A 16-CpG-based Prognostic Signature for Colorectal Cancer

Shuo Chen
Nankai University
Yan Wang
Shanghai Pudong New Area People's Hospital
Lin Zhang
Nankai University
Mingyue Xu
Nankai University
Boxue Wang
Nankai University
Yinan Su
Nankai University
Xipeng Zhang (zhangxipeng18212@outlook.com)
Nankai University https://orcid.org/0000-0002-9433-9368

Research article

Keywords: colorectal cancer, CpG, methylation, risk score

Posted Date: May 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-28805/v1

Abstract

Background: To develop a CpG-based prognostic prediction model to provide survival risk prediction for colorectal cancer. Differential methylation analysis was performed on 309 colorectal cancer and 38 adjacent cancer specimens from the Cancer Genome Atlas (TCGA).

Results: 2113 hypermethylation sites as well as 723 hypomethylation sites were screened out and 16 related CpG methylation loci were further identified. The risk score was calculated based on the methylation sites identified and utilized as an independent prognostic variable for multivariate Cox regression prediction model, which was further optimized by the independent prognostic factors (including stage and risk score).

Conclusion: This study has identified several potential prognostic biomarkers and established a CpG-based prognostic prediction model for colorectal cancer, which provides a valuable reference for future clinical research.

Introduction

Colorectal cancer, a predominant malignant tumor in the gastrointestinal tract, is the leading cause of cancer incidence and death worldwide. Moreover, its incidence and mortality rates have shown a rising trend[1]. The early symptoms of the patients are not obvious, following the changes in bowel habits, hematochezia, diarrhea, diarrhea and constipation alternating, local abdominal pain, while the late stage shows systemic symptoms such as anemia, weight loss and so on [2–4]. Colorectal cancer mostly occurs in people aged 40–70 years, but it has also been found in young people, which means that the disease has a tendency to become younger. In current statistics, colorectal cancer is more common in men. The incidence ratio of male to female is about 2:1[5]. Its morbidity and mortality are second only to gastric cancer, esophageal cancer as well as primary liver cancer in the digestive system malignant tumors[6]. Like other malignant tumors, the cause of the disease is still unclear and can occur in any part of the colon or rectum. The disease can spread to other tissues and organs through lymphatic, blood circulation even direct spread[7]. At present, the diagnosis can be confirmed by clinical manifestations, X-ray barium enema or fiberoptic colonoscopy. The key to treatment is early detection, timely diagnosis and surgical cure[8, 9].

It is well known that DNA methylation is abnormal in the majorities of cancers, including colorectal cancer [10–12]. Tumor DNA methylation has two general changes compared with normal cells of the same tissue type: demethylation in many regions of the genome is coordinated with *de novo* methylation of certain specific CpG islands. More notable, however, are the changes that occur on a wide range of CpG islands, which are often unmethylated in every tissue[13]. Although early observations suggested that this occurred primarily at the promoters of tumor suppressor genes as a result of growth selection, it now appears that this is a widely programmed process, possibly based on the relatively commonly used polyclonal compound targeting mechanism[14]. Vertebrate CpG islands are short, discrete DNA

sequences that are rich in GC and CpG, and are mainly unmethylated, which is significantly different from the average genomic pattern[15]. Meanwhile, CpG islands are considered as the sites of transcription initiation, including thousands of sites beyond the currently annotated promoters[16–19]. More and more studies demonstrate that abnormal DNA methylation of CpG islands plays an important role in the occurrence and development of tumors. Its characteristics such as the high stability of biological samples, sensitivity to tumor environmental factors, and ease of detection makes it very suitable as a clinically applicable biomarker for colorectal cancer diagnosis and prognosis.

In term of the complexity of the colorectal cancer prognosis, the current prediction model is not mature and needs to be updated for more targets. In address these issues, we constructed an innovative prognostic model for colorectal cancer based on 16 specific related CpG methylation locis. To make the prediction model more accurate, we did a multi-time evaluation. In general, these models can help predict the overall survival of colorectal cancer patients effectively, and can provide tremendous help for colorectal cancer patients.

Material And Methods

Study population

In this study, 347 colorectal cancer (CRC) samples were obtained from TCGA, which consisted of 309 tumor samples and 38 paracancerous samples that from a total of 296 patients. There were 287 CRC patients had complete survival information and their detailed clinicopathological characteristics were provide in Table 1.

Table 1
Clinical characteristics of 287 patients of COAD from TCGA.

