Construction and Validation of a Coagulation Factor-Related Prognostic Model for Colorectal Cancer Based on the Public Database

Hong-kai Xu  
First Affiliated Hospital of Wenzhou Medical University

Shao-liang Han  (slhan88@126.com)  
First Affiliated Hospital of Wenzhou Medical University

Hao-feng Lu  
First Affiliated Hospital of Wenzhou Medical University

Rui-shuai Dai  
First Affiliated Hospital of Wenzhou Medical University

Sai-yi Han  
Wenzhou Medical University

Wei-dong Xie  
First Affiliated Hospital of Wenzhou Medical University

Article

Keywords: colorectal cancer, coagulation factor, prognostic model, NMF, subtype

Posted Date: January 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2382656/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

**Background:** Colorectal cancer is one of the most common malignant cancers in the world, which is a serious threat to human health with increasingly diagnosed cases every year. It has been reported that coagulation factors play an important role in various cancers. However, the role of coagulation factor-related genes in colorectal cancer remains unknown.

**Methods:** Gene expression data with clinical information of colorectal cancer samples were downloaded from the TCGA (The Cancer Genome Atlas) database and Gene Expression Omnibus (GEO) database, respectively. The coagulation factor-related prognostic model was constructed based on univariate, LASSO, and multivariate Cox regression analysis. In addition, colorectal cancer patients were classified into different subtypes according to non-negative matrix factorization (NMF) analysis. The nomogram and calibration curves were plotted to validate the accuracy of the coagulation factor-related prognostic model. Finally, the proportion of the infiltrating immune cells in different risk groups was analyzed by using immune cell infiltration.

**Results:** Seven coagulation factor-related genes were screened out to establish a prognostic model. The risk score of each colorectal cancer sample was calculated by the product of each prognostic coagulation factor-related gene with prognostic value and the corresponding gene expression of each prognostic coagulation factor-related gene. Patients with colorectal cancer were classified into high- and low-risk groups according to the median risk score. Survival curves indicated that colorectal cancer patients in the high-risk group had a worse prognosis both in the training set, internal validation set, and external validation set. Colorectal cancer patients were divided into three subtypes (subtype C1, subtype C2, and subtype C3) according to the optimal number of clusters. The nomogram we established was accurate to predict the overall survival of colorectal cancer patients. The Sankey plot suggested that colorectal cancer patients in the subtype C2 and low-risk group had a better prognosis. Finally, immune cell infiltration analysis indicated that macrophages might play an important role in the development of colorectal cancer.

**Conclusion:** The coagulation factor-related prognostic model was established based on STIM1, PLCB1, MAPK12, F2RL2, C8G, C9, and ADCY5. The colorectal cancer patients were divided into three subtypes, including subtype C1, subtype C2, and subtype C3. These findings might provide novel therapeutic strategies for the treatment of patients with colorectal cancer.

Introduction

Colorectal cancer is one of the most common malignant cancers in the world, which is a serious threat to human health with increasingly diagnosed cases every year[1]. Worldwide, more than half of colorectal cancer patients developed colorectal cancer associated with liver metastases[2]. Recently, various modalities have been developed to treat patients with colorectal cancer, including colonoscopy screening, surgery, radiotherapy, chemotherapy, and immunotherapy[3]. Previous studies showed that colorectal
cancer therapy remains largely dependent on disrupting protooncogenes and tyrosine-protein kinase receptors, especially in unresectable tumors[4]. In addition, colorectal cancer patients with advanced stages have a significantly lower 5-year survival rate than patients with early-stage [5]. Therefore, novel and effective therapies for the treatment of colorectal cancer are urgently needed.

It has been reported that coagulation plays an important role in maintaining hemostasis, which is associated with innate immunity as it promotes host-pathogen defenses via inhibiting pathogen dissemination and facilitating pathogen killing[6, 7]. Platelets, endothelial cells, and coagulation factors are the most significant factors that participated in coagulation[8]. Previous studies demonstrated that many biomarkers related to coagulation disorder showed significant prognostic relevance in a variety of cancers[9–13]. The patients with malignant tumors is in a hypercoagulable and hyperfibrinolytic state, which is caused by the enhancement of coagulation function and the production of massive fibrin in the early stages. In addition, the degree of coagulation disorders associated with tumor invasion, metastasis, and the prognosis of patients[14]. Therefore, it is necessary to explore the mechanism of coagulation factor-related genes in patients with colorectal cancer.

