Construction of a Prognostic Model Based on Methylation-Related Genes in Patients with Colon Adenocarcinoma

ZhenDong Liu¹, YuYang Xu¹, Shan Jin², Xin Liu¹, BaoChun Wang¹

¹Department of General Surgery, Hainan General Hospital, Hainan Affiliated Hospital of Hainan Medical University, Haikou, Hainan Province, People’s Republic of China; ²Department of Anesthesiology, Second Affiliated Hospital of Hainan Medical University, Haikou, Hainan Province, People’s Republic of China

Correspondence: BaoChun Wang, Department of General Surgery, Hainan General Hospital, Hainan Affiliated Hospital of Hainan Medical University, Haikou, 570311, People’s Republic of China, Tel +86-13876180006, Email Drlance9@163.com

Purpose: Colon adenocarcinoma (COAD) is the second leading cause of death in the world, and the new incidence rate ranks third among all cancers. Abnormal DNA methylation is related to the occurrence and development of tumors. In this study, we aimed to identify genes associated with abnormal methylation in COAD.

Methods: COAD transcriptome data, methylation data and clinical information were downloaded from the TCGA database and GEO database. The differentially expressed genes (DEGs) and methylated genes (DMGs) were analyzed and identified in COAD. PCA analysis was applied to divide COAD into subtypes, and the survival and immune cell infiltration of each subtype were evaluated. Cox and LASSO analyses were performed to construct COAD risk model. GSEA was used to evaluate the enrichment pathways. The Kaplan–Meier was used to analyze the difference in survival. ROC curve was plotted to evaluate the accuracy of the model, and GSE17536 was used to verify the accuracy of the risk model. The risk model is combined with the clinicopathological characteristics of COAD patients to perform multivariate Cox regression analysis to obtain independent risk factors and draw nomograms.

Results: In total, 4564 DEGs and 1093 DMGs were screened, among which 298 were found to be overlapping genes. For 220 of these overlapping genes, the methylation was significantly negatively correlated to expression levels. An optimal signature from 4 methylated biomarkers was identified to construct the prognostic model.

Conclusion: Our study identified 4 methylated biomarkers in the COAD. Then, we constructed the risk model to provide a theoretical basis and reference value for the research and treatment of COAD.

Keywords: colorectal cancer, methylated genes, immune cell infiltration, WGCNA, prognosis

Introduction

Colorectal cancer (CRC) is currently the second most lethal tumor in the world, the incidence of CRC has increased to the third among all cancers.¹ It is generally believed that environmental and chronic intestinal inflammation are the risk factors of CRC.² Most cases of CRC are considered as a result of stepwise accumulation of mutations, suppression of tumor suppressor genes and activation of oncogenes.³ Colon adenocarcinoma (COAD) is the most common subtype of CRC. Significant progresses have been made in early diagnosis, surgery, radiotherapy and chemotherapy in the past decades;⁴ however, the prognosis of colorectal cancer patients is still poor, and the 5-year survival rate is low. Therefore, diagnostic markers and therapeutic targets of COAD are needed to further explore to provide new methods of the treatment and intervention for patients.

Recently, epigenetic alterations have been presumed significant contributors to cancer development.⁵ DNA methylation is the most widely studied epigenetic alteration in cancer, which can regulate gene expression, and involved in early stages of tumorigenesis. Epigenetic markers are thought to be interesting biomarkers for the diagnosis, prognosis and prediction of treatment response of cancer.⁶⁷ It is generally believed that epigenetic changes in CRC can be manifested earlier than genetic
changes. It makes the new epigenetic therapy an attractive treatment strategy by targeting to inhibit DNA methylation genes. However, there is insufficient evidence for the use of methylated genes in routine clinical practice for CRC currently, and further explorations are needed to discover more useful diagnostic biomarkers.

The purpose of this study was to find abnormal methylation genes in CRC and evaluate their prognostic value in colon cancer.

**Materials and Methods**

**Patient Samples**

Ten tumor tissues and their adjacent normal tissues were isolated and collected from patients who were diagnosed with COAD at the Department of General Surgery, Hainan General Hospital. Participants were enrolled upon obtaining their written consent. This study acquired the approval from the ethics committee of Hainan General Hospital (NO. 2022 719).

**Data Source**

The clinical data, transcription and methylation profiles of 370 colon adenocarcinoma (COAD) patients were downloaded from TCGA database and were used as training set. GSE17536 was downloaded from GEO database and was used as external validation sets. A summary of clinical background data is presented in Table 1.