		Total (N = 287)	Training set (N = 98)	Testing set one (N = 97)	Testing set two (N = 99)	Pvalue
Characteristics	Groups	Number	Number	Number	Number	
Sex	Male	155	48	56	54	0.6718
	Female	132	50	41	45	
Age at diagnosis	Median	66	67.5	65	67	0.9966
	Range	31–90	31–90	34–90	34–88	
Pathological TNM stage	I	43	12	17	16	0.9784
	II	110	43	35	35	
	III	83	28	30	27	
	IV	41	12	13	16	
	Unknown	10	3	2	5	
Vital Status	Alive	219	75	76	74	0.949
	Dead	68	23	21	25	

Differential Methylation Analysis

Differential methylation analysis was performed between tumor and paracancerous samples by using the minfi package in R language, after removing the CpG sites containing missing values. We then removed the methylation sites with an average of less than 0.02 by calculating the average of the methylation levels. Besides, methylation sites located on autosomes whose FDR values less than 0.05 and absolute values of the difference of average β values greater than 0.4 were identified as differential methylation CpGs.

Construction Of Prognostic Model

The 300 samples with complete survival information were randomly grouped according to the ratio of 2:1 as the training set and the validation set, respectively. Univariate Cox-regression analysis was used to screen for CpGs that were significantly related to CRC prognosis, with a threshold of $P < 0.05$. Then LASSO Cox-regression analysis was performed on the training set data to screen CpG set for the prognostic model of colorectal cancer. A risk score could be obtained for every CRC patient as follows

$$Riskscore = \sum_{i=1}^{n} Coef_i * x_i$$

Coefi was the risk coefficient of each CpG calculated by the LASSO-Cox model, and Xi was the methylation level of each CpG. CRC patients were classified into a high-risk group and low-risk group according to the cut-off value obtained from the R package “survminer”.

Establishment And Analysis Of Nomogram Prognosis Model

All independent prognostic factors identified by multivariate Cox were used to establish a nomogram to predict the probability of 1-year-OS, 3-years-OS, and 5-years-OS in CRC patients. The calibration curve of the nomogram was drawn to observe the relationship between the predicted probability of nomogram and the actual incidence. We then compared the performance of nomograms that including one or all prognostic factors in predicting CRC OS by using a time-dependent ROC curve.

Survival Analysis

The Kaplan-Meier method was used to estimate the overall survival rate of CRC patients. Wilcoxon signed-rank test was applied to comparing prognostic differences among multiple groups. Significance of the difference in survival rates between different groups was tested using log-rank with the cutoff of p value < 0.05.

Statistical analysis

Multivariate Cox regression model was used to analyze independence of risk score. Chi-square test was used in Tabel 1 to analyze the clinical features of CRC patients. Statistical analyses were performed using R software v3.5.2. P value < 0.05 was considered significant in all of the above.

Results

Prognosis-related CpGs

We obtained a total of 2,836 differential methylation CpGs (DMCs) in CRC tumor tissues compared with adjacent normal tissues, including 2,113 hyper-methylation and 723 hypo-methylation ones. Figure S1A illustrated the differential methylation landscape of all CpGs, and Figure S1B showed the methylation level of the 2,836 DMCs in CRC tumor and adjacent normal tissue samples as a heatmap.

Crc Prognostic Model

We used the univariate Cox regression analysis on 287 CRC patients and identified 53 DMCs significantly associated with OS. The coefficients of those 53 CpGs was illustrated in Figure S2A. LASSO-Cox was further established and determined 16 optimal prognostic genes, including cg01129320, cg01992382, cg03904639, cg05317090, cg06084210, cg09170112, cg09492451, cg09918510, cg11712188, cg12740527, cg14675211, cg15265085, cg16834823, cg18237607, cg18624636 and cg19611175. Besides, the Lambda value to turn parameter in the LASSO-Cox was in Figure S2B. And the risk score was as follow:

Risk Score = 1.1175 × β Value of cg01129320 + 0.049 × β Value of cg01992382 + 0.2382 × βValue of cg03904639 + 0.4814 × βValue of cg05317090 + 0.8733 × βValue of cg06084210 + 0.5062 × βValue of cg09170112 + 0.4175 × βValue of cg09492451 + 0.0327 × βValue of cg09918510 + 1.6943 × βValue of cg11712188 + 1.4206 × βValue of cg12740527 + 2.9668 × βValue of cg14675211 + 0.4418 × βValue of cg15265085 + 0.0137 × βValue of cg16834823 + 0.557 × βValue of cg18237607 + 0.4412 × βValue of cg18624636 + 0.9536 × βValue of cg19611175.

