# Evaluating the performance of a plasma dual-target test developed based on sense-antisense and dual-MGB probe technique for colorectal cancer detection

**Yanteng Zhao**
zyt19891006@126.com

The First Affiliated Hospital of Zhengzhou University

**Zhijie Wang**
Changhai Hospital, Second Military Medical University, Naval Medical University

**Qiuning Yu**
The First Affiliated Hospital of Zhengzhou University

**Xin Liu**
The First Affiliated Hospital of Zhengzhou University

**Xue Liu**
The First Affiliated Hospital of Zhengzhou University

**Shuling Dong**
The First Affiliated Hospital of Zhengzhou University

**Xianping Lv**
The First Affiliated Hospital of Zhengzhou University

**Yu Bai**
Changhai Hospital, Second Military Medical University, Naval Medical University

**Shaochi Wang**
The First Affiliated Hospital of Zhengzhou University

**Research Article**

**Keywords:** colorectal cancer, cell-free DNA, methylation, early detection, methylation-specific PCR

**Posted Date:** March 17th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2605730/v1

**Additional Declarations:** No competing interests reported.

# Abstract

## Background

Detecting colorectal cancer (CRC) via blood-based methylation tests shows good patient compliance and convenience, but some use to fail due to the low abundance of plasma cfDNA fragments. To address this issue, we designed this study to identify potential markers and enhance their performance to detect CRCs using sense-antisense and dual-MGB probe (SADMP) technique.

## Methods

The study was conducted in three steps: identifying eligible methylation markers in our discovery set, developing assay using the sense-antisense and dual-MGB probe (SADMP) technique, and evaluating the test performance for CRC detection in training and validation cohorts.

## Results

Findings of the discovery step indicated that adenoma and cancer samples exhibited similar methylation profiles and both had lower methylation levels than normal samples. Hypermethylated *NTMT1* and *MAP3K14-AS1* were recognized as the most promising candidate markers. The SADMP technique showed an ability to improve methylation signals by 2-fold than single-strand and single-MGB probe techniques. The MethyDT test, incorporating the SADMP technique, obtained an average sensitivity of 84.47% for CRC detection, higher than any single target alone, and without significant attenuation in specificity (average specificities of 91.81% for *NTMT1* and 96.93% for *MAP3K14-AS1* vs. 89.76% for MethyDT). For early (I-II) and late- (III-IV) stage CRC, the sensitivities were 82.61% and 88.64%, respectively. Meanwhile, the test performance was independent of patient age and gender.

## Conclusion

The MethyDT test incorporating the SADMP technique exhibits a higher sensitivity to perceive methylation signals and may serve as a promising noninvasive tool for CRC detection.

## Background

Most early-stage colorectal cancers (CRCs) are curable, especially for precancerous lesions, adenomas and polyps, which can be removed at the time of diagnosed by colonoscopy. Therefore, early detection of CRC is a highly valuable task. Several tests based on cell-free DNA (cfDNA) methylation have been developed and showed good performance for CRC detection. Epi proColon [1] is the first blood-based test used for early detection of CRC, which was developed based on methylated cfDNA of Septin9. Currently, the updated version of Epi proColon ® 2.0 CE obtained an improved sensitivity of 74.8%-81% and specificity of 96.3%-99% in a prospective cohort study compared to the first generation [2]. The Chinese version of Epi proColon simplified sample processing with a single reaction system in larger volume (60 ul) instead of a 2/3 algorithm (20 ul for three runs) and showed a sensitivity of 73% and specificity of 94.5% [3]. However, the accuracy of Epi proColon remains unsatisfactied, which was much lower than that of fecal DNA tests [4]. Moreover, plasma methylated Septin9 is a non-CRC specific marker that showed an ability to detect multiple cancer types, including hepatocellular carcinoma [5], gastric cancer [6], cervical cancer [7], and breast cancer [8].

Detecting methylation signal of cell-free tumor DNA (ctDNA) in plasma is challenging due to reasons of low abundance of ctDNA fragments, releasing by multiple organs or tissues, and DNA damage by bisulfite conversion. Therefore, eligible markers are essential, and proper detection techniques are fundamental to ensure their excellent performance [9]. The classical PCR-based methylation detection techniques are often developed based on a single strand BS-DNA, leaving the information of the other strand unused [10]. Besides, only one Taqman MGB probe is usually designed to provide fluorescent signals. Theoretically, designing primers for both sense and antisense strand DNA simultaneously or using multiple MGB probes will improve the sensitivity of a marker to detect methylation signals by enhancing fluorescent signals. Sarah Ø. Jensen et al. [11] made the first attempt to design a pair of primers for both sense and antisense strands, thus improving the performance of three targets for CRC detection. Meanwhile, the dual-strand technique was also successfully applied for methylated *HOXA9* in ovarian cancer (OV) to improve the detection sensitivity for OV [12].

Multi-MGB probe technique is rarely reported in previous studies. In order to enhance the ability of candidate markers to detect low-abundance ctDNA methylation signals in plasma, we attempted to apply dual-strand and dual-MGB probe techniques simultaneously, which we called the sense-antisense and dual-MGB probe (SADMP) technique, to develop a novel CRC plasma test. In this study, we first integrated various methylation datasets from public databases and identified a group of most promising candidate markers. Then the SADMP technology was used to develop a test for CRC detection. Finally, the test performance was comprehensively assessed in our recruited training and validation cohorts.

## Methods

# Data preparation

We collected 13 methylation datasets from public databases. The selected datasets met three criteria: 1) were generated by Illumina HumanMethylation 450k BeadChip and had the raw IDAT files, 2) the sample size was greater than 10, and 3) consisted of CRC or adenoma or adjacent normal samples. All the IDAT files were then processed using minfi tool [13] to obtain methylation β values. They were integrated as a single dataset (n = 1165), which we defined as discovery set for candidate markers identification.

Level 3 methylation data of 31 cancer types were retrieved from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). Normal adjacent tissue (NAT) and primary tumor tissue were retained, corresponding to 710 and 8258 samples, respectively. The 31 cancer types consist of ACC (n = 79), BLCA (n = 412), BRCA (n = 778), CESC (n = 306), CHOL (n = 36), CRC (n = 379), DLBC (n = 48), ESCA (n = 183), GBM (n = 137), HNSC (n = 523), KICH (n = 65), KIRC (n = 312), KIRP (n = 271), LGG (n = 513), LIHC (n = 374), LUAD (n = 456), LUSC (n = 364), MESO (n = 87), OV (n = 10), PAAD (n = 183), PCPG (n = 178), PRAD (n = 495), SARC (n = 257), SKCM (n = 104), STAD (n = 393), TGCT (n = 133), THCA (n = 503), THYM (n = 124), UCEC (n = 418), UCS (n = 57), UVM (n = 80). The data of colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) were merged as one CRC cohort, of which 379 were tumors and 45 were matched normal tissues (**Supplemental table 1**). The other three datasets collected from Gene Expression Omnibus (GEO) were used as validation sets to verify the methylation status of candidate differentially methylated CpGs (DMCs) (**Supplemental table 2**). The methylation data of GSE48684 [14], GSE40279 [15] and GSE122126 [16] cohorts, generated by the same platform of Illumina HumanMethylation BeadChip, were downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). GSE48684 cohort contained 105 suitable samples, of which 41 normal and 64 CRC samples were used in this study (**Supplemental table 3**). The GSE40279 cohort were whole blood cell (WBC) samples collected from 656 healthy individuals (**Supplemental table 4**). The GSE122126 cohort consisted of 101 samples, but only three CRC and 12 normal plasma samples were retained.

## Sample collection

This study is a case-control study, which enrolled 742 cases, including 30 CRC tissue and 712 plasma samples from the First Affiliated Hospital of Zhengzhou University between April 2022 and June 2022 (**Supplemental table 5**). Tissue samples were obtained from the preserved formalin-fixed paraffin-embedded (FFPE) sections collected from surgical patients. Plasma samples were collected in two steps. The first step (assay development step) recruited 211 participants, of which ten were excluded as they did not fulfill the included criteria. The rest 201 participants included 55 healthy blood donors, 71 CRC patients, and 75 individuals with other intestinal diseases. In step two (assay assessment), the validation cohort consisted of 511 participants after excluding twenty individuals whose pathology information were missing, including 117 healthy blood donors, 62 interfering diseases (8 other cancers and 54 non-intestinal diseases), 20 polyps, 40 adenomas, 120 CRCs, and 152 intestinal diseases (but not diagnosed as CRC). This study is a subproject of the Clinical Study of Pan-cancer DNA Methylation Test in plasma (ClinicalTrials ID: NCT05685524), which has been approved by the ethics committee of the First Affiliated Hospital of Zhengzhou University (approval number: 2022-KY-0631-002). All participants signed an informed consent form and were told the results. Patients with polyps, adenomas, or CRC were further confirmed by histopathological examinations. The included CRC patients were required to meet the following criteria: 1) did not receive radiotherapy or chemotherapy or surgery, 2) no other diseases or at least no other physical abnormalities, and 3) ages larger than 18. All CRC patients were classified as I, II, III, and IV stages according to the American Joint Cancer Committee (AJCC) staging system.