**Acquirement of Differentially Expressed Genes (DEGs) and Aberrantly Methylated Genes**

The microarray data of methylation and expression were analyzed by R software after normalization.

<table>
<thead>
<tr>
<th>Table 1 Characteristics of COAD Patients in Our Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
</tr>
<tr>
<td>&lt;70</td>
</tr>
<tr>
<td>≥70</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td><strong>T Stage</strong></td>
</tr>
<tr>
<td>T1</td>
</tr>
<tr>
<td>T2</td>
</tr>
<tr>
<td>T3</td>
</tr>
<tr>
<td>T4</td>
</tr>
<tr>
<td>TX</td>
</tr>
<tr>
<td><strong>N Stage (pN)</strong></td>
</tr>
<tr>
<td>NO</td>
</tr>
<tr>
<td>N1</td>
</tr>
<tr>
<td>N2</td>
</tr>
<tr>
<td>NX</td>
</tr>
<tr>
<td><strong>M Stage</strong></td>
</tr>
<tr>
<td>MO</td>
</tr>
<tr>
<td>MI</td>
</tr>
<tr>
<td><strong>TNM Stage</strong></td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>II</td>
</tr>
<tr>
<td>III</td>
</tr>
<tr>
<td>IV</td>
</tr>
<tr>
<td>X</td>
</tr>
</tbody>
</table>
R package “limma” was used to identify differentially expressed mRNAs between normal and COAD samples (| logFC | >1, \( P_{\text{adj}} < 0.05 \)). Similarly, “limma” package was used to screen aberrantly methylated genes (| logFC | >0.2, \( P_{\text{adj}} < 0.05 \)). Thereafter, the differentially expressed mRNAs and the aberrantly methylated genes were intersected.

**Consensus Clustering in COAD Patients**

Consensus clustering, which helped us determine the best grouping scheme, was performed using the ConcensusClusterPlus package in R. The cumulative distribution function (CDF) reflects the value distribution of the consensus matrix for different \( k \) values. The optimal value of \( k \) was defined as the smallest increase in the area under the cumulative distribution function (CDF) curve. The heatmap corresponding to consensus clustering was generated using R heatmap package (http://CRAN.R-project.org/package=heatmap).

**Survival Analyses**

Kaplan–Meier plots were used to illustrate overall survival among COAD subtypes defined by DNA methylation profiles. Significant differences among the clusters were assessed using the Log rank test. Survival analysis was performed using the survival package in R. The relationship between clinical and biological characteristics with DNA methylation clustering was analysed using the chi-square test.

**Immune Cell Infiltration Analysis**

We analyzed RNA-seq data from the 370 patients in CIBERSORTx, then compared the degree of infiltration of 22 immune cells between the two clusters using the Wilcoxon test. Pearson correlations between immune cells were calculated using the Corrplot package.

The stromal scores, immune scores and estimate stores in TCGA COAD dataset were calculated using ESTIMATE algorithm to estimate the tumor purity for each COAD sample.

**GO Annotation and KEGG Pathway Enrichment Analysis**

In order to know more about the molecular functions of DEGs in COAD, GO annotation and KEGG enrichment analysis were conducted with the aid of the “clusterProfiler” package. GO analysis tends to explain gene function from three aspects: molecular function, biological process, and cellular components, while the KEGG pathway enrichment analysis is prone to describe gene function in the genomic and molecular levels and show the correlated genes (\( P_{\text{adj}} < 0.05 \)).

**Establishment and Validation of the Prognostic Model**

The DEGs from the TCGA-COAD dataset were subjected to univariate Cox regression analysis. Genes with \( P_{\text{adj}} < 0.05 \) were regarded as the genes correlated with COAD prognosis. Afterwards, Lasso Cox regression analysis was performed to select gene signatures in the TCGA-COAD dataset using the R packages “glmnet” (https://cran.r-project.org/web/packages/glmnet/index.html). Four powerful prognostic genes were selected through optimal value of lambda (\( \lambda \)). A risk score was calculated for each sample, which was defined as a linear combination of expression values of individual normalized genes weighted by their estimated Cox model regression coefficients. Patients were divided into high-risk group and low-risk group according to the average risk score. Hallmarks and KEGG pathways significantly enriched in high-risk group were analyzed by GSEA. The Kaplan–Meier survival and ROC curves were drawn to evaluate the predictive value of the prognostic gene signature using the R packages “survival” and “survival ROC”, respectively. The risk model was tested in the external validation sets.