Risk score was a good assessment of patient survival prognosis

Patients were divided into high and low risk groups based on cut-off = 6.01. Kaplan-Meier curve showed that high-risk group had significantly longer survival compared to low-risk groups. Besides, we found that the AUCs of the 1-, 3-, and 5-year OS in the training set were 0.895, 0.89, and 0.959, respectively; the AUCs of the 1-, 3-, and 5-year OS in the first training set were 0.625, 0.659, and 0.74; the AUCs of OS at 1 year, 3 years, and 5 years in the second training set were 0.615, 0.565, and 0.65 respectively; the AUCs of OS at 1 year, 3 years, and 5 years in the whole training set were 0.709, 0.697, and 0.793, indicating that risk score could better predict patients 1-, 3-, and 5-year survival rates (Fig. 1A-1D). Collectively, the above results indicated that the established prognostic model according to these 16 CpG sites performed well in terms of survival prognosis.

Risk Score Was Independent Of Other Prognostic Factors

The 300 samples were grouped according to age, gender, and stage and their risk scores were calculated. The results revealed that there was a significant difference in risk scores between different ages, however, no significant difference could be obtained in genders and stages (Fig. 2A-2C). Next, multivariate Cox regression was established and proved the strong independence of our prognostic model including age, gender, and stage factors through the survival package in R (Fig. 2D). In addition to the risk score, stage was also an independent prognostic factor.

Nomogram could better predict the 1-year 3-year 5-year survival of patients.

The nomogram constructed using two independent prognostic factors, stage and risk score to predict the 1-, 3-, and 5-year OS of CRC patients was in Fig. 3A. Besides, the calibration chart to assess the accuracy of the nomogram was displayed in Fig. 3B, which proved the nomogram might infer estimates of slightly

higher or lower actual survival probability. Moreover, the 1-year, 3-year, 5-year AUC of the combined model was higher compared with a single factor, indicating that a nomogram established with all independent prognostic factors could better predict the patient's 1-year 3-year 5-year survival than risk score and stage (Fig. 4).

Discussion

The prognosis of colorectal cancer is critical, which mainly depends on early diagnosis and timely surgical treatment as well as effective prognostic prediction models. At present, researchers have been trying to find breakthroughs in the prevention, early screening, diagnosis and treatment of colorectal cancer, and have made many progress in these areas, but the prediction of its prognosis is still insufficient[20–22]. In this study, we performed differential methylation analysis on COAD samples and paracancerous control samples to reveal differential methylation sites, and then identified 16 methylation sites associated with prognosis. Furthermore, we established an innovative prognostic prediction model for colorectal cancer based on these specific loci.

It is well known that alterations in DNA methylation occur in cancer, including hypomethylation of oncogenes and hypermethylation of tumor suppressor genes. We screened 2836 differential expressed methylation sites, 2113 hypermethylation sites and 723 hypomethylation sites from 309 cancer samples and 38 adjacent cancer samples in this study (See Figure S1 for the results). It is worth noting that majority of sites were hypermethylated in the colorectal cancer. This raises a very interesting question: the overall level in tumor tissue may be hypermethylated, but it is hypomethylated at certain important sites. Since DNA methylation represents a molecular mechanism related to gene inhibition, it is believed that methylation in cancer may promote tumor phenotype by inhibiting genes that are initially active in the source tissues, especially those related to tumor suppressor factors, such as Rb, P53 etc[23, 24]. On the other hand, the other hypomethylation sites are potentially associated with oncogene-directed methylation-associated gene-repression pathways, taking Myc, Ras and Src as examples[25–27]. A study by Irizarry etc show that most methylation alterations in colorectal cancer occur in 'CpG island shores', with hypermethylation enriched closer to the associated CpG islands[28]. Moreover, Guo etc demonstrate that epigenomic changes in DNA CpG methylation are closely associated with the local inflammatory response from colorectal cancer[29]. Here we identified 16 specifc expressed CpG methylation sites, which associated with 9 different genes (CCDC48, CSNK1A1L, GDNF, HYDIN, IRX5, LONRF2, NALCN, SLC16A12, TNXB). Based on our search, several genes including CCDC48, LONRF2 and TNXB have not been well studied in the colorectal cancer, which provides a very valuable starting point for future research.

In addition to the comprehensive analysis of CpG methylation site, another feature of this study is to establish the risk scores and obtain the optimal cutoff value. According to the cutoff value, patients can be effectively divided into low risk group and high risk group (See Fig. 1 for details). Previously, several studies have developed the prognostic prediction model. Most of them were constructed by competing nostic information and expression of the lncRNAs, miRNAs,

and mRNAs in colorectal cancer specimen[30–32]. Here, in this study, the prognostic prediction model was established according to all independent prognostic factors (stage, risk scores), which are based on the colorectal cancer specific CpG methylation site. This provides an alternative model which may be critical to progress in the prognosis prediction of colorectal cancer.

