SUPPLEMENTARY MATERIAL

Exploring the DNA methylome of Korean patients with colorectal cancer consolidates the clinical implications of cancer-associated methylation markers

Author names: Sejoon Lee[*1], Kil-yong Lee[*2], Ji-Hwan Park[*3,4], Duck-Woo Kim[5], Heung-Kwon Oh[5], Seong-Taek Oh[2], Jongbum Jeon[3], Dongyoon Lee[3], Soobok Joe[3], Hoang Bao Khanh Chu[6], Jisun Kang[6], Jin-Young Lee[6], Sheehyun Cho[6], Hyeran Shim[6], Si-Cho Kim[6], Hong Seok Lee[6], Young-Joon Kim[6], Jin Ok Yang[†3], Jaeim Lee[†2], Sung-Bum Kang[†5]

[*]These authors contributed equally to this work.

†corresponding author

Affiliation:

[1]Precision Medicine Center, Seoul National University Bundang Hospital, Seongnam 13620, Republic of Korea

[2]Department of Surgery, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Uijeongbu 11765, Republic of Korea

[3]Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea

[4]Department of Bioscience, University of Science and Technology (UST), Daejeon 34113, Republic of Korea

[5]Department of Surgery, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam 13620, Republic of Korea

[6]Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, Republic of Korea

Corresponding Author's Information

Sung-Bum Kang, Tel.: +82-31-787-7087; Fax: +82-31-787-4077; E-mail:

kangsb@snubh.org
Jaeim Lee, Tel.: +82-31-820-5020; Fax: +82-31-847-2041; E-mail:
lji96@catholic.ac.kr
Jin Ok Yang, Tel: +82-42-879-8550; Fax: +82-42-879-8519; E-mail:
joy@kribb.re.kr

This file includes:

Supplementary Results
Supplementary Materials and Methods
Supplementary Figures l to 13
Supplemental Tables 1 to 8
: Supplemental Table 2 is provided in additional Excel spreadsheet.
Supplementary References

Supplementary Results

Identification of potential CRC diagnostic markers commonly used for diverse ethnic groups

To ensure the reliability of our data, we examined the similarity in the differential methylation patterns with previously established public CRC methylome. To this end, we conducted a comparative analysis of the Korean methylation profiles and TCGA methylation profiles of the patients (consisting of 404 tumor samples and 45 normal samples) assigned as colon adenocarcinoma (COAD) or rectum adenocarcinoma (READ). For the 15,968 probes (6,244 hypermethylaed and 9,724 hypomethylated positions), which were included in both the Illumina Infinium EPIC array and Illumina Infinium Human DNA Methylation 450K BeadChip (TCGA 450K) array platforms, we observed the analogous differential methylation patterns in the TCGA CRC dataset (**Supplementary Figure 10A**; See the details in **Supplementary Materials and Methods**). Additionally, when we compared the mean methylation differences for a total of 298,581 probes included in the both array platforms, we could also observe a robust correlation between the two datasets (**Supplementary Figure 10B**; Pearson's correlation coefficient: 0.948, $p < 0.0001$). All these results demonstrated the reliability of the Korean CRC methylome, mitigating potential biases introduced by variations in array platforms. Moreover, our findings from this methylome dataset could be expanded to the patients from other ethnic groups.

Based on the similarity between the two methylome datasets, we tried to identify potential CRC diagnostic markers, which could be used for other ethnic groups, rather than Korean ethnicity. To this end, we first selected 15,968 DMPs that were included in the both array platforms. By applying a Lasso regularization with a logistic function to the selected DMPs, we prioritized 21 key methylation markers (10 hypermethylated and 11 hypomethylated markers, **Supplementary Figure 11**) which enabled the classification of the tumor samples from adjacent normal tissue samples in Korean patients with CRC (See the details in **Supplementary Materials and Methods**). After constructing a prediction model for CRC with the methylation levels of these 21 positions, we confirmed the methylation patterns of the markers and tested the robustness and reproducibility of the model on the TCGA CRC dataset. Notably, the 10 hypermethylated and 11 hypomethylated markers showed similar

hyper- and hypomethylation patterns in TCGA CRC dataset, respectively (**Supplementary Figure 12A and Supplementary Table 8**). Moreover, the test of the model on TCGA CRC dataset yielded impressive predictive metrics: precision at 0.995, recall at 0.963, an overall accuracy of 0.962, and an area under the curve (AUC) of 0.960 (**Supplementary Figure 12B, C**). As an orthogonal validation of these markers, we also confirmed the 3 hypermethylation and 4 hypomethylation patterns from another studies, which conducted whole-genome bisulfite sequencing of CRC samples (**Supplementary Figure 13**) (1-4).

Supplementary Materials and Methods

Clinical specimens from CRC patients

In this study, we performed methylome profiling of the tumor and adjacent normal tissues from Korean patients with CRC. The two hospital datasets used in this study comprised 344 samples from the Seoul National University Bundang Hospital (BUNDANG; 165 samples) and The Catholic University Uijeong St. Mary's Hospital (SUNGMO; 179 samples). Of these, 235 tumor samples were from BUNDANG (130) and SUNGMO (105) and 109 normal samples were from BUNDANG (35) and SUNGMO (74).

Methylation microarray analysis

Genomic DNA (gDNA) was isolated from the tumor and adjacent normal tissues using the PureLink™ Genomic DNA Mini Kit (Invitrogen, Waltham, MA, USA), and its quality was checked using a NanoDrop® (ND-2000, Waltham, MA, USA) and agarose gel electrophoresis (1% gel; run conducted at 100 V for 30 min). Intact gDNA was diluted to 50 ng/μl based on Quant-iT Picogreen (Invitrogen, Waltham, MA, USA) quantitation and subjected to bisulfite conversion using the EZ DNA Methylation Kit (ZymoResearch, USA). Subsequently, the converted gDNA was amplified up to 1,000-fold through whole-genome amplification and then hybridized to the Infinium MethylationEPIC BeadChip (V1; WG-317-1001, Illumina, San Diego, CA, USA) following the manufacturer's recommended protocol. After completing the single-base extension in the Te-Flow chamber, the BeadChip was imaged using the iScan System (SY-101-1001, Illumina, San Diego, CA, USA) to produce raw data in the IDAT format.

Preprocessing the raw data by normalization, batch correction, and probe filtration

The EPIC array dataset was processed using the *minfi(v1.36)* pipeline (5). Initially, the raw intensities of 865,859 probes were extracted from the Cy3-green and Cy5-red channels of the raw .IDAT files. We evaluated the quality of the methylome data by inspecting the overall distribution of beta values and control strip plots, which

included the bisulfite conversion efficiency, extension quality, and specificity (**Supplementary Figure 2**). We then applied subset-quantile within array normalization (SWAN) (6) to correct technical discrepancies between type I and type II probes within each array. Next, we addressed the known batch effects specific to each EPIC array batch type by using the surrogate variable analysis (SVA) tool in conjunction with the *combat* method (7), followed by the removal of the 1,049 probes with the high batch bias. For downstream analysis, we filtered out sex-mismatched samples (11 samples) and excluded additional probes based on several dependencies for the further analysis. The excluded probes were methylation data of sex chromosomes (19,179 probes), known single nucleotide polymorphism (SNP) sites (161,078 probes) according to the genome annotations of the EPIC array, and poor-performing sites (1,881 probes) with their p values of probe detection ratio 〉 0.01. Additionally, for each probe, we calculated the difference between maximum and minimum beta values across all samples and excluded the 92,600 probes with the absolute differences 〈 0.1 for further analysis. Finally, 609,046 probe methylation beta values from 228 tumor and 105 normal samples (**Supplementary Table 1** and **Supplementary Figure 1**) were used for downstream analysis. Of note, 103 tumor and normal samples were obtained from the same patients. In this process, we compared the distribution of beta values between the raw and processed probes via principal component (PC) analysis (PCA), which revealed sex- and batch-related biases in the raw data (**Supplementary Figure 3** and **4**).