## Differential methylation analysis

Samples of the discovery set were classified into normal (NAT, mucosa), adenoma, and cancer (colon and rectal cancer) groups according to their disease status. We performed differential methylation analysis using the rank-sum test for three comparisons, cancer vs. normal, adenoma vs. normal, and cancer vs. adenoma. Significant DMCs are defined as the FDR < 0.05 and the fold change ranking the largest top 1% of all probes. DMCs were then divided into hyper- or hypo- DMCs if they showed high or low methylation levels in tumor or adenoma compared to normal. LASSO regression is implemented in the R package 'glmnet' to reduce the number of features. β values of 710 NATs and 376 CRCs in TCGA, used as independent variables and response variables, respectively, are input to the 'cv.glmnet' method with parameters alpha = 1 and family = binomial. We repeated LASSO regression 100 times and counted the frequencies of probes with non-zero coefficients in the regressions.

## Developing the dual-strand and dual-MGB probe technique

DNA extraction and bisulfite treatment were performed according to the instructions described in the previous study [31]. Primer design and MSP system are described in supplemental methods. Assay parameters were estimated using the standard cure experiments. To address the challenge of extremely low abundance of cfDNA fragments in plasma, we developed a technique called sense-antisense strand and dual-MGB probe (SADMP) technique to enhance the sensitivity of methylation-specific PCR to detect methylated cfDNA signals in plasma samples. This technique boosted the methylation signal approximately 2-fold than the baseline level in our assessment (see **Supplemental methods**).

## Statistical analysis

Data processing and analysis in this study were performed on R software (version 4.1.0). The 'glm' function was used to fit the logistic regression model with a parameter of 'family = binomial'. Receiver operating characteristic (ROC) curve analysis was conducted using the 'pROC' package, and the area under the ROC curve (AUC) was then calculated to assess the test classification performance. The optimal sensitivity and specificity of the test were estimated when Youden's index reached maximal. Sensitivity and specificity were calculated as following formulas:

$$sensitivity = \frac{\mathrm{Truepositive}}{\mathrm{Truepositive} + \mathrm{Falsenegative}}$$

$$specificity = \frac{\mathrm{Truenegative}}{\mathrm{Truenegative} + \mathrm{Falsepositive}}$$

And the Youden index = sensitivity + specificity-1.

Rank-sum test and Kruskal test were performed for the comparisons between two groups and the comparisons between multiple groups, respectively. The Chi-square test was used for comparisons between categorical variables. Other statistical methods used in this study were described in the corresponding results.

## Results

# Study design and participant characteristics

The flowchart of this study is showing in Fig. 1 which consists of three steps. In the first step, candidate markers were identified in our discovery set and validated in independent sets. The discovery set includes a total of 1165 suitable samples from 13 datasets with 452 normal, 168 adenomas and 545 CRCs. The independent validation sets are TCGA and GSE48684 with 421 samples (NAT = 45, CRC = 376) and 105 samples (NAT = 41, CRC = 64), respectively. Methylation status of candidate CpGs were then confirmed by Sanger sequencing. In the second step, an MSP system was established to detect the methylation signals of candidate markers. This step included designing appropriate primers, optimizing qPCR amplification system, and evaluating the assay's technical parameters. In the third step, the performance of the developed assay was assessed in training and validation sets. The primary indicators, including sensitivity, specificity, and AUC, were estimated in step 3.

# Landscape of the methylation patterns of the discovery set

Overall, adenoma and cancer samples showed lower methylation levels than normal samples (Fig. 2A). The methylation density curves of the three groups exhibited bimodal distributions (Fig. 2B), corresponding to hyper- and hypo-methylation peaks, respectively. While the hyper-methylation peaks of adenoma and tumor were lower than normal, indicating higher methylation levels in normal samples. Using t-SEN to analyze and visualize the structure of discovery set, we observed significant differences between normal and cancer samples, while adenomas overlapped with both normal and cancer samples (Fig. 2C). We then selected the top 1% of most variable probes to cluster the discovery set (K-means algorithm), and results showed that both normal and cancer samples clustered with each other. However, adenomas were separated into two subgroups with a noticeable difference, showing high (Methy-H) and low (Methy-L) methylation status and closer distance to cancer and normal samples, respectively (Fig. 2D). Further analysis revealed that tubular adenomas accounted for the largest proportion of Methy-L adenomas (40.35%, 23/57), while villous adenomas accounted for the largest proportion of Methy-H adenomas (64.38%, 47/73). To exclude the bias between different datasets, we added dataset as a strata factor and conducted fisher's test separately for the Methy-H and Methy-L adenomas, and no significant differences were obtained ($P > 0.05$).

# Identification of candidate DMCs

Using the discovery set, we attempted to identify cfDNA methylation markers for CRC detection. First, we analyzed the DMCs between cancer, adenoma, and normal samples. The three comparisons (cancer vs. normal, adenoma vs. normal, and cancer vs. adenoma) yielded 3000, 3051, and 1545 DMCs, respectively. Interestingly, these DMCs were dominated by hyper-DMCs (Fig. 3A). Further investigations revealed that most DMCs were distributed in upstream regions of genes (5'UTR, TSS1500, TSS500, and 1stExon), except for cancer vs. adenoma where DMCs were mainly located in gene bodies (Fig. 3B). We observed an extremely high proportion of overlapping DMCs between cancer vs. normal and adenoma vs. normal, significantly higher than cancer vs. adenoma (Fig. 3C). To obtain the most appropriate DMCs, we focused on those overlapped DMCs between cancer vs. normal and adenoma vs. normal (2237 probes). We then evaluated the methylation levels of the 2237 DMCs on 32 cancer types of TCGA. Eligible DMCs were those with methylation levels < 0.2 on other cancer types (non-CRC), >=0.55 on CRC, and < 0.15 on normal samples (n = 710), which gave us 75 accessible probes (Fig. 3D). LASSO regression was used to reduce the number of DMCs, with eight presented in 100 replicates with non-zero coefficients (Fig. 3E). Finally, in WBC samples from healthy individuals, we found that seven DMCs showed pretty low methylation levels except for cg17892556 (Fig. 3F). Therefore, the 7 DMCs are recognized as the most promising markers.

## The methylation levels of MTNT1 and MAP3K14-AS1 in validation sets

We select two of the seven DMCs, cg14015706 and cg08247376, locating in the first exon of *NTMT1* and 200 bp of the upstream *MAP3K14-AS1* transcription start site, respectively (**Supplemental methods**), for the following study because they are in CpG-enriched regions, ideal for designing MSP primers and MGB probes. In TCGA dataset, the two probes showed frequently hypermethylated events in cancer samples compared to normal samples (median β = 0.62 [1st quantile − 3rd quantile: 0.55–0.69] vs 0.037 [1st quantile − 3rd quantile: 0.031–0.042] for cg14015706 and 0.59 [1st quantile − 3rd quantile: 0.48–0.68] vs 0.067 [1st quantile − 3rd quantile: 0.057–0.081] for cg08247376) (Fig. 4A). Meanwhile, their methylation levels did not show strong correlations between patient age and methylation level in TCGA CRC cohort (Pearson's correlation coefficient = 0.13 for cg14015706 and 0.061 for cg08247376) (**Supplemental Fig. 1**). In GSE48684 dataset, significantly higher methylation levels for both probes were also observed in cancer samples than in normal samples (Fig. 4B). Similarly, they showed pretty low methylation levels in 710 adjacent normal samples, with median β values of 0.028 [1st quantile − 3rd quantile: 0.033–0.041] for cg14015706 and 0.053 [1st quantile − 3rd quantile: 0.081–0.18] for cg08247376 (Fig. 4C). Similar results were observed in 656 healthy WBC samples (Fig. 4D). The median β values of cg14015706 and cg08247376 were 0.058 [1st quantile − 3rd quantile: 0.051–0.065] and 0.073 [1st quantile − 3rd quantile: 0.067–0.079] respectively. The GSE122126 dataset consisted of three CRC and 13 healthy plasma samples. We found that both probes exhibited hypermethylated in CRC plasmas and hypomethylated in healthy plasmas (Fig. 4E), indicating a high consistency of the methylation status between tissues and plasmas.

