medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Early Diagnosis and Prognostic Prediction of Colorectal Cancer through Plasma Methylation Regions

Lingqin Zhu[1]*, Lang Yang[2,3]*, Fangli Men[4]*, Jianwei Yu[5]*, Shuyang Sun[1], Chenguang Li[1], Xianzong Ma[3], Junfeng Xu[2,3], Yangjie Li[3], Ju Tian[6], Xin Wang[3], Hui Xie[3], Qian Kang[1], Linghui Duan[3], Xiang Yi[7], Wei Guo[8], Xueqing Gong[8], Ni Guo[4], Youyong Lu[9#], Joseph Leung[10#], Yuqi He[1, 3#], Jianqiu Sheng[2,3#]

1. Department of Gastroenterology, Beijing Chest Hospital, Beijing Tuberculosis and Thoracic Tumor Research Institute, Laboratory for Clinical Medicine, Capital Medical University, No. 9 Beiguan Street, Tongzhou District, Beijing 101149, China
2. Senior Department of Gastroenterology, The First Medical Center of Chinese People's Liberation Army General Hospital, No. 28 of Fuxing Road, Haidian District, Beijing 100853, China;
3. Department of Gastroenterology, Seventh Medical Center of Chinese PLA General Hospital, No.5 Nanmencang, Dongcheng District, Beijing 100700, China;
4. Department of Gastroenterology, Dongying People's Hospital, Nan Yi Lu 317, Dongying District, Dongying 257000, China
5. Department of Gastroenterology，Longyan First Affiliated Hospital of Fujian Medical University, 105 Jiuyi Road (north), Longyan 364000, China
6. Department of Hepatobiliary and Pancreatic Surgery, The First Medical Center of Chinese PLA General Hospital, Institute of Hepatobiliary Surgery and Key Laboratory of Digital Hepetobiliary Surgery of Chinese PLA General Hospitall, No. 28 of Fuxing Road, Haidian District, Beijing 100853, China;
7. Mega Genomics Limited, 401 Health Work, North Garden Road, Haidian District, Beijing 100083, China;
8. Shanghai Yingce Biotechnology Co., Ltd, Room 202, 203, 2nd Floor, No. 4, Lane 288, Mingnan Road, Songjiang District, Shanghai 201613, China;
9. Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital/Institute, No. 52 Fucheng Road, Haidian District, Beijing 100142, China
10. Section of Gastroenterology, Sacramento VA Medical Center, Mather, California

*These authors contributed equally to this work.

Word count of the manuscript text (3365)

#Correspondence to:
Email: youyonglu@hsc.pku.edu.cn
Prof. Youyong Lu, MD, Ph.D: Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital/Institute, No. 52 Fucheng Road, Haidian District, Beijing 100142, China;
Email: ercpmaster@hotmail.com
Prof. Joseph Leung, Section of Gastroenterology, Sacramento VA Medical Center, 10535 Hospital Way, 111/G, Mather, CA 95655. Tel: (916)366-5339, Email: jwleung@ucdavis.edu
Email: endohe@163.com
Prof. Yuqi He, MD, Ph.D: Department of Gastroenterology, Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Thoracic Tumor Research Institute, No. 9

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Beiguan Street, Tongzhou District, Beijing 101149, China; Department of Gastroenterology, the 7th Medical Center of Chinese PLA General Hospital, No.5 Nanmencang, Dongcheng District, Beijing 100700, China;
Email: jianqiu@263.net
Prof. Jianqiu Sheng, MD, Ph.D: Department of Gastroenterology, the 7th Medical Center of Chinese PLA General Hospital, No.5 Nanmencang, Dongcheng District, Beijing 100700, China. Phone: +86-10-66721299; Fax: +86-10-66721024;

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

The methylation of plasma cell-free DNA (cfDNA) has emerged as a valuable diagnostic and prognostic biomarker in various cancers including colorectal cancer (CRC). Currently, there are no biomarkers that serve simultaneously for early diagnosis and prognostic prediction in CRC patients. Herein, we developed a plasma panel (27 DMRs, differential methylated regions) and validated its superior performance across CRC diagnosis and prognosis prediction in an independent cohort. We first conducted a preliminary screening of 119 CRC tissue samples to identify CRC-specific methylation features. Subsequently, a CRC-specific methylation panel was developed by further filtering 161 plasma samples. Then machine learning algorithms were applied to develop diagnosis and prognosis models using cfDNA samples from 51 CRC patients and 33 normal controls. The diagnosis model was tested in a cohort consisting of 30 CRC, 37 advanced adenoma (AA), and 14 healthy plasma samples, independently validated in a cohort consisting of 18 CRC, 91 NAA, 23 AA and 34 healthy plasma samples. In the tissue external validation cohort (GSE48684), the cfDNA methylation diagnosis model conducted with the panel, have the area under the curve (AUC) reached 0.983, and for the plasma cfDNA model in the external validation cohort, the sensitivities for NAA, AA and CRC 0 -II are 48.4%. 52.2% and 66.7% respectively, with a specificity of 88%. Additionally, the panel was applied to patient staging and metastasis, performing well in predicting CRC distant metastasis (AUC = 0.955) and prognosis (AUC = 0.867). Using normal samples as control, the changes in methylation score in both tissue and plasma were consistent across different lesions, although the degree of alterations varied with severity. The methylation scores vary between paired tissue and blood samples, suggesting distinct

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

mechanisms of migration from tumor tissue to blood for the 27 DMRs. Together, Our cfDNA methylation models based on 27 DMRs can identify different stages of CRC and predict metastasis and prognosis, ultimately enabling early intervention and risk stratification for CRC patients.

Keywords: Colorectal cancer; cell-free DNA Methylation; Diagnosis prediction; Prognosis prediction

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Colorectal cancer (CRC) is the third most prevalent cancer globally and ranks as the fourth leading cause of cancer death, accounting for 9.4% of cancer-related mortalities [1]. When patients were diagnosed with CRC, over half of their family (62.9%) faced financial burdens [2, 3]. Extensive research underscores that the survival rates of individuals diagnosed with advanced CRC (Stage III or IV) witness a significant decrement [4, 5]. Early detection and removal of precancerous lesions remain the most effective strategy to prevent CRC-associated mortality [6]. While colonoscopy is regarded as the gold standard for CRC detection, its limitations including invasiveness, suboptimal patient compliance and risk of complications including intestinal perforation, warrant further consideration [7]. Guaiac-based fecal occult blood test (gFOBT), fecal immunochemical test (FIT), carcino-embryonic antigen (CEA), and multi-target stool DNA (mt-sDNA) testing, are constrained in clinical application due to the lower sensitivity[8, 9].

Recent research demonstrated a profound correlation between CRC and the development of genetic and epigenetic alterations. During early stage of CRC development, epigenetic modifications surpass the frequency of gene mutations, indicating their potential as diagnostic biomarkers in screening for colon polyps and cancer[10]. Circulating tumor DNA (ctDNA) in cell-free DNA (cfDNA) primarily come from apoptotic and necrotic tumor cells, carrying cancer-specific epigenetic alterations [11]. Notably, blood cfDNA methylation emerges as a promising cancer screening pathway because of its early appearance in tumorigenesis and abundant signal density [12].

The value of cfDNA methylation in early diagnosis, detection of recurrence, molecular

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

subtyping and prognostic prediction of CRC has been proven[13]. However, there are still reported limitations. One study found a diagnostic panel for CRC demonstrated great performance, but it was derived from CRC tissue and normal blood leukocyte methylation data, potentially introducing bias due to the inconsistent sample types [14]. Another study addressed the potential of cfDNA methylation in patient risk stratification, but the requirement for blood sampling every three months led to poor patient compliance[15].

