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

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;

Abstract

Cell-free DNA (cfDNA) methylation is a valuable biomarker in various cancers including colorectal cancer (CRC), but marker for both early diagnosis and prognostic prediction remain a critical unmet need. Here, we report the development of a 27-DMR (differentially methylated regions) plasma panel with dual diagnostic and prognostic functions. We first identified CRC-specific methylation features from 119 tissue samples and 161 plasma samples. Machine learning algorithms were then applied to develop diagnosis and prognosis models using the plasma samples in training cohort. In the tissue external validation cohort (GSE48684), the cfDNA methylation diagnosis model conducted with the panel, achieved an area under the curve (AUC) of 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%. Beyond its diagnostic capabilities, the methylation panel was also a powerful predictor of distant metastasis (AUC = 0.955) and patient 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 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

Introduction

Colorectal cancer (CRC) is the third most prevalent malignancy and the fourth leading cause of cancer-related mortality worldwide, accounting for 9.4% of all cancer deaths [1]. When patients were diagnosed with CRC, over half of their family (62.9%) faced financial burdens [2, 3]. The survival rates of individuals diagnosed with advanced CRC (Stage III or IV) witness a significant decrement [4, 5], underscoring early detection and removal of precancerous lesions remain the most effective strategy to prevent CRC-associated mortality [6]. Although colonoscopy is the gold standard for CRC detection, its invasiveness, suboptimal patient compliance and risk of complications including intestinal perforation limit its widespread effectiveness [7]. Other non-invasive tests, such as 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 by insufficient sensitivity, particularly for early-stage lesions [8, 9].

Recent research demonstrated a profound correlation between CRC and the development of genetic and epigenetic alterations. During early stage of tumorigenesis, 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), a component of cell-free DNA (cfDNA) released from apoptotic and necrotic tumor cells, carries cancer-specific epigenetic alterations [11]. Consequently, cfDNA methylation emerges as a promising biomarker for cancer screening due to its early onset and high signal density [12].

While the utility of cfDNA methylation in early diagnosis, surveillance, molecular subtyping

and prognostic prediction of CRC has been proven, current approaches have limitations [13]. For instance, a diagnostic panel has shown excellent performance but was developed using inconsistent sample types (e.g., CRC tissue vs. normal blood leukocyte methylation data), introducing potential bias [14]. Another study addressed the potential of cfDNA methylation in patient risk stratification, but required blood sampling every three months, which can hamper patient compliance and clinical feasibility [15].

In this study, we aimed to develop and validate a plasma-based biomarker panel, initially identified from CRC tissues, for the comprehensive diagnosis, metastatic assessment, and prognostic prediction of CRC. We performed a meticulous, multi-stage screening process starting with CRC and normal tissues, followed by refinement using plasma samples. Crucially, by utilizing a substantial cohort of paired tissue and plasma samples from the same patients, we minimized potential confounding biases. The resulting 27 differential methylated regions (DMRs) panel was designed to, from a single blood test, ascertain the presence of CRC, distinguish lesion stages, identify distant metastasis, and stratify patients by risk. Such a multi-functional tool holds significant promise for guiding the clinical management of CRC, from early detection to personalized treatment strategies.

Materials and Methods

Study design and samples

This study was designed to first identify CRC-specific DNA methylation markers and then to develop and validate models for CRC diagnosis, metastasis, and prognosis. The marker discovery phase utilized data from TCGA public database and clinically collected tissue and plasma samples. 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 screening cohort was obtained from fifteen pairs of tissue samples sequenced with the Roche Nimble Gen Seq Cap Epi method from the Seventh Medical Center of PLA General Hospital. The 89 tissue samples and 77 plasma samples used for marker screening were collected between June to 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 December 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 July 2022 to December 2022 at the First Hospital of Longyan, Fujian Medical University. An additional public tissue validation cohort (GSE48684) was obtained from the Gene Expression Omnibus (GEO). Detailed patient inclusion and exclusion criteria, as well as methods for DNA extraction, library construction, and targeted bisulfite sequencing, are provided in the Supplementary Materials.

CRC-specific methylation markers selection

An initial marker pool was generated by integrating TCGA methylation data with our 15 pairs

of tissue methylation sequencing results. This informed the design of a targeted panel, which was then screened in the 89 samples (13 normal tissues and 38 paired CRC tissues), identifying 404 DMRs. Then, the candidates were further evaluated in 77 plasma samples from 13 Normal, 15 non-advanced adenoma (NAA), 12 advanced adenoma (AA) and 37 CRC participants. 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 correlation between the methylation levels of the 27 DMRs and the expression of their associated genes was assessed using Spearman's rank correlation.

Construction of the CRC diagnosis model

A diagnostic model was developed using a logistic regression algorithm trained on methylation data from the training cohort. The model was built using five-fold cross-validation, optimizing for the minimum binary deviance. Its performance was evaluated by Receiver Operating Characteristic (ROC) curve analysis and the optimal cutoff value for the resulting methylation score was determined using the maximum Youden index. 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

A metastasis prediction model was constructed using the methylation score from the diagnostic model combined with patients' clinical metastasis status. Similarly, a prognostic model was developed by integrating the methylation score with patient survival data. The performance of both models was evaluated using ROC analysis, with optimal cutoffs

determined by the maximum Youden index.

Statistical analysis

All statistical analyses were performed 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 by ROC analysis using the "roc" function from the R package pROC, generating the area under the ROC curve (AUC) with 95% confidence interval (CI). Survival analyses were conducted using Kaplan-Meier (KM) curves with the log-rank test and hazard ratios were calculated using Cox proportional-hazards models.

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-s4.

Methylation markers selection

We employed a multi-stage funneling strategy to identify a robust panel of CRC-specific methylation markers, starting from a broad analysis of public data and progressively refining the list through validation in our own tissue and plasma samples. Analyzing the 450k chip methylation data from TCGA, we identified the top 500 differentially methylated CpG sites (DMCs) between CRC and control tissues ($\Delta\beta > 0.50$, FDR < 0.05), 477 hypermethylated sites ($\Delta\beta > 0.20$, FDR < 0.05) located in the TSS or First Exon regions of significantly downregulated genes (log2FC < -1, FDR < 0.05), and 499 sites that were hypermethylated in CRC ($\Delta\beta > 0.20$) but not in five other major cancers (esophageal, gastric, lung, liver, and breast cancer; $\Delta\beta \leq 0.20$). The union of these three approaches, supplemented with known

CRC-related methylation sites from existing literature, yielded an initial list of 1,438 candidate DMCs (Supplementary Figure 1A). Unsupervised clustering of these sites revealed distinct methylation patterns between normal and tumor tissues in the TCGA cohort (Figure s1B). To validate these findings in our patient population, we performed methylation sequencing (SeqCap Epi) on 15 paired tissue samples (5 advanced adenoma pairs and 10 CRC pairs). From this analysis, we selected the top 1,400 DMCs based on their statistical significance ($\Delta\beta > 0.20$, FDR < 0.05) (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 across 89 tissues (13 normal tissues and 38 paired CRC 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). To identify markers detectable non-invasively, we assessed the 404 DMRs in 77 plasma samples (13 Normal, 15 NAA, 12 AA, 37 CRC) and narrowed the list to 142 DMRs that showed significant differential methylation in plasma (Wilcoxon test; $\Delta\beta > 0.01$, FDR < 0.05) (Figure s3E-F). Finally, to select the most robust and minimal signature for diagnostic modeling, we applied a Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis to the training cohort methylation data. The analysis involved 500 random iterations, and we selected markers that appeared in more than 50 iterations and yielded a final panel of 27 top-performing DMRs for plasma-based CRC detection. Detailed information on these 27 DMRs is provided in Table s5. Correlation

analysis using TCGA data confirmed that the methylation levels of the majority of these 27 DMRs were significantly correlated with the expression of their associated genes, underscoring their biological relevance (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 3C). In the training cohort, the ROC curve analysis showed an AUC value of 0.928 [95% confidence interval (CI): 0.874–0.981] (Figure 3D). 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 3E-H). The sensitivities of the CRC diagnosis model in the training cohort for stages 0-II, III-IV, and all CRC stages were 80%, 92%, and 86%, respectively, with a specificity of 85% (Figure 3I). 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). Compared to the traditional tumor marker CEA, the model demonstrated higher sensitivity in detecting early-stage colorectal cancer and adenomas (Table s7). 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 CRC metastasis and prognosis models

To investigate the clinical relevance of our methylation signature, we first correlated the plasma methylation scores with clinicopathological parameters. The scores were significantly associated with tumor stage and metastasis ($P < 0.001$, Figure 5A-B) but not with age, sex, or lesion location ($P > 0.05$, Figure s6A-C). Based on these associations, we developed models to predict metastasis and prognosis using the 27-DMR panel. The metastasis model demonstrated excellent performance, achieving an AUC of 0.969 (95% CI: 0.926-1) in the training cohort and 0.955 (95% CI: 0.878-1) in the validation cohort (Figure 5C). Correspondingly, patients with metastatic disease (M1) had significantly higher methylation scores than those without (M0) and exhibited shorter overall survival (OS) (Figure s6D-F). The prognosis model also showed high accuracy, with AUCs of 0.883 (95% CI: 0.778-0.988) in the training set and 0.867 (95% CI = 0.728-1) in the validation set (Figure 5D). Using an

optimal cutoff, the prognosis model effectively stratified patients into high- and low-risk groups. Kaplan-Meier analysis revealed that patients in the high-risk group had significantly worse OS (Figure 5E). Critically, this high-risk group encompassed nearly all patients who developed distant metastases or died during follow-up (Figure 5F-I). To confirm the score's independent prognostic value, we performed a multivariate Cox regression analysis, which showed that the 27-DMR methylation score was an independent predictor of OS (Table S6). These findings establish that our models can successfully stratify patients, identifying those who may require more aggressive treatment.

Distinct Release Dynamics of Methylation Markers from Tissue to Plasma

To understand the biological basis for the model's performance characteristics, we compared methylation scores between paired tissue and plasma samples across the spectrum of disease progression. We found that in tissue, the 27-DMR signature was already strongly present at the precancerous adenoma (AA) stage (Figure 6A). However, in plasma, the signal intensity increased in a stepwise manner, becoming most prominent in advanced-stage CRC (Figure 6B). This discrepancy suggests that the release of cfDNA markers into circulation is stage-dependent. Indeed, a subset of DMRs was detectable in plasma even in early-stage disease, and these markers were associated with genes involved in cellular secretion pathways (Figure 6C, s6G). In contrast, other DMRs were only detected in the plasma of patients with advanced CRC. This suggests a dual mechanism of release: limited, active secretion of certain markers from early lesions, followed by massive, passive release from widespread cell death in larger, necrotic tumors. This dynamic explains why the diagnostic sensitivity of the plasma test is highest for advanced-stage disease.