In this study, we constructed the coagulation factor-related prognostic model based on univariate, LASSO, and multivariate Cox regression analysis. In addition, we classified the colorectal cancer patients into different subtypes according to NMF clustering analysis. We plotted the nomogram and calibration curves to validate the accuracy of the coagulation factor-related prognostic model. Finally, we performed the immune cell infiltration analysis to evaluate the proportion of the infiltrating immune cells in different risk groups.

Methods

Data acquisition

We downloaded the transcriptome RNA sequencing data and corresponding clinicopathologic information of colorectal cancer samples from The Cancer Genome Atlas (TCGA) database. Meanwhile, we downloaded RNA sequencing data and survival information of 1048 colorectal cancer samples (GSE40967) from the Gene Expression Omnibus (GEO) database.

Extraction of coagulation factor-related gene expression

The expression of coagulation factor-related genes in LUAD was extracted in the expression matrix profiles of the training set and test sets. Next, the coagulation factor-related gene expression of the training set and test sets were normalized before further analysis.

Identification of prognostic coagulation factor-related genes

Firstly, the coagulation factor-related genes with prognostic value were obtained based on the univariate Cox regression analysis. Next, we performed Lasso Cox regression analysis to minimize prediction error
and remove potential collinearity among coagulation factor-related genes. Subsequently, the multivariate Cox regression analysis was utilized to screen out the prognostic coagulation factor-related genes.

Construction and validation of the prognostic coagulation factor-related model

According to the univariate, LASSO, and multivariate Cox regression analysis, we obtained an equation to calculate the risk score of each colorectal cancer patient in the training set as follows:

\[ \text{Riskscore} = \sum (X_1Y_1 + X_2Y_2 + \cdots + X_iY_i) \]

In this formula, \( X \) represents the prognostic value of each prognostic coagulation factor-related gene. \( Y \) represents the expression of each corresponding prognostic coagulation factor-related gene. \( i \) represents the number of prognostic coagulation factor-related genes. Subsequently, all the colorectal cancer patients were assigned to the high- and low-risk group based on the median value of the risk score. The GSE40967 cohort downloaded from the GEO database was selected as an external validation set. Consistent with the training set, colorectal cancer patients in the external validation set were also divided into high- and low-risk groups according to the median risk score. The Kaplan–Meier survival curves were plotted using the “survival” package in RStudio software (version 4.1.2) to visualize the differences in prognosis between high- and low-risk groups. Using the R package “survivalROC”, we mapped the ROC curves to access the accuracy of the prognostic coagulation factor-related model.

NMF classification of colorectal cancer patients

We performed NMF clustering analysis using the “NMF” package in R software (version 4.1.2) to divide colorectal cancer patients into different molecular subgroups based on the coagulation factor-related gene expression. Survival curves were plotted to visualize the prognosis of colorectal cancer patients in different molecular subgroups. Additionally, the ESTIMATE algorithm was performed to explore the relationship between different molecular subgroups and 22 immune cells.

Establishment of the nomogram for colorectal cancer patients

We utilized the “rms” package to build a nomogram including clinical characteristics and the risk score, which can predict the overall survival of colorectal cancer patients. In addition, we evaluated the accuracy of the nomogram with calibration curves.

Correlation analysis of clinical characteristics

We employed the R package “pheatmap” to visualize the correlation of risk score with clinical characteristics between low- and high-risk colorectal cancer patients. In addition, Chi-square tests were used to analyze each significant clinical characteristic.

Mutations in the high- and low-risk group

We utilized R package “maftool” to visualize the tumor mutation frequency (TMF) of colorectal cancer patients in high-risk and low-risk, respectively.
Immune cell infiltration analysis

To quantify the immune infiltration levels in colorectal cancer samples, the Cell-type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) algorithm was performed to calculate the abundance of 22 tumor-infiltrating immune cells.

GO enrichment analysis

To elucidate underlying mechanisms of the differentially expressed coagulation factor-related genes between high- and low-risk groups in biological process, GO enrichment analysis was performed to explore enriched GO terms with R packages “clusterProfiler,” “enrichplot,” and “ggplot2”. In addition, GO terms can be classified into three aspects: biological processes (BPs), cellular components (CCs) and molecular functions (MFs).

Statistical analysis

All the statistical analyses were performed with RStudio (version 4.1.2). p < 0.05 was considered as statistically significant.