In this study, we aimed to develop plasma biomarkers derived from CRC tissues that exhibit superior performance in the diagnosis, metastasis and prognostic prediction for CRC. We meticulously screened CRC and normal tissues, and refining our selection within plasma samples. More importantly, the substantial proportion of paired tissue and plasma samples (from the same patient), effectively minimized potential data bias. The final resulting panel comprised 27 differential methylated regions (DMRs), with the area under the curve (AUC) reached 0.983 in the tissue validation cohort. Given that DNA methylation markers in plasma primarily stem from tumor tissues, these 27 DMRs exhibit significant potential in CRC plasma diagnostics. With a single blood test, we can ascertain the presence of CRC, distinguish the specific stages of the lesion (advanced adenoma, CRC 0-II, CRC III-IV), help identify distant metastasis, and achieve optimal risk stratification for CRC patients. This test is of clinical significance and holds promise in guiding the diagnosis and treatment for CRC.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Materials and Methods

Study design and samples

Specific DNA methylation markers for CRC were identified through analysis of TCGA public database data, alongside collected tissue and plasma samples. Subsequently, diagnosis, metastasis, and prognosis models for CRC were established and validated. The 450k chip methylation data encompassing colorectal, esophageal, gastric, lung, liver, and breast cancers, and CRC transcriptome sequencing data were downloaded from The Cancer Genome Atlas (TCGA) database. The methylation data of tissue validation cohort was obtained from Fifteen pairs of tissue samples from the Seventh Medical Center of PLA General Hospital were sequenced with the Roche Nimble Gen Seq Cap Epi method. The 89 tissue samples and 77 plasma samples used for marker screening were collected between June and December 2020 at the Seventh Medical Center of PLA General Hospital and Dongying People's Hospital. Participants for model establishment and validation were enrolled from June 2020 to July 2022 in the Seventh Medical Center of PLA General Hospital, Dongying People's Hospital. The paeticipants for external validation cohort were enrolled from June 2021 to December 2022 at the First Hospital of Longyan, Fujian Medical University. Inclusion and exclusion criteria are outlined in the Supplementary Materials. Detailed methods for DNA extraction, library construction, targeted bisulfite sequencing, and methylation data processing are provided in the Supplementary Materials.

CRC-specific methylation markers selection

Combining TCGA methylation data with our 15 pairs of tissue methylation sequencing results, probes were designed for further screening in the 89 samples (13 normal tissues and 38 paired

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

CRC tissues), 404 DMRs were identified. The detailed methods were listed in supplementary methods. Then, the methylation levels of 77 plasma samples from 13 Normal, 15 non-advanced adenoma (NAA), 12 advanced adenoma (AA) and 37 CRC participants, were analyzed. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis was used to filter markers, and 27 regions that appeared in more than 50 iterations were chosen as plasma diagnostic markers. The relationship between the methylation levels of the 27 DMRs and the transcription levels of their respective genes was compared by Spearman correlation coefficients.

Construction of the CRC diagnosis model

The methylation data from the training cohort underwent machine learning, culminating in the establishment of a five-fold cross-validated model characterized by binary deviance minimization standards. The diagnosis model was constructed using logistic regression, and its performance was evaluated through Receiver Operating Characteristic (ROC) curve analysis. The optimal cutoff value was determined using the maximum Youden index. The methylation data of tissue validation cohort (GSE48684) was downloaded from GEO (Gene Expression Omnibus) dataset. Using the methylation scores derived from the diagnosis model, the CRC staging (AA, CRC 0-II, CRC III-IV) was predicted.

Construction of the CRC metastasis and prognosis models

Utilizing the methylation scores derived from the diagnosis model, in conjunction with patient pathological parameters (metastasis), the plasma CRC metastasis model was constructed. The performance of this model underwent evaluation through ROC analysis, with the optimal cutoff value determined via the maximum Youden index. Furthermore, employing the same

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

methodology, the plasma prognosis model was established and validated, based on the methylation scores derived from the CRC diagnosis model and survival information from the patients.

Statistical analysis

All statistical analyses were conducted using SPSS 26.0 or R 4.1.3. Two-sided tests were used for p-values, with differences deemed statistically significant at $P < 0.05$. Model performance was assessed through ROC analysis using the "roc" function from the R package pROC, generating AUC and a 95% confidence interval (CI). Kaplan-Meier (KM) curves were used for survival curve estimation, with comparisons made through the log-rank test and hazard ratios determined by Cox regression.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Patient characteristics

To delineate DNA methylation biomarkers specific to CRC, 119 tissues and 77 plasma samples (originating from the same patients with tissues) were collected for methylation sequencing analysis, narrowing down to 142 DMRs. Then the 27 DMRs were selected with the sequencing data of 84 plasma samples. Following this, CRC diagnosis, metastasis, and prognosis models were constructed using plasma samples from 84 individuals (51 CRC, 33 Normal). Validation was then carried out in an independent cohort (30 CRC, 37 AA, 14 Normal). The CRC diagnosis model was further tested in an external validation cohort (18 CRC 0-II, 91 NAA, 23 AA, 34 Normal). Detailed information about the study design is illustrated in Figure 1, and comprehensive patient characteristics are summarized in Table s2-4.

Methylation markers selection

Analyzing the 450k chip methylation data from TCGA, the differential methylated sites (DMCs) in CRC were identified (Figure s1A). Combining with additional CRC-related methylation sites reported in the literature, 1438 DMCs were selected for validation. The heat map depicted the methylation levels of 1438 DMCs, revealing distinct methylation patterns between Normal and Tumor. (Figure s1B). The top 1400 DMCs were carefully selected based on the results of DNA methylation sequencing of 15 paired tissue samples (Figure s2A-C). The distribution of the DMCs was predominantly observed in introns and promoters, signifying substantial implications for gene expression and cellular functions (Figure s2D). Merging the 1438 DMCs with the 1400 DMCs for probe design, 404 DMRs were discerned

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

across 89 tissues (Figure s3A). The different methylation patterns between Normal and Tumor were also quite evident (Figure s3B-C). These DMRs correspond to genes that play pivotal roles in biological processes such as cell-cell adhesion, digestive system development, and cAMP and cGMP pathways (Figure s3D). Subsequently, 142 regions remained after sequencing 77 plasma samples (Figure s3E-F). The final 27 DMRs were selected after analyzing methylation data of the training cohort samples. Detailed information of the 27 DMRs is provided in Table s5. Correlation analysis revealed the methylation levels of the majority of the 27 DMRs are correlated with the transcription levels of their respective genes (Figure s4).

Development and validation of the CRC diagnosis model

Using methylation data from 27 DMRs in tissues, we applied machine learning to construct a diagnosis model for CRC and the AUC reached a high value of 0.994 (Figure 2A). In the external tissue independent validation cohort (GSE48684), the AUC for CRC is 0.983 and for AA is 0.966. (Figure 2B-C). The sensitivities for AA and CRC were 95% and 94%, respectively (Figure 2D-E). Besides, the plasma-based diagnosis model for CRC was constructed with training cohort and validated in an external independent validation cohort, the 27 DMRs exhibited distinctive methylation patterns between CRC and normal individuals (Figure 3A-B). The methylation scores of the model were significantly elevated in the CRC group compared to the normal and AA groups ($P < 0.001$), and they increased with advance tumor staging (Figure 3H). In the training cohort, the ROC curve analysis showed an AUC value of 0.928 [95% confidence interval (CI): 0.874–0.981] (Figure 3C). The sensitivities of the CRC diagnosis model in the training cohort for stages 0-II, III-IV, and all CRC stages

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

were 80%, 92%, and 86%, respectively, with a specificity of 85% (Figure 3I). ROC curve analysis in the validation cohort showed AUC values for AA, stages 0-II, III-IV, and all CRC stages as 0.714 (0.550–0.878), 0.890 (0.749–1.000), 0.967 (0.899–1.000), and 0.929 (0.827–1.000). (Figure 3D-G). The CRC diagnosis model demonstrated a specificity of 93% in the validation cohort, with sensitivities for AA, stages 0-II, III-IV, and all CRC stages of 43%, 67%, 100%, and 83%, respectively (Figure 3I). We further tested the early diagnostic performance of the model in an external validation cohort, and the results showed that the model achieved a sensitivity of 52% for AA and 48% for NAA (Figure 4A-E),

The plasma diagnosis model also serves to discern the specific staging of CRC patients. ROC curve analysis on the validation cohort revealed an AUC of 0.813 (0.656–0.970) for discriminating between CRC 0-II and CRC III-IV, and an AUC of 0.799 (0.693–0.905) for distinguishing between AA and CRC (Figure s5A-C). Kaplan-Meier survival curves revealed a significant reduction in overall survival (OS) for CRC III-IV patients compared to CRC 0-II patients identified by this diagnosis model (Figure s5D-E).