Discussion

Liquid biopsy of cfDNA has become an ideal clinical detection method because it is minimally invasive and easily sampled. However, a key challenge in liquid biopsy is confirming that cfDNA markers truly originate from tumor tissue. We addressed this by implementing a rigorous, multi-stage screening strategy that began with TCGA data, was validated in our own CRC tissue samples, and was finally optimized in plasma. This ensures that our 27-DMR signature is not only statistically powerful but also biologically rooted in CRC pathobiology. In this study, we implemented a rigorous, multi-stage screening strategy that began with integrating TCGA tissue data with self-collected tissue methylation data to identify CRC-related methylation sites, was validated in our own CRC tissue samples, and was finally optimized in plasma. This ensures that our 27-DMR signature is not only statistically powerful but also biologically rooted in CRC pathobiology. Our resulting diagnostic model, built upon these tissue-of-origin-confirmed markers, demonstrates significant potential for early diagnosis, staging, metastasis and prognostic assessment.

A critical benchmark for any new CRC screening tool is its performance against existing methods, especially in detecting precancerous lesions. For instance, the FDA-approved circulating methylated SEPT9 DNA (mSEPT9) offers a specificity of 91.5% but provides limited sensitivities of 11.2%, 35.0%, 63.0%, 46.0%, and 77.4% for AA and CRC stages I-IV, respectively [(16)]. While more recent cfDNA methylation models have improved overall CRC detection, their ability to detect precancerous lesions can still be modest; one such study reported a sensitivity of 33.3% for adenomas [(17)]. In this context, our compact 27-marker plasma model demonstrates a highly competitive performance, achieving an overall

sensitivity of 83.3% and a specificity of 92.9% for CRC. Most notably, our model shows a substantial improvement in detecting high-risk precancerous lesions, with sensitivities for advanced adenomas (AA) and non-advanced adenomas (NAA) reaching 52.2% and 48.4%, respectively. This represents a nearly five-fold increase in sensitivity for advanced adenomas compared to the mSEPT9 test [16, 17]. This enhanced capability positions our model as a promising new tool for CRC screening in high-risk populations, enabling earlier intervention before the onset of malignancy.

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.

Beyond early detection, cfDNA analysis is a valuable tool for risk stratification and monitoring for recurrence in CRC [20]. However, many current approaches require serial

postoperative testing to track disease progression [15, 21]. A key advantage of our model is its ability to provide significant prognostic information from a single, preoperative blood sample. Our findings reveal that the 27-DMR methylation score progressively increases with CRC stage advancement, a feature with direct clinical utility [22, 23]. Firstly, this allows for effective preoperative staging, as the model can distinguish early-stage (0–II) from late-stage (III–IV) CRC with high accuracy (AUC = 0.813), offering a simple, non-invasive method to help guide initial treatment planning. Secondly, and perhaps more significantly, the preoperative score is a powerful predictor of patient prognosis. It was strongly associated with distant metastasis (AUC = 0.955), enabling the stratification of patients into high- and low-risk groups before surgery even begins. This capability is particularly valuable as it provides crucial prognostic insights that can inform perioperative management—such as identifying patients who may require more aggressive treatment or intensive surveillance—from a single baseline measurement, potentially reducing the reliance on continuous postoperative monitoring.

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]. The clinical performance of our panel is underpinned by the biological functions of its associated genes, many of which are known to be involved in CRC pathogenesis. 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, 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]. The correlation we observed between the methylation levels of our 27 DMRs and the transcription levels of their respective genes in TCGA datasets further supports their functional relevance. 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.

Despite the promising results, we acknowledge several limitations in this study. Firstly, although obtained from multiple institutions, the overall sample size is relatively small. Therefore, the cfDNA methylation model should be further validated in larger, prospective clinical trials including a larger and more balanced set of precancerous lesions. Secondly, a prospective study is necessary to directly compare or combine the cfDNA methylation model with clinically established protein markers such as CEA and CA19-9, and Fecal Immunochemical Test (FIT) to determine its ultimate clinical utility.

In conclusion, our study introduces a novel, tissue-validated 27-DMR cfDNA methylation signature that effectively serves three critical clinical needs: early detection of CRC, including high-risk precancerous lesions; preoperative staging; and prognostic risk stratification for metastasis. As a robust, convenient, and cost-effective detection method, the preoperative application of this biomarker panel has the potential to facilitate more informed clinical decision-making and significantly improve the management of patients with colorectal cancer.

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

The data has been deposited in the National Genomics Data Center (NGDC) and is accessible under the accession number OMIX009128. All of the data supporting this work will be made available from the corresponding author upon reasonable request.

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.

407 REFERENCES

408 1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer
409 Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185
410 Countries. CA Cancer J Clin. 2021;71(3):209-49. Epub 2021/02/05. doi: 10.3322/caac.21660. PubMed
411 PMID: 33538338.
412 2. Ghazal LV, Abrahamse P, Ward KC, Morris AM, Hawley ST, Veenstra CM. Financial Toxicity and Its
413 Association With Health-Related Quality of Life Among Partners of Colorectal Cancer Survivors. JAMA
414 network open. 2023;6(4):e235897. Epub 2023/04/07. doi: 10.1001/jamanetworkopen.2023.5897.
415 PubMed PMID: 37022684; PubMed Central PMCID: PMCPMC10080378 National Cancer Institute (NCI)
416 during the conduct of the study. Mr Abrahamse reported receiving grants from the National Institutes
417 of Health during the conduct of the study and outside the submitted work. No other disclosures were
418 reported.
419 3. Herriges MJ, Shenhav-Goldberg R, Peck JI, Bhanvadia SK, Morgans A, Chino F, et al. Financial
420 Toxicity and Its Association With Prostate and Colon Cancer Screening. Journal of the National
421 Comprehensive Cancer Network : JNCCN. 2022;20(9):981-8. Epub 2022/09/09. doi:
422 10.6004/jnccn.2022.7036. PubMed PMID: 36075394.
423 4. Olivier T, Prasad V. Molecular testing to deliver personalized chemotherapy recommendations:
424 risking over and undertreatment. BMC medicine. 2022;20(1):392. Epub 2022/11/10. doi:
425 10.1186/s12916-022-02589-6. PubMed PMID: 36348413; PubMed Central PMCID: PMCPMC9644653.
426 5. Patel SG, Karlitz JJ, Yen T, Lieu CH, Boland CR. The rising tide of early-onset colorectal cancer: a
427 comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early
428 detection. The lancet Gastroenterology & hepatology. 2022;7(3):262-74. Epub 2022/01/30. doi:
429 10.1016/s2468-1253(21)00426-x. PubMed PMID: 35090605.
430 6. He X, Hang D, Wu K, Nayor J, Drew DA, Giovannucci EL, et al. Long-term Risk of Colorectal Cancer
431 After Removal of Conventional Adenomas and Serrated Polyps. Gastroenterology. 2020;158(4):852-
432 61.e4. Epub 2019/07/16. doi: 10.1053/j.gastro.2019.06.039. PubMed PMID: 31302144; PubMed
433 Central PMCID: PMCPMC6954345.
434 7. Hayman CV, Vyas D. Screening colonoscopy: The present and the future. World J Gastroenterol.
435 2021;27(3):233-9. Epub 2021/02/02. doi: 10.3748/wjg.v27.i3.233. PubMed PMID: 33519138; PubMed
436 Central PMCID: PMCPMC7814366.
437 8. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget stool
438 DNA testing for colorectal-cancer screening. The New England journal of medicine.
439 2014;370(14):1287-97. Epub 2014/03/22. doi: 10.1056/NEJMoa1311194. PubMed PMID: 24645800.
440 9. Young GP, Symonds EL, Allison JE, Cole SR, Fraser CG, Halloran SP, et al. Advances in Fecal Occult
441 Blood Tests: the FIT revolution. Digestive diseases and sciences. 2015;60(3):609-22. Epub 2014/12/11.
442 doi: 10.1007/s10620-014-3445-3. PubMed PMID: 25492500; PubMed Central PMCID:
443 PMCPMC4366567.
444 10. Okugawa Y, Grady WM, Goel A. Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers.
445 Gastroenterology. 2015;149(5):1204-25.e12. Epub 2015/07/29. doi: 10.1053/j.gastro.2015.07.011.
446 PubMed PMID: 26216839; PubMed Central PMCID: PMCPMC4589488.
447 11. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients.
448 Nature reviews Cancer. 2011;11(6):426-37. Epub 2011/05/13. doi: 10.1038/nrc3066. PubMed PMID:
449 21562580.