Thereafter, multivariate Cox regression analyses were performed to identify independent prognostic factors for COAD patients. A nomogram was constructed to predict 1-, 3-, and 5-year survival in COAD patients according to independent prognostic factors. Nomogram performance was assessed using calibration curves and ROC curves.

**RNA Extraction and cDNA Synthesis**

Total RNAs of human cells were extracted using the Trizol reagent (Invitrogen, Carlsbad, CA, USA) according to the instructions of the manufacturer and treated with RQ1 DNase (Promega, Madison, WI, USA) to remove DNA. The
quality and quantity of the purified RNA were determined by measuring the absorbance at 260 nm and 280 nm (A260 and A280) using a SmartSpec Plus Spectrophotometer (Bio-Rad Laboratories, Inc., Hercules, CA, USA). RNA integrity was further verified by electrophoresis using a 1.5% agarose gel. All RNA samples were stored at −80 °C for future use. Reverse transcription reactions were carried out using ReverTra Ace qPCR RT Kit (TOYOBO Life Science, Shanghai, China), according to the manufacturer’s instructions.

**Quantitative Real-Time PCR (qRT-PCR)**

Expression levels of Bal1, ERBB4, MPP2 and PCDH9 gene were detected by qRT-PCR. For BAI1, the primers used were as follows: 5’-GTCCTCAGAACCACCCCATATTG-3’ (forward), 5’-GTCCTCTCAATCTCTAACCTGTC-3’ (reverse). For ERBB4, the primers used were as follows: 5’-TTTCGGAATTTGAGGAAATGG-3’ (forward), 5’-GAAACTCCTTTGCCCCCTCGTA-3’ (reverse). For MPP2, the primers used were as follows: 5’-GAGACGTTCAGAGCCCCTG-3’ (forward), 5’-ACCAGCTCCAGGTTTGTC-3’ (reverse). For PCDH9, the primers used were as follows: 5’-ATTGGATAGAGAGCGCAGAAGAGTCC-3’ (forward), 5’-ACCAGCTCCAGGTTTGTC-3’ (reverse). Actin was used as an internal reference control and its detection was performed with the following primers: Forward: 5′-TGGACTTCGAGCAAGAGATG-3′ and reverse: 5′-GAAGGAAGGCTGGAAGAGTG-3′.

The qRT-PCR was performed on a Bio-Rad S1000 with Bestar SYBR GreenRT-PCR Master Mix (TOYOBO). PCR conditions consisted of denaturing at 95 °C for 1 min, and 40 cycles of denaturing at 95 °C for 15s followed by annealing and extension at 60 °C for 30s. Relative gene expression was calculated using the Livak and Schmittgen 2−ΔΔCt method, normalized with the reference gene Actin. PCR amplifications were performed in triplicate for each sample.

**Results**

Identification of the Differentially Expressed Genes and the Aberrantly Methylated Genes in COAD

Based on the mRNA expression data and the methylation data obtained from the TCGA-COAD dataset, a total of 4564 differentially expressed mRNAs (Figure 1A) and 1093 abnormal methylation sites (Figure 1B) were obtained. Thereafter, 298 genes were identified by intersecting differentially expressed mRNAs and the aberrantly methylated genes (Figure 1C). Among them, the expressions of 220 genes were negatively correlated with methylated level.

Consensus Clustering of COAD Identified Distinct DNA Methylation Prognosis Subtypes with Different Prognosis and Immune Microenvironment

These 220 methylated DEGs were used for consensus clustering to obtain distinct DNA methylation prognostic molecular subtypes of COAD (Figure 2A–C). Finally, we created a heatmap of the consensus matrix when k = 2 (Figure 2D). Meanwhile, principal component analysis (PCA) showed that patients in the cluster 2 were significantly
Figure 2. Identification of the DNA methylation prognosis subtypes based on 220 methylated DEGs.

