 0.509, 0.603 and 0.631, respectively. The results show that the most obvious prognostic effect is in 5 years. Comparing the AUC values of survival in each stage of the validation set and the training set, the difference between the two AUC values in 5 years is the smallest, and the result is more accurate. Finally, the model is validated. Figure 9 shows the ROC curves of the training set TCGA and the validation set GSE72968 sample data. The ROC plot is a graph reflecting the relationship between sensitivity and specificity (Figure 9).

**Table 1 Index and 11 DEGs as prognostic indicators.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSN1</td>
<td>-0.000625538</td>
</tr>
<tr>
<td>ERI1</td>
<td>-0.003511305</td>
</tr>
<tr>
<td>ETF1</td>
<td>-0.000545896</td>
</tr>
<tr>
<td>GLMN</td>
<td>-0.012356453</td>
</tr>
<tr>
<td>LRRC8A</td>
<td>0.001503675</td>
</tr>
<tr>
<td>MAPKAPK3</td>
<td>-0.002917185</td>
</tr>
<tr>
<td>NOP14</td>
<td>-0.000267735</td>
</tr>
</tbody>
</table>
4 Discussion and Conclusion

The morbidity and mortality of CRC remain high despite advances in surgical and medical treatment of CRC\(^{[39]}\). Successful screening techniques and reducing the risk of CRC are critical to helping reduce CRC mortality. Understanding the etiology and mechanisms of colorectal cancer progression is extremely important for improving the survival rate of colorectal cancer patients. Recently, the rapidly developing microarray technology has been widely used to compare the expression levels of genes and has been used to predict disease progression to make an accurate assessment of prognosis\(^{[40]}\).

In this study, two microarray datasets (GSE20916, GSE33133) were analyzed to identify DEGs in CRC tissue samples. A total of 5267 and 4233 DEGs were identified in the two datasets. GSE20916 included 2847 up-regulated genes and 2420 down-regulated genes. GSE33133 identified a total of 4233 differential genes, including 2104 up-regulated genes. The intersection of up-regulated genes in the two datasets was obtained by Venn Analysis as 1058 DEGs. Subsequently, GO enrichment analysis and pathway analysis were performed. Among the 1058 genes, 992 genes with survival and clinical information in TCGA were screened. Then, survival analysis was performed based on Cox regression and lasso regression analysis models from 992 genes, and a prognostic risk model was obtained. Eleven DEGs were identified as potential prognostic markers. The results of the model show that the time period with the most obvious prognostic effect is 5 years, and the AUC value is the highest. The verification results of the ROC curve are consistent with the model results of the survival analysis. The survival curve showed that LRRC8A, PCAT6, PLA2G15, SRD5A1, TIGD1 may be oncogenes, and DSN1, ERI1, EIT1, GLMN, MAPKAPK, NOP14 may be tumor suppressor genes. Among them, PLA2G15 and NOP14 have the most significant effect. It can be speculated that DSN1, ERI1, EIT1, GLMN, MAPKAPK, NOP14 are effective in inhibiting the further development and lethality of CRC, and the most obvious period is 5 years.
NOP14 encodes Nucleolar Protein, it is an 875 amino group protein that is highly conserved in eukaryotes and is involved in many biological processes such as DNA repair and replication and cell cycle control. Recent studies support the involvement of NOP14 in cancer initiation and progression. NOP14 expression has been reported in various cancers such as Melanoma and breast cancer, bladder cancer and pancreatic cancer\(^{[41][42][43]}\), for example, NOP14 inhibits the Wnt/\(\beta\)-catenin signaling pathway, thereby reducing the proliferation and metastasis of melanoma and breast cancer, thereby acting as a tumor suppressor role of genes. Likewise, in bladder cancer, NOP14 attenuated miR-502-5p-mediated inhibition of tumor cell migration and proliferation. Nonetheless, the role of NOP14 depends on the specific type of cancer. In pancreatic cancer, NOP14 is highly expressed and has been shown to promote tumor cell growth and invasion by stabilizing mutant P53, which inhibits P21 expression by inducing miR-17-5p. Apart From that, as a recent analysis reports, increased expression of NOP14 is associated with improved prognosis in CRC\(^{[44]}\). This suggests that it inhibits tumor development in CRC, consistent with the findings of this study. Another study found that the mRNA and protein levels of NOP14 were significantly upregulated in CRC tissues, which suggested that NOP14 could serve as a potential biomarker for CRC. Moreover, its high expression improves the prognosis of CRC patients by inhibiting signaling pathways involved in tumor growth and promoting immune response\(^{[45]}\).