450 12. Zeng C, Stroup EK, Zhang Z, Chiu BC, Zhang W. Towards precision medicine: advances in 5-
451 hydroxymethylcytosine cancer biomarker discovery in liquid biopsy. Cancer communications (London,
452 England). 2019;39(1):12. Epub 2019/03/30. doi: 10.1186/s40880-019-0356-x. PubMed PMID:
453 30922396; PubMed Central PMCID: PMCPMC6440138.
454 13. Malla M, Loree JM, Kasi PM, Parikh AR. Using Circulating Tumor DNA in Colorectal Cancer:
455 Current and Evolving Practices. Journal of clinical oncology : official journal of the American Society of
456 Clinical Oncology. 2022;40(24):2846-57. Epub 2022/07/16. doi: 10.1200/jco.21.02615. PubMed PMID:
457 35839443; PubMed Central PMCID: PMCPMC9390824.
458 14. Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, et al. Circulating tumor DNA methylation profiles
459 enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Science translational
460 medicine. 2020;12(524). Epub 2020/01/03. doi: 10.1126/scitranslmed.aax7533. PubMed PMID:
461 31894106.
462 15. Mo S, Ye L, Wang D, Han L, Zhou S, Wang H, et al. Early Detection of Molecular Residual Disease
463 and Risk Stratification for Stage I to III Colorectal Cancer via Circulating Tumor DNA Methylation. JAMA
464 oncology. 2023;9(6):770-8. Epub 2023/04/20. doi: 10.1001/jamaoncol.2023.0425. PubMed PMID:
465 37079312; PubMed Central PMCID: PMCPMC10119774 Genomics (Shanghai) Ltd. Dr H. Wang
466 reported being an employee of Singlera Genomics (Shanghai) Ltd and having a patent pending for
467 202180000598.2. Dr Liu reported being an employee of Singlera Genomics (Shanghai) Ltd and having
468 a patent pending for 202180000598.2. No other disclosures were reported.
469 16. Church TR, Wandell M, Lofton-Day C, Mongin SJ, Burger M, Payne SR, et al. Prospective
470 evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut.
471 2014;63(2):317-25. Epub 2013/02/15. doi: 10.1136/gutjnl-2012-304149. PubMed PMID: 23408352;
472 PubMed Central PMCID: PMCPMC3913123.
473 17. Zhao F, Bai P, Xu J, Li Z, Muhammad S, Li D, et al. Efficacy of cell-free DNA methylation-based
474 blood test for colorectal cancer screening in high-risk population: a prospective cohort study.
475 Molecular cancer. 2023;22(1):157. Epub 2023/09/29. doi: 10.1186/s12943-023-01866-z. PubMed
476 PMID: 37770864; PubMed Central PMCID: PMCPMC10538018.
477 18. Cho SS, Park JW, Kang GH, Kim JH, Bae JM, Han SW, et al. Prognostic Impact of Extramural
478 Lymphatic, Vascular, and Perineural Invasion in Stage II Colon Cancer: A Comparison With Intramural
479 Invasion. Diseases of the colon and rectum. 2023;66(3):366-73. Epub 2022/03/26. doi:
480 10.1097/dcr.0000000000002339. PubMed PMID: 35333785.
481 19. Hanrahan V, Currie MJ, Gunningham SP, Morrin HR, Scott PA, Robinson BA, et al. The angiogenic
482 switch for vascular endothelial growth factor (VEGF)-A, VEGF-B, VEGF-C, and VEGF-D in the adenoma-
483 carcinoma sequence during colorectal cancer progression. The Journal of pathology. 2003;200(2):183-
484 94. Epub 2003/05/20. doi: 10.1002/path.1339. PubMed PMID: 12754739.
485 20. Zhou H, Zhu L, Song J, Wang G, Li P, Li W, et al. Liquid biopsy at the frontier of detection,
486 prognosis and progression monitoring in colorectal cancer. Molecular cancer. 2022;21(1):86. Epub
487 2022/03/27. doi: 10.1186/s12943-022-01556-2. PubMed PMID: 35337361; PubMed Central PMCID:
488 PMCPMC8951719.
489 21. Reinert T, Henriksen TV, Christensen E, Sharma S, Salari R, Sethi H, et al. Analysis of Plasma Cell-
490 Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer. JAMA oncology.
491 2019;5(8):1124-31. Epub 2019/05/10. doi: 10.1001/jamaoncol.2019.0528. PubMed PMID: 31070691;
492 PubMed Central PMCID: PMCPMC6512280 Balcioglu, Hafez, Goel, Rabinowitz, Billings, Swenerton,

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

536 expression of TMEM240 are potential early-onset biomarkers for colorectal cancer detection, poor
537 prognosis, and early recurrence prediction. Clinical epigenetics. 2020;12(1):67. Epub 2020/05/14. doi:
538 10.1186/s13148-020-00855-z. PubMed PMID: 32398064; PubMed Central PMCID: PMCPMC7218647.
539 29. Yang R, Zhang G, Dong Z, Wang S, Li Y, Lian F, et al. Homeobox A3 and KDM6A cooperate in
540 transcriptional control of aerobic glycolysis and glioblastoma progression. Neuro-oncology.
541 2023;25(4):635-47. Epub 2022/10/11. doi: 10.1093/neuonc/noac231. PubMed PMID: 36215227;
542 PubMed Central PMCID: PMCPMC10076951.
543 30. Hata S, Pastor Peidro A, Panic M, Liu P, Atorino E, Funaya C, et al. The balance between KIFC3 and
544 EG5 tetrameric kinesins controls the onset of mitotic spindle assembly. Nat Cell Biol. 2019;21(9):1138-
545 51. Epub 2019/09/05. doi: 10.1038/s41556-019-0382-6. PubMed PMID: 31481795.
546 31. Zhang T, Chen S, Peng Y, Wang C, Cheng X, Zhao R, et al. NOVA1-Mediated SORBS2 Isoform
547 Promotes Colorectal Cancer Migration by Activating the Notch Pathway. Frontiers in cell and
548 developmental biology. 2021;9:673873. Epub 2021/10/26. doi: 10.3389/fcell.2021.673873. PubMed
549 PMID: 34692669; PubMed Central PMCID: PMCPMC8531477.
550 32. Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Scott Shaffer J, et al. Quantitative evidence for early
551 metastatic seeding in colorectal cancer. Nature genetics. 2019;51(7):1113-22. Epub 2019/06/19. doi:
552 10.1038/s41588-019-0423-x. PubMed PMID: 31209394; PubMed Central PMCID: PMCPMC6982526.
553 33. Wang H, Gong P, Chen T, Gao S, Wu Z, Wang X, et al. Colorectal Cancer Stem Cell States
554 Uncovered by Simultaneous Single-Cell Analysis of Transcriptome and Telomeres. Advanced science
555 (Weinheim, Baden-Wurttemberg, Germany). 2021;8(8):2004320. Epub 2021/04/27. doi:
556 10.1002/advs.202004320. PubMed PMID: 33898197; PubMed Central PMCID: PMCPMC8061397.
557 34. Feng D, Gao P, Henley N, Dubuissez M, Chen N, Laurin LP, et al. SMOC2 promotes an epithelial-
558 mesenchymal transition and a pro-metastatic phenotype in epithelial cells of renal cell carcinoma
559 origin. Cell death & disease. 2022;13(7):639. Epub 2022/07/23. doi: 10.1038/s41419-022-05059-2.
560 PubMed PMID: 35869056; PubMed Central PMCID: PMCPMC9307531.
561 35. Lee Y, Dho SH, Lee J, Hwang JH, Kim M, Choi WY, et al. Hypermethylation of PDX1, EN2, and MSX1
562 predicts the prognosis of colorectal cancer. Experimental & molecular medicine. 2022;54(2):156-68.
563 Epub 2022/02/17. doi: 10.1038/s12276-022-00731-1. PubMed PMID: 35169223; PubMed Central
564 PMCID: PMCPMC8894425 interests.
565 36. Wan J, Su Y, Song Q, Tung B, Oyinlade O, Liu S, et al. Methylated cis-regulatory elements mediate
566 KLF4-dependent gene transactivation and cell migration. eLife. 2017;6. Epub 2017/05/30. doi:
567 10.7554/eLife.20068. PubMed PMID: 28553926; PubMed Central PMCID: PMCPMC5466421.
568

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) Methylation scores of Normal、AA and CRC individuals generated from the diagnosis model in the training and validation cohort (The dashed line represents the cutoff value). (D)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. (E) 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. (F-H) In the validation cohort, the performance of the diagnosis model in differentiating early-stage CRC (0-II) from normal individuals (F); in

differentiating advanced-stage CRC (III-IV) from normal individuals (G); in differentiating AA from normal individuals (H). (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.

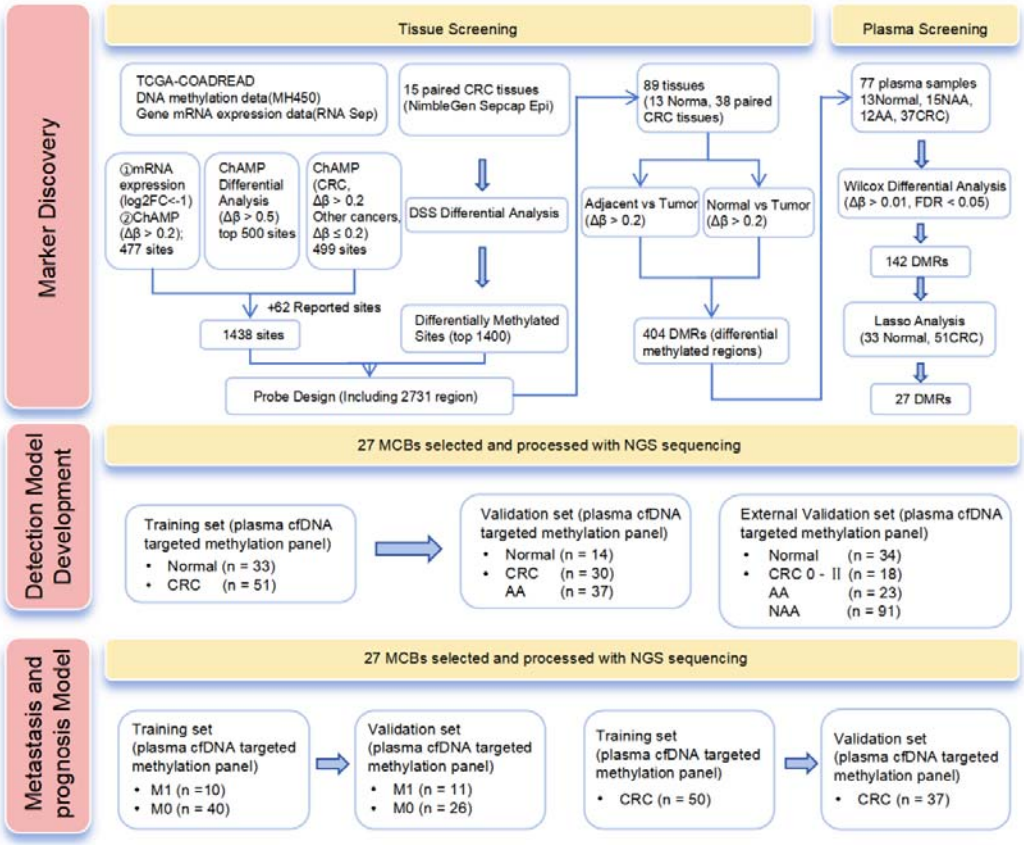
Figure 4. External validation of plasma diagnosis model. (A-C) In the external validation cohort, the performance of the diagnosis model in differentiating early-stage CRC (0-II) from normal individuals (A); in differentiating AA from normal individuals (B); in differentiating NAA from normal individuals (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

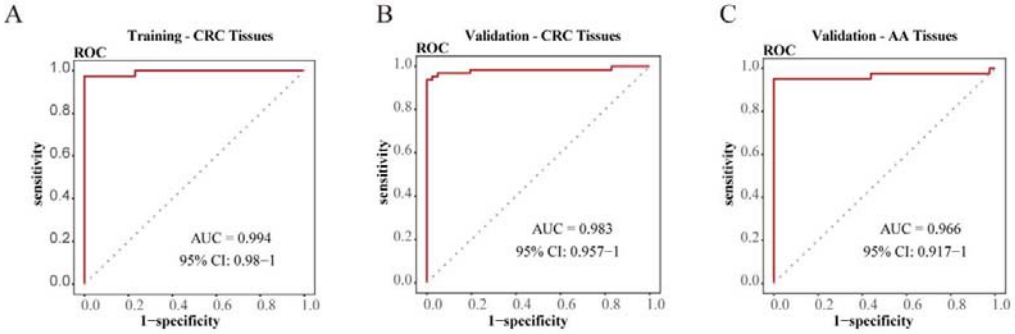
between the high-risk CRC group and low-risk CRC group of prognosis model in the training cohort and validation cohort. (F, H) Metastasis and prognosis prediction of the training and validation patitients (n = 88). The black dotted line at the cutoff value divides the patients into high-risk and low-risk 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.