Development and validation of the CRC metastasis and prognosis models

Through an analysis of the relationship between the methylation scores in the diagnosis model of CRC patients' plasma samples and clinical-pathological parameters, we observed that methylation scores are associated with metastasis and staging ($P < 0.001$, Figure 5A-B), but not with age, gender, and lesion location ($P > 0.05$, Figure s6A-C). This suggests the possible role of 27 DMRs in metastasis and prognosis prediction. Therefore, we developed plasma-based CRC metastasis and prognosis models based on the 27 CRC-specific DMRs. In the training and validation cohorts, the metastasis model demonstrated high AUCs of 0.969 (95%

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

CI: 0.926-1) and 0.955 (95% CI: 0.878-1), respectively (Figure 5C). Methylation scores of M1-CRC individuals identified from the metastasis model were significantly higher than M0-CRC individuals (Figure s6D). Additionally, the M1-CRC patients exhibited shorter OS than M0-CRC patients. (Figure s6E-F). For the prognosis model, ROC curves demonstrated excellent performance in the training cohort (AUC = 0.883, 95% CI: 0.778-0.988) and validation cohort (AUC = 0.867, 95% CI = 0.728-1). (Figure 5D). Furthermore, Kaplan-Meier survival curves revealed a significant reduction in OS for the high-risk group of CRC patients identified by this prognosis model (Figure 5E). Based on the cutoff value determined by the algorithm, we divided CRC patients into high-risk and low-risk groups. Notably, almost all patients who developed distant metastases and those who had died belonged to the high-risk group (Figure 5F, H). We further characterized the survival status distribution and metastasis status between the two groups. As expected, the high-risk group had a higher proportion of deceased individuals, while non-metastatic patients were more prominent in the low-risk group (Figure 5G, I). These results indicate that the metastasis and prognosis models successfully identified patients who require further treatment. Multivariate regression analysis revealed a substantial correlation between the methylation score of the 27 DMRs and OS, indicating methylation score as an independent prognostic factor for CRC (Table s6). These findings underscore the considerable potential of the prognosis model based on the 27 DMRs specific to CRC in predicting the prognosis and conducting risk stratification for CRC patients.

Changes in methylation scores of tissue and plasma samples across different populations

Collectively, the CRC diagnosis model, established based on the methylation levels in these

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

27 DMRs within tissues, exhibits robust performance. However, the plasma-based diagnosis model exhibited reduced performance in AA and CRC 0-II compared to CRC III-IV. Therefore, we analyzed the methylation scores of the 27 DMRs generated from diagnosis models in paired tissue and plasma samples. Notably, in tissues, the methylation scores are ranked as AA > CRC > Normal (Figure 6A), suggesting significant alterations in the methylation of the 27 DMRs at the onset of precancerous lesions. In plasma, methylation scores gradually increase with CRC progression (CRC III-IV > CRC 0-II > AA > Normal, Figure 6B). Further analysis of individual DMR methylation levels in the blood revealed discernible distinctions in certain DMRs during both AA and CRC 0-II stages, while variances in another subset of DMRs became detectable exclusively during CRC III-IV stages (Figure 6C). We then conducted a gene analysis of DMRs identifiable in early-stage CRC blood on the Metascope website. Our findings revealed their association with processes such as "secretion by cell" and "cell-cell adhesion." (Figure s6G) Consequently, we hypothesize that although alterations of 26 DMRs were observed in AA tissues, only a selected few are actively released into the bloodstream through cell secretion by tumor cells. The majority of these alterations are likely to be identified in the blood only after the occurrence of apoptosis or necrosis in tumor cells. These results elucidate the inconsistency in the performance of tissue and plasma diagnosis models and the differences in diagnostic efficacy for different CRC stages using the same model.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

Liquid biopsy of cfDNA has become an ideal clinical detection method because it is minimally invasive and easily sampled. However, due to the low concentration of ctDNA in bodily fluids and the heterogeneity of tumor cells, many DNA markers for CRC currently lack certainty whether they originate from CRC tissues. In this study, we integrated TCGA tissue data with self-collected tissue methylation data to identify CRC-related methylation sites. Probes were designed and applied in paired tissue and plasma samples to further screen the methylation sites. We successfully determined 27 DMRs originating from CRC tissues, which were subsequently used to construct diagnostic model. In the validation cohorts of tissue and plasma samples, the diagnosis models based on the 27 DMRs we established could effectively differentiate samples between CRC and normal participants, in addition, the plasma diagnosis model could distinguish the different stages of CRC. Furthermore, the transfer prognosis model, established based on diagnosis model scores, could effectively demonstrate whether the CRC had metastasized, separating CRC patients into high-risk and low-risk groups. Several blood-based methylation biomarker candidates have been proposed for early detection of CRC. For instance, the FDA-approved circulating methylated SEPT9 DNA (mSEPT9) demonstrates sensitivities of 11.2%, 35.0%, 63.0%, 46.0%, and 77.4% for AA and CRC stages I-IV, respectively, with a specificity of 91.5% [16]. In a recent study, a cfDNA methylation-based CRC screening model has a sensitivity of 86.4% and a specificity of 90.7%, utilizing 149 markers derived from blood samples [17]. In our study, 27 markers were selected through a layered screening process from tissue and plasma, undergoing marker selection, model development, and validation to ensure model robustness. Then the superior

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

performance of our plasma diagnosis model was validated, with a sensitivity of 83.3% and specificity of 92.9%. Notably, the sensitivity for AA and NAA reached 52.2% and 48.4% respectively, far surpassing the 11.2% sensitivity of mSEPT9 and the 33.3% reported in another blood screening study.[16, 17]. In summary, our diagnostic model performs well in detecting precancerous lesions of CRC and has the potential to become a screening method for high-risk populations of CRC.

In plasma, we observed the methylation scores of 27 DMRs gradually increase with CRC progression. It is possibly due to less vascular infiltration in early CRC [18, 19]. On the other hand, the ctDNA detected in the blood of early-stage CRC is largely derived from tumor cells actively secreting into the bloodstream, making it challenging to be precisely captured with current detection technologies due to the limited quantity. Conversely, ctDNA in the blood of late-stage CRC primarily emanated from apoptotic and necrotic tumor cells, resulting in a larger quantity that facilitates easier detection. However, in reality, the methylation scores of the 27 DMRs have indeed exhibited noticeable changes in AA and early cancerous tissues. Additionally, in the tissue diagnosis model using the 27 DMRs, the sensitivity was 94% and the specificity was 100%. Therefore, we have grounds to believe that the 27 DMRs detected in plasma originate from CRC tumor tissues. With the implementation of more sensitive detection methods, the performance of these 27 DMRs in early CRC plasma diagnosis is likely to be further enhanced.

Additionally, cfDNA contributes to risk stratification and early recurrence detection in CRC[20]. However, current methods rely on continuous blood cfDNA testing by patients [15, 21]. This study discovered that the preoperative cfDNA methylation level of 27 DMRs is linked to

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

distant metastasis and is valuable for predicting the prognosis of CRC, consistent with certain prior research. [22, 23]. Our model predicted AUCs of 0.955 and 0.867 for distant metastasis and prognosis in CRC, respectively. This underscores the potential utility of the preoperative application of this cfDNA methylation model for risk stratification, serving as an effective tool to improve the perioperative management of CRC patients.

The treatment strategies for CRC differ significantly across various stages. Molecular stratification approaches for CRC patients are on the rise, and concurrently, the clinical application of biomarkers to determine treatment decisions is gradually gaining traction [24]. Studies have demonstrated the precise identification of T1 CRC patients at risk of lymph node metastasis using a specific set of miRNAs, thereby potentially mitigating unnecessary overtreatment [25, 26]. Our CRC diagnosis model, with an AUC of 0.813 for distinguishing CRC 0-II from CRC III-IV, could also serve as a potent, straightforward, and cost-effective preoperative screening/detection method, guiding patients to select more appropriate treatment plans.

DNA methylation regulates gene transcription, guiding the progression from normal mucosa to AA and ultimately CRC. This process involves silencing tumor suppressor genes and activating oncogene transcription [27]. Gene transcription levels correlate notably with methylation levels at specific sites, consistent with our research findings. Additionally, several genes housing the identified 27 DMRs have been partially explored in CRC. Some genes contribute to tumor growth. For instance, high methylation and diminished expression of transmembrane protein 240 (TMEM240) regulate CRC cell proliferation, predicting poor prognosis [28]. The transcription factor homeobox A3 (HOXA3) activates aerobic glycolysis,

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

promoting tumor growth [29]. KIFC3 controls mitotic spindle assembly initiation [30]. Moreover, several genes are implicated in tumor metastasis, such as NOVA alternative splicing regulator 1 (NOVA1), which promotes CRC migration by activating the Notch pathway [31]. Protein tyrosine phosphatase receptor type T (PTPRT) contributes to early CRC dissemination [32]. SPARC-related modular calcium binding 2 (SMOC2) serves as the distinctive signature of cancer stem cells (CSCs) in CRC and promotes epithelial-to-mesenchymal transition (EMT) [33, 34]. Increased methylation and expression of pancreatic and duodenal homeobox 1 (PDX1) facilitate CRC invasion and migration[35]. However, the mechanisms by which alterations in specific DNA methylation impact gene expression are intricate. Certain transcription factors selectively recognize sequences with methylated CpG (mCpG) and influence the expression of multiple genes[36]. Further exploration of the potential functional mechanisms of these DNA methylation markers may deepen our understanding of the molecular processes underlying CRC development, offering promising therapeutic targets.