To conclude, this article study CpG methylation sites from TCGA data base. Through the comparison between colorectal cancer and paracancerous control samples, 16 CpG methylation sites show specific methylation manners, which indicates their potential functions for the colorectal cancer. With these sites, a critical prognostic prediction model has been developed. Overall we shed light on questions and challenges posed by the colorectal cancer, and we establish an innovative prediction model which can provide great help for future understanding of colorectal cancer.

Abbreviations

CRC	colorectal cancer
DMCs	differential methylation CpGs

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The dataset supporting the conclusions of this article is available in the TCGA (http://tcgaportal.org/).

Competing interests: The authors declare that they have no competing interests.

Funding: This research was funded by the Project of Integrated Traditional Chinese and Western Medicine, Chinese Medicine Department, Tianjin Municipal Health Committee **(grant number** 2019120). **The funders had no role in the design or implementation of this study or the drafting of the manuscript.**

Authors' contributions: SC, YW put forward the ideas of this article, written this article and analysed the data. LZ, YX helped for acquisition of data and analysis and interpretation of data. XW, NS, PZ helped for revising the manuscript All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

References

1. Gupta N, Kupfer SS, Davis AM. Colorectal Cancer Screening. JAMA. 2019;321(20):2022–3.
2. [illegible]nmation and colorectal cancer. Cancer. 2018;124(1):13.

3. Wenzel C. [Intra-tumour heterogeneity in colorectal cancer]. Pathologe. 2019;40(5):527–8.
4. Russell J. Management of colorectal cancer patients with brain metastases. ANZ J Surg. 2018;88(3):126.
5. The Lancet O. Colorectal cancer: a disease of the young? Lancet Oncol. 2017;18(4):413.
6. Dickson I. Colorectal cancer: Engineered colons for cancer research. Nat Rev Gastroenterol Hepatol. 2016;13(9):500.
7. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453.
8. Tejpar S, Stintzing S, Ciardiello F, Tabernero J, Van Cutsem E, Beier F, Esser R, Lenz HJ, Heinemann V. Prognostic and Predictive Relevance of Primary Tumor Location in Patients With RAS Wild-Type Metastatic Colorectal Cancer: Retrospective Analyses of the CRYSTAL and FIRE-3 Trials. JAMA Oncol. 2017;3(2):194–201.
9. Venook AP. Right-sided vs left-sided colorectal cancer. Clin Adv Hematol Oncol. 2017;15(1):22–4.
10. Asano N, Takeshima H, Yamashita S, Takamatsu H, Hattori N, Kubo T, Yoshida A, Kobayashi E, Nakayama R, Matsumoto M, et al. Epigenetic reprogramming underlies efficacy of DNA demethylation therapy in osteosarcomas. Sci Rep. 2019;9(1):20360.
11. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019;20(10):590–607.
12. Jung SE, Shin KJ, Lee HY. DNA methylation-based age prediction from various tissues and body fluids. BMB Rep. 2017;50(11):546–53.
13. Sabit H, Abdel-Ghany SE, OA MS, Mostafa MA, El-Zawahry M. Metformin Reshapes the Methylation Profile in Breast and Colorectal Cancer Cells. Asian Pac J Cancer Prev. 2018;19(10):2991–9.
14. Puccini A, Berger MD, Naseem M, Tokunaga R, Battaglin F, Cao S, Hanna DL, McSkane M, Soni S, Zhang W, et al. Colorectal cancer: epigenetic alterations and their clinical implications. Biochim Biophys Acta Rev Cancer. 2017;1868(2):439–48.
15. Fuso A, Lucarelli M. CpG and Non-CpG Methylation in the Diet-Epigenetics-Neurodegeneration Connection. Curr Nutr Rep. 2019;8(2):74–82.
16. Jia M, Gao X, Zhang Y, Hoffmeister M, Brenner H. Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. Clin Epigenetics. 2016;8:25.
17. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. Cancer Res. 2016;76(12):3446–50.
18. Kurdyukov S, Bullock M. **DNA Methylation Analysis: Choosing the Right Method**. *Biology (Basel)* 2016, 5(1).
19. Wang Q. CpG methylation patterns are associated with gene expression variation in osteosarcoma. Mol Med Rep. 2017;16(1):901–7.