Figure 1


Marker Discovery
Tissue Screening
Plasma Screening
TCGA-COADREAD
DNA methylation deta(MH450)
Gene mRNA expression data(RNA Seq)
15 paired CRC tissues
(NimbleGen Sepcap Epi)
89 tissues
(13 Norma, 38 paired CRC tissues)
77 plasma samples
13Normal, 15NAA, 12AA, 37CRC)
①mRNA expression (log2FC<-1) ②ChAMP (Δβ > 0.2); 477 sites
ChAMP Differential Analysis (Δβ > 0.5) top 500 sites
ChAMP (CRC, Δβ > 0.2 Other cancers, Δβ ≤ 0.2) 499 sites
DSS Differential Analysis
Adjacent vs Tumor (Δβ > 0.2)
Normal vs Tumor (Δβ > 0.2)
Wilcox Differential Analysis (Δβ > 0.01, FDR < 0.05)
142 DMRs
+62 Reported sites
1438 sites
Differentially Methylated Sites (top 1400)
404 DMRs (differential methylated regions)
Lasso Analysis (33 Normal, 51CRC)
27 DMRs
Probe Design (Including 2731 region)
Detection Model Development
27 MCBs selected and processed with NGS sequencing
Training set (plasma cfDNA targeted methylation panel)
• Normal (n = 33)
• CRC (n = 51)
Validation set (plasma cfDNA targeted methylation panel)
• Normal (n = 14)
• CRC (n = 30)
AA (n = 37)
External Validation set (plasma cfDNA targeted methylation panel)
• Normal (n = 34)
• CRC 0 - II (n = 18)
AA (n = 23)
NAA (n = 91)
Metastasis and prognosis Model
27 MCBs selected and processed with NGS sequencing
Training set (plasma cfDNA targeted methylation panel)
• M1 (n =10)
• M0 (n = 40)
Validation set (plasma cfDNA targeted methylation panel)
• M1 (n = 11)
• M0 (n = 26)
Training set (plasma cfDNA targeted methylation panel)
• CRC (n = 50)
Validation set (plasma cfDNA targeted methylation panel)
• CRC (n = 37)

634 Figure 2



D

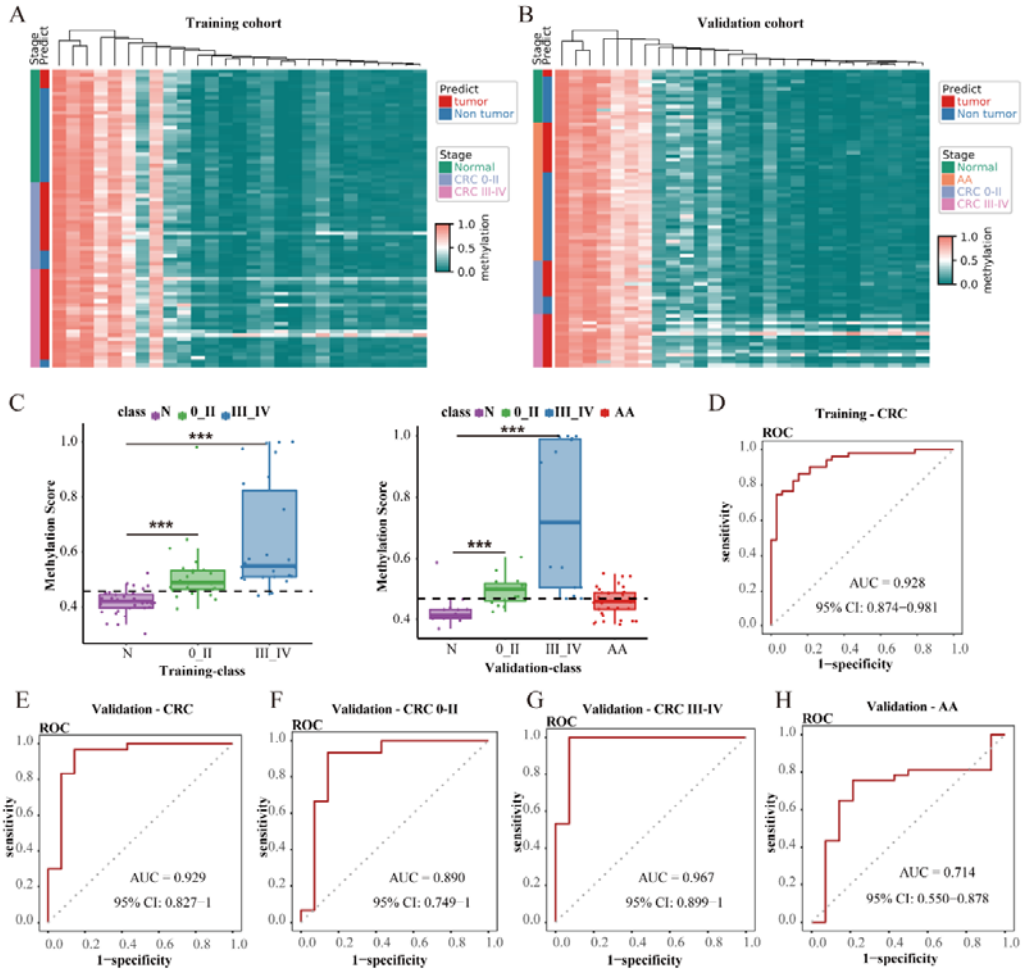
Training cohort		Predicted			
	Total	Negative	Positive	Sensitivity	Specificity
Normal	13	13	0		1
Tumor	38	1	37	0.97	