## Validation of the methylation status of MTNT1 and MAP3K14-AS1 by Sanger sequencing

Pyrosequencing was performed for 30 CRC tissues and 30 normal controls to confirm the methylation status of the target regions of *MTNT1* and *MAP3K14-AS1*. The *MTNT1* marker contained two amplification regions in sense and antisense strands, covering 10 and 3 key CpG sites, respectively. The *MAP3K14-AS1* marker contained one amplification region in the antisense strand, covering six key CpG sites. We successfully obtained the sequencing results of amplified products from the *NTMT1* antisense strand assay on 25 normal and 29 cancer tissues. Overall, these CpG sites were widely methylated in cancer samples but unmethylated in normal samples (Fig. 5A, **Supplemental table 6**). Though the methylation status of several CpG sites in the *NTMT1* sense-strand was missing due to sequencing failure, we still observed frequently methylated events in cancer samples (Fig. 5B, **Supplemental table 6**). Hypermethylated events were also found for the amplified products of *MAP3K14-AS1* antisense-strand assay in cancer samples but not in normal samples (Fig. 5C, **Supplemental table 6**). Moreover, each CpG site of the three target regions showed significantly higher methylated frequency in cancer samples than in normal samples (**Supplemental table 7**).

# Performance of the test in training cohort

According to the standard curves of *NTMT1* and *MAP3K14-AS1* (**Supplemental Fig. 4**), we first estimated the cfDNA input of these two genes in plasma samples in training set. The median copy numbers of *NTMT1* and *MAP3K14-AS1* were 71.76 [1st – 3rd quantile: 11.29–177.54] and 37.17 [1st – 3rd quantile: 1.637–73.096] on CRC samples, both significantly higher than those of healthy and other non-CRC samples (Fig. 6A). In addition, both copy numbers exhibited an increasing trend in the CRC samples from stage I to IV (**Supplemental Fig. 5**). Meanwhile, CRC samples showed much lower Ct values than non-CRC and healthy samples (Fig. 6B). ROC curve analysis was conducted to assess the ability of the two genes to discriminate CRC samples from non-CRC samples, with Ct values and disease status passed as the parameters of predictor and response. We obtained AUC values of 0.87 (95% CI: 0.81–0.93) and 0.77 (95% CI: 0.70–0.82) for *NTMT1* and *MAP3K14-AS1*, respectively (Fig. 6C&D). Since single gene provided limited methylation information, we then attempted to combine the two markers, and the combined assay was then named MethDT test.

Two strategies were adopted to obtain a better appropriate combination algorithm for MethyDT test. Strategy 1 was to construct a logistic regression model, and the estimated AUC value was 0.89 (95% CI: 0.85–0.94) with optimal sensitivity and specificity of 83.10% and 89.23%, respectively (Fig. 6E). Strategy 2 was the 1/2 algorithm, where a positive measurement was determined when the Ct of any single marker was less than its corresponding threshold. The optimal Ct cutoff values were 49.73 and 48.36 for *NTMT1* and *MAP3K14-AS1*, respectively, when Youden’s index achieved maximal (**Supplemental table 8&9**). At these thresholds, the two target sensitivities were 78.87% and 54.93%, with specificities of 91.54% and 97.69% (**Supplemental table 10**). Interestingly, strategy 2 obtained an equal sensitivity and specificity as strategy 1. Since strategy 2 is much simpler for examining physicians to interpret the test results in clinical practice, we adopted the 1/2 algorithm as the combination algorithm for MethyDT test. After the algorithm was fixed, the MethyDT test obtained sensitivities of 79.17% and 91.30% for early- (I-II) and late- (III-IV) stage CRCs (**Supplemental table 11**). Additionally, no significant variations were observed for the MethyDT test sensitivity in detecting CRC patients with different ages and sex (**Supplemental table 11**).

# Performance of the test in validation cohort

Preliminary results of the training set suggested that MethyDT test showed an improved sensitivity for CRC detection compared to single target alone. We then assessed the test performance in an independent validation set. The estimated copy numbers of *NTMT1* and *MAP3K14-AS1* in plasma samples in validation set were the highest in CRC samples (**Supplemental Fig. 6A&B**) and did not show significant variations across different stages (**Supplemental Fig. 6C&D**). Overall, when using all non-CRC samples (healthy donors, interfering diseases, polyps, adenomas, and intestinal diseases) as control, the sensitivity and specificity of *NTMT1* for CRC detection were 75.83% and 92.07%, while they were 64.17% and 96.16% for *MAP3K14-AS1* (Table 1). According to the fixed algorithm in training set, the sensitivity and specificity of MethyDT test were 85.83% and 90.28%, better than those of any single target (Table 1). When using interfering diseases and healthy donors as controls, the test specificities were 87.96% and 95.73%, respectively (Table 1). The positive prediction rate of MethyDT test was 73.05% (95%CI: 65.73% ~ 80.37%) when non-CRC samples were control, but improved to 75.74% (95%CI: 68.53% ~ 82.94%) and 95.37% (95%CI: 91.41% ~ 99.33%) when interfering diseases and healthy donors were controls (Table 1). The negative prediction rates for non-CRCs, interfering diseases and healthy donors were 95.41% (95%CI: 93.27% ~ 97.54%), 93.41% (95%CI: 90.38% ~ 96.44%) and 86.82% (95%CI: 80.98% ~ 92.66%) respectively. For early- and late-stage CRCs, the sensitivities were 82.61%, 88.64% (**Supplemental table 12)**. Meanwhile, the MethyDT test did not show significantly different sensitivities in detecting CRC patients with different ages and sex in validation set (**Supplemental table 12**). For adenomas and polyps, the MethyDT test obtained positive detection rates of 30.00% (12/40) and 10.00% (2/20) (**Supplemental table 13**).

Table 1
The performance of MethyDT test in validation set.

| Target | Sensitivity (95%CI) | Specificity (95%CI) | PPV (95%CI) | NPV (95%CI) | Accuracy (95%CI) | Comparisons |
|---|---|---|---|---|---|---|
| NTMT1 | 75.83 (89.39 ~ 94.75) | 92.07 (68.17 ~ 83.49) | 74.59 (66.86 ~ 82.32) | 92.54 (89.93 ~ 95.16) | 88.26 (85.47 ~ 91.05) | CRC vs non-CRC |
| MAP3K14-AS1 | 64.17 (94.26 ~ 98.07) | 96.16 (55.59 ~ 72.75) | 83.7 (76.15 ~ 91.24) | 89.74 (86.83 ~ 92.64) | 88.65 (85.90 ~ 91.40) | CRC vs non-CRC |
| MethyDT | 85.83 (87.35 ~ 93.22) | 90.28 (79.59 ~ 92.07) | 73.05 (65.73 ~ 80.37) | 95.41 (93.27 ~ 97.54) | 89.24 (86.55 ~ 91.92) | CRC vs non-CRC |
| NTMT1 | 75.83 (86.62 ~ 93.67) | 90.15 (68.17 ~ 83.49) | 77.12 (69.54 ~ 84.70) | 89.49 (85.88 ~ 93.11) | 85.79 (82.34 ~ 89.23) | CRC vs interfering disease |
| MAP3K14-AS1 | 64.17 (92.74 ~ 97.77) | 95.26 (55.59 ~ 72.75) | 85.56 (78.29 ~ 92.82) | 85.86 (81.94 ~ 89.77) | 85.79 (82.34 ~ 89.23) | CRC vs interfering disease |
| MethyDT | 85.83 (84.10 ~ 91.81) | 87.96 (79.59 ~ 92.07) | 75.74 (68.53 ~ 82.94) | 93.41 (90.38 ~ 96.44) | 87.31 (84.02 ~ 90.60) | CRC vs interfering disease |
| NTMT1 | 75.83 (93.29 ~ 99.87) | 96.58 (68.17 ~ 83.49) | 95.79 (91.75 ~ 99.83) | 79.58 (72.95 ~ 86.21) | 86.08 (81.67 ~ 90.48) | CRC vs healthy |
| MAP3K14-AS1 | 64.17 (95.94 ~ 100) | 98.29 (55.59 ~ 72.75) | 97.47 (94.00 ~ 100) | 72.78 (65.85 ~ 79.72) | 81.01 (76.02 ~ 86.01) | CRC vs healthy |
| MethyDT | 85.83 (92.06 ~ 99.39) | 95.73 (79.59 ~ 92.07) | 95.37 (91.41 ~ 99.33) | 86.82 (80.98 ~ 92.66) | 90.72 (87.02 ~ 94.41) | CRC vs healthy |

# Discussion

It is generally believed that early detected CRC patients can be treated more straightforwardly and have better prognoses. Several stool DNA-based tests have been provided, showing excellent performance in detecting CRCs at their early stages [18–21]. Blood sampling is more acceptable than stool sampling, but the blood-based tests are less reported and usually exhibited lower sensitivities than stool-DNA tests, ranging from 47–87% [21]. This study presented a systemic pipeline for the methylation markers discovery, test development and evaluation in training and validation sets. The developed MethyDT test creatively utilized a sense-antisense and dual-MGB probe (SADMP) technique, showing an enhanced ability to detect methylation signals in plasma samples. After a comprehensive evaluation, the test obtained an overall sensitivity and specificity of 85.83% and 90.28% respectively, for CRC detection at ten milliliters of blood (2 ~ 3 ml plasma).