20. Datta S, Chowdhury S, Roy HK. Metabolism, microbiome and colorectal cancer. Aging. 2017;9(4):1086–7.
21. Little KY, Kirkman JA, Duncan GE. Beta-adrenergic receptor subtypes in human pineal gland. J Pineal Res. 1996;20(1):15–20.
22. Morales S, Monzo M, Navarro A. Epigenetic regulation mechanisms of microRNA expression. Biomol Concepts. 2017;8(5–6):203–12.
23. Uxa S, Bernhart SH, Mages CFS, Fischer M, Kohler R, Hoffmann S, Stadler PF, Engeland K, Muller GA. DREAM and RB cooperate to induce gene repression and cell-cycle arrest in response to p53 activation. Nucleic Acids Res. 2019;47(17):9087–103.
24. Engeland K. Cell cycle arrest through indirect transcriptional repression by p53: I have a DREAM. Cell Death Differ. 2018;25(1):114–32.
25. Frankson R, Yu ZH, Bai Y, Li Q, Zhang RY, Zhang ZY. Therapeutic Targeting of Oncogenic Tyrosine Phosphatases. Cancer Res. 2017;77(21):5701–5.
26. Hameed DA, Yassa HA, Agban MN, Hanna RT, Elderwy AM, Zwaita MA. Genetic aberrations of the K-ras proto-oncogene in bladder cancer in relation to pesticide exposure. Environ Sci Pollut Res Int. 2018;25(22):21535–42.
27. McFadden DG, Politi K, Bhutkar A, Chen FK, Song X, Pirun M, Santiago PM, Kim-Kiselak C, Platt JT, Lee E, et al. Mutational landscape of EGFR-, MYC-, and Kras-driven genetically engineered mouse models of lung adenocarcinoma. Proc Natl Acad Sci U S A. 2016;113(42):E6409–17.
28. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.
29. Guo Y, Wu R, Gaspar JM, Sargsyan D, Su ZY, Zhang C, Gao L, Cheng D, Li W, Wang C, et al. DNA methylome and transcriptome alterations and cancer prevention by curcumin in colitis-accelerated colon cancer in mice. Carcinogenesis. 2018;39(5):669–80.
30. Huang QR, Pan XB. Prognostic lncRNAs, miRNAs, and mRNAs Form a Competing Endogenous RNA Network in Colon Cancer. Front Oncol. 2019;9:712.
31. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010–22.
32. Sidaway P. Colorectal cancer: Genomic landscape of mCRC revealed. Nat Rev Clin Oncol. 2018;15(3):134–5.