Identification of DMPs

To identify differentially methylated positions (DMPs) between the tumor and normal samples, we applied an *F*-test by using the *dmpfinder* function (8) from the *minfi* package (5). The *p* values of the *F*-test were adjusted to *q*-values by using Benjamini-Hochberg (9) procedure. We identified the DMPs as the ones with i) the absolute difference in the mean beta values between the tumor and normal samples 〉 0.15 and ii) the *q*-values 〈 1×10^{-6}. The hyper- and hypomethylated positions in tumors were determined as the DMPs with the difference in the mean beta values 〉 0.15 and 〈 -0.15, respectively. For annotation of genomic regions, we used EPIC array manual 1.05B (https://support.illumina.com/array/array_kits/infinium-

methylationepic-beadchip-kit/downloads.html) (TSS1500: 1500 base pairs to 200 base pairs upstream of the transcription start site [TSS]; TSS200: 200 base pairs upstream of the TSS to the TSS; Shore: 2 kb from each end of the island, Shelf: from 2 to 4 kb from the CpG island; Open sea: outside of CpG islands, shores, and shelves).

We then performed enrichment analysis for each genomic annotation (*e.g.*, CpG island and open-sea regions and TSS1500 and first exon regions in **Figure 1E**) and by calculating the odds ratio for hyper- and hypomethylated DMPs. To compute the enrichment significance, we estimated an empirical null distribution of the odds ratio by performing random sampling experiments 10,000 times. Briefly, in each experiment, probes with sizes that were same as those of the hyper- or hypomethylated positions were randomly sampled, and the odds ratio was measured. For each genomic annotation, the p values for the odds ratio were calculated using the empirical distributions by the one-tailed test.

Functional enrichment analysis of GOBPs and KEGG pathways

The enrichment analysis was performed for the hyper- and hypomethylated positions annotated with genomic regions (at TSS1500, TSS200, 5'-UTR, first exon, body, and 3'-UTR) by using DAVID software (10). For each genomic region, we first obtained the GenBank accession IDs linked to the individual DMPs. The GOBPs from GOBP FAT and KEGG pathways represented by the accession IDs were identified as the ones with the enrichment $p < 0.05$ and the number of genes > 4. Moreover, to further examine the effectiveness of the enrichment p values of the hyper- and hypomethylated positions, we selected a set of negative control positions at each region, which were not differentially methylated, as the positions with i) the absolute difference in the mean beta values between the tumor and normal samples < 0.05 and ii) the p values > 0.9. We also calculated the enrichment p values of the GOBPs and KEGG pathways for the negative control positions at each genomic region. For visualization in the heat map, the enrichment p value was converted into a Z-score by $Z = N^{-1}(1 - p)$, where $N^{-1}(\cdot)$ denotes the inverse standard normal distribution.

Comparative analysis of the methylome profiles between Korean CRC and TCGA CRC

Among a total of the 609,046 probes in the Illumina Infinium EPIC array, 298,581 probes were also included in Illumina Infinium Human DNA Methylation 450K BeadChip (TCGA 450K) array platform, which was used for TCGA CRC cohort (consisting of 404 tumor samples and 45 normal samples). For these overlapped probes, we computed the mean differences of the methylation levels between tumor and adjacent normal tissues in the TCGA dataset. We then assessed the similarity of the mean differences between two CRC cohorts by calculating Pearson's correlation coefficient. Similarly, among the 38,607 DMPs identified from the Korean CRC methylome, we found that 15,968 probes were included in the TCGA 450K array and then also measured the similarity of the mean differences of methylation levels.

Identification of methylation markers for a predictive modeling of CRC diagnosis

To construct a predictive model for CRC diagnosis, we selected the methylation markers from the 15,968 DMPs, which were the probes included in the TCGA 450K array, by applying a feature selection methodology based on Lasso regularization (11) coupled with a logistic regression function. Briefly, by using all the beta values of the DMPs from tumor and adjacent normal tissues of Korean patients with CRC, we iteratively ran Lasso modeling 200 times. Among the 15,968 DMPs, we determined 21 probes, which had non-zero coefficients in at least 50% of 200 runs, as the methylation markers used for the prediction of the disease. Subsequently, a new logistic regression model was developed by using the beta values of the 21 methylation markers in the methylome data and clinical information of the Korean cohort to predict the occurrence of CRC. To evaluate the robustness and reproducibility of the constructed prediction model, we applied the model into the independent dataset, namely the TCGA CRC methylome.

Clustering of the tumor samples based on the CIMP markers

Among the CIMP probe set (4,327 probes) derived from 258 previously identified CIMP gene markers (12), we selected 1,470 highly variable sites with their absolute

value of standard deviation 〉 0.15. For 228 tumor samples, we performed K-means clustering 100 times on the beta values of the selected CIMP marker probes. The tumor samples were categorized into three groups, and each group was classified as CIMP-H, CIMP-L, or non-CIMP based on the respective mean methylation level for each group.

Association and enrichment analysis of clinicopathological characteristics with the CIMP status

To investigate the associations between CIMP status and clinicopathological characteristics, we employed various statistical methods tailored to the type and distribution of the data. For categorical clinical variables, such as sex, and, location, a Chi-square test was performed to assess the independence between CIMP status and the variables, excluding the AJCC stage, T-stage, differentiation, and MSI status. Since the four variables had at least one categories with fewer than five samples, we performed a Fisher's exact test. For the clinicopathological variables with the p values of the significance 〈 0.05, we further conducted an enrichment analysis of clinicopathological characteristics for CIMP status, by calculating the expected frequencies, and standardized residuals from a contingency table. We determined the significantly enriched clinicopathological variables (i.e., when the observed frequency significantly exceeded the expected frequency) as the ones with their p values of the enrichment significance 〈 0.05 and standardized residuals 〉 1.5.

Regarding continuous variables, we examined the significance of the mean differences of age, CIMP markers, and *MLH1* methylation levels across the three CIMP groups by applying an analysis of variance (ANOVA) with Sidak correction as a post hoc test (13). For the relapse-free survival analysis, used log-rank test (14).

Identification of novel CIMP marker candidates from the Korean CRC methylation profiles

To identify the novel CIMP marker candidates, we selected 680 probes from the 7,824 hypermethylated positions in the tumor samples according to the following

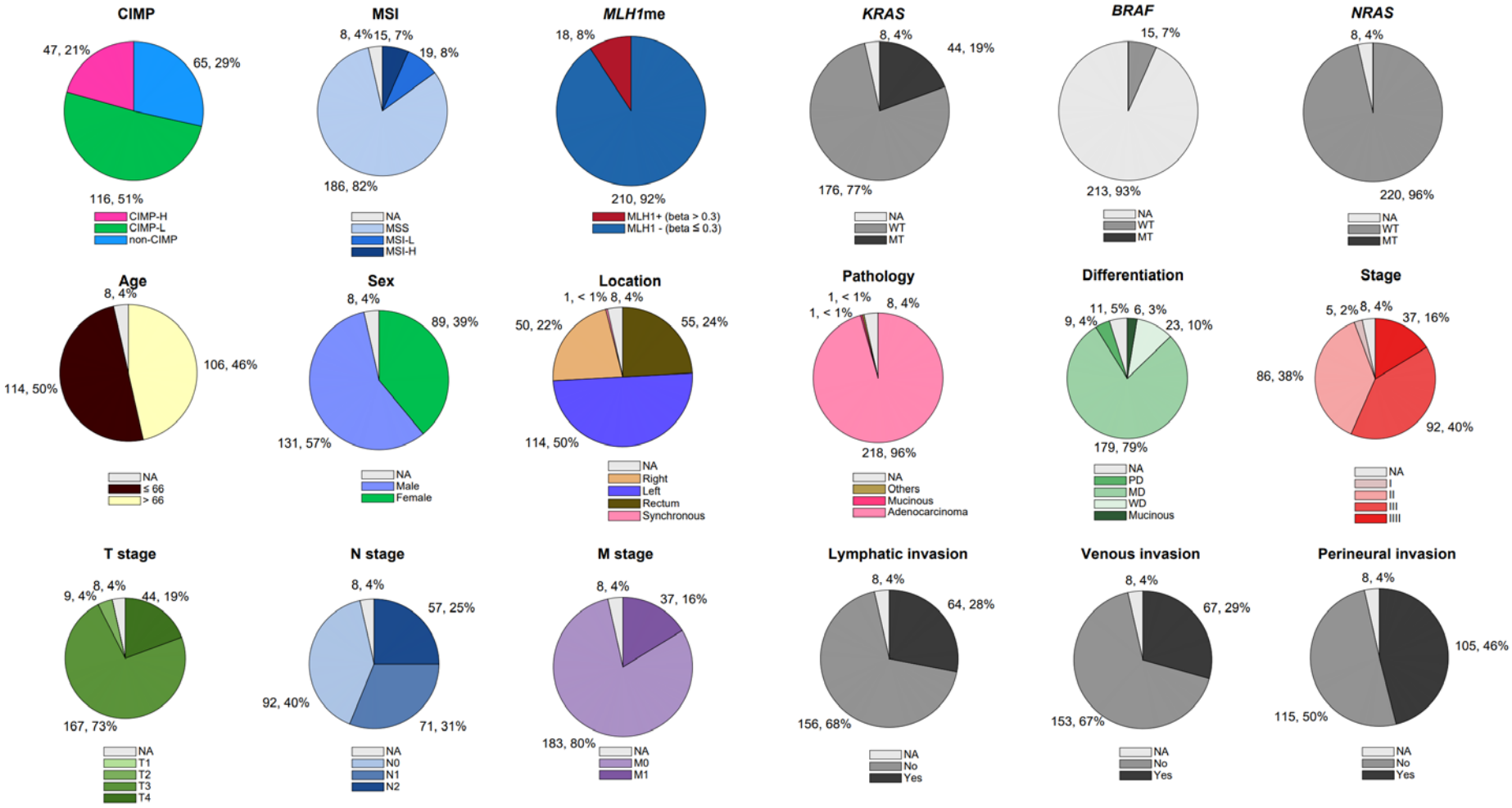
criteria: high variability in the methylation levels (standard deviation 〉 0.2) and annotations to CpG island region. To test whether the selected probes show the similar stratification performance to that of the CIMP markers, we performed K-means clustering on the beta values of the selected probes, and obtained three clusters (C1 - C3). We measured the similarity between the stratification of the Korean patients by using the selected probes and the CIMP stratification by calculating how many patients in CIMP-High, CIMP-Low, and non-CIMP were belonged to each of C1 - C3.