**Notes:**
(A) Consensus clustering matrix fork = 2. 
(B) CDF plot of the consensus score (k = 2–9). 
(C) Delta area curve of the sample. The Delta area curve of the consensus clustering indicates the relative change in area under the cumulative distribution function (CDF) curve for each category number k compared with k − 1. The horizontal axis represents the category number k, and the vertical axis represents the relative change in area under the CDF curve. 
(D) Consensus matrix, the heat map of sample clustering when k = 2. 
(E) PCA based on clusters 1 and 2 results. 
(F) Kaplan–Meier analysis of survival rate for clusters 1 and 2. 
(G) Differential immune infiltrates in clusters 1 and 2 of TCGA-COAD dataset. The proportion of 22 immune infiltrates for each case is estimated using the ESTIMATE algorithm. 
(H) Correlation matrix followed by unsupervised hierarchical clustering in 22 immune subsets. Pearson correlation coefficients (R) were calculated. Correlation coefficients were plotted with negative correlation (blue), positive correlation (red), and R = 0 (white). 
(I) Comparison of the (I) stromal, (J) immune and (K) estimate scores between clusters 1 and 2.
different from patients in the cluster 1 (Figure 2E). Kaplan–Meier survival analysis showed that the outcome of COAD patients in two clusters was significantly different (Figure 2F).

Comparing the degree of infiltration of 22 kinds of immune cells, we detected that the abundance of most immune cell types, except T cell CD4 naive, was significantly higher in cluster 2 than those in cluster 1 (Figure 2G). Monocytes, T cells gamma delta and eosinophils showed a positive correlation (Figure 2H). Also, the cluster 2 was found to have the higher stromal, immune and estimate scores (Figure 2I–K).

**GO Annotation and KEGG Pathway Enrichment Analysis**

To gain more insight into the molecular mechanism of the 220 methylated DEGs underlying the initiation and progression of COAD, GO annotation and KEGG pathway enrichment analysis were carried out using the “clusterProfiler” package. The most enriched GO terms of 220 genes were regulation of cation channel activity, embryonic organ morphogenesis, and glycosaminoglycan binding (Figure 3A), and the KEGG analysis revealed that the genes were significantly enriched in the signaling pathways involved in neuroactive ligand–receptor interaction, calcium signaling pathway (Figure 3B).

**Establishment and Validation of the Prognostic Model**

Firstly, the 220 methylated DEGs were subjected to univariate Cox regression analysis, and 29 genes with \( P_{\text{adj}} < 0.05 \) were identified to be significantly correlated with OS (Table 2). Then, lasso-penalized Cox analysis was performed to further narrow the mRNAs (Figure 4A and B). A 4-gene signature about BAI1, ERBB4, MPP2, and PCDH9 was identified based on the optimal value of \( \lambda \). Spearman correlation analysis showed that 4 genes have strong correlation with each other (Figure 4C). The risk score was calculated as follows: \( \text{Risk score} = (0.008 \times \text{BAI1}) + (0.054 \times \text{ERBB4}) + (0.062 \times \text{MPP2}) + (0.085 \times \text{PCDH9}) \). The samples were classified into high-risk and low-risk groups according to the median risk score. Analyzing the correlation between clinical traits and risk scores, we found that the severity of traits is positively correlated with risk scores (Figure 4D–H).

Then, we analyzed the distribution of clinicopathological variables and the expression of four genes in the high- and low-risk groups (Figure 5A). The high-risk and low-risk groups were distinguishable in both the N stage and stage groups. The expression levels of 4 genes in the high-risk groups are higher than those in the low-risk groups (Figure 5B–E). Survival analysis showed that the lower survival outcome of COAD patients was related to higher risk score (Figure 5F–G). The Kaplan–Meier analysis showed a significant difference in the outcome of the patients between the high-risk group and the low-risk group (Figure 5H). The area under the ROC curve (AUC) for the model was 0.8 in TCGA (Figure 5I). In the GSE17536 cohort, Kaplan–Meier analysis also showed a significant difference of disease-specific survival (DSS) (Figure 5J) and overall survival (OS) (Figure 5K) in the two groups (high-risk and low-risk).