The colon lesions (adenoma and CRC) display lower methylation levels overall, except in regulatory regions, as shown by the fact that tumor and adenoma had more DMCs in promoters than normal, which has been reported in previous study [22]. Previous studies have focused on the CpG Island Methylator Phenotype (CIMP) found in CRC. In this study, it seems that adenoma can be divided into two subclasses, methy-H and methy-L based on the global methylation levels, where they have similar methylation profiles to CRC and normal, respectively. Moreover, we also found that tubular adenomas are common in methy-L subclass, while villous adenomas are more often in methy-H subclass. A few studies implied that CIMP is rarely found in tubular adenomas, but frequently in tubulovillous and villous adenomas [23], which is confirmed in this study. This study also found a large proportion of overlapping DMCs between cancer vs normal and adenoma vs normal, indicating that many CpGs have undergone aberrant methylation events at the adenoma phase (precancerous lesion) during the developing sequence of normal-adenoma-CRC, which provides robust evidence for discovering the methylation markers for CRC early detection.

One of the challenges of blood-based tests is accurately detecting the target DNA fragments derived from intended tumor tissues. Currently, the origins of cfDNA in blood are still poorly understood, although they have been used in many areas, including drug assistance [24], recurrence monitoring [25], and cancer diagnosis [26]. Usually, cfDNA in blood system is thought to be released by apoptotic cells or necrotic cells or positively secreted by some activated cells [27]. The complicated origin of cfDNA makes blood-based tests more susceptible to interfering diseases, leading to a high false-positive rate. Therefore, specificity is a critical indicator for a blood-based test, and a high specificity can reduce the false-positive measurements caused by other non-CRC diseases. In the marker discovery step, adjacent normal samples from according 32 cancer types in TCGA database were used to control the low methylation levels of candidate markers in other tissues, which effectively attenuated the interference of unintended cfDNAs derived from other tissues or organs. In assay development, we designed highly selective MSP primers that did not show normal amplification curves even when unmethylated DNAs were used as templates at $10^7$ copies. In assay assessment, MethyDT test achieved a specificity of 90.28% when interfering diseases and healthy individuals were grouped as normal controls. For interfering cases, the test showed a positive detection rate of less than 10%. These results suggested that MethyDT test had an excellent ability to discriminate CRC from other diseases. In addition, the methylation levels of candidate markers in whole blood cells were limited to no more than 0.1, which ensured a low methylation background noise.

The test sensitivity is associated with the amount of DNA input. Theoretically, methylated signals in larger blood amount are more likely to be detected because of the availability of more cfDNA templates. However, accessible blood amount is often limited in clinical practice due to participants body conditions or other factors. In this study, 10 ml of blood (approximately 2–3 ml of plasma) was drawn from participants for performing MethyDT test. The estimated average median copy numbers in CRC samples between training and validation sets were 92.44 [1st-3rd quantile: 12.45–186.99] for *NTMT1* and 46.62 [1st-3rd quantile. 1.64–85.97] for *MAP3K14-AS1*. Since the lowest detection limits of the two markers were 10 and 5 copies/ul, which is lower than their estimated input copies, we thought that the current blood amount is sufficient for MethyDT test to detect CRC samples without the risk of missing measurements due to insufficient DNA input.

The application of SADMP technique also contributed to the improved sensitivity of MethyDT test. The dual-strand technique was first used for ctDNA methylation detection and had been proved enhancing the markers' performance in previous studies [11, 12], which was observed in this study too. Simultaneously detecting the methylation signals of *NTMT1* sense- and antisense-strand allowed the MSP Ct value of dual-strand assay to shift forward by one compared to single-strand assay. As a result, the detection limit of *NTMT1* assay reached ten copies which were lower than any single strand assay. Meanwhile, two MGB probes located downstream of forward and reverse primers of *MAP3K14-AS1, respectively*, were designed in the current study. During PCR strand extension, the polymerase enzymes cleaved the 5-primer sequence of probes and released two fluorescent groups. The dual-MGB probe technique would theoretically double the fluorescent signals when both probes share the same channel, leading to an earlier Ct value similar to that of the dual-strand technique. Serial dilution experiments confirmed the superiority of dual-MGB probes over one MGB probe. These results suggested that applying the SADMP technique can be a feasible strategy to enhance the detection sensitivity of candidate markers.

Two combination algorithms were adopted to evaluate the MethyDT test performance in training set, and both suggested greater AUC values and higher sensitivities for the combined markers than any single marker. However, the MethyDT test showed a decreased specificity compared to both single markers (from 91.54% and 97.69–89.23%), which was also observed in other studies [28, 29]. In validation set, using the locked algorithm, the test achieved an overall sensitivity of 85.36% and specificity of 90.28%. The specificity improved to 95.73% when healthy individuals were selected as control, comparable to that of *SEPT9* [2]. These data demonstrated the robust performance of MethyDT test for CRC detection.

The current test utilized a 1/2 algorithm instead of a logistic regression model for several reasons, though they showed the same sensitivity and specificity in the training set. First, in clinical practice, the 1/2 algorithm allowed the examination staff to determine the measurements according to Ct values reported by the device directly, facilitating the interpretation of detection results. Second, the 1/2 algorithm can avoid outputting ambiguous results near the threshold of predicted probabilities by a logistic regression model. Since the logistic regression model predicts the probability of each sample being CRC, the probability cutoff is a critical parameter. Therefore, samples near different probability cutoff values will be determined with opposite results. Third, the 1/2 algorithm provided a redundancy strategy because approximately 65% of CRC cases were detected positively by both markers (**Supplemental table 14&15**).

Early diagnosis or screening techniques are essential to improve patient survival time when curable treatments are available. Studies have shown that the 5-year survival rate of early detected CRC is almost 90%, while it was only 20% for advanced CRC [30]. In validation set, the MethyDT test sensitivity was

82.61% for early-stage CRC detection, slightly lower than that of late-stage (stage III-IV, 88.64%), but without significant variation between them. Notably, the MethyDT test obtained a positive detection rate of 30.00% (12/40) for advanced adenomas, significantly higher than for polyps and other interfering diseases, implying its ability to detect the CRC precancerous lesions. Although the detected adenomas will lead to a high false-positive rate, it is meaningful in clinical practice because it provides a risk warning for individuals before the adenomas progress to CRC, and they should undertake ongoing follow-ups in the future.

The current study has some limitations that may hamper the interpretation of these results. 1) The training and validation sets were retrospective cohorts, and most CRC patients exhibited symptoms. While for asymptomatic subjects, the higher proportion of early-stage CRCs and precancerous lesions may result in a lower sensitivity than reported here. Besides, the CRC patient age deviates from healthy individuals, which may impact the test accuracy. 2) Participants in this study were enrolled from a single center. The patients enrolled in this study represented a subset of CRC, which may bias these results. 3) The two markers used in this study were identified from the methylation profiles of CRC tissue samples, not representing the methylation characteristics of cfDNA. Therefore, several eligible cfDNA methylation markers can be missed. 4) MethyDT test showed relatively lower sensitivity for early-stage CRC detection, especially for precancerous adenomas. Further improvement is needed in the future. 5) The dual-strand and dual-MGB probe techniques are able to enhance the sensitivity of MethyDT test for methylation signals detection, but they are not applicable for all candidate markers. The dual-strand technique may be attempted when both sense and antisense strands are suitable for designing MSP primers, while the multiple MGB probe technique is limited by the amplicon length, which is usually less than 100 bp.

# Conclusion

In summary, the findings of this study reveal similar methylation profiles between adenoma and CRC, but the methylation patterns in adenomas show differences depending on the pathology (tubular, tubule-villous, or villous). The application of SADMP technique effectively enhanced the sensitivity of MethyDT test to detect methylation signals in plasma samples. The overall sensitivity and specificity of MethyDT test reached 85.83% and 90.28%, respectively, indicating its potential utility as a noninvasive CRC detection tool. However, it should be noted that many issues still need to be addressed in the future.

# Abbreviations

CRC: colorectal cancer; cfDNA, cell-free DNA; MSP: methylation-specific PCR; MGB: minor groove binder; TCGA: The Cancer Genome Atlas; SADMP: sense-antisense and dual-MGB probe; Ct: cycle threshold; MethyDT: methylated dual-target; ROC: receiver operating characteristic; AUC: area under the ROC curve.

# Declarations

## Ethics approval and consent to participate

This study is a subproject of the Clinical Study of Pan-cancer DNA Methylation Test in plasma (ClinicalTrials ID: NCT05685524), which has been approved by the ethics committee of the First Affiliated Hospital of Zhengzhou University (approval number: 2022-KY-0631-002). The study was conducted under the relevant guidelines and regulations of the Declaration of Helsinki. All participants signed an informed consent form and were told the test results.

### Consent for publication

Not Applicable.

### Availability of data and materials

The datasets (GSE77954, GSE101764, GSE131013, GSE164811, GSE193535, GSE129364, GSE139404, GSE107352, GSE75546, GSE77965, GSE199057, GSE68060, GSE48684, GSE40279 and GSE122126) supporting the conclusions of this article are available in the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). The E-MTAB-6450 data is available in EMBL biostudies (https://www.ebi.ac.uk/biostudies/arrayexpress). The source code and intermediate data can be accessed from git-hub (https://github.com/amsinfor/CRC-methylation).