This study has certain limitations. Firstly, although obtained from multiple institutions, the small sample size was insufficient. Therefore, the cfDNA methylation model should be further validated in large-sample trials in the future. Secondly, a prospective study is necessary to compare or combine the cfDNA methylation model with clinically commonly used markers such as CEA and CA19-9.

In conclusion, our cfDNA methylation model based on 27 DMRs can identify different stages of CRC, predict metastasis and prognosis, and ultimately achieve early intervention and risk stratification for CRC patients. The preoperative application of our DNA methylation biomarker as a robust, convenient, and cost-effective detection method can contribute to

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

making more informed clinical decisions and improving the perioperative management of CRC patients.

AUTHOR CONTRIBUTIONS

Yuqi He and Jianqiu Sheng conceived and designed the study. Lang Yang, Fangli Men, Jianwei Yu, Xianzong Ma, Junfeng Xu, Yangjie Li, Ju Tian, Hui Xie, Qian Kang, Linghui Duan, Xiang Yi, Wei Guo, Ni Guo collected the samples and curated the date. Xueqing Gong, Lingqin Zhu, Lang Yang, Fangli Men, Jianwei Yu, Shuyang Sun, Chenguang Li, Xianzong Ma, Junfeng Xu, Yangjie Li, Xiang Yi, Wei Guo, Ni Guo, Youyong Lu, Joseph Leung, Yuqi He and Jianqiu Sheng analysed the data. Lingqin Zhu wrote the manuscript with the assistance of Youyong Lu, Joseph Leung, Yuqi He and Jianqiu Sheng.

ACKNOWLEDGEMENTS

This research was funded by grants from the National Natural Science Foundation of China (Grant no. 82273245), the Capital's Funds for Health Improvement and Research (Grant no. 2022-1-5082), and the Beijing Natural Science Foundation (Grant no. 7212107), Sponsored by Dongying City Natural Science Foundation (Grant no. 2023ZR026), Sponsored by Longyan City Science and Technology Plan Project (Grant no. 2022LYF17082).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All of the data supporting this work will be made available from the corresponding author upon reasonable request.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Seventh Medical Center of PLA General Hospital (Approval No. 2016-70, 2020-78), Dongying People's Hospital (Approval No. DYYW-2019-002-01), and the First Hospital of Longyan, Fujian Medical University (Approval No. 2021-k0001). Informed consents were obtained from all participants involved in the study.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49. Epub 2021/02/05. doi: 10.3322/caac.21660. PubMed PMID: 33538338.

2. Ghazal LV, Abrahamse P, Ward KC, Morris AM, Hawley ST, Veenstra CM. Financial Toxicity and Its Association With Health-Related Quality of Life Among Partners of Colorectal Cancer Survivors. JAMA network open. 2023;6(4):e235897. Epub 2023/04/07. doi: 10.1001/jamanetworkopen.2023.5897. PubMed PMID: 37022684; PubMed Central PMCID: PMCPMC10080378 National Cancer Institute (NCI) during the conduct of the study. Mr Abrahamse reported receiving grants from the National Institutes of Health during the conduct of the study and outside the submitted work. No other disclosures were reported.

3. Herriges MJ, Shenhav-Goldberg R, Peck JI, Bhanvadia SK, Morgans A, Chino F, et al. Financial Toxicity and Its Association With Prostate and Colon Cancer Screening. Journal of the National Comprehensive Cancer Network : JNCCN. 2022;20(9):981-8. Epub 2022/09/09. doi: 10.6004/jnccn.2022.7036. PubMed PMID: 36075394.

4. Olivier T, Prasad V. Molecular testing to deliver personalized chemotherapy recommendations: risking over and undertreatment. BMC medicine. 2022;20(1):392. Epub 2022/11/10. doi: 10.1186/s12916-022-02589-6. PubMed PMID: 36348413; PubMed Central PMCID: PMCPMC9644653.

5. Patel SG, Karlitz JJ, Yen T, Lieu CH, Boland CR. The rising tide of early-onset colorectal cancer: a comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. The lancet Gastroenterology & hepatology. 2022;7(3):262-74. Epub 2022/01/30. doi: 10.1016/s2468-1253(21)00426-x. PubMed PMID: 35090605.

6. He X, Hang D, Wu K, Nayor J, Drew DA, Giovannucci EL, et al. Long-term Risk of Colorectal Cancer After Removal of Conventional Adenomas and Serrated Polyps. Gastroenterology. 2020;158(4):852-61.e4. Epub 2019/07/16. doi: 10.1053/j.gastro.2019.06.039. PubMed PMID: 31302144; PubMed Central PMCID: PMCPMC6954345.

7. Hayman CV, Vyas D. Screening colonoscopy: The present and the future. World J Gastroenterol. 2021;27(3):233-9. Epub 2021/02/02. doi: 10.3748/wjg.v27.i3.233. PubMed PMID: 33519138; PubMed Central PMCID: PMCPMC7814366.

8. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget stool DNA testing for colorectal-cancer screening. The New England journal of medicine. 2014;370(14):1287-97. Epub 2014/03/22. doi: 10.1056/NEJMoa1311194. PubMed PMID: 24645800.

9. Young GP, Symonds EL, Allison JE, Cole SR, Fraser CG, Halloran SP, et al. Advances in Fecal Occult Blood Tests: the FIT revolution. Digestive diseases and sciences. 2015;60(3):609-22. Epub 2014/12/11. doi: 10.1007/s10620-014-3445-3. PubMed PMID: 25492500; PubMed Central PMCID: PMCPMC4366567.

10. Okugawa Y, Grady WM, Goel A. Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers. Gastroenterology. 2015;149(5):1204-25.e12. Epub 2015/07/29. doi: 10.1053/j.gastro.2015.07.011. PubMed PMID: 26216839; PubMed Central PMCID: PMCPMC4589488.

11. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nature reviews Cancer. 2011;11(6):426-37. Epub 2011/05/13. doi: 10.1038/nrc3066. PubMed PMID: 21562580.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

12. Zeng C, Stroup EK, Zhang Z, Chiu BC, Zhang W. Towards precision medicine: advances in 5-hydroxymethylcytosine cancer biomarker discovery in liquid biopsy. Cancer communications (London, England). 2019;39(1):12. Epub 2019/03/30. doi: 10.1186/s40880-019-0356-x. PubMed PMID: 30922396; PubMed Central PMCID: PMCPMC6440138.

13. Malla M, Loree JM, Kasi PM, Parikh AR. Using Circulating Tumor DNA in Colorectal Cancer: Current and Evolving Practices. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2022;40(24):2846-57. Epub 2022/07/16. doi: 10.1200/jco.21.02615. PubMed PMID: 35839443; PubMed Central PMCID: PMCPMC9390824.

14. Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Science translational medicine. 2020;12(524). Epub 2020/01/03. doi: 10.1126/scitranslmed.aax7533. PubMed PMID: 31894106.

15. Mo S, Ye L, Wang D, Han L, Zhou S, Wang H, et al. Early Detection of Molecular Residual Disease and Risk Stratification for Stage I to III Colorectal Cancer via Circulating Tumor DNA Methylation. JAMA oncology. 2023;9(6):770-8. Epub 2023/04/20. doi: 10.1001/jamaoncol.2023.0425. PubMed PMID: 37079312; PubMed Central PMCID: PMCPMC10119774 Genomics (Shanghai) Ltd. Dr H. Wang reported being an employee of Singlera Genomics (Shanghai) Ltd and having a patent pending for 202180000598.2. Dr Liu reported being an employee of Singlera Genomics (Shanghai) Ltd and having a patent pending for 202180000598.2. No other disclosures were reported.