**Figure 3** Gene ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the 220 methylated DEGs.

**Notes:** (A) GO annotation; (B) KEGG pathway enrichment.
GSEA Analysis

We analyzed the hallmarks and KEGG pathways enriched in the high-risk group by GSEA. Hallmarks such as adipogenesis, apical junction, hypoxia, myogenesis, and notch signaling were found to be significantly enriched in the high-risk group (Figure 6A). Additionally, KEGG pathways, including adherens junction, colorectal cancer, pancreatic cancer, ERBB, and

Table 2 Univariate Cox Regression Analysis of Overall Survival for 29 Methylated DEGs

<table>
<thead>
<tr>
<th>ID</th>
<th>HR</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCDH9</td>
<td>1.305635503</td>
<td>0.000416363</td>
</tr>
<tr>
<td>MPP2</td>
<td>1.389700255</td>
<td>0.000944191</td>
</tr>
<tr>
<td>SH3GL3</td>
<td>1.66570761</td>
<td>0.001690292</td>
</tr>
<tr>
<td>ERBB4</td>
<td>1.533967257</td>
<td>0.00238687</td>
</tr>
<tr>
<td>BAI1</td>
<td>1.339050065</td>
<td>0.002693955</td>
</tr>
<tr>
<td>DPY19L2</td>
<td>1.286991265</td>
<td>0.004343696</td>
</tr>
<tr>
<td>SPOCK3</td>
<td>1.269237521</td>
<td>0.00862801</td>
</tr>
<tr>
<td>GPR26</td>
<td>1.515352433</td>
<td>0.010395582</td>
</tr>
<tr>
<td>SUSD5</td>
<td>1.20318346</td>
<td>0.01462343</td>
</tr>
<tr>
<td>THBS4</td>
<td>1.094211413</td>
<td>0.014733892</td>
</tr>
<tr>
<td>SYPL2</td>
<td>1.23181808</td>
<td>0.016285655</td>
</tr>
<tr>
<td>CLDN11</td>
<td>1.18579082</td>
<td>0.016526704</td>
</tr>
<tr>
<td>PDE1B</td>
<td>1.206703294</td>
<td>0.023462408</td>
</tr>
<tr>
<td>LRRCD7</td>
<td>1.287364371</td>
<td>0.023538227</td>
</tr>
<tr>
<td>MAP6</td>
<td>1.197648896</td>
<td>0.024951089</td>
</tr>
<tr>
<td>TNFAIP8L3</td>
<td>1.23212855</td>
<td>0.027177005</td>
</tr>
<tr>
<td>GAP43</td>
<td>1.139091339</td>
<td>0.030803494</td>
</tr>
<tr>
<td>HCG4</td>
<td>1.217710765</td>
<td>0.031797059</td>
</tr>
<tr>
<td>HAPLN4</td>
<td>1.250325133</td>
<td>0.036725183</td>
</tr>
<tr>
<td>CLEC4G1</td>
<td>1.361106247</td>
<td>0.039393969</td>
</tr>
<tr>
<td>GRIN2A</td>
<td>1.188587806</td>
<td>0.039606636</td>
</tr>
<tr>
<td>P2RX2</td>
<td>1.442913531</td>
<td>0.041288082</td>
</tr>
<tr>
<td>SV2B</td>
<td>1.143512853</td>
<td>0.04381782</td>
</tr>
<tr>
<td>GSTM2</td>
<td>1.17805037</td>
<td>0.045839375</td>
</tr>
<tr>
<td>JPH4</td>
<td>1.218200089</td>
<td>0.04613808</td>
</tr>
<tr>
<td>PACSIN1</td>
<td>1.219576637</td>
<td>0.047952613</td>
</tr>
<tr>
<td>RNF180</td>
<td>1.209748005</td>
<td>0.048280595</td>
</tr>
<tr>
<td>KCNK12</td>
<td>1.620793438</td>
<td>0.048418966</td>
</tr>
<tr>
<td>TCEAL2</td>
<td>1.115879088</td>
<td>0.049142325</td>
</tr>
</tbody>
</table>
WNT signaling pathway, were also significantly enriched in the high-risk group (Figure 6B). These findings suggest potential roles for the four prognostic genes in the progression, cell fate and immune responses of COAD.

**Construction of the Nomogram in COAD**

Next, we performed multivariate analyses to detect independent prognostic factors. Age, gender, histology, distant metastasis, lymph node metastasis (LNM), stage, cancer status and risk score were included into multivariate analysis. The result showed that age, LNM, stage, cancer status and risk score were significantly associated with prognosis (Figure 6C), indicating that they are independent prognostic factors in COAD. Thereafter, we constructed a nomogram to predict the 1-, 3-, and 5-year survival of COAD patients based on independent prognostic factors (Figure 6D). The calibration curves for 1-, 3-, and 5-year showed that the nomogram-predicted probability of survival was close to the actual survival (Figure 6E–G). ROC curves further demonstrated the high accuracy of the nomogram in predicting the survival of COAD patients that the AUC was 0.778, 0.81 and 0.836 for 1-, 3-, and 5-year (Figure 6H–J).
Figure 5 Prognostic risk scores correlated with clinicopathological features.