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	

635
636

Figure 3

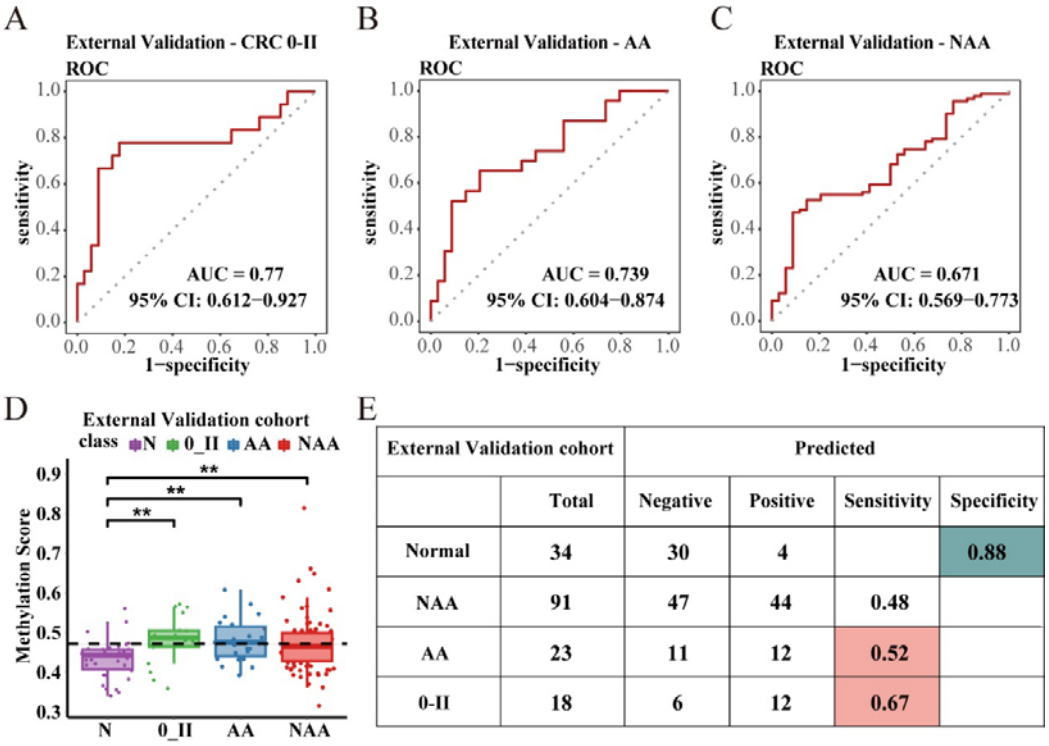


I

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	

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	

Figure 5

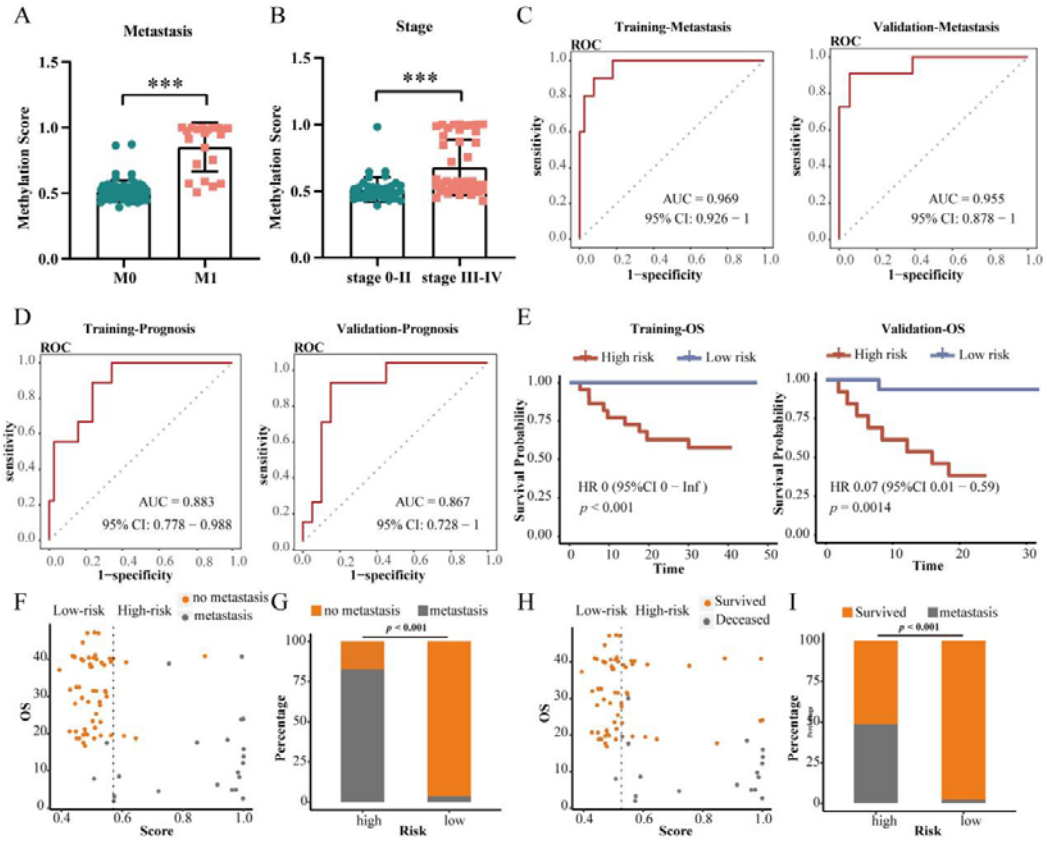
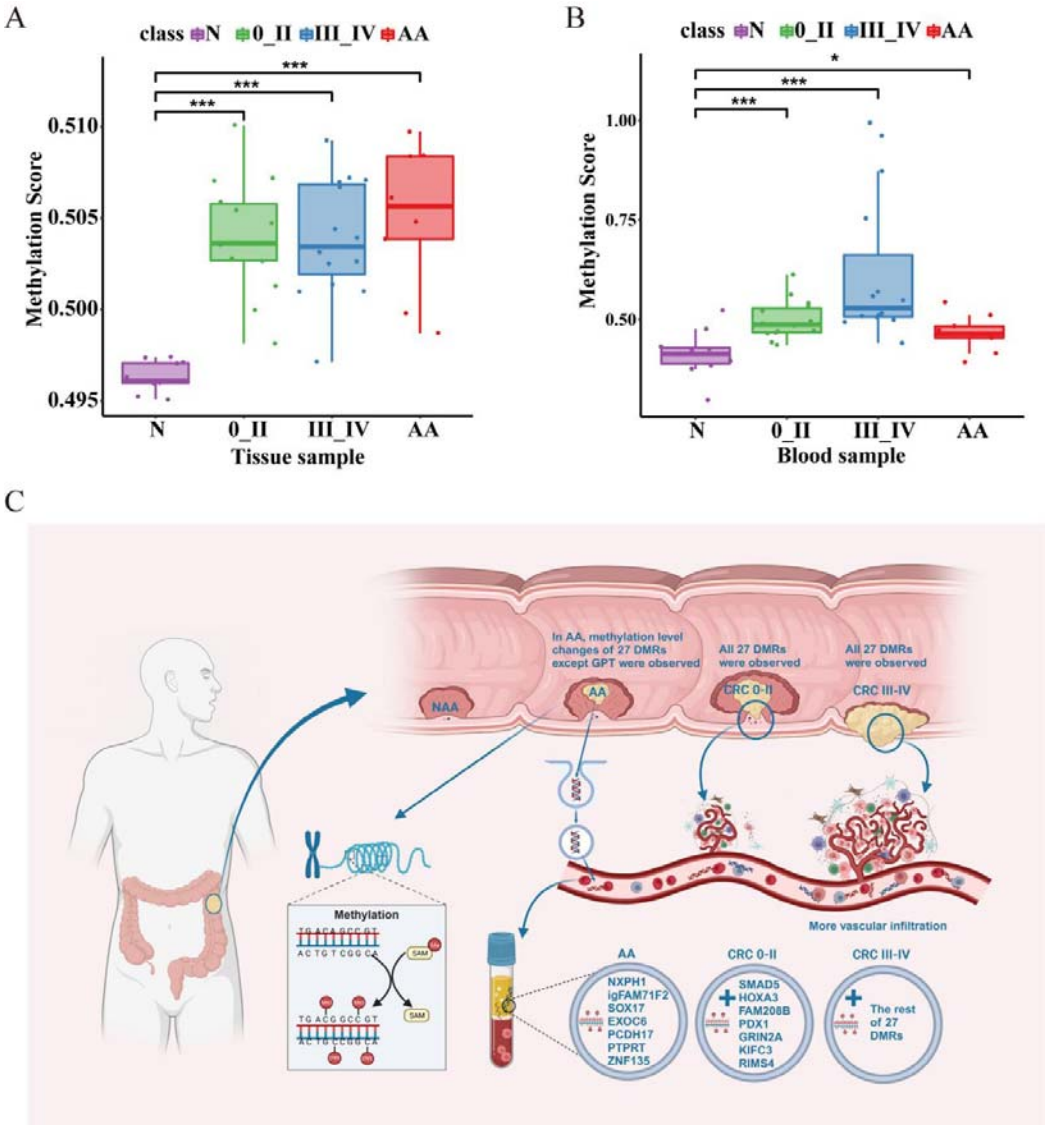

A
Metastasis
Methylation Score
***
M0
M1
B
Stage
Methylation Score
***
stage 0-II
stage III-IV
C
Training-Metastasis
ROC
sensitivity
AUC = 0.969
95% CI: 0.926 − 1
1−specificity
Validation-Metastasis
ROC
sensitivity
AUC = 0.955
95% CI: 0.878 − 1
1−specificity
D
Training-Prognosis
ROC
sensitivity
AUC = 0.883
95% CI: 0.778 − 0.988
1−specificity
Validation-Prognosis
ROC
sensitivity
AUC = 0.867
95% CI: 0.728 − 1
1−specificity
E
Training-OS
High risk
Low risk
Survival Probability
HR 0 (95%CI 0 − Inf)
p < 0.001
Time
Validation-OS
High risk
Low risk
Survival Probability
HR 0.07 (95%CI 0.01 − 0.59)
p = 0.0014
Time
F
Low-risk
High-risk
no metastasis
metastasis
OS
Score
G
no metastasis
metastasis
p < 0.001
Percentage
high
low
Risk
H
Low-risk
High-risk
Survived
Deceased
OS
Score
I
Survived
metastasis
p < 0.001
Percentage
high
low
Risk

Figure 6


A
class N 0_II III_IV AA
Methylation Score
0.510
0.505
0.500
0.495
***
***
***
N
0_II
III_IV
AA
Tissue sample
B
class N 0_II III_IV AA
Methylation Score
1.00
0.75
0.50
*
***
***
N
0_II
III_IV
AA
Blood sample
C
In AA, methylation level changes of 27 DMRs except GPT were observed
All 27 DMRs were observed
All 27 DMRs were observed
NAA
AA
CRC 0-II
CRC III-IV
Methylation
SAM
More vascular infiltration
AA
NXPH1
igFAM71F2
SOX17
EXOC6
PCDH17
PTPRT
ZNF135
CRC 0-II
SMAD5
HOXA3
FAM208B
PDX1
GRIN2A
KIFC3
RIMS4
CRC III-IV
The rest of 27 DMRs

Online Supplementary Material

Contents

Supplementary Methods

Supplementary Figure s1-s6

Supplementary Figure 1(Figure s1). Screening for CRC-specific DMCs based on TCGA data.
Supplementary Figure 2(Figure s2). DMCs selection in tissues.
Supplementary Figure 3(Figure s3). Selection of methylation markers.
Supplementary Figure 4(Figure s4). The relationship between the methylation levels of the 27 DMRs and the transcription levels of their respective genes in TCGA datasets.
Supplementary Figure 5(Figure s5). Development and validation of a plasma diagnosis model for discerning the specific staging of CRC patients.
Supplementary Figure 6(Figure s6). The construction of the metastasis model.

Supplementary Table s1-s6

Table s1 Explanation of specific phrase and abbreviation
Table s2 Patient characteristics of the 15 paired tissues
Table s3 Patient characteristics of the 64 participants (Tissues)
Table s4 Patient characteristics of the blood cohort
Table s5. Annotation information for 27 regions
Table s6 Univariate and multivariate analysis of methylation scores and clinical parameters.
Table s7 Comparison of diagnostic performance between the methylation score and CEA for early-stage colorectal cancer and adenomas in the external validation cohort.

Supplementary Methods

The inclusion and exclusion criteria

Inclusion criteria included individuals aged over 18, with healthy control subjects having negative results from colonoscopy or pathology, and no history of other cancers. Patients with colorectal cancer and precancerous lesion were required to have a confirmed diagnosis with colonoscopy and/or pathology, and without concurrent malignancies involving other systems or organs. Blood was collected from participants before any form of treatment (surgery, radiotherapy, or chemotherapy).

CRC-specific methylation markers selection

Using the following methodology, tissue-specific methylation markers were identified for CRC. Initially, ChAMP differential analysis software was used to analyze CRC methylation data (450k chip methylation data) from TCGA, filtering out differentially methylated sites between CRC and control ($\Delta\beta>0.50$, FDR<0.05). The top 500 sites, contingent on the extent of differential methylation, were earmarked as candidate marker sites. Merging TCGA CRC methylation data with transcriptome data, 477 hypermethylated sites ($\Delta\beta>0.20$, FDR<0.05) in the TSS and First Exon regions of downregulated genes (log2FC<−1 and FDR<0.05) were selected. DNA methylation data encompassing CRC, esophageal cancer, gastric cancer, lung cancer, liver cancer, and breast cancer were retrieved from the TCGA database. Using the ChAMP differential analysis software, 499 CRC-specific methylation sites were identified (CRC, $\Delta\beta>0.20$, FDR<0.05; for other cancers, $\Delta\beta\leq0.20$). The amalgamation of the outcomes from the aforementioned three stages, supplemented by the inclusion of previously reported CRC-related methylation sites from the literature, yielded a list of 1438 potential differential methylated CpG sites (DMCs). Subsequently, DNA methylation of 15 pairs of tissue samples from Chinese CRC patients (comprising 5 pairs of advanced adenoma tissues and 10 pairs of CRC tissues) were sequenced by Roche Nimble Gen's SeqCap Epi method, which encompasses approximately 80Mb of Roche methylation probes, covering 5.5 million methylation sites, including all sites from the Illumina 450K chip. The top 1400 DMCs were selected based on adjusted *P* value ($\Delta\beta > 0.20$, FDR < 0.05). These 1400 DMCs were then amalgamated with the previously identified 1438 sites, and custom-designed probes were synthesized. By testing the 89 samples (13 normal tissues and 38 paired CRC tissues) with the special probes, the DMCs between tumors and adjacent tissues, as well as between cancer and normal tissues were identified ($\Delta\beta > 0.20$, FDR < 0.05). Through the intersection of these identified sites and the amalgamation of those in close proximity, 404 CRC-specific DMRs were ultimately identified.