DSN1 is a protein-coding gene that is part of the MIS12 kinetochore complex and is required for proper kinetochore assembly and cell cycle progression\(^{[46]}\). It is critical for malignant transformation associated with chromosome 20q amplification in CRA (colorectal adenomas). Elevated DSN1 expression was positively correlated with chr20 amplification, high CIN index, and SKA3 overexpression while DSN1 knockdown resulted in slightly increased G2/M phase arrest but did not affect cell growth rate or apoptosis in both CRC cell lines\(^{[47]}\). These results suggest that overexpression of DSN1 contributes to the development of colorectal cancer. In addition, a recent study reports that overexpression of DSN1 is associated with poor survival in patients with colorectal cancer, gastric cancer, liver cancer, and breast cancer, and can be used as a prognostic biomarker for various tumor diseases\(^{[48][49][50]}\).

Lysosomal phospholipase A2 (PLA2G15) is a transacylase that recognizes short-chain ceramides as fatty acyl receptors, it is 50% identical to lecithin cholesterol acyltransferase; these enzymes share close structural homology\(^{[51]}\). Besides, PLA2G15 has broad substrate recognition capacity for a variety of glycerophospholipids, including oxidized phospholipids, it catabolizes pulmonary surfactant and processes lipid antigens for CD1 protein
Moreover, PLA2G15 is a ubiquitous enzyme that is uniquely characterized by its subcellular localization to lysosomes and late endosomes and it is the first and only lysosomal phospholipase A identified to date\(^2\). Although initially characterized as a member of the PLA2 family, recent work has been supported by its unique properties as transacylases, clearly demonstrating the activities of PLA1 and PLA2. Additionally, according to a recent study, knockdown of NFATC3-PLA2G15 reduces the expression of markers of epithelial-mesenchymal transition (EMT) such as vimentin, twist, and fibronectin in colo-320 cells when studying CRC samples and adjacent normal tissue, mRNA expression, and increased mesenchymal-epithelial transition markers of E-cadherin, claudin-1, and FOXC2. Similarly, the NFATC3-PLA2G15 FT-containing line, NFATC3-PLA2G15 knockdown also inhibited invasion, colony-forming ability, and cell proliferation\(^3\), and NFATC3-PLA2G15 fusion transcript levels were elevated in 19 pairs of CRC tumor and adjacent normal tissue samples\(^5\). These evidences suggests that NFATC3-PLA2G15 may promote CRC development by enhancing EMT invasion and proliferation, consistent with the results of this study that PLA2G15 might be the considerable prognostic biomarkers of CRC.

The results of this study may have clinical implications for the prognosis of CRC. However, the current study has certain limitations. First of all, the AUC value obtained by the survival analysis is not particularly large, between 0.5 and 0.7, and the effect is not particularly significant. Second, among the 11 DEGs screened, the potential roles of some genes are still unknown, and further experimental studies are needed.

In conclusion, 11 DGEs were identified as significantly associated with OS in CRC patients by performing the above-mentioned analysis of GEO and TCGA. These 11 genes can be used as new independent prognostic biomarkers, among which, the most significant ones are NOP14 and PLA2G15, which can be used to predict the clinical outcome of CRC patients. However, further analysis of the role and underlying mechanisms of gene expression in CRC is required to re-establish its specific role in CRC prognosis.
Acknowledgments
Not applicable.

Funding
Not applicable.

Competing interests
The authors have no conflicts of interest to declare.

Ethics statement
Our study did not require an ethical board approval because it did not contain human or animal trials.

References:


Figure Legends

Figure 1 The distribution of differences in gene expression levels
(A) Volcano plot showing expression data of CRC and normal tissue in GSE20916 microarray profiles. (B) Volcano plot showing expression data in CRC and normal tissue in GSE33133. The x-axis is the average difference between CRC and normal samples (log2 fold change differential expression fold, according to the positive and negative log2 fold change value to determine whether the expression of these genes increases or decreases), the larger the gene difference is distributed on the x-axis at both ends top; the y-axis represents log-transformed P-values converted to -log10.