16. Church TR, Wandell M, Lofton-Day C, Mongin SJ, Burger M, Payne SR, et al. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut. 2014;63(2):317-25. Epub 2013/02/15. doi: 10.1136/gutjnl-2012-304149. PubMed PMID: 23408352; PubMed Central PMCID: PMCPMC3913123.

17. Zhao F, Bai P, Xu J, Li Z, Muhammad S, Li D, et al. Efficacy of cell-free DNA methylation-based blood test for colorectal cancer screening in high-risk population: a prospective cohort study. Molecular cancer. 2023;22(1):157. Epub 2023/09/29. doi: 10.1186/s12943-023-01866-z. PubMed PMID: 37770864; PubMed Central PMCID: PMCPMC10538018.

18. Cho SS, Park JW, Kang GH, Kim JH, Bae JM, Han SW, et al. Prognostic Impact of Extramural Lymphatic, Vascular, and Perineural Invasion in Stage II Colon Cancer: A Comparison With Intramural Invasion. Diseases of the colon and rectum. 2023;66(3):366-73. Epub 2022/03/26. doi: 10.1097/dcr.0000000000002339. PubMed PMID: 35333785.

19. Hanrahan V, Currie MJ, Gunningham SP, Morrin HR, Scott PA, Robinson BA, et al. The angiogenic switch for vascular endothelial growth factor (VEGF)-A, VEGF-B, VEGF-C, and VEGF-D in the adenoma-carcinoma sequence during colorectal cancer progression. The Journal of pathology. 2003;200(2):183-94. Epub 2003/05/20. doi: 10.1002/path.1339. PubMed PMID: 12754739.

20. Zhou H, Zhu L, Song J, Wang G, Li P, Li W, et al. Liquid biopsy at the frontier of detection, prognosis and progression monitoring in colorectal cancer. Molecular cancer. 2022;21(1):86. Epub 2022/03/27. doi: 10.1186/s12943-022-01556-2. PubMed PMID: 35337361; PubMed Central PMCID: PMCPMC8951719.

21. Reinert T, Henriksen TV, Christensen E, Sharma S, Salari R, Sethi H, et al. Analysis of Plasma Cell-Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer. JAMA oncology. 2019;5(8):1124-31. Epub 2019/05/10. doi: 10.1001/jamaoncol.2019.0528. PubMed PMID: 31070691; PubMed Central PMCID: PMCPMC6512280 Balcioglu, Hafez, Goel, Rabinowitz, Billings, Swenerton,

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Aleshin, Lin, and Zimmermann, Messrs Sethi, Srinivasan, Olson, and Dashner, and Ms Navarro, reported receiving support from Natera Inc outside the submitted work. Dr Billings reported receiving support from Trovagene, OmniSeq, MissionBio, and Metastat outside the submitted work. Dr Zimmermann has a pending patent for Provisional. Dr Lindbjerg Andersen reported receiving grants from Novo Nordisk Foundation, Danish Council for Strategic Research, Danish Council for Independent Research, and Danish Cancer Society during the conduct of the study. No other disclosures were reported.

22. Tie J, Wang Y, Cohen J, Li L, Hong W, Christie M, et al. Circulating tumor DNA dynamics and recurrence risk in patients undergoing curative intent resection of colorectal cancer liver metastases: A prospective cohort study. PLoS medicine. 2021;18(5):e1003620. Epub 2021/05/04. doi: 10.1371/journal.pmed.1003620. PubMed PMID: 33939694; PubMed Central PMCID: PMCPMC8128260 following competing interests: BV, KWK, & NP are founders of, and hold equity in Thrive Earlier Detection and Personal Genome Diagnostics. KWK & NP are on the Board of Directors of, and consultants to, Thrive Earlier Detection. KWK & BV are consultants to Sysmex, Eisai, Personal Genome Diagnostics and CAGE Pharma and hold equity in CAGE Pharma. KWK, BV, and NP are consultants to and hold equity in NeoPhore. BV is a consultant to and holds equity in Catalio Capital Management. NP is an advisor to and holds equity in CAGE Pharma. The companies named above, as well as other companies, have licensed previously described technologies related to the work described in this paper from Johns Hopkins University. BV, KWK, and NP are inventors on some of these technologies. Licenses to these technologies are or will be associated with equity or royalty payments to the inventors as well as to Johns Hopkins University. CT and the University are entitled to royalty distributions related to technology licensed to Thrive Earlier Detection. CT is a consultant to Thrive. The terms of all these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies. All other authors declare no competing interests.

23. Kobayashi S, Nakamura Y, Taniguchi H, Odegaard JI, Nomura S, Kojima M, et al. Impact of Preoperative Circulating Tumor DNA Status on Survival Outcomes After Hepatectomy for Resectable Colorectal Liver Metastases. Annals of surgical oncology. 2021;28(8):4744-55. Epub 2021/01/05. doi: 10.1245/s10434-020-09449-8. PubMed PMID: 33393041.

24. Sveen A, Kopetz S, Lothe RA. Biomarker-guided therapy for colorectal cancer: strength in complexity. Nature reviews Clinical oncology. 2020;17(1):11-32. Epub 2019/07/11. doi: 10.1038/s41571-019-0241-1. PubMed PMID: 31289352; PubMed Central PMCID: PMCPMC7577509.

25. Miyazaki K, Wada Y, Okuno K, Murano T, Morine Y, Ikemoto T, et al. An exosome-based liquid biopsy signature for pre-operative identification of lymph node metastasis in patients with pathological high-risk T1 colorectal cancer. Molecular cancer. 2023;22(1):2. Epub 2023/01/08. doi: 10.1186/s12943-022-01685-8. PubMed PMID: 36609320; PubMed Central PMCID: PMCPMC9817247.

26. Wada Y, Shimada M, Murano T, Takamaru H, Morine Y, Ikemoto T, et al. A Liquid Biopsy Assay for Noninvasive Identification of Lymph Node Metastases in T1 Colorectal Cancer. Gastroenterology. 2021;161(1):151-62.e1. Epub 2021/04/06. doi: 10.1053/j.gastro.2021.03.062. PubMed PMID: 33819484; PubMed Central PMCID: PMCPMC10360659.

27. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. Trends in genetics : TIG. 2021;37(11):1012-27. Epub 2021/06/15. doi: 10.1016/j.tig.2021.05.002. PubMed PMID: 34120771.

28. Chang SC, Liew PL, Ansar M, Lin SY, Wang SC, Hung CS, et al. Hypermethylation and decreased

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

expression of TMEM240 are potential early-onset biomarkers for colorectal cancer detection, poor prognosis, and early recurrence prediction. Clinical epigenetics. 2020;12(1):67. Epub 2020/05/14. doi: 10.1186/s13148-020-00855-z. PubMed PMID: 32398064; PubMed Central PMCID: PMCPMC7218647.

29. Yang R, Zhang G, Dong Z, Wang S, Li Y, Lian F, et al. Homeobox A3 and KDM6A cooperate in transcriptional control of aerobic glycolysis and glioblastoma progression. Neuro-oncology. 2023;25(4):635-47. Epub 2022/10/11. doi: 10.1093/neuonc/noac231. PubMed PMID: 36215227; PubMed Central PMCID: PMCPMC10076951.

30. Hata S, Pastor Peidro A, Panic M, Liu P, Atorino E, Funaya C, et al. The balance between KIFC3 and EG5 tetrameric kinesins controls the onset of mitotic spindle assembly. Nat Cell Biol. 2019;21(9):1138-51. Epub 2019/09/05. doi: 10.1038/s41556-019-0382-6. PubMed PMID: 31481795.

31. Zhang T, Chen S, Peng Y, Wang C, Cheng X, Zhao R, et al. NOVA1-Mediated SORBS2 Isoform Promotes Colorectal Cancer Migration by Activating the Notch Pathway. Frontiers in cell and developmental biology. 2021;9:673873. Epub 2021/10/26. doi: 10.3389/fcell.2021.673873. PubMed PMID: 34692669; PubMed Central PMCID: PMCPMC8531477.

32. Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Scott Shaffer J, et al. Quantitative evidence for early metastatic seeding in colorectal cancer. Nature genetics. 2019;51(7):1113-22. Epub 2019/06/19. doi: 10.1038/s41588-019-0423-x. PubMed PMID: 31209394; PubMed Central PMCID: PMCPMC6982526.