Results

Construction of the prognostic coagulation factor-related model

In this study, colorectal cancer samples downloaded from TCGA database were randomly assigned into the training set and internal validation set. As shown in SupplementaryTable 1, the coagulation factor-related genes with prognostic value were screened through univariate Cox regression analysis. We then performed in LASSO regression analysis, 13 coagulation factor-related genes were screened based on the univariate Cox regression analysis (Fig. 1A and Fig.1B). Through multivariate Cox regression analysis, we established the prognostic coagulation factor-related model with STIM1, PLCB1, MAPK12, F2RL2, C8G, C9 and ADCY5 (Figure 1C). The risk score of each colorectal cancer patient can be calculated using the formular as follows:

Risk score = STIM1 × 0.14035 + PLCB1 × 0.37716 + MAPK12 × 0.45132 + F2RL2 × (-0.49693) + C8G × 0.03306 + C9 × 3.34435 + ADCY5 × 0.66410

According to the median value of the risk score, colorectal cancer patients were divided into high- and low-risk groups in the training set. The risk plot, survival status, and the heatmap of the 7 prognostic coagulation factor-related genes were shown in Figures 2A, C, E. In the survival curve, patients with colorectal cancer in the high-risk group had a worse prognosis than those in the low-risk group (Figure 2G). According to the ROC analysis, the AUC for 1-, 2-, and 3-year OS were 0.767, 0.812, and 0.768, respectively (Figure 2G). The abovementioned results indicated that the prognostic model had good performance to predict the prognosis of patients with colorectal cancer in the training set.
Validation of the prognostic coagulation factor-related model in the internal and external validation sets

In this study, we validated the prognostic coagulation factor-related model in the internal and external validation sets. Consistent with the training set, patients with colorectal cancer were assigned into high- and low-risk groups (Figure 2B, Figure 3A). The survival status indicated that colorectal cancer patients in the high-risk group had poor prognosis (Figure 2D, Figure 3C). The heatmaps were presented to visualize the distribution of the 7 prognostic coagulation factor-related genes between different groups (Figure 2F, Figure 3D). In addition, patients with colorectal cancer in the high-risk group had lower overall survival probability than those in the low-risk group (Figure 2H, Figure 3B). According to the ROC analysis, the AUCs for 1-, 2-, and 3-year OS were 0.639, 0.663, and 0.621 in the internal validation set, while the AUCs for 1-, 2-, and 3-year OS were 0.617, 0.600, and 0.571 in the external validation set (Figure 2J, Figure 3E). Taken together, our model is accurate to predict the prognosis of patients with colorectal cancer in the internal and external validation sets.

NMF classification of colorectal cancer patients

In this study, the expression characteristics of coagulation factor-related genes in colorectal cancer using NMF clustering algorithm were also explored. The results indicated that k = 3 appeared to be an optimal selection for sorting the colorectal cancer patients into three subtypes (subtype C1, subtype C2, and subtype C3) (Figure 4A). Additionally, the survival curve demonstrated that there were significant differences in overall survival time among subtype C1, subtype C2, and subtype C3 (p< 0.05, Figure 4B). The subtype C2 was significantly gathered in Natural killer T cell, subtype C3 were significantly gathered in Type 1 T helper cell, plasmacytoid dendritic cell (Figure 4C).

Independent prognostic factor of the coagulation factor-related model

Then, the univariate and multivariate independent prognostic analysis were conducted to verify whether the risk score could be an independent prognostic factor for the overall survival of patients with colorectal cancer. The results suggested that the risk score could be an independent prognostic factor for the overall survival of patients with colorectal cancer in the training set (Figures 5A, B).

Establishment of nomogram and calibration curves

To better stratify colorectal cancer patients with different prognoses, the nomogram was conducted to predict the 1-, 3-, 5-year survival of patients with colorectal cancer, including age, gender, stage, T stage (Figure 6A). In addition, calibration curves suggested that the predicted overall survival of colorectal cancer patients was in good agreement with the actual overall survival of colorectal cancer patients at 1, 3, and 5 years (Figures 6B-D). Based on those above mentioned results, we could conclude that our proposed nomogram was accurate to predict the overall survival of colorectal cancer patients.

Clinical correlation analysis of clinical characteristics
In this work, the differences of the clinical characteristics between high- and low-risk groups were analyzed. As shown in Figure 7A, stage, T stage, and M stage were significantly different between the high- and low-risk groups. As shown in Figure 7B, the proportion of M0 patients was mostly distributed in the low-risk group. Stage II patients, Stage III patients, and Stage IV patients were almost equally distributed in both high- and low-risk groups (Figure 7C). T1 patients and T3 patients were almost equally distributed in both high- and low-risk groups (Figure 7D).