Finally, we determined 16 novel CIMP marker candidates from the 680 probes as the probes with their mean differences of methylation levels 〉 0.2 and p values of a pairwise *T*-test 〈 0.0001 in the following comparisons: (i) CIMP-H versus CIMP-L groups and (ii) CIMP-L versus non-CIMP groups.

Calculation of methylation levels of the promoter-like region

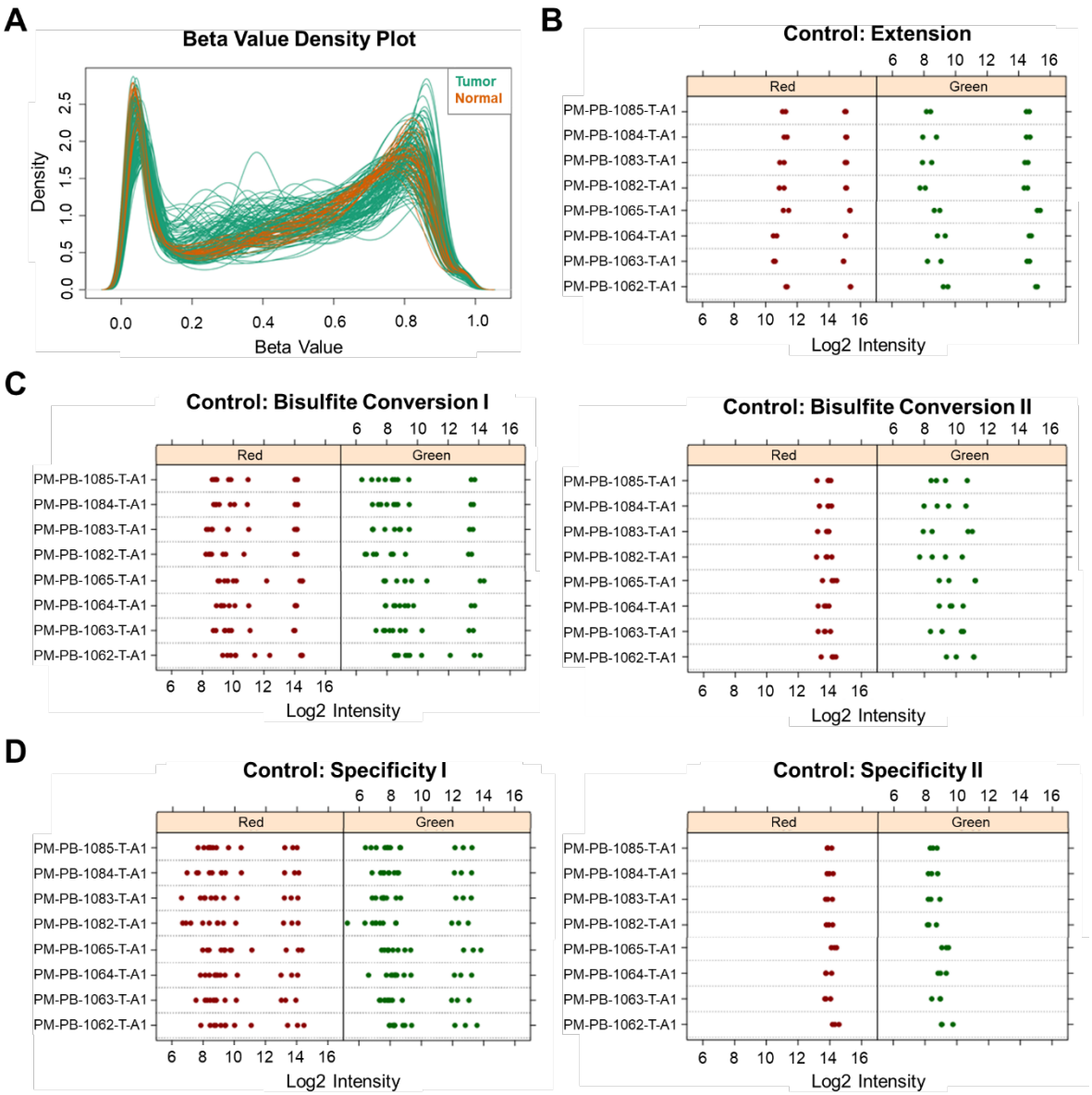
For each gene, we computed the methylation levels of the promoter-like region by averaging the beta values of all the probes annotated as the promoter-like regions (TSS1500, TSS200, 5' UTR, and first exon).

Supplementary Figures and Tables



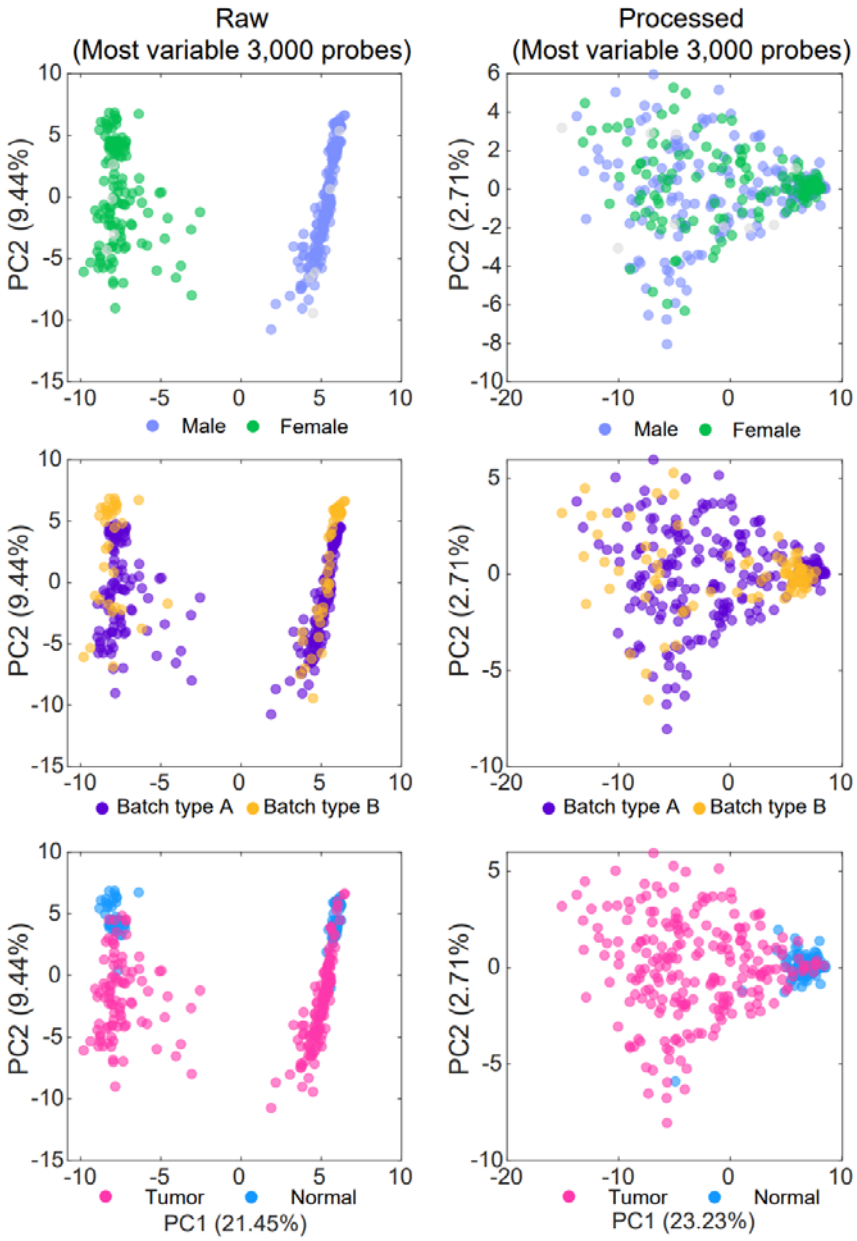
Supplementary Figure 1. Clinicopathological characteristics of the 228 patients with CRC. For each of the 18 clinical characteristics, distribution of 228 Korean CRC patients are shown as a pie chart including the proportion of not applicable