Figure 2 Intersection of up-regulated DEGs
Intersection of all up-regulated DEGs in the expression data of GSE20916 and GSE33133.

Figure 3 GO cluster analysis of DEGs
Biological processes, Cellular component and Molecular functions of up DEGs

Figure 4 KEGG pathway analysis

Figure 5 Lasso Cox regression variable filtration
(A) Cross-validation. Screening genes with prognostic efficacy from 88 genes, sample size N = 88, these genes are highly co-expressed in the sample, that is, there is collinearity between variables, and 11 genes are obtained, see Table 1. (B) Multivariate Cox regression analysis to construct a prognostic model and risk scoring. The abscissa is the log λ value, and the ordinate is the evaluation index corresponding to each λ value. The error bar is used to show the evaluation index of multiple models. Mean + standard error, you can see that there are two vertical dotted lines in the figure, the dotted line on the left corresponds to the best λ value of the evaluation index, that is, lambda.min, the larger the c-index value, the better, and the smaller the deviance value, the better; The dotted line on the right represents the λ value of the model whose evaluation index is within 1 standard error of the optimal value, namely lambda.1se.
Figure 6 Kaplan-Meier OS curves


Figure 7 PPI plot of DEG

Differential protein interaction network diagram of up-regulated genes.

Figure 8 PPI plot of DEG after selection

The differential protein network diagram of up-regulated genes that can form an interaction relationship, and the single differential protein that cannot form an interaction relationship no longer appears.

Figure 9 Receiver operating characteristic curve (ROC curve)

(A) ROC curve of ROC sample data with training set TCGA. (B) ROC curve of the validation set GSE72968 sample data. The X-axis of the abscissa is the specificity, also known as the false positive rate (false positive rate: 1 - Specificity), the closer the X-axis is to zero, the higher the accuracy; the Y-axis of the ordinate is called the sensitivity, also known as the true positive rate (Sensitivity: Sensitivity), the larger the Y-axis, the better the accuracy. According to the position of the curve, the whole graph is divided into two parts. The area under the curve is called AUC (Area Under Curve), which is used to indicate the prediction accuracy. The higher the AUC value, the larger the area under the curve, indicating the prediction The higher the accuracy. The closer the curve is to the upper left corner (the smaller the X, the larger the Y), the higher the prediction accuracy. The area under the ROC curve, AUC, is normally between 0.5 and 1. In the case of AUC > 0.5, the closer the AUC is to 1, the better the diagnostic effect.

Note: From Figure 9, it can be concluded that both the training set and the validation set are 60 months, that is, the area under the ROC curve in the five-year period is the largest, and the AUC value is the highest, which is consistent with the results of the survival analysis and conforms to the above conjecture.
GEO database from the GSE20916 and GSE33133 datasets

Functional enrichment analysis

Survival analysis

Oncogenes

Tumor suppressor genes

Colorectal cancer

NOP14 and PLA2G15 can be used to predict clinical outcomes in patients with colorectal cancer.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSN1</td>
<td>-0.000625538</td>
</tr>
<tr>
<td>ERI1</td>
<td>-0.003511305</td>
</tr>
<tr>
<td>ETF1</td>
<td>-0.000545896</td>
</tr>
<tr>
<td>GLMN</td>
<td>-0.012356453</td>
</tr>
<tr>
<td>LRRC8A</td>
<td>0.001503675</td>
</tr>
<tr>
<td>MAPKAPK3</td>
<td>-0.002917185</td>
</tr>
<tr>
<td>NOP14</td>
<td>-0.000267735</td>
</tr>
<tr>
<td>PCAT6</td>
<td>0.001982782</td>
</tr>
<tr>
<td>PLA2G15</td>
<td>0.004877215</td>
</tr>
<tr>
<td>SRD5A1</td>
<td>0.001475701</td>
</tr>
<tr>
<td>TIGD1</td>
<td>0.023072475</td>
</tr>
</tbody>
</table>