Subsequently, the methylation levels of 77 plasma samples (13 Normal, 15 NAA, 12 AA, 37 CRC) were assessed using the Wilcox Differential Analysis ($\Delta\beta > 0.01$, FDR < 0.05). To further reduce methylation markers, 500 random iterations were performed, extracting 75% of the samples from the training set each time. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis was employed to filter markers. Ultimately, 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.

DNA extraction from tissues and plasma

DNA extraction from AA, CRC, and healthy colon tissues was carried out using the TIANamp Genomic DNA Kit (DP304, TIANGEN, Beijing, China). Before DNA extraction, tissue pathology evaluation confirmed the presence of tumor cells in CRC tissue samples. We collected 10 milliliters of whole blood from each participant. First, they were centrifuged at 1400×g for 25 minutes at 4□ temperature. Then, the upper layer was transferred and

centrifuged again at 14000 ×g for 10 minutes at 4□ temperature to obtain plasma. The MagMAX™ Cell-Free DNA Isolation Kit (A29319, Thermo Fisher, Waltham, MA, USA) was utilized for cfDNA extraction from plasma, with special attention to avoiding repeated freeze-thaw cycles to prevent cfDNA degradation. The concentration of the isolated cfDNA was analyzed by fluorometric quantification using the Qubit Fluorometer 4.0 (Thermo Fisher, Waltham, MA, USA) and the Qubit dsDNA HS Assay Kit (Thermo Fisher). The fragment size profiles of cfDNA were assessed using the 2100 bioanalyzer (Agilent Inc., Santa Clara, CA, USA) with the Agilent High-Sensitivity DNA Kit. CfDNA with a yield < 5 ng or exhibiting quality issues was excluded.

Library construction

Up to 1000 ng of gDNA was fragmented with Covaris M220 and used to construct an indexed library using KAPA HyperPrep Kit, PCR free. Then follow the steps of the EZ DNA Methylation Lightning Kit (Zymo Research; Irvine, CA) for the bisulfite conversion of the DNA library. Up to 5 ng of plasma cfDNA was subjected to bisulfite conversion using the EZ DNA Methylation Lightning Kit (Zymo Research; Irvine, CA), converted cfDNA was used to prepare indexed sequencing library using Accel-NGS Methyl-Seq DNA library preparation kits (Swift BioSciences; Ann Arbor, MI).

Targeted bisulfite sequencing using the SeqCap Epi Enrichment System

Libraries were quantified using the Qubit dsDNA HS Assay Kit before pooling 3 (cfDNA) or 4 (gDNA) individual libraries per capture of SeqCap Epi CpGiant probes from Roche Nimblegen. Hybridization and capture were performed using the HyperCap Target Enrichment Kit from Roche. Pooled libraries were sequenced on an Illumina Hiseq X Ten using paired-end 150-bp reads.

Targeted bisulfite sequencing using the Twist Target Enrichment System

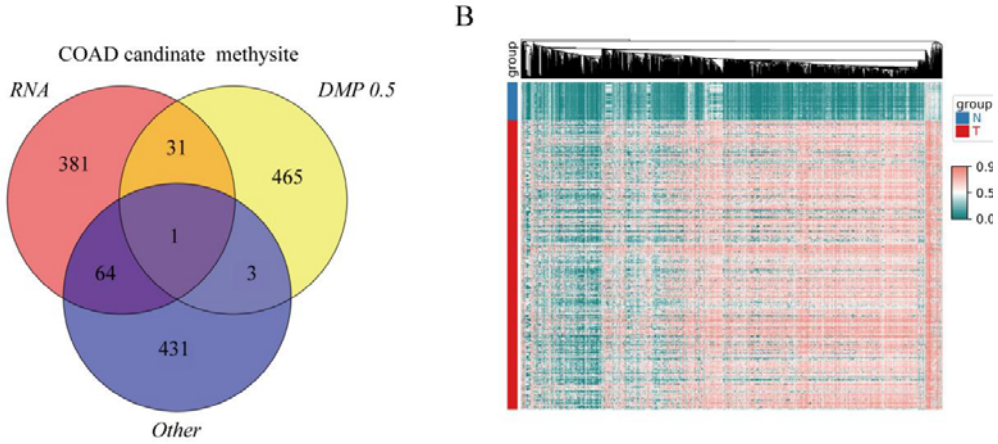
Libraries were quantified using the Qubit dsDNA HS Assay Kit before pooling 4 or 8 individual libraries per capture of Twist Human Methylome Panel or Twist Methyl Custom Panel respectively. Hybridization and capture were performed using the Fast Hybridization Target Enrichment kit from Twist Bioscience. Pooled libraries were sequenced on an Illumina Novaseq 6000 or Hiseq X Ten using paired-end 150-bp reads.

Methylation data processing

The raw image data files acquired through high-throughput sequencing were subjected to base calling using the bcl2fastq software, resulting in the conversion to raw sequencing data stored in the fastq file format. Rigorous quality control measures were applied to the on-machine raw data, facilitating the exclusion of reads of inferior quality. Bismark (Krueger, 2011, invoking Bowtie2 at the core) was used for the alignment analysis of methylation data with the reference genome. After the acquisition of alignment results, Bismark (Krueger, 2011) was once again utilized for the identification of methylation sites. Differential methylation analysis for paired samples was conducted employing the paired t-test method, while unpaired samples underwent differential methylation analysis utilizing the wilcoxon rank sum method.