**Mutation analysis between high- and low-risk group**

Patients with colorectal cancer in the high-risk group had markedly higher frequencies of TP53, TTN and MUC16 mutations (Figure 8A). Consistent with the high-risk group, colorectal cancer patients had markedly higher frequencies of TP53, TTN and MUC16 mutations in the low-risk group (Figure 8B). The Sankey plot suggested that colorectal cancer patients in the subtype C2 and low-risk group had a better prognosis (Figure 8B).

**Immune cell infiltration analysis**

In the training set, the percentage of 22 immune cells in each colorectal cancer sample was shown in Supplementary figure 1A. The heatmap showed that macrophages dominated whether in the high-risk and low-risk groups (Supplementary figure 1B). The proportion of T cells CD4 memory activated and Macrophages M1 were significantly higher in the low-risk group (Supplementary figure 1C).

**GO functional enrichment analysis of differentially expressed genes between high- and low-risk group**

As shown in Supplementary Figure 2, GO functional enrichment analysis revealed that the muscle contraction, immunoglobulin complex, and antigen binding were the enriched pathways in the aspects of BP, CC, and MF, respectively.

**Discussion**

The incidence of colorectal cancer is increasing year by year, and despite treatments have improved in recent years, recurrences and metastases still remain challenging to treat. In addition, colorectal cancer display molecular heterogeneity during tumourigenesis and therapeutic treatment[15], which might explain why patients with colorectal cancer had poor prognosis and limited therapeutic strategies. It is reported that patients with cancer often complicate with coagulation and fibrinolysis activation, which is associated with a higher risk of invasion, metastasis, and poorer long-term outcomes[16–18]. However, there are few studies investigating the role of coagulation factor-related genes in colorectal cancer. Thus, studying coagulation factor-related genes in colorectal cancer might be useful to plan the implementation of treatment protocols for patients with colorectal cancer.

In this study, we comprehensively analyzed the relationship between coagulation factor-related genes and colorectal cancer. And the coagulation factor-related prognostic model was established based on seven prognostic coagulation factor-related genes using univariate, LASSO, and multivariate Cox regression.
analyses. Subsequently, patients with colorectal cancer were classified into high- and low-risk groups based on the median value of the risk score. Survival curves indicated that colorectal cancer patients in the high-risk group had a worse prognosis both in the training set, internal validation set, and external validation set. ROC curves indicted that our model is accurate to predict the prognosis of patients with colorectal cancer. In addition, colorectal cancer patients were divided into three subtypes with the optimal number of clusters according to NMF clustering analysis. The proportion of infiltrating immune cells in three subtypes were significantly different, which demonstrated that coagulation factor-related genes are involved in the tumor microenvironment. According to abovementioned results, we can conclude that the classification of colorectal cancer patients and the establishment of coagulation factor-related prognostic model might provide novel therapeutic strategies for the treatment of patients with colorectal cancer.

We constructed the prognostic coagulation factor-related model based on STIM1, PLCB1, MAPK12, F2RL2, C8G, C9 and ADCY5. In addition, it is reported that the overexpression of STIM1 can promote the progression of colorectal cancer[19]. Previous studies showed PLCB1 overexpression can cause liver tumor cells to proliferate, contributing to the poor prognosis of liver cancer[20]. PLCB1 has also been detected in colon cancer, and its abnormal expression is closely linked to colon cancer[21]. It has been demonstrated that MAPK12 regulates cancer cell metastases [22]. Moreover, MAPK12 can also influence the risk and survival of patients with colorectal cancer by modulating the MAPK signaling pathway[23, 24]. F2RL2 has been reported to be a prognostic marker for glioma and breast cancer[25]. C8G plays a critical role in endothelial responses to inflammation and has potential utility as an S1PR2 inhibitor in treating inflammation[26]. The production of C9 was strongly upregulated by tumor necrosis factor alpha (TNF-α)[27]. ADCY5 is involved in elevating fasting glucose and increasing type 2 diabetes risks[28]. According to these results, we can conclude that the prognostic coagulation factor-related genes we identified were significant to predict the prognosis of patients with colorectal cancer.

According to the GO enrichment analysis, muscle contraction, immunoglobulin complex, and antigen binding were the most significant pathways enriched in the aspects of BP, CC, and MF, respectively. Among genes involved in muscle contraction, actin is involved in epithelial-to-mesenchymal transition (EMT) and is associated with cancer progression and metastasis[29]. Increasing evidence demonstrated that cancer cells express high levels of immunoglobulin[30]. Antigen binding residues located outside of the traditionally-defined CDRs but within these structural consensus regions make a significant contribution to antigen binding[31].