Notes: (A) Heatmap and clinicopathologic features of high- and low-risk groups. (B–E) Distribution of expression levels stratified by 4 genes. (F) Distribution of risk scores in high-risk group and low-risk group. Red point indicates case in high-risk group and green indicates low-risk case. (G) Distribution of survival status of patients in high-risk group and low-risk group. Green point represents alive and red point for death. **P < 0.01, ***P < 0.001. (H) Survival curve of outcome. Red line depicts the survival of high-risk patients and blue line for low-risk patients. (I) ROC Curve for risk score. (J) DSS curve and (K) OS curve for COAD (GSE17536).

Abbreviation: AUC, area under curve.
More and more evidence indicated great exploration potential in the pathogenesis of malignant tumors. The TCGA database covers 36 types of cancer and detailed clinical information of patients involved. The GEO database provides accession numbers for Gene Set Enrichment Analysis (GSEA) in high-risk patients and construction of the nomogram in COAD.

**Notes:**
(A) The enriched Hallmarks in high-risk group.
(B) The enriched KEGG pathways in high-risk group.
(C) Multivariate analysis was performed for assessing risk factors and constructing prognostic nomogram.
(D) The predicted 1-, 3-, and 5-year survival rates of COAD based on independent prognostic nomogram constructed using the risk score from prognostic signature and clinicopathological parameters.
(E–G) The calibration plot for internal validation of the nomogram. (H–J) ROC curves of the nomogram for 1-, 3-, and 5-year overall survival in COAD to evaluate the predictive performance of the nomogram.

**Discussion**

More and more evidence indicated great exploration potential in the pathogenesis of malignant tumors. The TCGA database covers 36 types of cancer and detailed clinical information of patients involved. The GEO database provides
online tools and faster data updates. The combination of these tools provides us a more convenient and efficient method to explore various malignant tumors.

The usual treatment of COAD is a comprehensive treatment strategy consisting of surgery, neoadjuvant therapy, transformation therapy and postoperative adjuvant therapy. However, the 5-year survival rates of COAD patients remain unsatisfactory. It is necessary for us to identify novel biomarkers to improve the management of COAD.

In this study, we identified 220 aberrantly expressed methylation genes, and cluster analysis indicated that there was a significant difference in survival and immune infiltration between the two clusters of patients. The increasing expression level in cluster 2 group has positive correlations with CD8+ T cell, CD4+ Memory Resting T cell, M0 macrophage, M1 macrophage and M2 macrophage, which are associated with tumor immune microenvironment. The infiltration of immune cells could be a promising prognostic biomarker for colorectal cancer.16 To further investigate the effect of methylation genes in biological processes and pathways, GO annotation and KEGG pathway enrichment analysis were performed. These 220 genes were enriched in signaling pathways like the regulation of cation channel activity, embryonic organ morphogenesis, glycosaminoglycan binding, cell fate commitment, synapse organization, collagen trimer, which were associated with the biological process of colonic epithelium and mucosa formation. Studies have revealed that cation channel activity and calcium signaling pathway were associated with the proliferation, migration, and survival of cancer cell.17–19 Glycosaminoglycan could promote the growth of colorectal cancer.20 In addition, DNA methylation and inhibition of gene expression are intimately related to the cellular differentiation process and colonic cancer formation.21,22 The embryonic organ morphogenesis and cell fate commitment signaling pathways clustered by 220 methylated genes in our study were also corroborated, so aberrantly expressed methylation genes may have effect on colon cancer through biological process.