### Competing interests

All authors declare no conflict of interest.

### Funding

This work was financially supported by Medical Science and Technology Research Plan Joint Construction Project of Henan Province (2018020121).

### Authors' contributions

SC W and Y B conceptualized and designed the study. YT Z developed the SADMP technique. ZJ W QN Y, and X L provided clinical samples and patient information. X L and SL D conducted cfDNA extraction and bisulfite converted. YT Z and ZJ W performed data analysis and prepared the manuscript. SC W and Y B reviewed the manuscript. XP L validated the study results.

### Acknowledgements

We thank Dr. Dihan Zhou and Dr. Kangkang wan for the assistance in MSP designing and data analysis in this work.

## References

1. Powrózek, T.; Krawczyk, P.; Kucharczyk, T.; Milanowski, J. Septin 9 promoter region methylation in free circulating dna—potential role in noninvasive diagnosis of lung cancer: preliminary report. *Med.*

*Oncol.* **2014,** *31*, 917.

2. Lamb, Y.N.; Dhillon, S. Epi procolon® 2.0 ce: a blood-based screening test for colorectal cancer. *Mol. Diagn. Ther.* **2017,** *21*, 225-232.
3. Sun, J.; Fei, F.; Zhang, M.; Li, Y.; Zhang, X.; Zhu, S.; Zhang, S. The role of msept9 in screening, diagnosis, and recurrence monitoring of colorectal cancer. *Bmc Cancer* **2019,** *19*, 450.
4. Tepus, M.; Yau, T.O. Non-invasive colorectal cancer screening: an overview. *Gastrointest. Tumors* **2020,** *7*, 62-73.
5. Li, B.; Huang, H.; Huang, R.; Zhang, W.; Zhou, G.; Wu, Z.; Lv, C.; Han, X.; Jiang, L.; Li, Y.; et al. Sept9 gene methylation as a noninvasive marker for hepatocellular carcinoma. *Dis. Markers* **2020,** *2020*, 6289063.
6. Cao, C.Q.; Chang, L.; Wu, Q. Circulating methylated septin 9 and ring finger protein 180 for noninvasive diagnosis of early gastric cancer. *Transl. Cancer Res.* **2020,** *9*, 7012-7021.
7. Jiao, X.; Zhang, S.; Jiao, J.; Zhang, T.; Qu, W.; Muloye, G.M.; Kong, B.; Zhang, Q.; Cui, B. Promoter methylation of sept9 as a potential biomarker for early detection of cervical cancer and its overexpression predicts radioresistance. *Clin. Epigenetics* **2019,** *11*, 120.
8. Matsui, S.; Kagara, N.; Mishima, C.; Naoi, Y.; Shimoda, M.; Shimomura, A.; Shimazu, K.; Kim, S.J.; Noguchi, S. Methylation of the sept9_v2 promoter as a novel marker for the detection of circulating tumor dna in breast cancer patients. *Oncol. Rep.* **2016,** *36*, 2225-2235.
9. Babayan, A.; Pantel, K. Advances in liquid biopsy approaches for early detection and monitoring of cancer. *Genome Med.* **2018,** *10*, 21.
10. Li, L.C.; Dahiya, R. Methprimer: designing primers for methylation pcrs. *Bioinformatics* **2002,** *18*, 1427-1431.
11. Jensen, S.Ø.; øgaard, N.; Nielsen, H.J.; Bramsen, J.B.; Andersen, C.L. Enhanced performance of dna methylation markers by simultaneous measurement of sense and antisense dna strands after cytosine conversion. *Clin. Chem.* **2020,** *66*, 925-933.
12. Faaborg, L.; Fredslund, A.R.; Waldstrøm, M.; Høgdall, E.; Høgdall, C.; Adimi, P.; Jakobsen, A.; Dahl, S.K. Analysis of hoxa9 methylated ctdna in ovarian cancer using sense-antisense measurement. *Clin. Chim. Acta* **2021,** *522*, 152-157.
13. Aryee, M.J.; Jaffe, A.E.; Corrada-Bravo, H.; Ladd-Acosta, C.; Feinberg, A.P.; Hansen, K.D.; Irizarry, R.A. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics* **2014,** *30*, 1363-1369.
14. Luo, Y.; Wong, C.J.; Kaz, A.M.; Dzieciatkowski, S.; Carter, K.T.; Morris, S.M.; Wang, J.; Willis, J.E.; Makar, K.W.; Ulrich, C.M.; et al. Differences in dna methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* **2014,** *147*, 418-429.
15. Hannum, G.; Guinney, J.; Zhao, L.; Zhang, L.; Hughes, G.; Sadda, S.; Klotzle, B.; Bibikova, M.; Fan, J.B.; Gao, Y.; et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **2013,** *49*, 359-367.

16. Moss, J.; Magenheim, J.; Neiman, D.; Zemmour, H.; Loyfer, N.; Korach, A.; Samet, Y.; Maoz, M.; Druid, H.; Arner, P.; et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease. *Nat. Commun.* **2018,** *9*, 5068.
17. Wang, Z.; Shang, J.; Zhang, G.; Kong, L.; Zhang, F.; Guo, Y.; Dou, Y.; Lin, J. Evaluating the clinical performance of a dual-target stool dna test for colorectal cancer detection. *J. Mol. Diagn.* **2022,** *24*, 131-143.
18. Imperiale, T.F.; Ransohoff, D.F.; Itzkowitz, S.H.; Levin, T.R.; Lavin, P.; Lidgard, G.P.; Ahlquist, D.A.; Berger, B.M. Multitarget stool dna testing for colorectal-cancer screening. *N. Engl. J. Med.* **2014,** *370*, 1287-1297.
19. Zhao, G.; Liu, X.; Liu, Y.; Li, H.; Ma, Y.; Li, S.; Zhu, Y.; Miao, J.; Xiong, S.; Fei, S.; et al. Aberrant dna methylation of sept9 and sdc2 in stool specimens as an integrated biomarker for colorectal cancer early detection. *Front. Genet.* **2020,** *11*, 643.
20. Oh, T.J.; Oh, H.I.; Seo, Y.Y.; Jeong, D.; Kim, C.; Kang, H.W.; Han, Y.D.; Chung, H.C.; Kim, N.K.; An, S. Feasibility of quantifying sdc2 methylation in stool dna for early detection of colorectal cancer. *Clin. Epigenetics* **2017,** *9*, 126.
21. Nassar, F.J.; Msheik, Z.S.; Nasr, R.R.; Temraz, S.N. Methylated circulating tumor dna as a biomarker for colorectal cancer diagnosis, prognosis, and prediction. *Clin. Epigenetics* **2021,** *13*, 111.
22. Luo, Y.; Wong, C.J.; Kaz, A.M.; Dzieciatkowski, S.; Carter, K.T.; Morris, S.M.; Wang, J.; Willis, J.E.; Makar, K.W.; Ulrich, C.M.; et al. Differences in dna methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* **2014,** *147*, 418-429.
23. Kakar, S.; Deng, G.; Cun, L.; Sahai, V.; Kim, Y.S. Cpg island methylation is frequently present in tubulovillous and villous adenomas and correlates with size, site, and villous component. *Hum. Pathol.* **2008,** *39*, 30-36.
24. Morris, V.K.; Strickler, J.H. Use of circulating cell-free dna to guide precision medicine in patients with colorectal cancer. *Annu. Rev. Med.* **2021,** *72*, 399-413.
25. Sanz-Garcia, E.; Zhao, E.; Bratman, S.V.; Siu, L.L. Monitoring and adapting cancer treatment using circulating tumor dna kinetics: current research, opportunities, and challenges. *Sci. Adv.* **2022,** *8*, i8618.
26. Roy, D.; Tiirikainen, M. Diagnostic power of dna methylation classifiers for early detection of cancer. *Trends Cancer* **2020,** *6*, 78-81.
27. Bronkhorst, A.J.; Ungerer, V.; Holdenrieder, S. The emerging role of cell-free dna as a molecular marker for cancer management. *Biomol Detect Quantif* **2019,** *17*, 100087.
28. Zhao, G.; Li, H.; Yang, Z.; Wang, Z.; Xu, M.; Xiong, S.; Li, S.; Wu, X.; Liu, X.; Wang, Z.; et al. Multiplex methylated dna testing in plasma with high sensitivity and specificity for colorectal cancer screening. *Cancer Med.* **2019,** *8*, 5619-5628.
29. Bagheri, H.; Mosallaei, M.; Bagherpour, B.; Khosravi, S.; Salehi, A.R.; Salehi, R. Tfpi2 and ndrg4 gene promoter methylation analysis in peripheral blood mononuclear cells are novel epigenetic noninvasive biomarkers for colorectal cancer diagnosis. *J. Gene. Med.* **2020,** *22*, e3189.