information. For the individual categories of each characteristics, the numbers and percentages of patients are described in the chart. For location, right-sided locations include ascending, cecum, hepatic flexure, transverse, and left-sided locations include descending, rectosigmoid, sigmoid, and splenic flexure. CIMP: 5'-C-phosphate-G-3' island methylator phenotype. MSI: microsatellite instability; MSS: microsatellite stability; MSI-H: high microsatellite instability; MSI-L: low microsatellite instability; MT: mutation; WT: wild type; WD: well-differentiated; MD: moderately differentiated; PD: poorly differentiated; Yes: Presence of cancer cells in lymph vessels or in blood vessels or surrounding nerves; No: Absence of cancer cells in lymph vessels or in blood vessels or surrounding nerves; NA, not applicable.

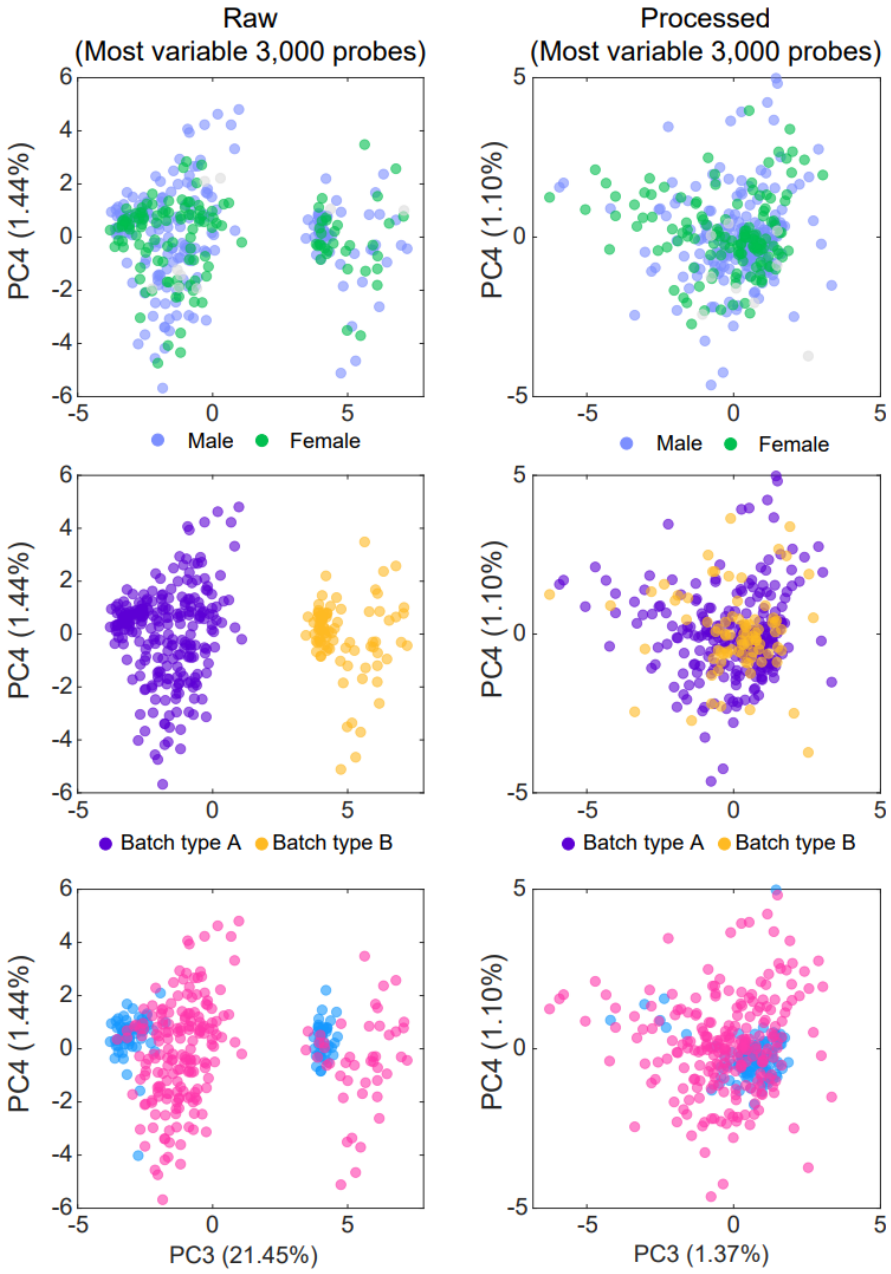


Supplementary Figure 2. Density plot of methylation beta values and control strip plots.