Figures

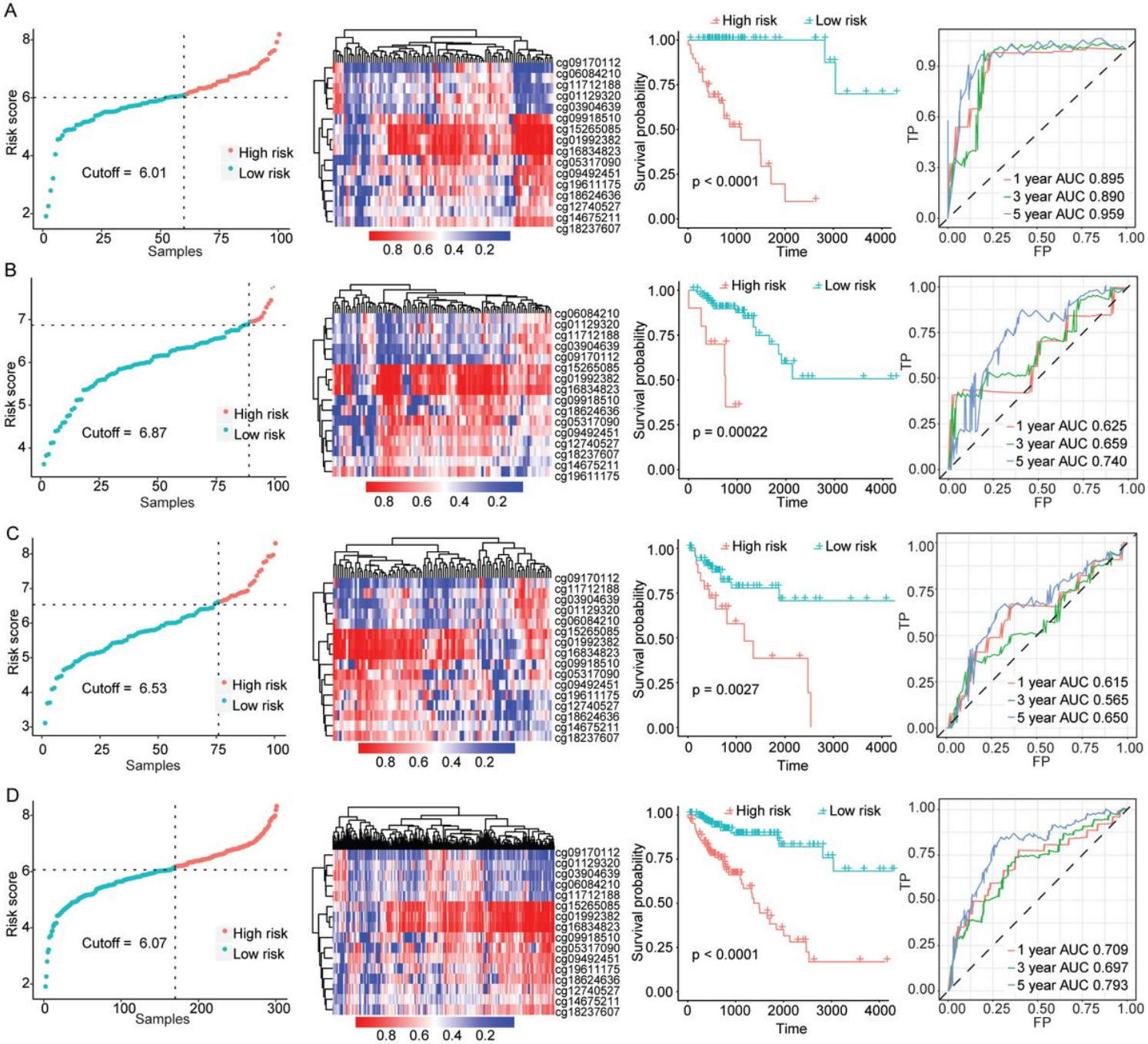

Figure 1

Evaluation for the prognostic ability of risk score. (A-D) The time-dependent ROC curves, the risk scores, the methylation heat map and the Kaplan Meier survival curves stratified by the optimal risk score of CRC samples from the TCGA training set, two testing sets, and the whole TCGA data set, respectively. The horizontal axis of the ROC curve is the false positive rate, and the vertical axis is the true positive rate. The horizontal axis of the risk score scatter plot is the sample data volume, the vertical axis is the risk score value, the color represents different groups, and the intersection point of the two groups is the optimal cutoff value. In the heat map, red represents high methylation levels and blue represents low methylation levels. The horizontal axis of Kaplan Meier survival curve is time in day, the vertical axis is

ent grouping.