30. Ladabaum, U.; Dominitz, J.A.; Kahi, C.; Schoen, R.E. Strategies for colorectal cancer screening. *Gastroenterology* **2020,** *158*, 418-432.
31. Li, R.; Qu, B.; Wan, K.; Lu, C.; Li, T.; Zhou, F.; Lin, J. Identification of two methylated fragments of an sdc2 cpg island using a sliding window technique for early detection of colorectal cancer. *Febs Open Bio* **2021,** *11*, 1941-1952.
32. Zhang, L.; Dong, L.; Lu, C.; Huang, W.; Yang, C.; Wang, Q.; Wang, Q.; Lei, R.; Sun, R.; Wan, K.; et al. Methylation of sdc2/tfpi2 and its diagnostic value in colorectal tumorous lesions. *Front. Mol. Biosci.***2021,***8*, 706754.

## Figures

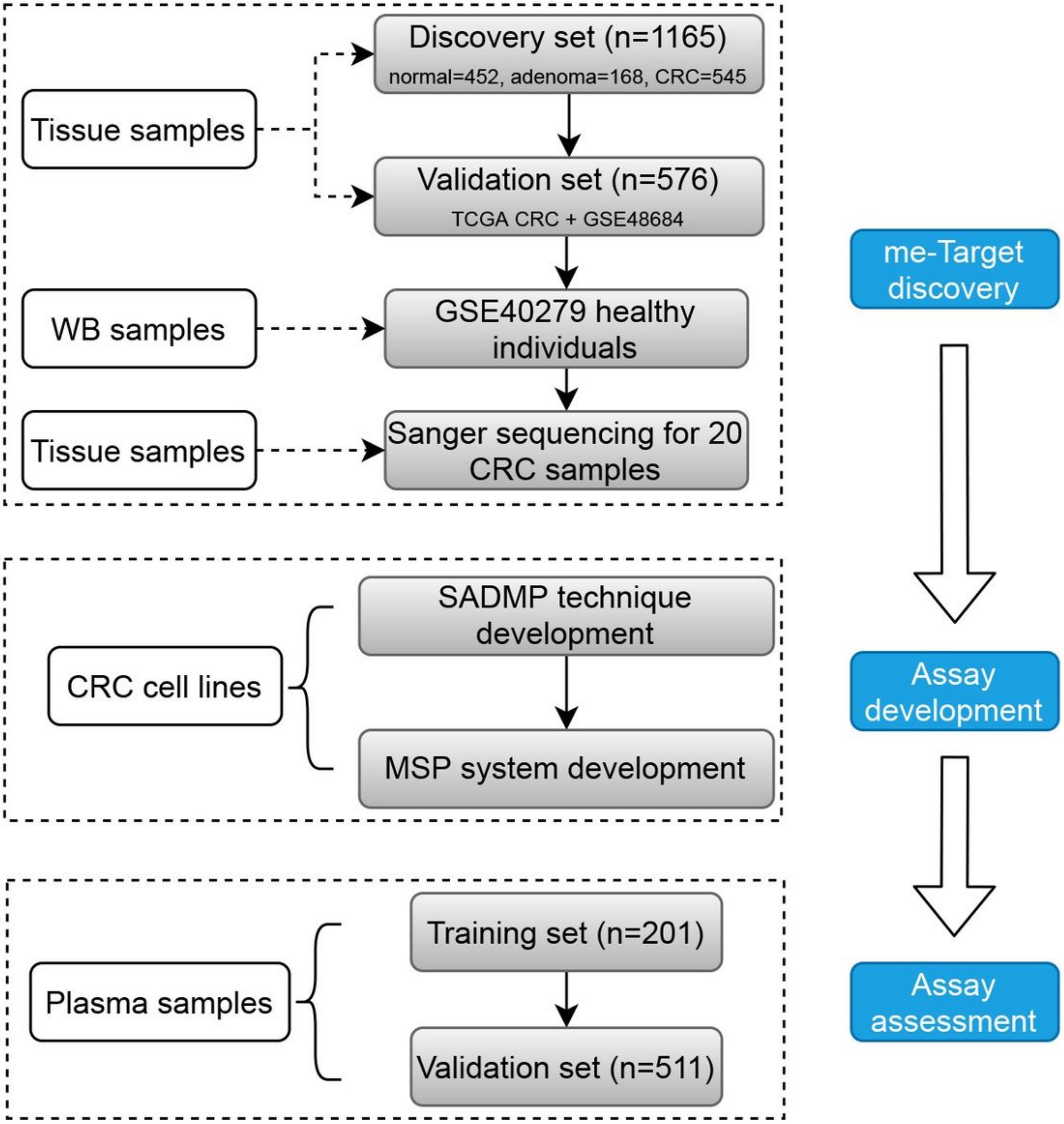
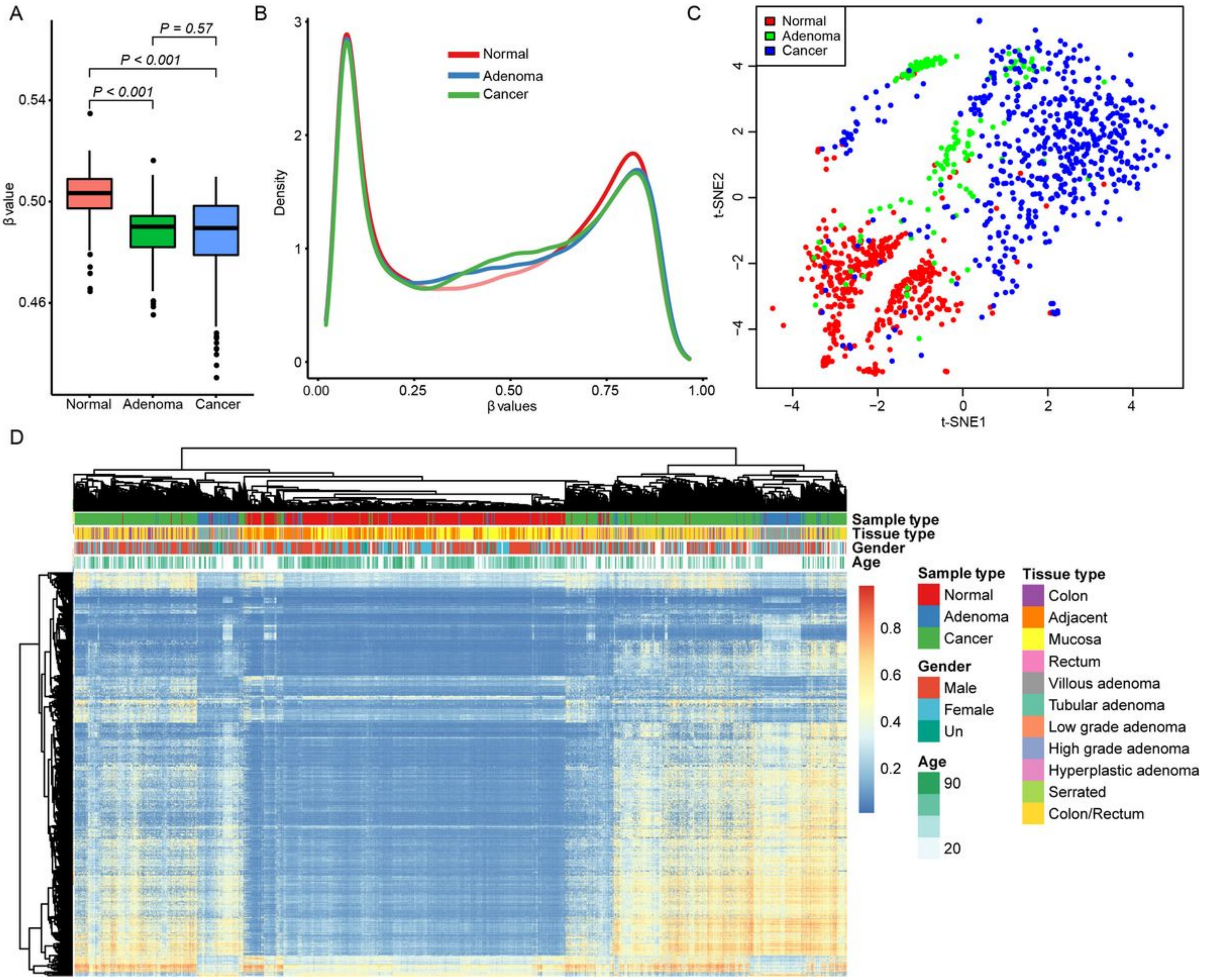