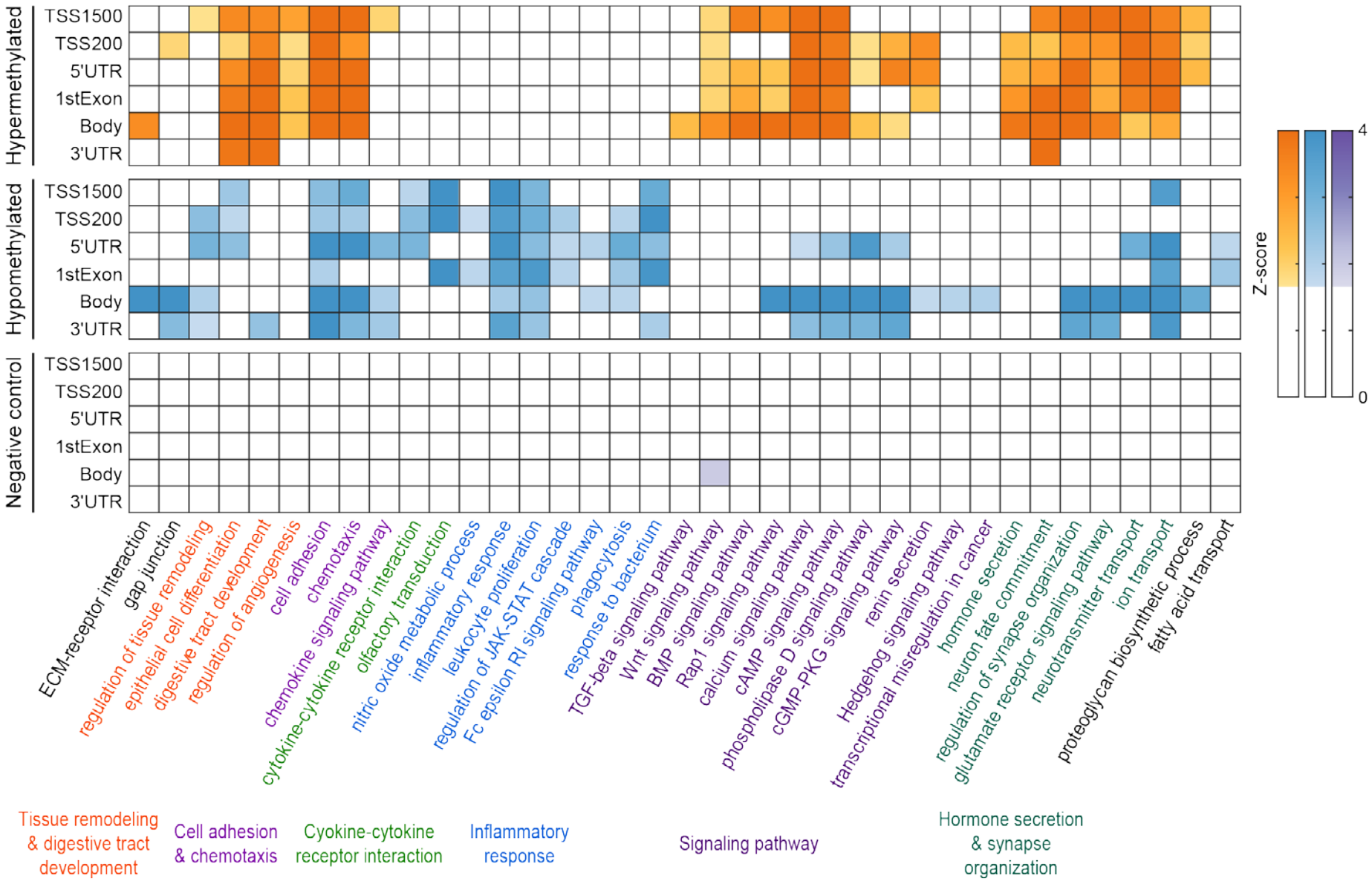
A. Density plot of the methylation beta values from individual samples (orange: normal samples; green: tumor samples). B - D. Examples of control strip plots representing extension efficiency (B), bisulfite conversion efficiency (C), and specificity (D).



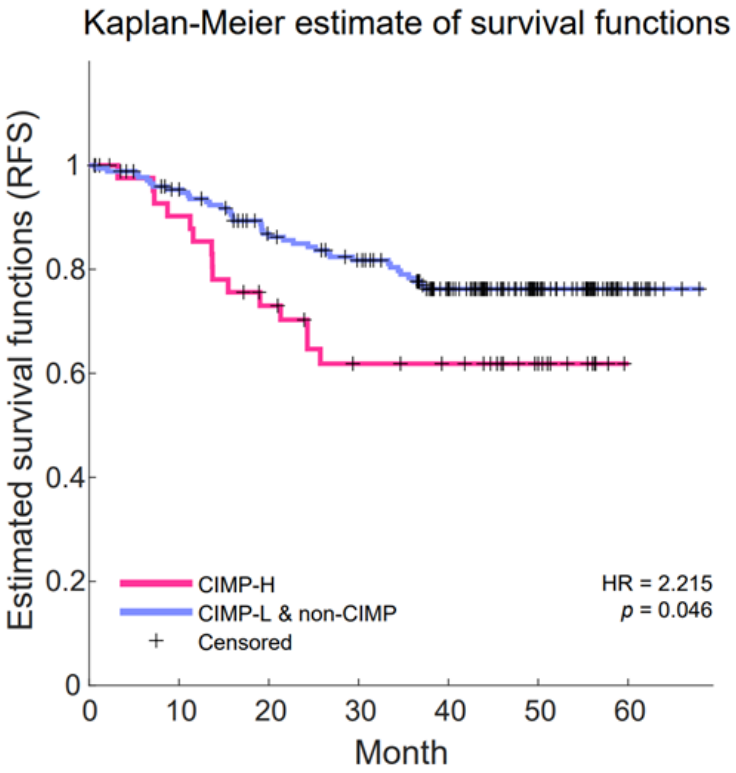
Supplementary Figure 3. Comparison of the methylome dataset before and after the bias correction. Among the total probes, we used the 3,000 probes with the largest variance of beta values across all samples for a principal component (PC) analysis of raw (before bias correction; left) and processed (after bias correction; right) datasets. The plots show the PC1 (x-axis) and PC2 (y-axis) with their explained variances. The individual samples in the plots were labeled according to sex (top: male and female), batch number (middle: batch types), and tumor status (bottom: tumor and normal).



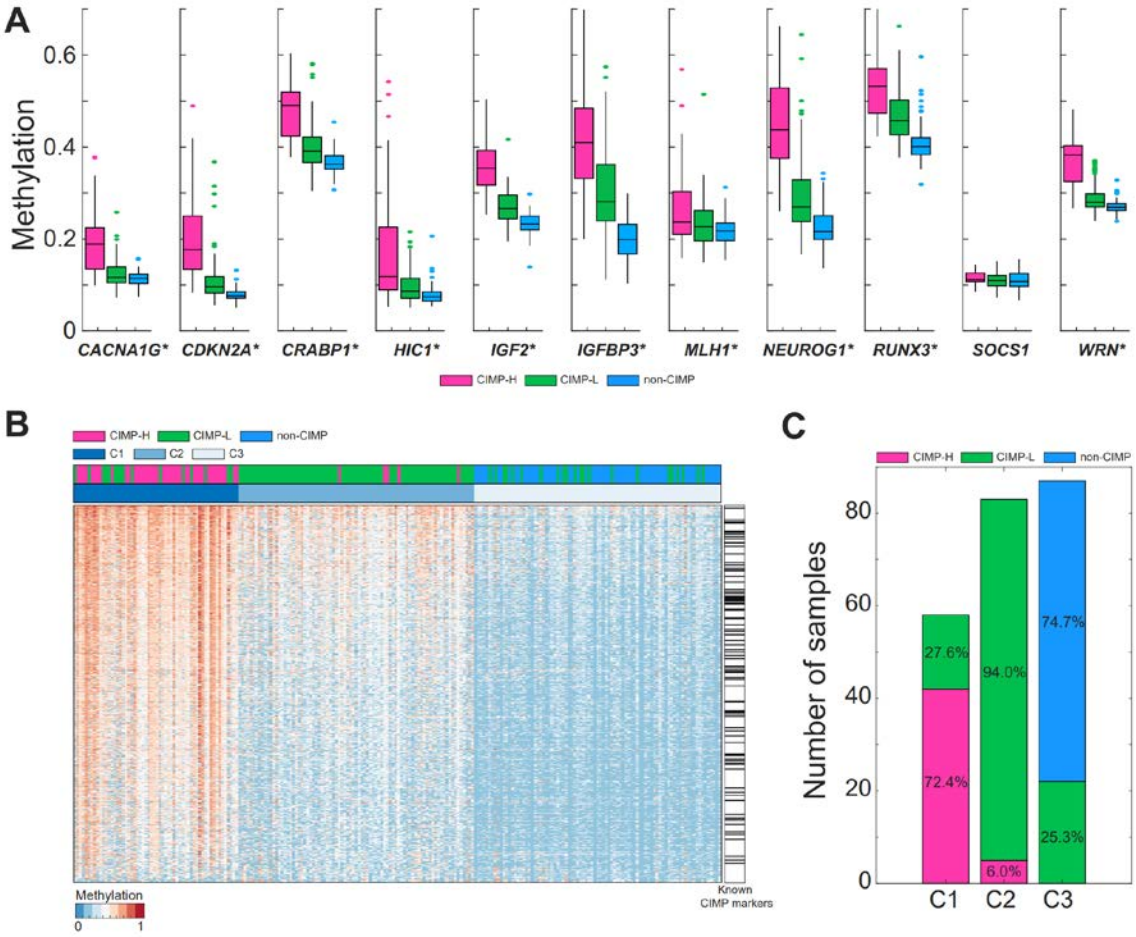
Supplementary Figure 4. Comparison of the methylome dataset before and after the bias correction. Among the total probes, we used the 3,000 probes with the largest variance of beta values across all samples for a principal component (PC) analysis of raw (before bias correction; left) and processed (after bias correction; right) datasets. The plots show the PC3 (x-axis) and PC4 (y-axis) with their explained variances. The individual samples in the plots were labeled according to sex (top: male and female), batch number (middle: batch types), and tumor status (bottom: tumor and normal).