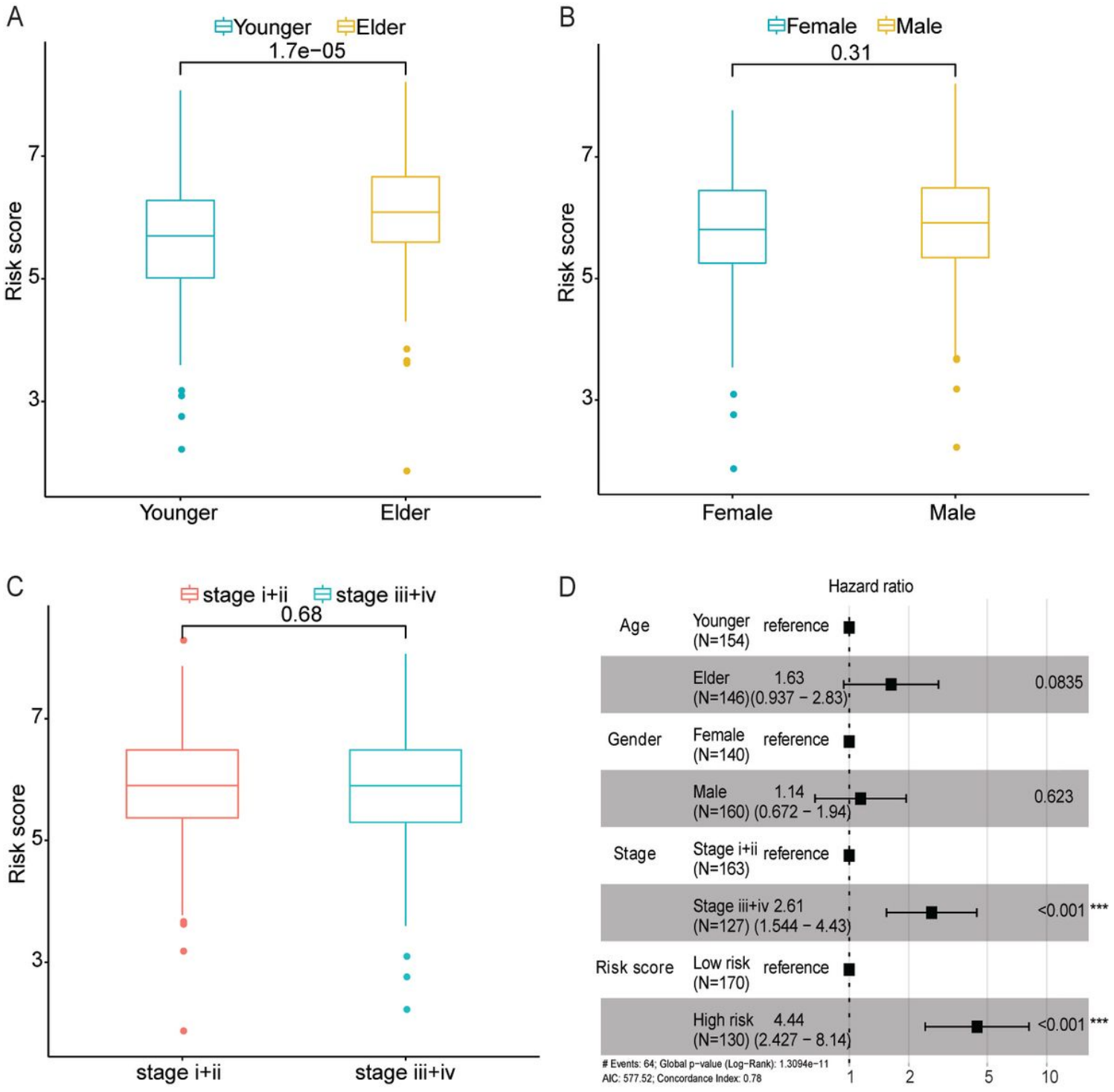

Figure 2

Relationship between risk scores and different clinicopathological features. (A-C) Box plot of risk scores between different groups. The horizontal axis is different groups, the vertical axis is risk scores, and different colored squares represent different sample groups. The value written on the line between the two groups is the P-value, and P<0.05 indicates a significant difference. (D) Multivariate Cox regression analysis forest plot.