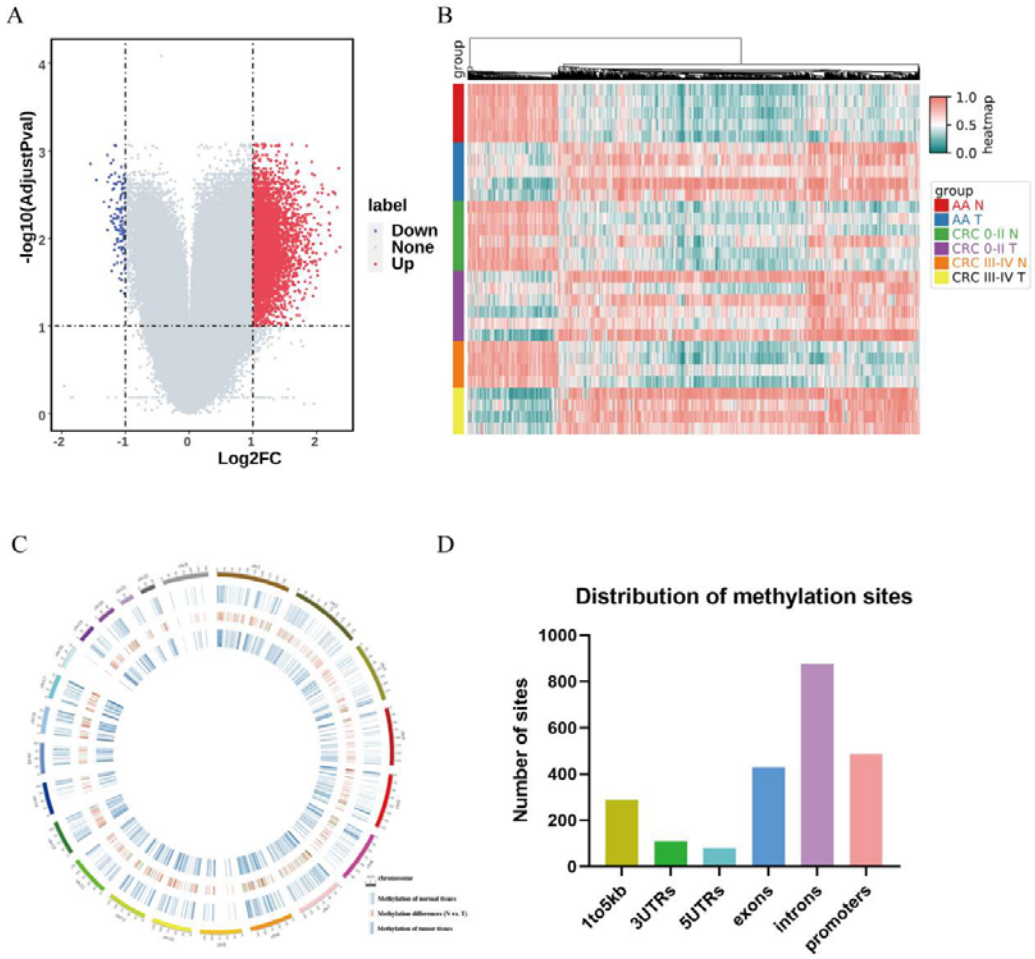
Supplementary Figures

Supplementary Figure 1



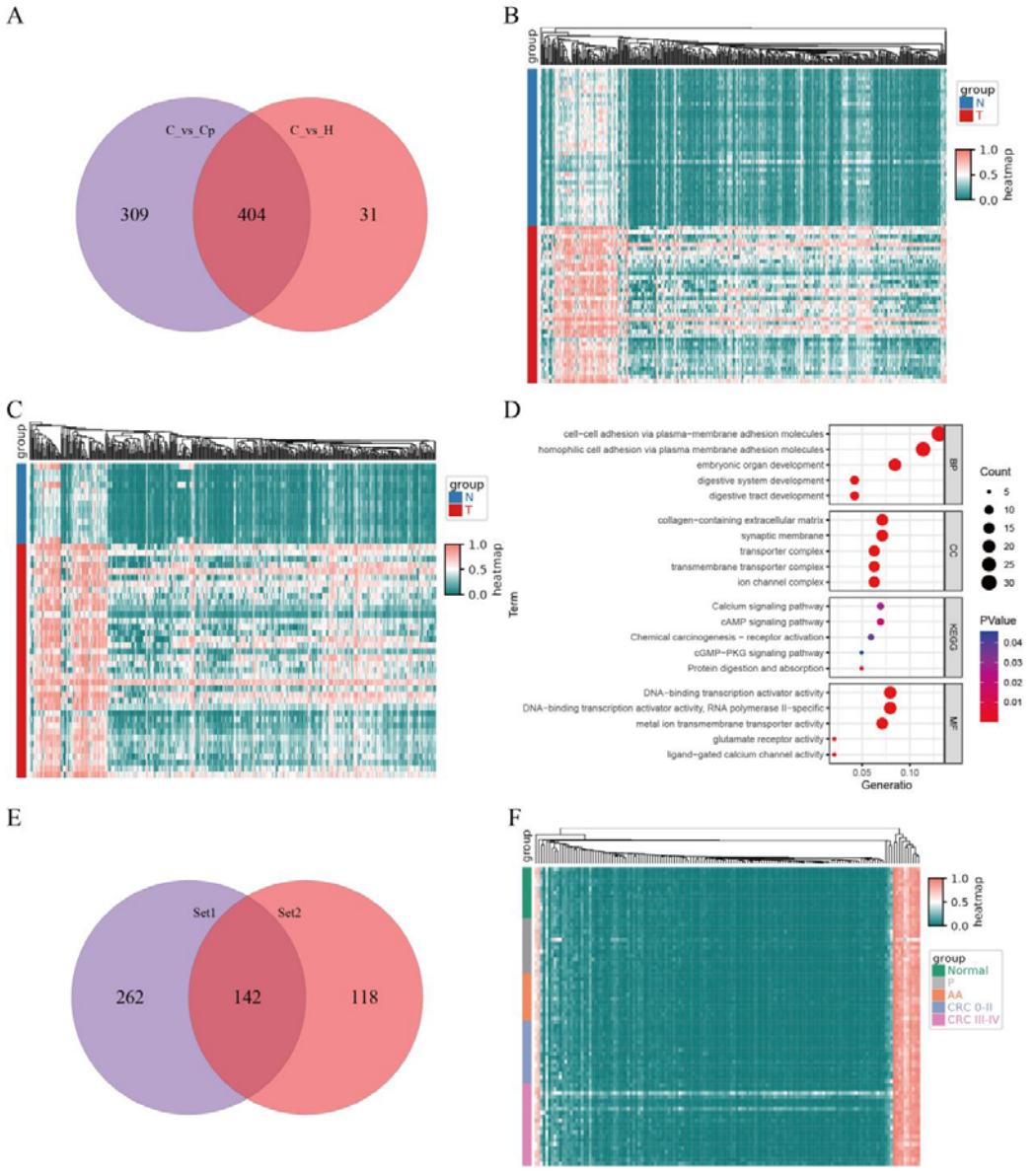
Supplementary Figure 1. Screening for CRC-specific DMCs based on TCGA data. (A) Venn diagram of the results from three DMCs filtering methods based on TCGA datasets. (B) Heatmap illustrating the methylation level of the 1438 methylated sites by unsupervised clustering. Abbreviations: DMCs, differential methylated CpG sites; CRC, colorectal cancer.

Supplementary Figure 2



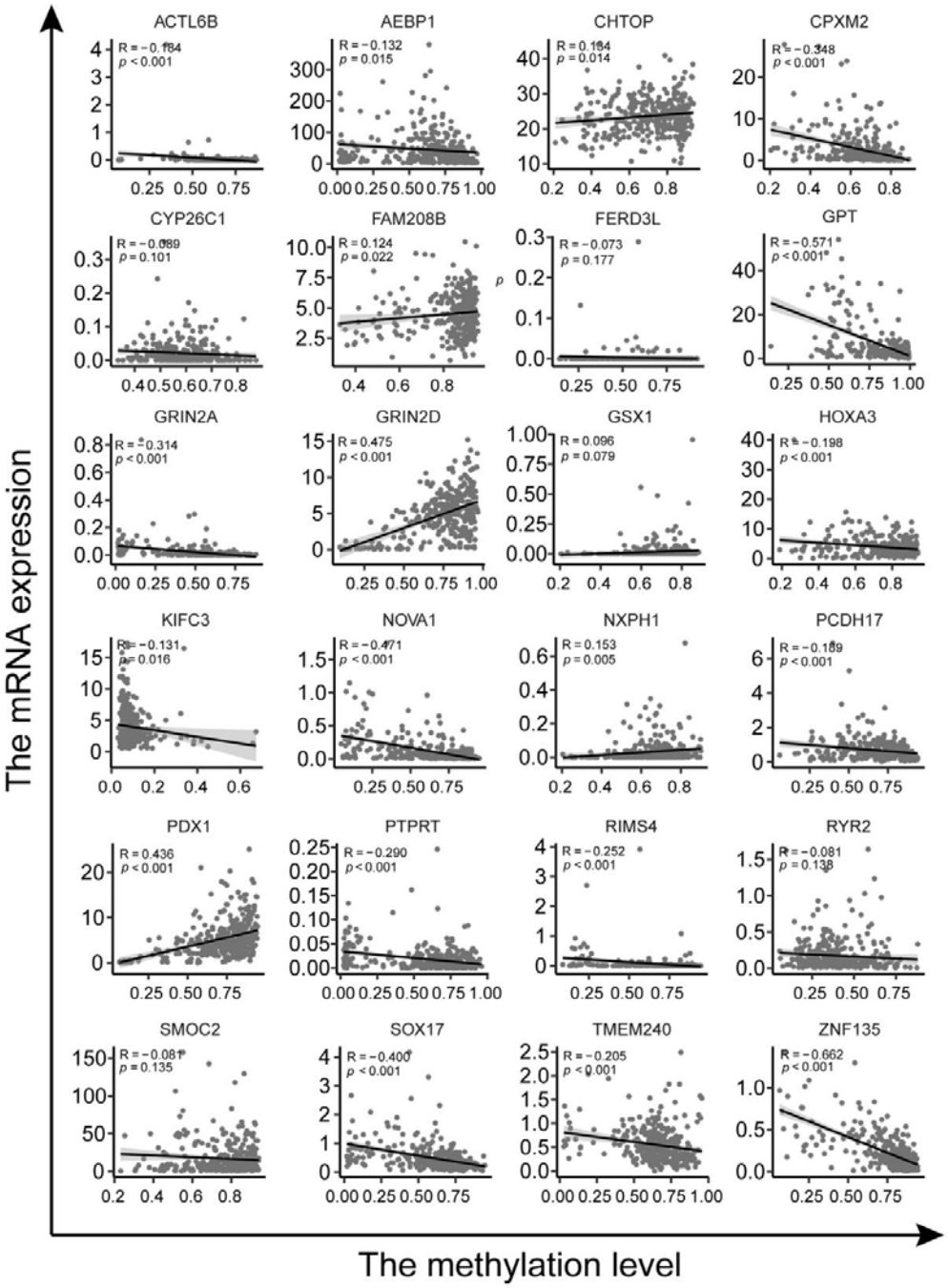
Supplementary Figure 2. DMCs selection in tissues we collected. (A) Volcano plot illustrates the CRC specific DMCs. (B) Heatmap analysis of the top 1400 DMCs between CRC tissues and adjacent tissues by unsupervised clustering. (C) Circular plot of 1400 DMCs depicting methylation level differences between CRC and adjacent tissues. (D) Distribution of 1400 DMCs in gene functional regions. Abbreviations: DMCs, differential methylated CpG sites; CRC, colorectal cancer.

Supplementary Figure 3



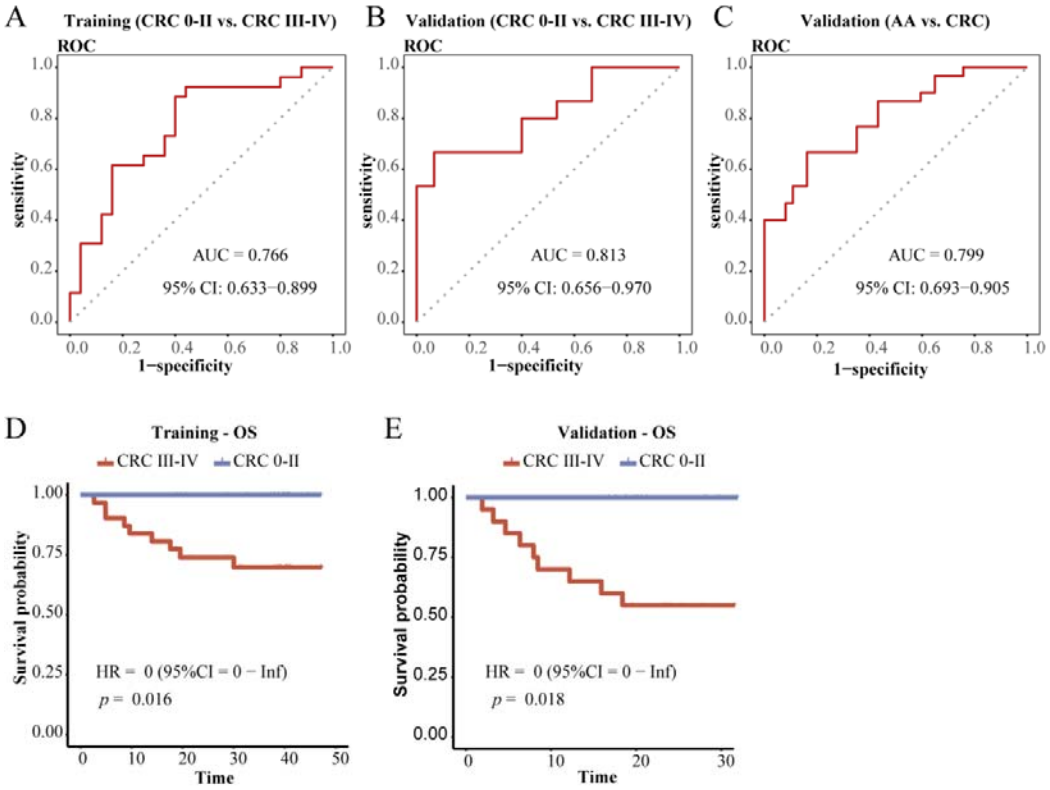
Supplementary Figure 3. Selection of methylation markers. (A) The overlap of DMRs between CRC and adjacent tissues, as well as between CRC and normal tissues. (B) Heatmap illustrating the methylation level of the 404 DMRs in CRC and adjacent tissues. (C) The methylation level of the 404 DMRs in CRC and normal tissues. (D) The gene enrichment analysis for the 404 regions suggests their involvement in processes like cell-cell adhesion, digestive system development, and signaling pathways such as cAMP and cGMP. (E) Based on the results of plasma methylation detection, 142 DMRs were subsequently chosen for further analysis. (F) Heatmap illustrating the methylation levels of the 142 DMRs in the blood samples from normal individuals, those with AA, and patients with CRC. Abbreviations: DMRs, differential methylated regions; AA, adenomatous polyps; CRC, colorectal cancer.