33. Wang H, Gong P, Chen T, Gao S, Wu Z, Wang X, et al. Colorectal Cancer Stem Cell States Uncovered by Simultaneous Single-Cell Analysis of Transcriptome and Telomeres. Advanced science (Weinheim, Baden-Wurttemberg, Germany). 2021;8(8):2004320. Epub 2021/04/27. doi: 10.1002/advs.202004320. PubMed PMID: 33898197; PubMed Central PMCID: PMCPMC8061397.

34. Feng D, Gao P, Henley N, Dubuissez M, Chen N, Laurin LP, et al. SMOC2 promotes an epithelial-mesenchymal transition and a pro-metastatic phenotype in epithelial cells of renal cell carcinoma origin. Cell death & disease. 2022;13(7):639. Epub 2022/07/23. doi: 10.1038/s41419-022-05059-2. PubMed PMID: 35869056; PubMed Central PMCID: PMCPMC9307531.

35. Lee Y, Dho SH, Lee J, Hwang JH, Kim M, Choi WY, et al. Hypermethylation of PDX1, EN2, and MSX1 predicts the prognosis of colorectal cancer. Experimental & molecular medicine. 2022;54(2):156-68. Epub 2022/02/17. doi: 10.1038/s12276-022-00731-1. PubMed PMID: 35169223; PubMed Central PMCID: PMCPMC8894425 interests.

36. Wan J, Su Y, Song Q, Tung B, Oyinlade O, Liu S, et al. Methylated cis-regulatory elements mediate KLF4-dependent gene transactivation and cell migration. eLife. 2017;6. Epub 2017/05/30. doi: 10.7554/eLife.20068. PubMed PMID: 28553926; PubMed Central PMCID: PMCPMC5466421.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure Legends

Figure 1. Flow diagram of the study.

Figure 2. Development and validation of tissue diagnosis model based on the 27 CRC-specific DMRs. (A) The use of ROC curve analysis to assess the performance of the tissue diagnostic model in differentiating CRC patients from normal individuals of the training cohort (51 tissues we collected). (B-C) The use of ROC curve analysis to assess the performance of the tissue diagnostic model in differentiating CRC patients from normal individuals(B) and distinguish AA from Normal subjects (C) in tissue validation cohort 1(GSE 48684 dataset). (D-E) The sensitivity and specificity of the diagnosis model in the tissue training(D) and validation cohort 1(E).

Figure 3. Development and validation of a plasma diagnosis model based on the 27 CRC-specific DMRs. (A-B) Heatmap illustrating the DMRs between CRC and advanced adenoma and healthy controls in the training (A) and validation cohort (B). (C) The use of ROC curve analysis to assess the performance of the diagnosis model in differentiating CRC patients from normal individuals of the training cohort. (D) The use of ROC curve analysis to evaluate the performance of the diagnosis model in differentiating CRC patients from normal individuals of the validation cohort. (E-G) In the validation cohort, the performance of the diagnosis model in differentiating early-stage CRC (0-II) from normal individuals (E); in differentiating advanced-stage CRC (III-IV) from normal individuals (F); in differentiating AA from normal individuals (G). (H) Methylation scores of Normal、 AA and CRC

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

individuals generated from the diagnosis model in the training and validation cohort (The dashed line represents the cutoff value). (I) The sensitivity and specificity of the diagnosis model stratified by stages in the training and validation cohort. Abbreviations: DMRs, differential methylated regions; ROC, receiver operating characteristic; AA, advanced adenoma; CRC, colorectal cancer.

Figure 4. External validation of plasma diagnosis model. (A-C) In the external validation cohort, the performance of the diagnosis model in differentiating early-stage CRC (0-Ⅱ) from normal individuals (A); in differentiating AA from normal individuals (B); in differentiating NAA from normal individuals (C). (D) Methylation scores of Normal、NAA、AA and CRC (0-Ⅱ) individuals generated from the diagnosis model in the external validation cohort (The dashed line represents the cutoff value). (E) The sensitivity and specificity of the diagnostic model for NAA, AA, and CRC in the external validation cohort.

Figure 5. Development and validation of a plasma-based metastasis model and a plasma-based prognosis model. (A-B) The methylation scores generated by the diagnosis model for CRC patients across various M stages (A) and different disease stages (B). (C) ROC curve analysis was performed to assess the performance of 27 DMRs methylation scores generated by the diagnosis model in distinguishing M1-CRC patients from M0-CRC patients in the training (AUC = 0.969) and validation (AUC = 0.955) cohorts. (D) ROC analysis evaluates the performance of 27 DMRs methylation scores generated by the diagnosis model in predicting the prognosis of CRC patients. (E) Kaplan-Meier survival curves comparing OS

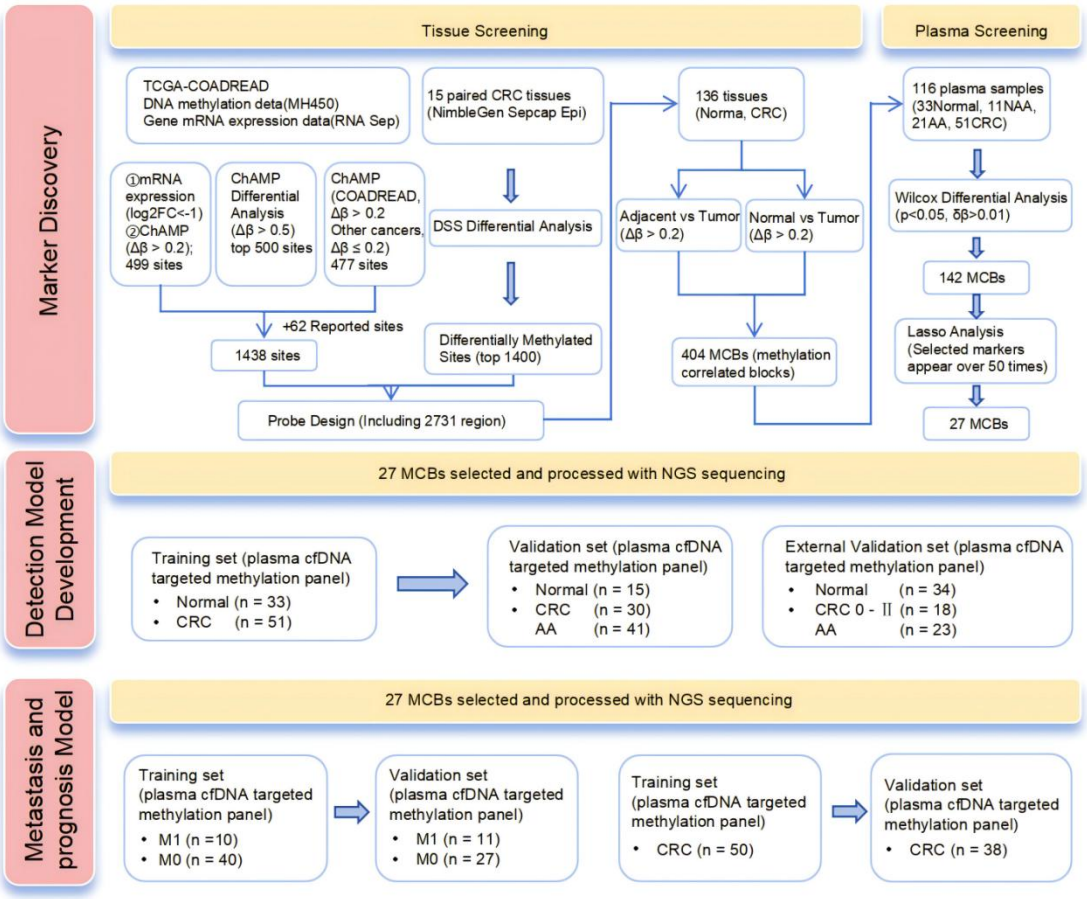
medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

between the high-risk CRC group and low-risk CRC group of prognosis model in the training cohort and validation cohort. (F, H) Metastasis and prognosis prediction of the training and validation patitients(n = 88). The black dotted line at the cutoff value divides the patients into high-risk and low-risk groups. Yellow circles and gray circles represent patients without distant metastasis and with distant metastasis, respectively (F). Yellow circles and gray circles represent patients survived and deceased, respectively (H). (G, I) The proportion of metastasis and deceased is higher in the high-risk group. The p-value was calculated using a two-sided Fisher's exact test. Abbreviations: DMRs, differential methylated regions; ROC, receiver operating characteristic; M1-CRC, colorectal cancer with distant metastasis; M0-CRC, colorectal cancer without distant metastasis; AUC, the area under the curve.