Hypermethylated
Hypomethylated
Negative control
TSS1500
TSS200
5'UTR
1stExon
Body
3'UTR
Z-score
4
0
ECM-receptor interaction
gap junction
regulation of tissue remodeling
epithelial cell differentiation
digestive tract development
regulation of angiogenesis
cell adhesion
chemotaxis
chemokine signaling pathway
cytokine-cytokine receptor interaction
olfactory transduction
nitric oxide metabolic process
inflammatory response
leukocyte proliferation
regulation of JAK-STAT cascade
Fc epsilon RI signaling pathway
phagocytosis
response to bacterium
TGF-beta signaling pathway
Wnt signaling pathway
BMP signaling pathway
Rap1 signaling pathway
calcium signaling pathway
cAMP signaling pathway
phospholipase D signaling pathway
cGMP-PKG signaling pathway
renin secretion
Hedgehog signaling pathway
transcriptional misregulation in cancer
hormone secretion
neuron fate commitment
regulation of synapse organization
glutamate receptor signaling pathway
neurotransmitter transport
ion transport
proteoglycan biosynthetic process
fatty acid transport
Tissue remodeling & digestive tract development
Cell adhesion & chemotaxis
Cyokine-cytokine receptor interaction
Inflammatory response
Signaling pathway
Hormone secretion & synapse organization

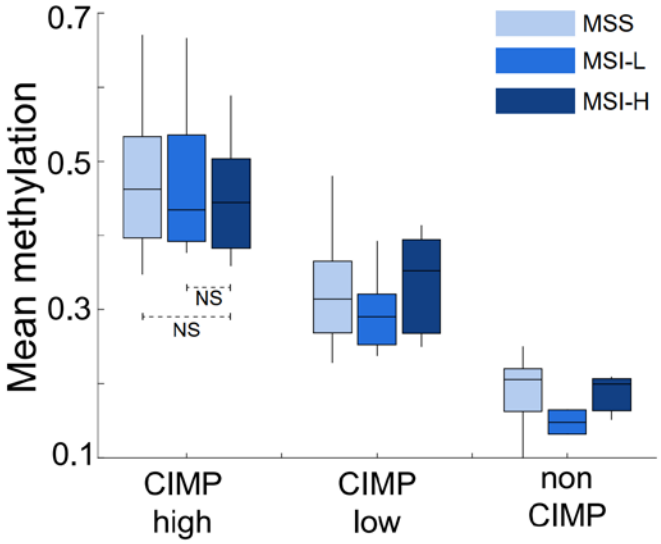
Supplementary Figure 5. Heat map showing the functional enrichment patterns of GOBPs and KEGG pathways by hyper- and hypomethylated, and negative control positions at genomic regions (TSS1500, TSS200, 5'-UTR, first exon, body, and 3'-UTR). Color bar, gradient of Z-score for the enrichment *p* value computed by using DAVID software. A set of negative control positions at each region were determined as the ones with i) the absolute difference in the mean beta values between the tumor and normal samples 〈 0.05 and ii) the *p* values 〉 0.9.



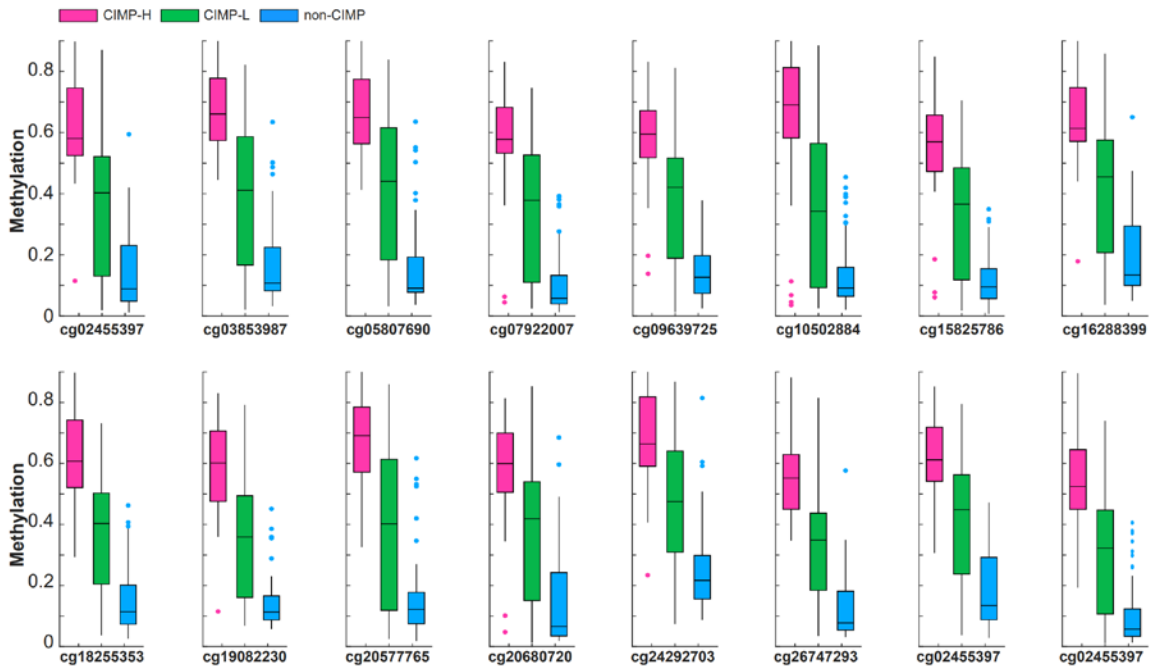
Supplementary Figure 6. Recurrence free survival analysis. Kaplan-Meier plot for CIMP-H and non CIMP-H groups. HR, *p* represent hazard ratio and significance of log-rank test, respectively.



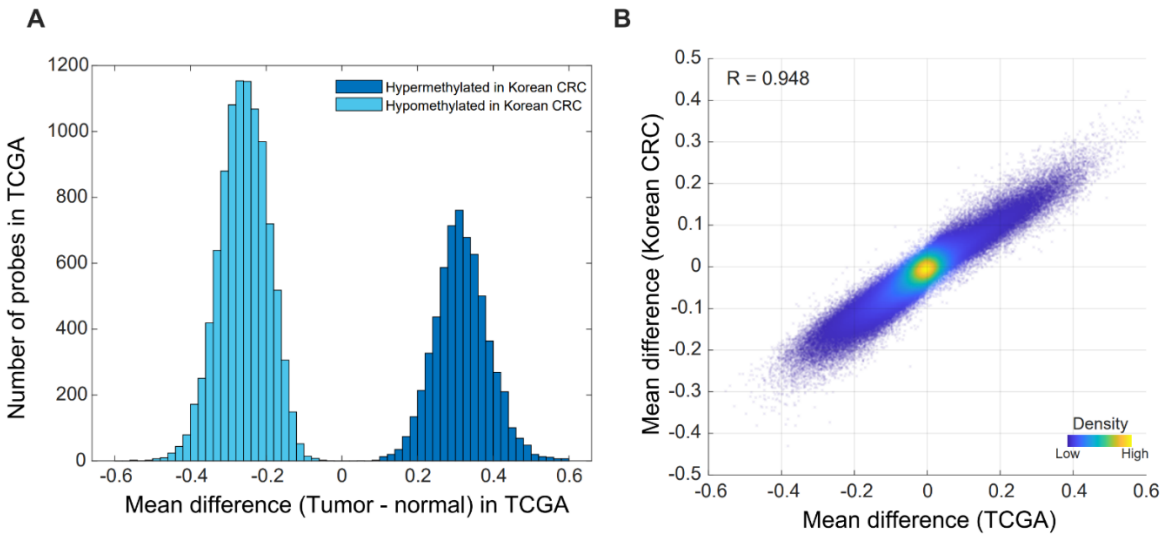
Supplementary Figure 7. Comparative analysis of hypermethylation patterns in Korean CRC according to CIMP categories. (A) Comparison with promoter methylation levels across the three CIMP subgroups: CIMP-H, CIMP-L, and non-CIMP. '*' denotes the significance *p* values (〈 0.05) of the mean difference between CIMP-H and non CIMP-H. (B) Clustering of the methylation profiles of 680 selected hypermethylated probes. Labels C1, C2, and C3 denote new cluster groups defined by these 680 hypermethylated probes. The right-hand black-and-white bar represents the categorization based on previously defined CIMP marker probes (C) Proportional representation of established CIMP categories within newly defined C1, C2, and C3 clusters. We found that C1, C2, and C3 were likely to be matched to CIMP-H, CIMP-L, and non-CIMP.