Supplementary Figure 4



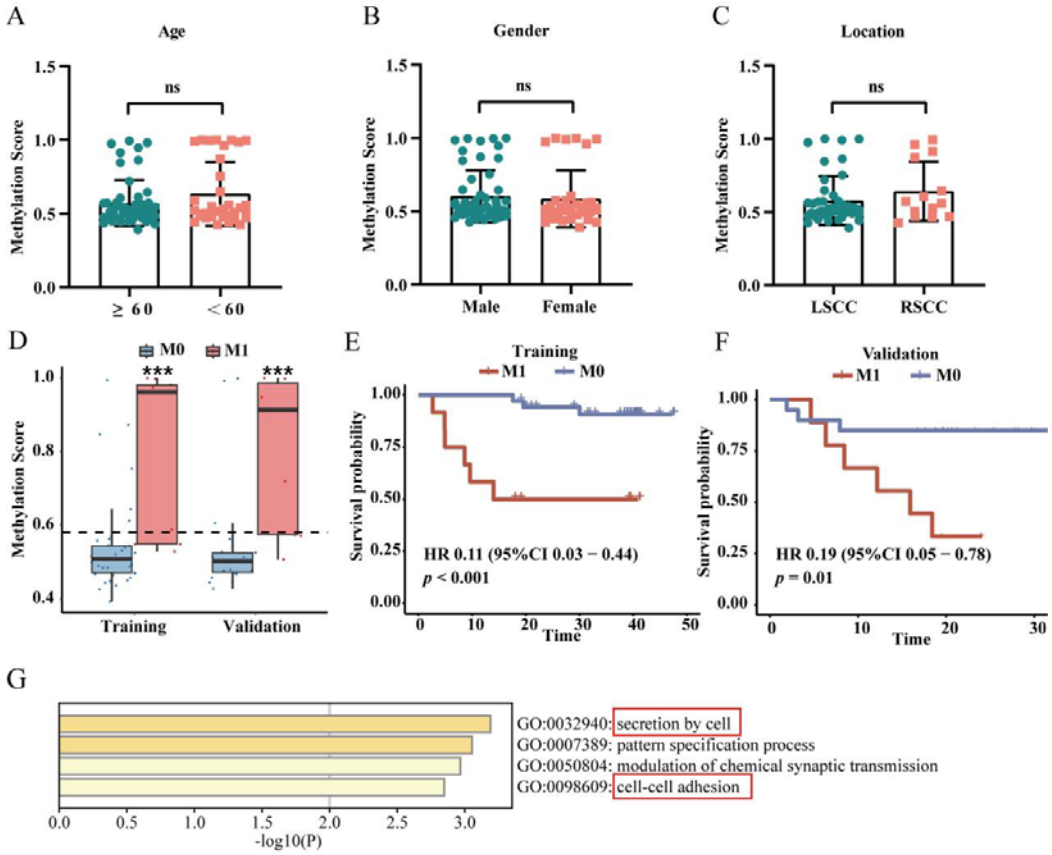
Supplementary Figure 4. The relationship between the methylation levels of the 27 DMRs and the transcription levels of their respective genes in TCGA datasets.

Supplementary Figure 5



Supplementary Figure 5. Development and validation of a plasma diagnosis model for discerning the specific staging of CRC patients. (A) The use of ROC curve analysis to assess the performance of the diagnostic model in differentiating CRC 0-II patients from CRC III-IV patients of the training cohort. (B-C) The use of ROC curve analysis to evaluate the performance of the diagnostic model in distinguishing CRC 0-II patients from CRC III-IV patients (B), AA from CRC patients (C) of the validation cohort. (D-E) Kaplan-Meier survival curves comparing OS between the CRC III-IV group and CRC 0-II group in the training cohort (D) and validation cohort (E).

Supplementary Figure 6



Supplementary Figure 6. The construction of the metastasis model. (A-C) The methylation scores generated from the diagnosis model for CRC patients with different age (A), gender (B) and location (C). (D) Methylation scores of M0-CRC and M1-CRC individuals generated from the metastasis model (The dashed line represents the cutoff value). (E-F) Kaplan-Meier survival curves comparing OS between the M1-CRC group and M0-CRC group of metastasis model in the training cohort (E) and validation cohort (F). (G) Functional enrichment analysis of genes associated with DMRs exhibiting methylation level changes in blood from patients with AA or CRC 0-II stages.

Supplementary Tables

Table s1 Explanation of specific phrase and abbreviation

Specific phrase	Basic annotation
Training cohort	The cohort of subjects used to construct the model.
Validation cohort	The cohort of subjects used to assess the model's performance.
Diagnosis model	The model capable of determining whether a subject has CRC
Prognosis model	The model capable of distinguishing between good and poor prognosis for CRC patients.
Metastasis model	The model determining the presence of metastasis in CRC patients
High-risk CRC group	A group comprising CRC patients with higher methylation scores (the results also indicated that higher methylation scores signify poorer prognosis)
Low-risk CRC group	A group comprising CRC patients with lower methylation scores (the results also indicated that lower methylation scores signify greater prognosis)
cfDNA	Cell-Free DNA
ctDNA	Circulating tumor DNA
DMCs	Differential methylated sites
DMRs	Differential methylation regions

Table s2 Patient characteristics of the 15 paired tissues

	15 paired tissues	
	AA (n = 5)	CRC (n = 10)
Age, years		
Mean ± SD	53.8 ± 10.6	59.7 ± 10.5
Gender		
Male	3	7
Female	2	3
Stage		
0/II	IA	6
III/IV	IA	4

Table s3 Patient characteristics of the 64 participants (Tissues)

	Tissue cohort			
	Normal (n = 13)	NAA (n = 8)	AA (n = 11)	CRC (n = 38)
Age, years				
Mean ± SD	51.5 ± 7.0	54.4 ± 7.5	60.0 ± 13.4	58.5 ± 10.1
Gender				
Male	6	6	6	24
Female	7	2	5	14
Stage				
0/II-no. (%)	IA	IA	IA	44.7
III/IV-no. (%)	IA	IA	IA	55.3

Table s4 Patient characteristics of the blood cohort

	Normal (n = 81)			CRC (n = 99)			AA (n =60)		NAA (n = 91)
	Training (n = 33)	Validation (n = 14)	External Validation (n =34)	Training (n = 51)	Validation (n = 30)	External Validation (n = 18)	Validation (n = 37)	External Validation (n = 23)	External Validation (n = 91)
Age, years									
Mean ± SD	55.6 ± 9.1	55.9 ± 7.7	53.2 ± 9.0	61.0 ± 10.3	62.6 ± 8.5	63.2 ± 10.6	58.1 ± 9.8	62.6 ± 9.0	57.1 ± 9.3
Gender									
Male	14	8	14	32	17	11	26	11	51
Female	19	6	20	19	13	7	11	12	40
Stage									
0/II-no. (%)	IA	IA	IA	25	15	18	IA	IA	IA
III/IV-no. (%)	IA	IA	IA	26	15	0	IA	IA	IA

Table s5. Annotation information for 27 regions

chr	start	end	geneAnno	gene_name	gene_type	function region
chr1	1475208	1475210	TMEM240	TMEM240	protein_coding	intron
chr1	153610799	153610801	CHTOP	CHTOP	protein_coding	CDS
chr1	237206583	237206584	RYR2	RYR2	protein_coding	intron
chr10	5735348	5735376	FAM208B	FAM208B	protein_coding	intron
chr10	94820109	94820110	EXOC6 859	EXOC6	protein_coding	downstream
chr10	125651033	125651035	CPXM2	CPXM2	protein_coding	CDS
chr13	28367740	28367742	GSX1	GSX1	protein_coding	CDS
chr13	28501365	28501367	PDX1 997	PDX1	protein_coding	downstream
chr13	58204974	58204975	PCDH17 969	PCDH17	protein_coding	promoter
chr14	27065973	27065975	NOVA1	NOVA1	protein_coding	intron
chr16	10276798	10276800	GRIN2A 187	GRIN2A	protein_coding	promoter
chr16	57836743	57836745	KIFC3	KIFC3	protein_coding	intron
chr19	48947559	48947561	GRIN2D	GRIN2D	protein_coding	3-UTR
chr19	58570418	58570428	ZNF135 179	ZNF135	protein_coding	promoter
chr2	132183099	132183101	MZT2A 39372	ig_MZT2A		intergenic
chr20	41818770	41818771	PTPRT 160	PTPRT	protein_coding	promoter
chr20	43439544	43439546	RIMS4 565	RIMS4	protein_coding	promoter
chr5	135529351	135529387	SMAD5 4916	SMAD5	protein_coding	downstream
chr6	168844068	168844070	SMOC2	SMOC2	protein_coding	intron
chr7	8481985	8482108	NXPH1	NXPH1	protein_coding	intron
chr7	19185085	19185144	FERD3L 41	FERD3L	protein_coding	promoter
chr7	27167314	27167365	HOXA3	HOXA3	protein_coding	intron
chr7	44143992	44143993	AEBP1	AEBP1	protein_coding	5-UTR
chr7	100253900	100253902	ACTL6B	ACTL6B	protein_coding	intron
chr7	128337724	128337726	FAM71F2 10795	ig_FAM71F2		intergenic
chr8	55370336	55370337	SOX17 158	SOX17	protein_coding	promoter
chr8	145729798	145729800	GPT	GPT	protein_coding	CDS

Table s6 Univariate and multivariate analysis of methylation scores and clinical parameters.

Variables	Univariate analysis		Multivariate analysis	
	HR	*P* value	HR	*P* value
Age (years)				
>60 vs ≤60	0.803 (0.326 - 1.978)	0.634	1.394 (0.539 - 3.605)	0.494
Gender				
Male vs Female	1.026 (0.404 - 2.608)	0.957	1.252 (0.474 - 3.308)	0.650
Location				
LSCC vs. RSCC	1.841 (0.611 - 5.547)	0.278	1.731 (0.557 - 5.380)	0.343
Methylation Score				
(High vs. Low)	221.554 (27.312 - 1797.252)	<0.001	283.583 (30.251 - 2658.403)	<0.001

Table s7 Comparison of diagnostic performance between the methylation score and CEA for early-stage colorectal cancer and adenomas in the external validation cohort.

External Validation cohort		Methylation score predicted			CEA predicted		
Sample	total	negative	positive	specificity/ sensitivity	negative	positive	specificity/ sensitivity
Normal	12	11	1	0.92	12	0	1
0-II	11	4	7	0.64	9	2	0.18
AA	11	4	7	0.64	9	2	0.18
NAA	28	17	11	0.39	27	1	0.04