Figure 6. Methylation scores of paired tissue and plasma samples. (A) The methylation scores of tissues from Normal, AA, CRC 0-II, CRC III-IV individuals. (B) The methylation scores of blood samples from Normal, AA, CRC 0-II, CRC III-IV individuals. (C) The mechanism diagram interpreting the methylation scores changes of tissue and plasma across different populations.

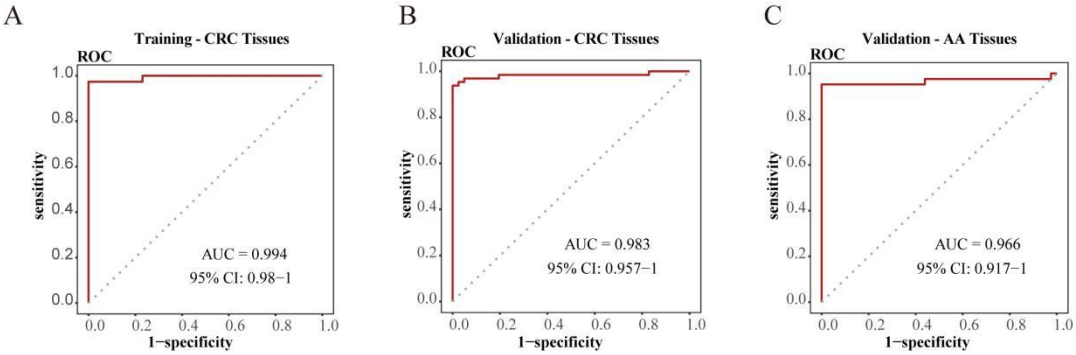
medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 1



medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2



D

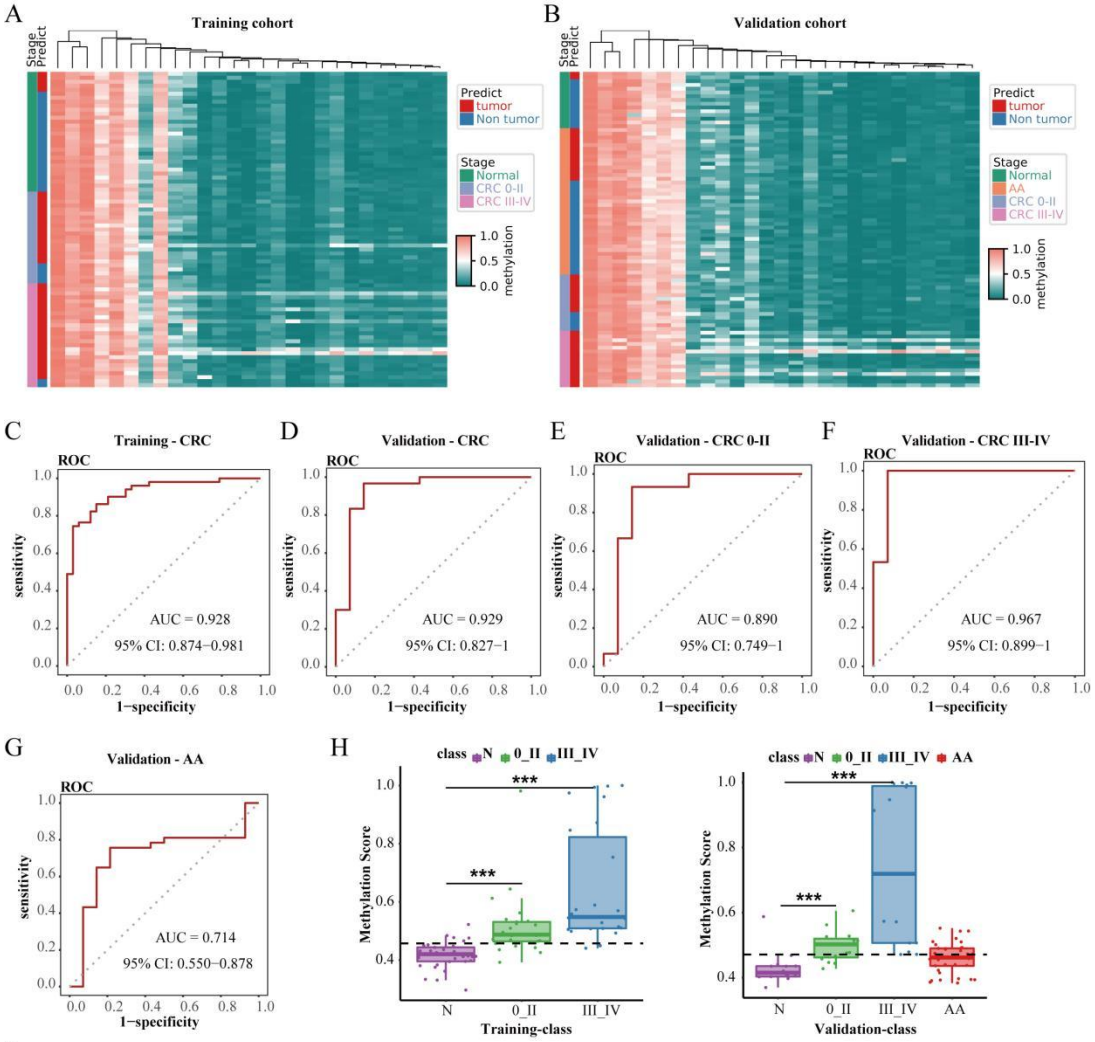
Training cohort		Predicted			
	Total	Negative	Positive	Sensitivity	Specificity
Normal	13	13	0		1
Tumor	38	1	37	0.97	

E

Validation cohort (GSE48684)		Predicted			
	Total	Negative	Positive	Sensitivity	Specificity
Normal	41	41	0		1.00
AA	42	2	40	0.95	
Tumor	64	4	60	0.94	

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3



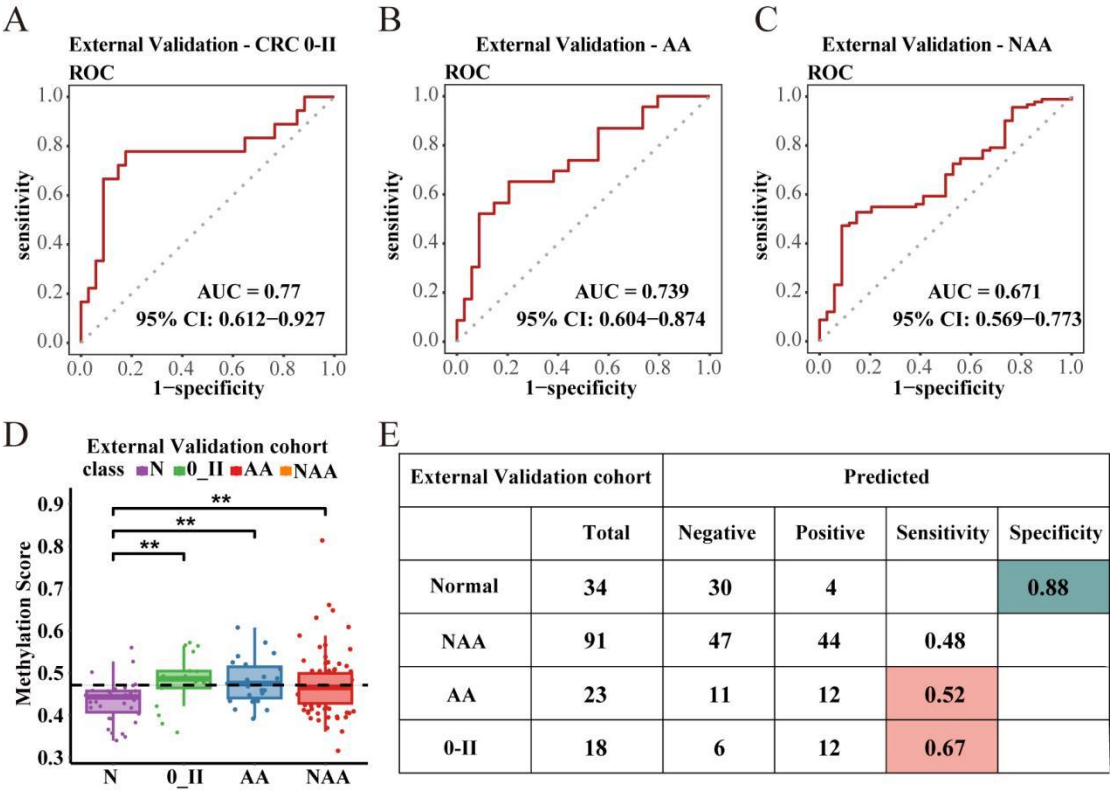
I

Training cohort		Predicted			
	Total	Negative	Positive	Sensitivity	Specificity
Normal	33	28	5		0.85
0-II	25	5	20	0.80	
III-IV	26	2	24	0.92	
Tumor	51	7	44	0.86	