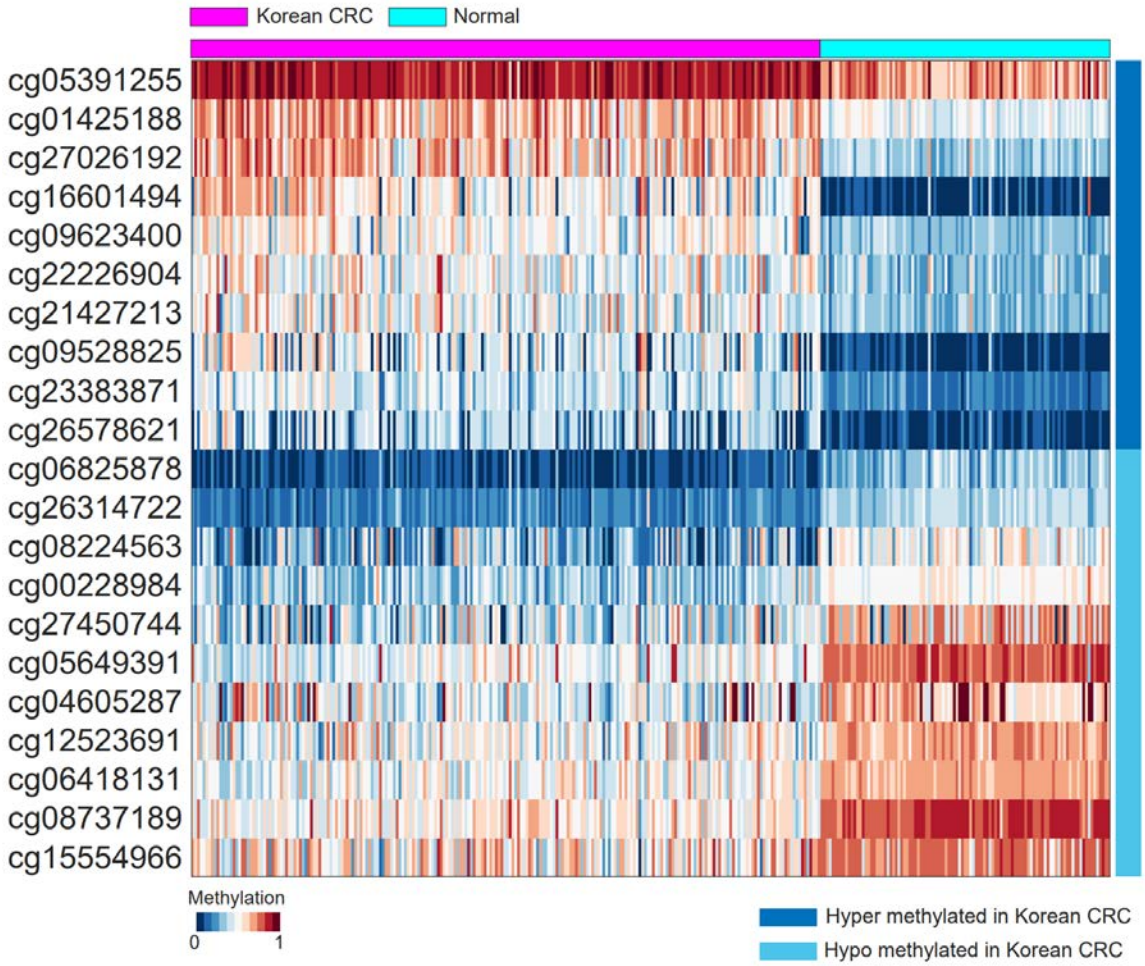
Supplementary Figure 8. Boxplot showing mean beta values of CIMP marker probes for the patient groups classified as their CIMP and microsatellite instability (MSI) statuses. The boxes display the lower, median and upper quartiles; the whiskers represent the minimum and maximum values. NS denotes not significant by one-way ANOVA with a post hoc test (Sidak correction).



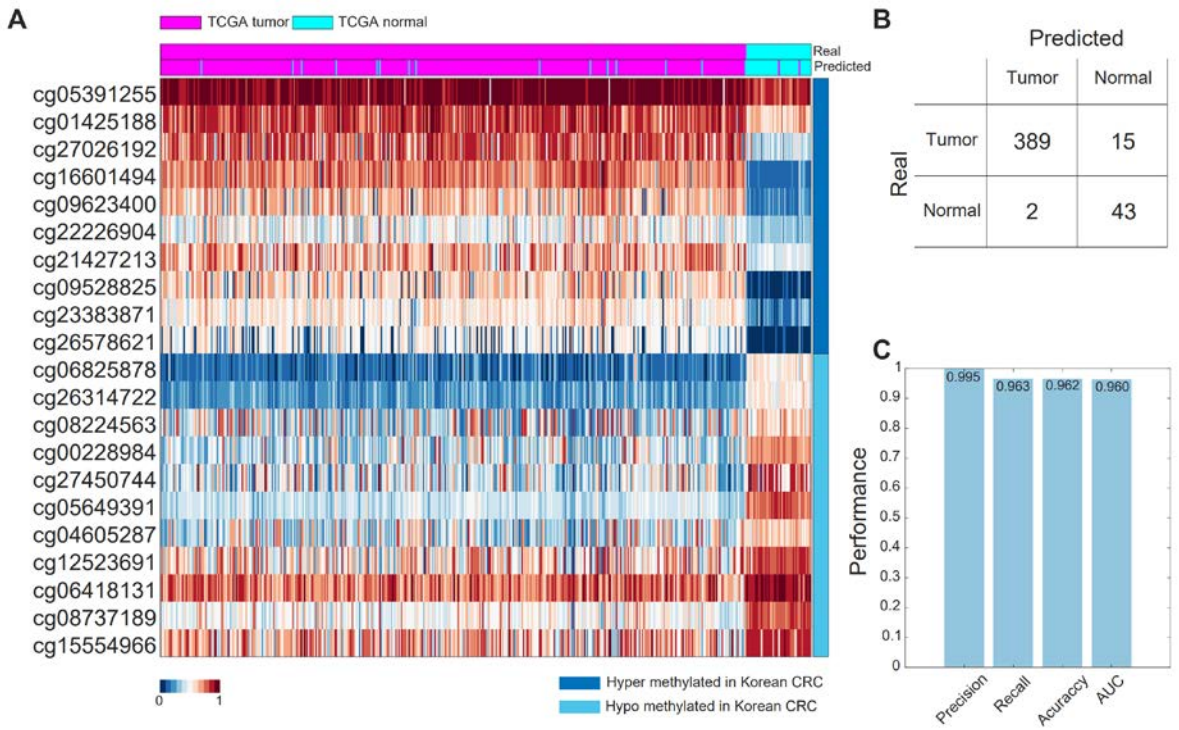
Supplementary Figure 9. Comparison with methylation levels of 16 cg probes across the three CIMP subgroups: CIMP-H, CIMP-L, and non-CIMP.



Supplementary Figure 10. Comparative Analysis of CRC methylomes between Korean and TCGA cohorts. (A) Distribution of methylation levels observed in the TCGA CRC dataset at differentially methylated positions (DMPs) in Korean CRC. (B) Correlation analysis between Korean and TCGA CRC focusing on the 298,581 overlapping probes; 'R' denotes Pearson's correlation coefficient.