Then, single-factor regression and LASSO regression algorithm were used to find characteristic genes, and 4-gene (BAI1, PCDH9, ERBB4 and MPP2) signature was identified. These genes were reported to be closely related to the prognosis of colon cancer. As a number of brain-specific angiogenesis inhibitor (BAI) family, BAI1 is an orphan GPCR-like receptor abundantly expressed in normal brain and has been suggested to play an important role in angiostasis by acting as a mediator of p53 signaling.23–25 It was reported that colorectal cancer was negatively correlated with BAI1 gene expression.23 The expression of PCDH9 is downregulated in tumor samples compared with normal tissues and suggested to be a potential biomarker for predicting the prognosis of patients with cerebral glial tumors.26 ERBB4 is over-expressed in colon cancer and enhances the survival and growth of cells driven by Ras and/or WNT signaling.27 The mutation of ERBB4 confers the risk of colorectal cancer, and this result suggested that ERBB4 would be a potentially novel biomarker for CRC susceptibility.28,29 MPP2 is a tumor suppressor gene and primarily participates in the adhesion of epithelial cell and immune infiltration.30,31 The expression levels of MPP2 were high in CRC tissues.32 It was reported that MPP2 is related to the 5-year survival of rate in colon cancer patients.33 Studies have also constructed a prognostic risk model to predict overall survival in COAD, with MMP2 being the hub gene in both.31,32 Similar results emerged in our study that low-risk group with BAI1, PCDH9, ERBB4 and MPP2 low-expression had a better overall survival outcome, although the expression of these 4 genes has no significant difference between cancerous and adjacent tissues (data not shown). In short, the methylation regulation mechanisms of these DEGs need to be further studied.

Of course, the specific role and mechanism of these genes in COAD need further experimental research. In our study, the potential mechanisms of four prognostic genes regulating colon cancer were analyzed by GSEA in high-risk group, and there were some hallmarks and signal pathways significantly enriched in high-risk group. Apical junction pathway is associated with metastatic potential, which could lead to aggressive clinical characteristics with shorter survival.34,35 Epithelial–mesenchymal transition (EMT) plays an essential role in cancer development,36 and previous reports showed that the apical junction pathway could increase metastatic potential through EMT as well as angiogenesis.34 In addition, GSEA results in our study revealed enrichment of both WNT and adherens junction signaling pathways in high-risk group. WNT signaling affected the proliferation and tumor formation invasion and metastasis of colon cancer cells,37,38 while adherens junction pathway involved in the canonical WNT signaling pathway.39 Notch signaling contributes to EMT of colon cancer and correlated to colonic inflammation.40,41 In summary, the abnormal expression of methylation genes may lead to the imbalance of these organisms which lead to worse survival.
Risk factors such as carcinoembryonic antigen (CEA) level, tumor site, TNM stage and other clinical pathological data have been reported related to prognosis for COAD, which have certain limitations as prognostic markers.\textsuperscript{42–44} Therefore, there needs further exploration of promising prognostic parameters to improve the survival and quality of life for COAD patients. In view of the important role of DNA methylation in colon cancer, it is necessary to pay more attention to the potential predictive of methylation-related genes for COAD prognosis.\textsuperscript{45} Several investigators have published related research. Xue et al identified 7 methylation-related genes as novel prognostic markers, which accurately classified patients into high- and low-risk groups with significantly different overall survival times.\textsuperscript{46} Zhou et al found that YTHDF3, KIAA1429, ALKBH5 and METTL3 were significantly correlated with OS of colon cancer and recto-sigmoid cancer (RSC).\textsuperscript{47} In our study, we constructed a four-gene risk prediction model, which combined age, LNM, stage, cancer status and risk score to facilitate personalized prediction. Time-dependent ROC indicated the high accuracy of the nomogram in predicting the survival of COAD patients.

In this study, we constructed a nomogram to predict COAD survival. However, there are still some limitations in our study. The prognostic symptom model needs to be further validated in other large independent samples to ensure the reliability. Functional experiments need to further reveal the possible mechanisms predicting the role of autophagy genes. In a word, we have provided a theoretical reference for the exploration of potential biomarkers for the diagnosis and prognosis of COAD in the future.

**Data Sharing Statement**

The gene expression, methylation, and clinical data are available from NCBI Gene Expression Omnibus (GEO: GSE17536) [https://www.ncbi.nlm.nih.gov/geo/] and the cancer genome database (TCGA: COAD) [https://portal.gdc.cancer.gov]. The RT-qPCR data are available from the corresponding author upon request.

**Ethics Approval and Consent to Participate**

The study was approved by the ethical committee of Hainan General Hospital (NO. [2022] 719). Written informed consent was obtained from individual. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

**Acknowledgments**

We would like to gratefully acknowledge Guangzhou Huayin Medical Laboratory Center. Ltd. for the kind technical help.

**Funding**

This research was supported by the Natural Science Foundation of Hainan (Grants No. 823RC562).

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