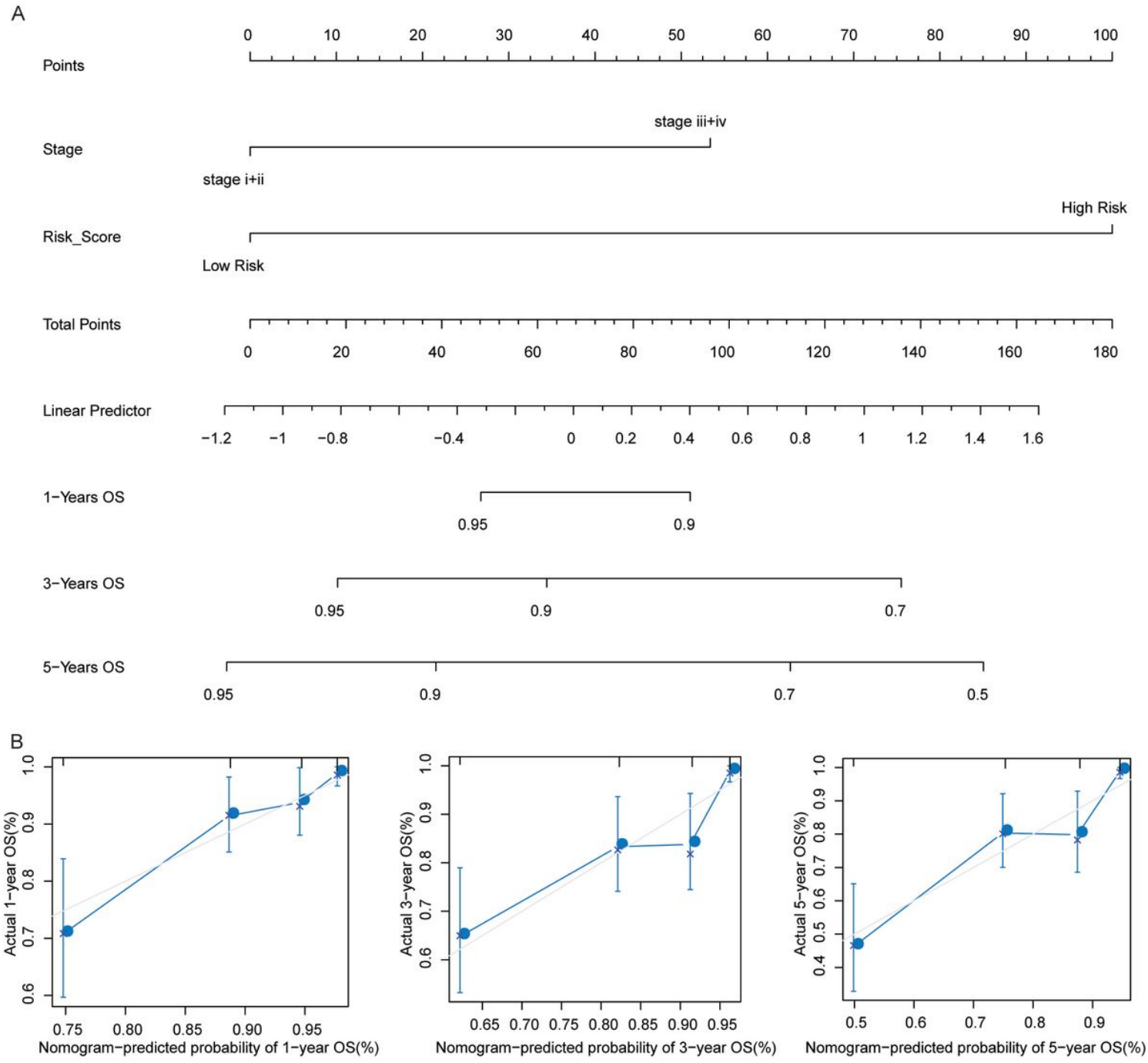

Figure 3

The prediction for overall survival in colorectal cancer patients by nomogram. (A) Constructed nomogram model. For each patient, three lines were drawn up to determine the points obtained from each factor. The sum of these points is on the total points axis, and a line is drawn down to determine the likelihood of overall survival for colorectal cancer patients at 1, 3, and 5 years. (B) Calibration chart for internal verification of the nomogram. The horizontal axis represents nomogram-predicted probability of overall survival, the vertical axis represents actual survival respectively.

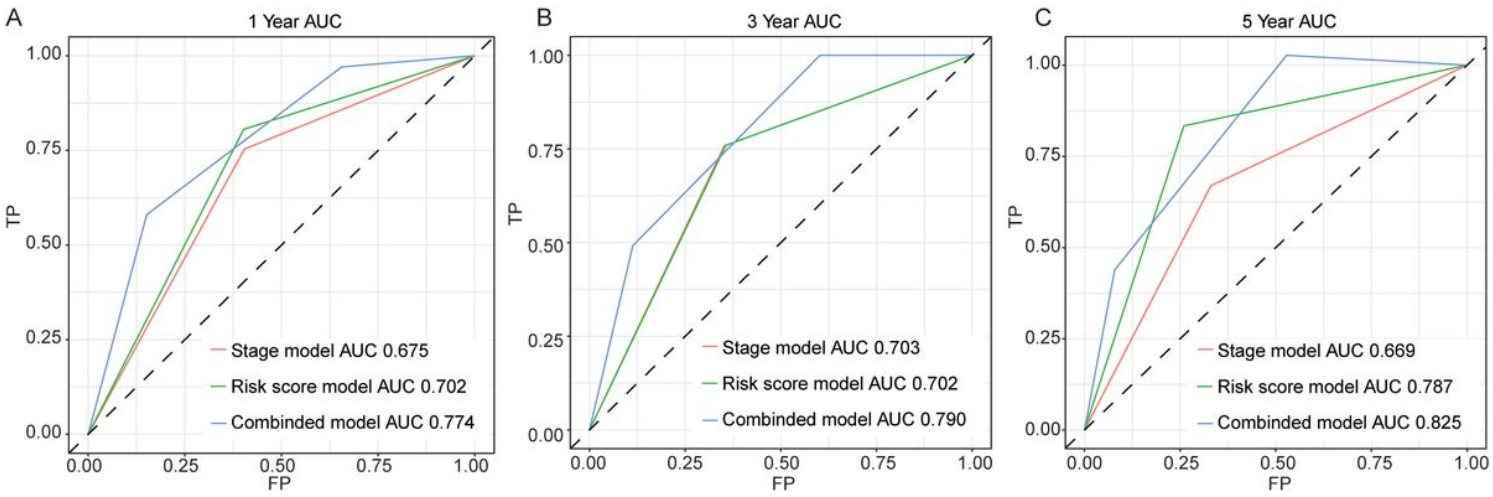

Figure 4

Time-dependent ROC curve of the nomogram model. (A-C) The time-dependent ROC curve of the 1-, 3-, and 5-year overall survival rates of nomogram models for colorectal cancer patients respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.tif
- FigureS2.tif