However, there were several limitation in this study, and the conclusions were based on bioinformatics analysis of public databases, lacking basic experiments in vivo and vitro.

**Conclusion**

The prognostic coagulation factor-related model was established based on STIM1, PLCB1, MAPK12, F2RL2, C8G, C9 and ADCY5. The colorectal cancer patients were divided into three subtypes, including
subtype C1, subtype C2, and subtype C3. These findings might provide novel therapeutic strategies for the treatment of patients with colorectal cancer.

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Availability of data and materials**

Any data and R script in this study can be obtained from the corresponding author upon reasonable request. The final manuscript was read and approved by all authors. In this study, publicly available datasets were analyzed. These are available on The Cancer Genome Atlas (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/).

**Competing interests**

The authors declare that the study was conducted without any financial relationships that could be considered as potential conflicts of interest.

**Authors’ contributions**

Shao-liang Han conceptualized and designed the study. Hong-kai Xu. Hao-feng Lu. and Rui-shuai Dai wrote the main manuscript text and Sai-yi Han. Wei-dong Xie prepared figures 1-8. Hong-kai Xu and Hao-feng Lu collected the data, completed experiment, and performed the analysis. All authors reviewed the manuscript.

**Acknowledgements**

We are grateful to the TCGA and GEO database for providing the platform and to the contributors for uploading their meaningful datasets.

**References**


**Figures**
Figure 1

Identification of coagulation factor-related genes by LASSO, multivariate Cox regression analysis in colorectal cancer. (A) Coefficient profiles of 24 coagulation factor-related genes. (B) 10-fold cross-validation for turning parameter (λ) selection. (C) Seven prognostic coagulation factor-related genes were screened out based on multivariate Cox regression analysis.
Figure 2

Construction of the coagulation factor-related prognostic model in the training set and validation of the coagulation factor-related prognostic model in the internal validation set. The risk plots of colorectal cancer patients in the training set (A), and internal validation set (B). The survival status of colorectal cancer patients in the training set (C), and internal validation set (D). The distribution of seven prognostic coagulation factor-related genes between the high- and low-risk group in the training set (E), and internal
validation set (F). The Kaplan–Meier analysis of overall survival between the high- and low-risk group in the training set (G), and internal validation set (H). ROC curves for predicting the 1-, 2- and 3-year survival of colorectal cancer patients in the training set (I), and internal validation set (J).

Figure 3

Validation of the coagulation factor-related prognostic model in the external validation set. (A) The risk plot of colorectal cancer patients in the external validation set. (B) The Kaplan–Meier analysis of overall survival between the high- and low-risk group in the external validation set. (C) The survival status of colorectal cancer patients in the external validation set. (D) The distribution of seven prognostic coagulation factor-related genes between the high- and low-risk group in the external validation set. (E) ROC curves for predicting the 1-, 2- and 3-year survival of colorectal cancer patients in the external validation set.
Patients with colorectal cancer were divided into three coagulation factor-related subtypes based on the NMF method. (A) Three subgroups were screened out as optimal values for consensus clustering. (B) The overall survival of colorectal cancer patients in three different coagulation factor-related subgroups. (C) Immune infiltration of colorectal cancer patients in three different coagulation factor-related subgroups.
Figure 5

Univariate (A) and multivariate (B) independent prognosis analysis of clinical characteristics and risk score.
Figure 6

Construction of nomograms to predict the overall survival of patients with colorectal cancer in the training set. (A) Nomogram including age, gender, stage, T stage, and risk score, for 1-, 3- and 5 years overall survival in patients with colorectal cancer. (B-D) Calibration curves of the nomogram for 1-, 3- and 5 years overall survival in patients with colorectal cancer.
Figure 7

The correlation analysis of clinical characteristics. (A) Heatmap of clinical characteristics. Clinical correlation plots of (B) M stage, (C) stage, (D) T stage.
Figure 8

Mutations in high- and low-risk groups. Overview of somatic mutations in the high- (A) and low-risk group (B). (C) Sankey plot reveals the relationship between coagulation factor-related subtypes and patients in the high-risk and low-risk groups. CFR cluster represents coagulation factor-related cluster.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.
• Supplementary11.jpg
• Supplementary21.jpg
• SupplementaryTable1.xlsx