Validation cohort		Predicted			
	Total	Negative	Positive	Sensitivity	Specificity
Normal	14	13	1		0.93
AA	37	21	16	0.43	
0-II	15	5	10	0.67	
III-IV	15	0	15	1	
Tumor	30	5	25	0.83	

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 4

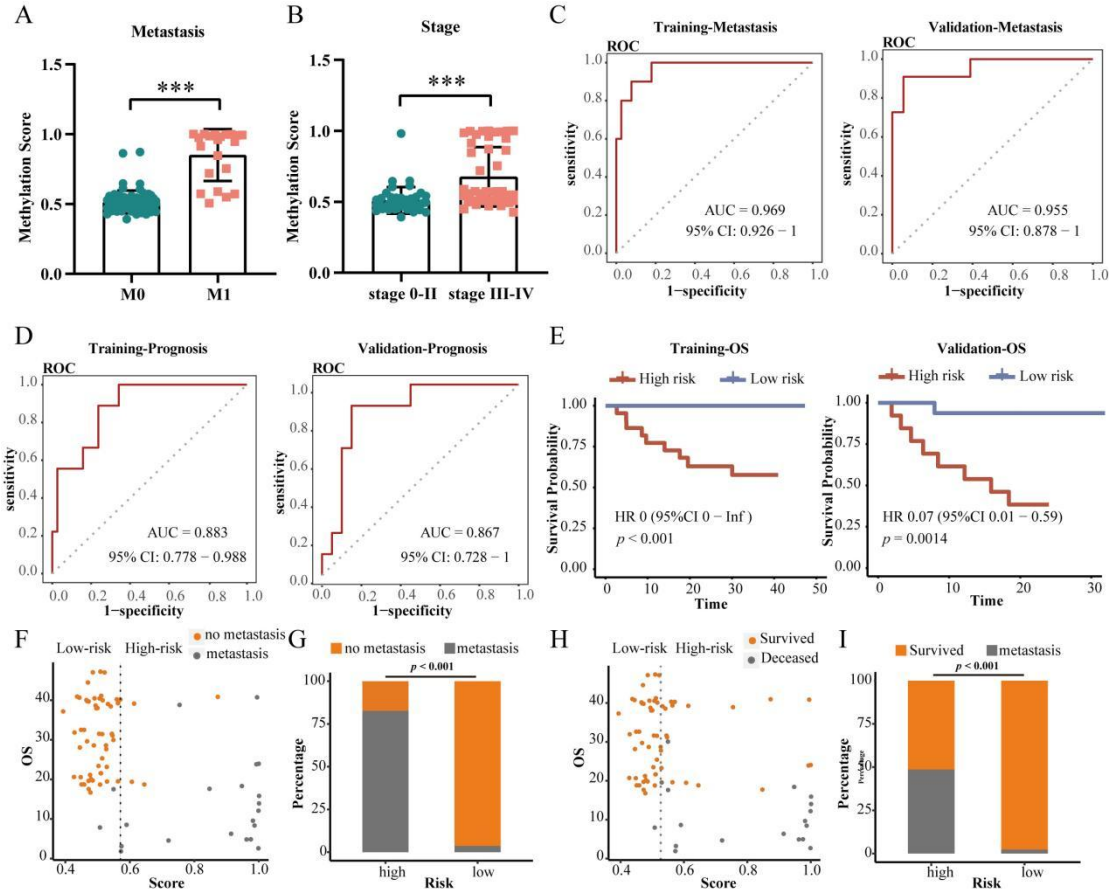


E

External Validation cohort		Predicted			
	Total	Negative	Positive	Sensitivity	Specificity
Normal	34	30	4		0.88
NAA	91	47	44	0.48	
AA	23	11	12	0.52	
0-II	18	6	12	0.67	

medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 5



medRxiv preprint doi: https://doi.org/10.1101/2024.11.28.24317652; this version posted December 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

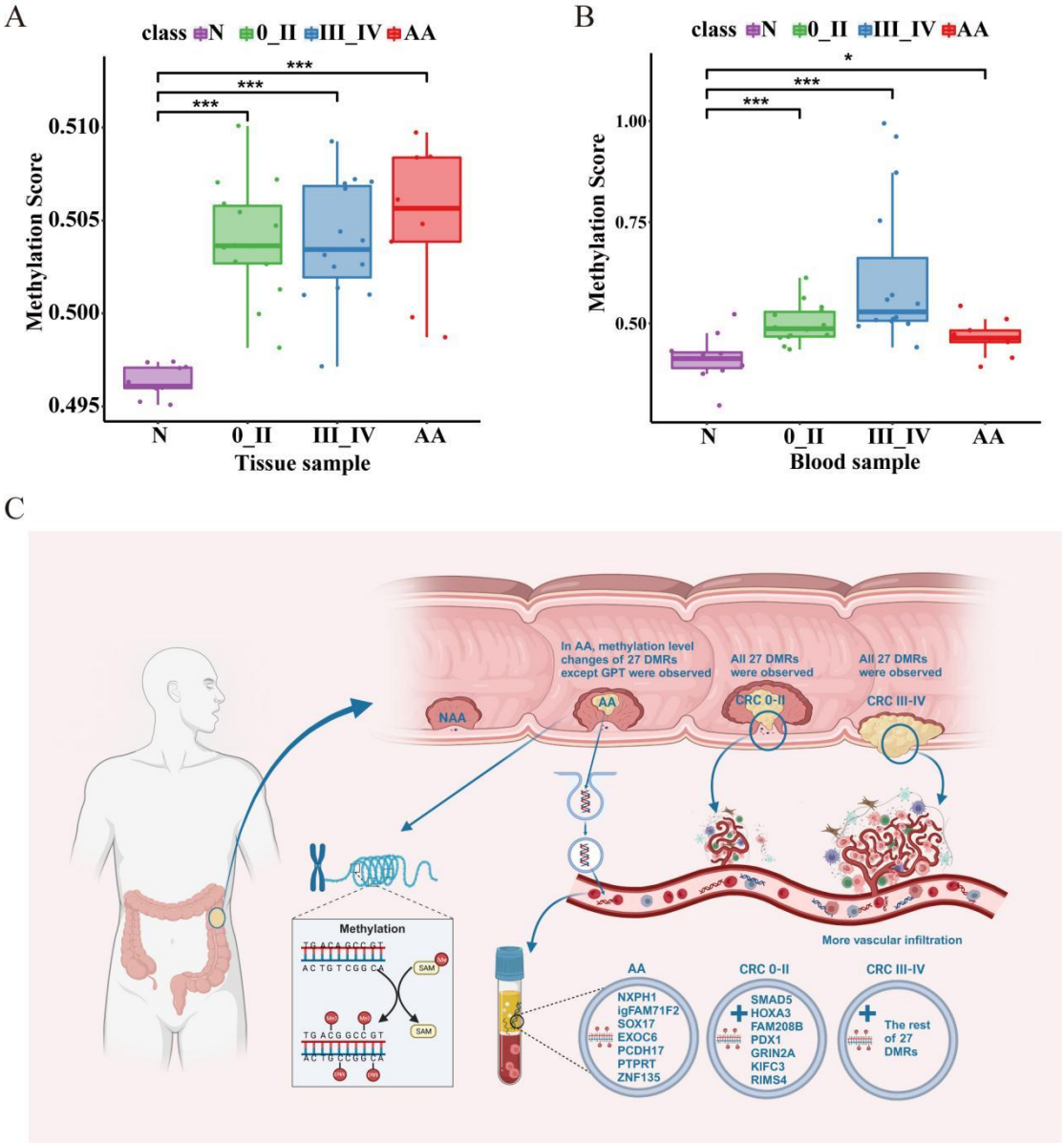



Figure 6