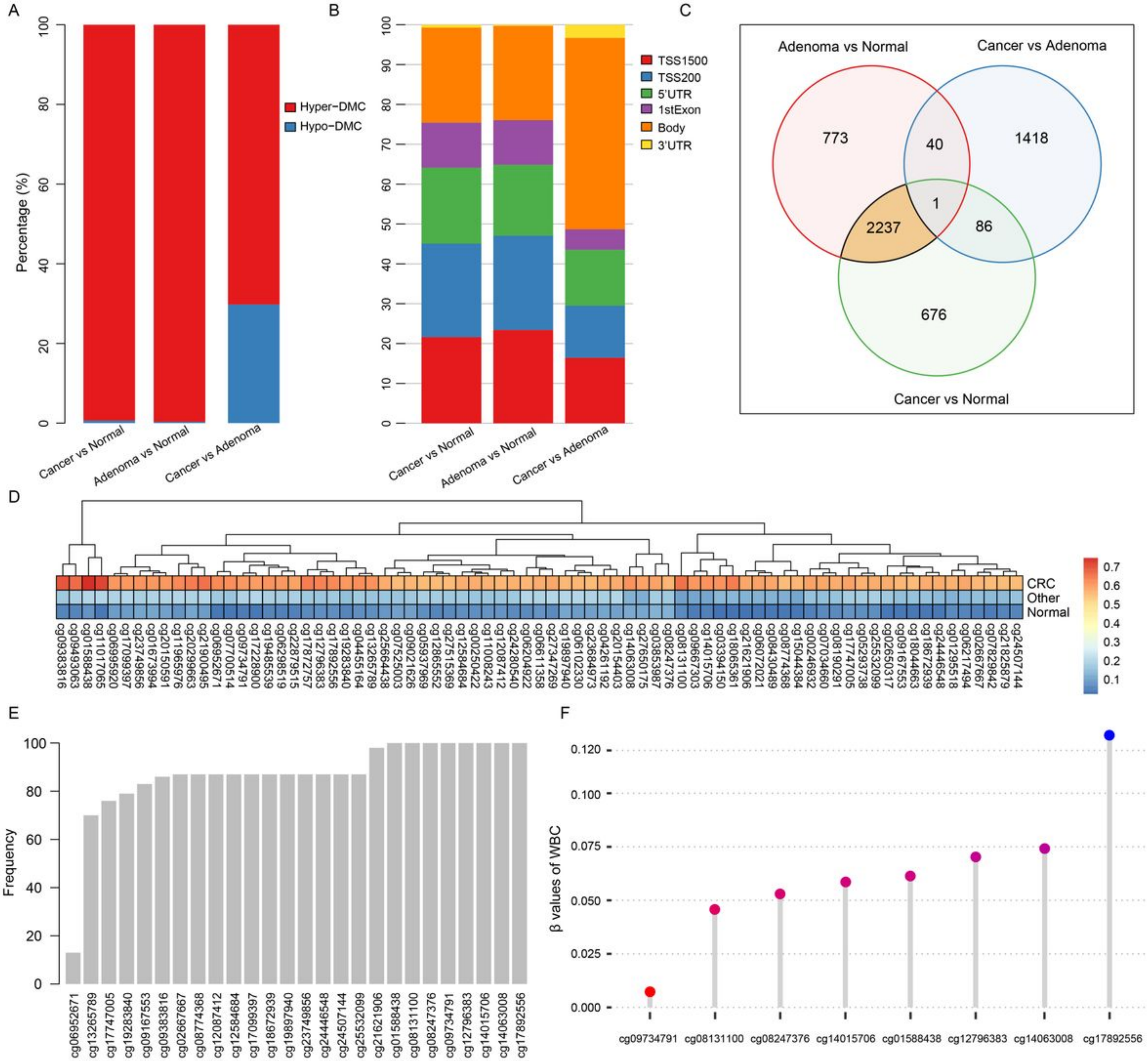
**Figure 1**

**The flowchart of this study.** The three steps were separated by three dashed boxes. CRC: colorectal cancer, WBC: whole blood cell, MSP: methylation-specific PCR. The CRC cell lines used in this study were Hacat and HT-29 as we described in previous studies [31,32].



## Figure 2

**Landscape of the methylation patterns of the discovery set. A**: boxplot showing the overall methylation levels of normal, adenoma and cancer samples. The average β value of all probes for each sample was calculated as the sample overall methylation level. P-values were estimated by Kruskal test. **B**: Density curves of probe methylation levels in normal, adenoma and cancer samples. **C**: t-SNE visualizing normal, adenoma and cancer samples in the discovery set. **D**: Heatmap showing the most variable probes between normal, adenoma and cancer samples.



## Figure 3

**Identification of candidate markers. A**: Percentage of hyper-DMC and hypo-DMC between the three comparisons. **B**: Percentage of DMC at different genomic regions between the three comparisons. **C**: Venn diagram showing DMCs between the three comparisons. **D**: Methylation values of 75 probes meeting the criteria on TCGA 31 cancer types. The other refers to 30 non-CRC cancer samples. Normal refers to 710 NATs samples. **E**: Frequency of probes with non-zero coefficient in 100 LASSO regressions. **F**: The average methylation levels of 8 probes on WBC samples.

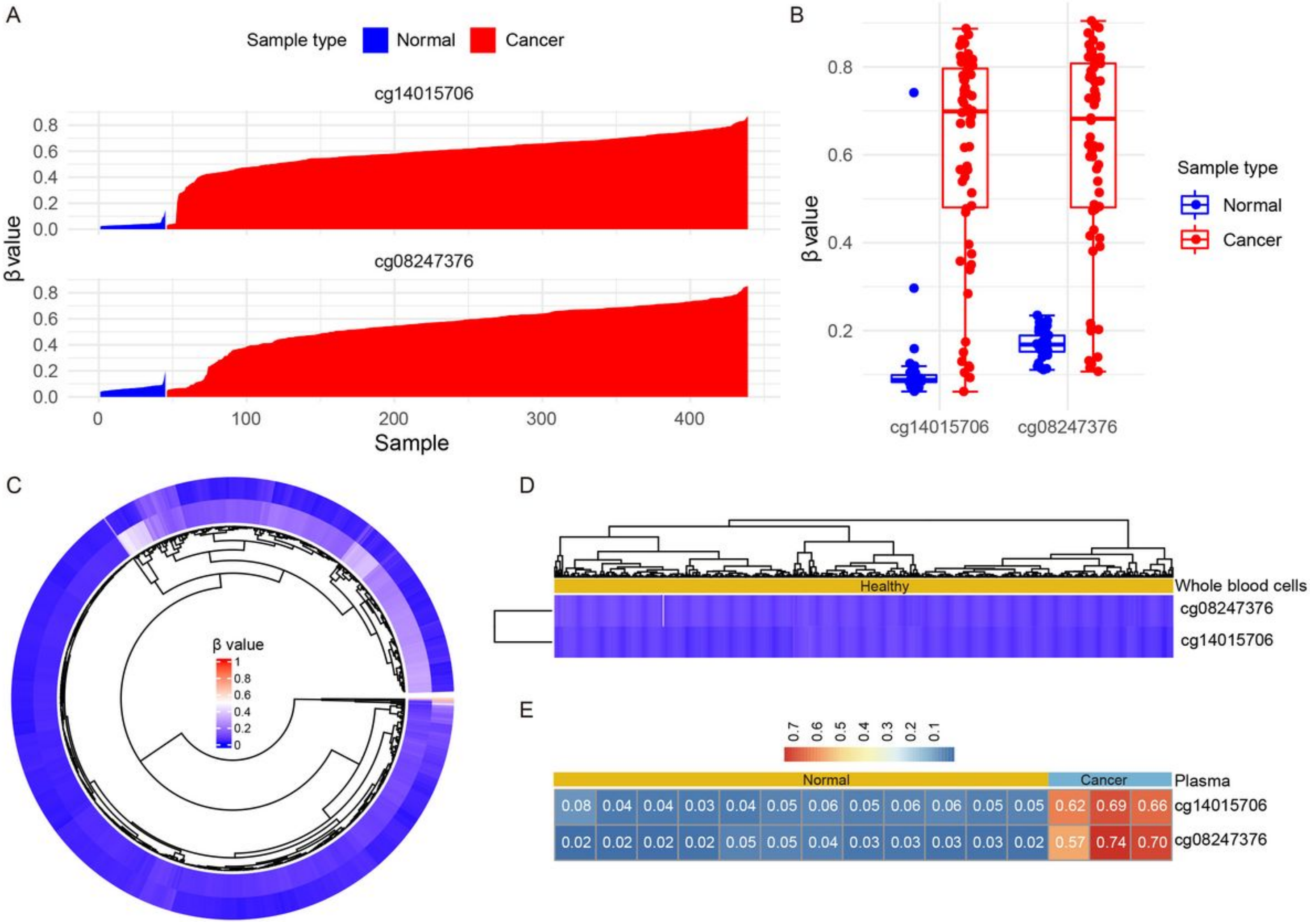
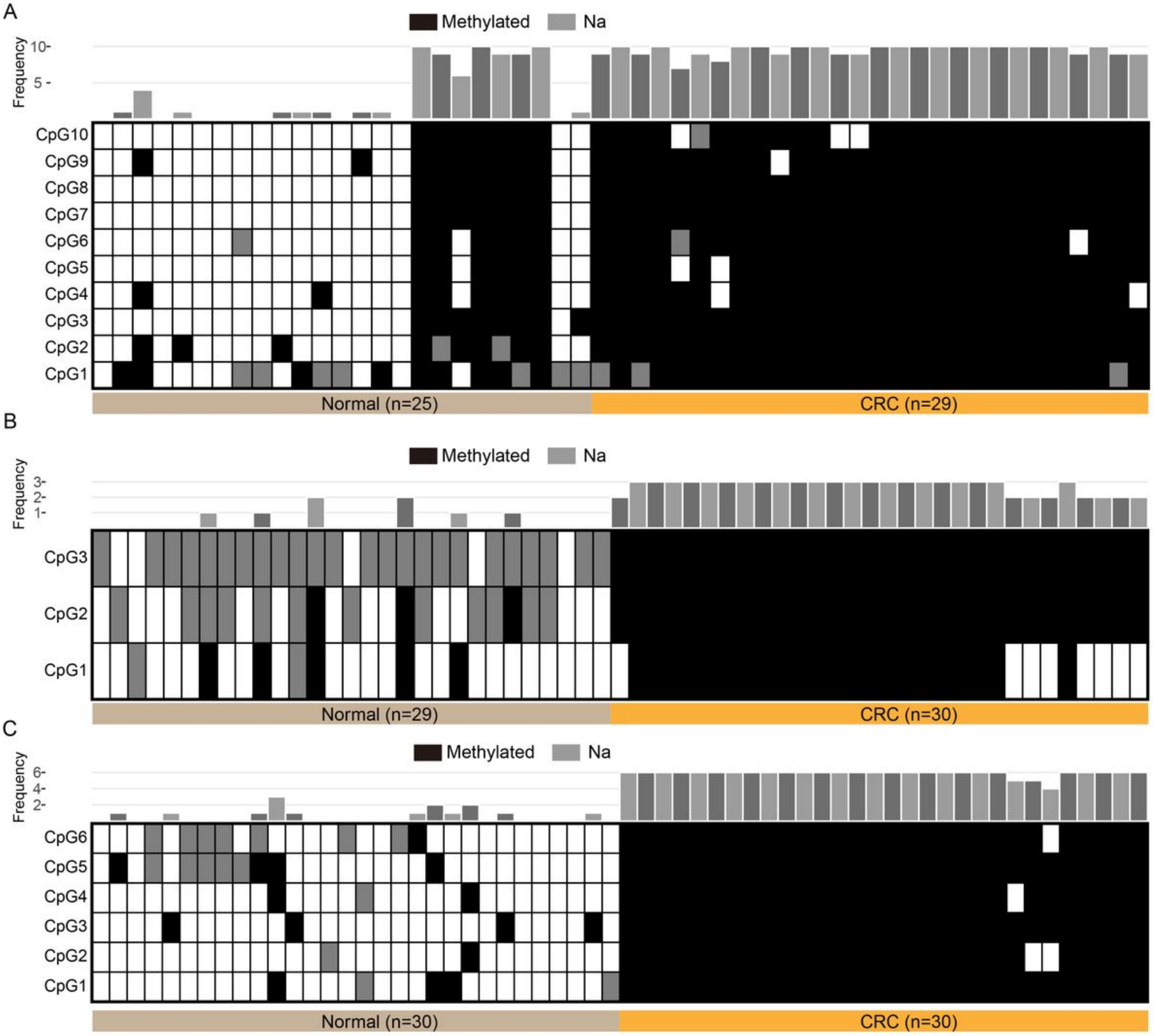


**Figure 4**

**Methylation profiles of the two candidate probes in different datasets. A:** The methylation profiles of cg14015706 and cg08247376 in normal and cancer samples in TCGA dataset. **B:** The methylationβ values of cg14015706 and cg08247376 between normal and cancer samples in GSE48684 dataset. **C:** The methylation profiles of cg14015706 and cg08247376 in 710 adjacent normal samples. The outer circle and inner circle indicated cg14015706 and cg08247376, respectively. **D:** The methylation profiles of cg14015706 and cg08247376 in whole blood cells collected from healthy individuals. **E:** The cfDNA methylation profiles of cg14015706 and cg08247376 in normal and cancer plasma samples. Numbers in the heatmap indicated the methylation β values.



## Figure 5

**The methylation status of candidate markers verified by Sanger sequencing. A-C** showing the methylation status of each CpG site in *MTNT1* sense (**A**), antisense (**B**) and *MAP3K14-AS1* antisense (**C**) amplicons between normal and cancer tissues. The three amplicons covered 10, 3 and 6 key CpGs, respectively. Methylated CpGs were presented in black cells. Gray cells indicated the methylation status were not determined due to failed sequencing.

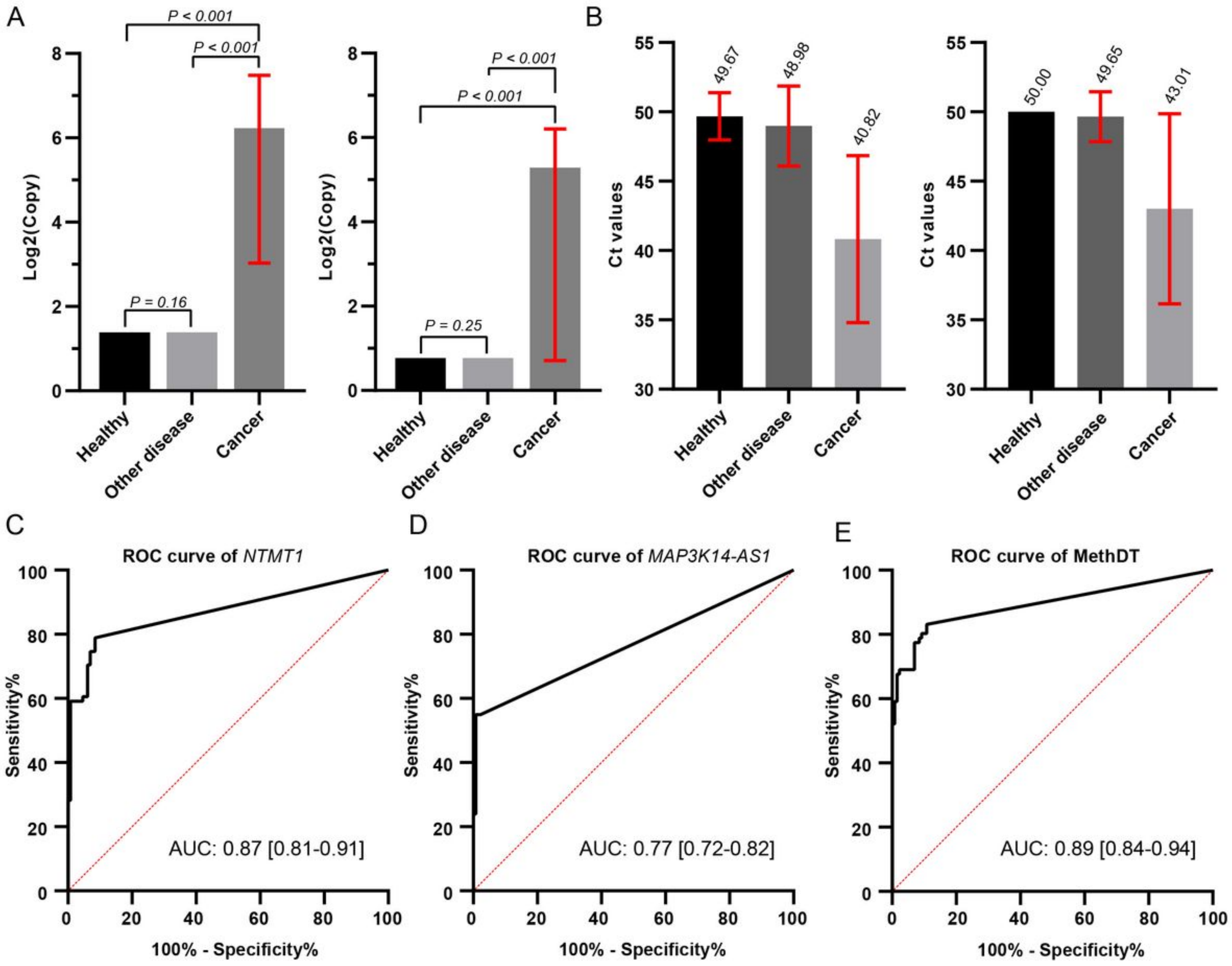


Figure 6

**The performance of candidate markers in training set. A**: The estimated copies of *NTMT1* and *MAP3K14-AS1* in different sample types. **B**: The MSP Ct values of *NTMT1* and *MAP3K14-AS1* in different sample types. **C-E**: ROC curves of *NTMT1* (**C**), *MAP3K14-AS1* (**D**) and MethyDT (**E**) tests. Error bars in **A** and **B** indicated 1$^{st}$ and 3$^{rd}$ quantiles.

## Supplementary Files

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

- supplementarymaterials.docx