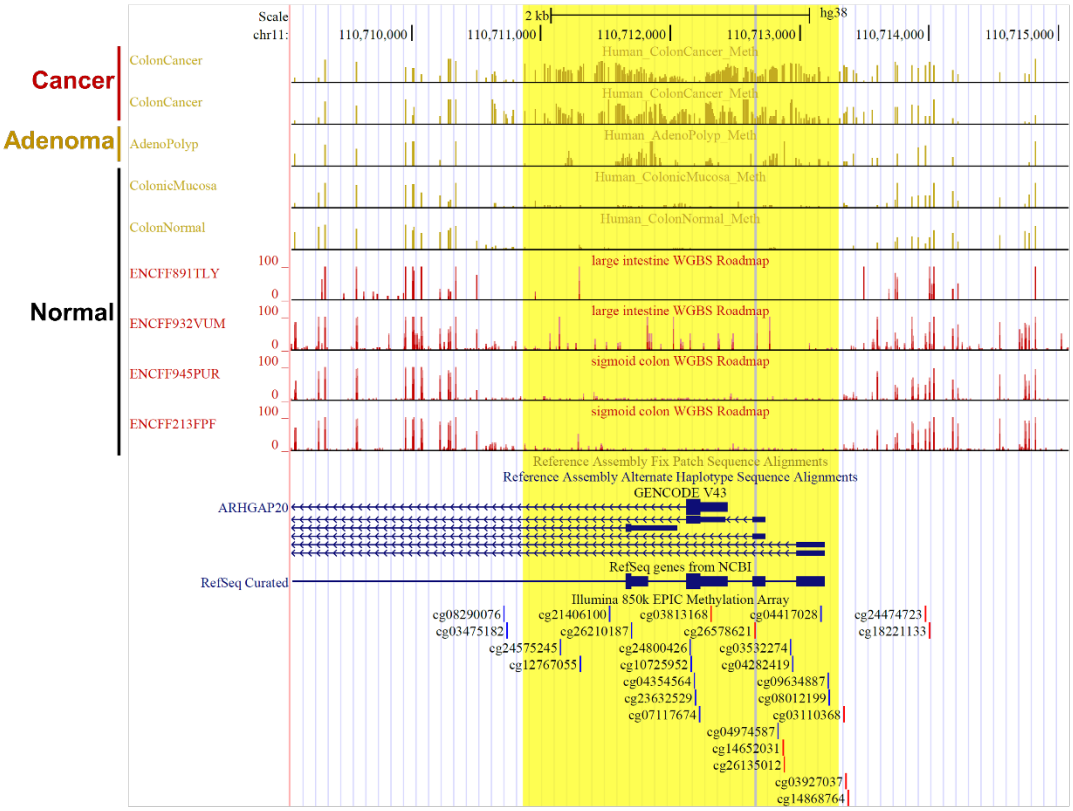


Supplementary Figure 11. Methylation Profiles of 21 selected probes for CRC diagnosis in a Korean cohort. The heatmap displays the methylation beta values for 10 hypermethylated and 11 hypomethylated probes. In the sample color bar, magenta represents CRC samples, and light blue indicates adjacent normal tissue. The probe color bar employs dark blue and sky blue to hypermethylated and hypomethylated probes in Korean CRC, respectively.

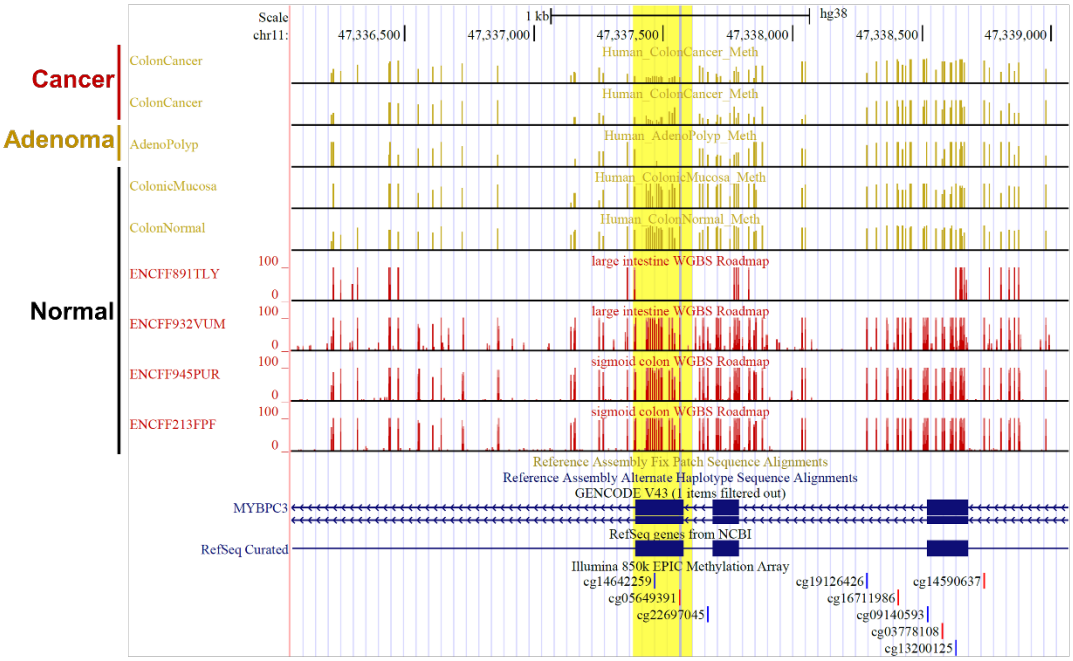


Supplementary Figure 12. Validation of CRC prediction in the TCGA Cohort using 21 diagnostic markers from Korean CRC. (A) Methylation profiles of the 21 selected diagnostic probes in the TCGA CRC cohort. In the sample color bar, magenta denotes CRC samples, and light blue represents adjacent normal tissues. 'Predicted' and 'Real' indicate model predictions and actual cancer annotations, respectively. The probe color bar utilizes dark blue and sky blue to signify hypermethylated and hypomethylated probes from the Korean CRC study, respectively. (B) Confusion matrix illustrating the counts of prediction of the CRC classifier model in the TCGA CRC cohort. (C) Performance metrics assessing the accuracy of CRC presence predictions in the independent TCGA cohort.

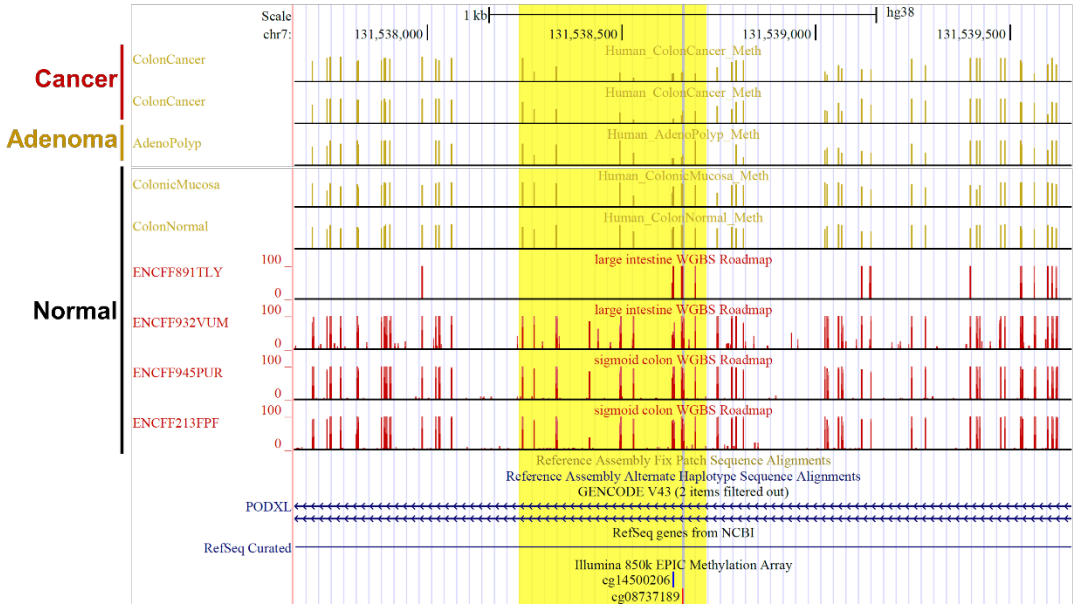
E


Scale
chr11:
2 kb
hg38
110,710,000
110,711,000
110,712,000
110,713,000
110,714,000
110,715,000
Cancer
Adenoma
Normal
ColonCancer
ColonCancer
AdenoPolyp
ColonicMucosa
ColonNormal
ENCFF891TLY
ENCFF932VUM
ENCFF945PUR
ENCFF213FPF
Human_ColonCancer_Meth
Human_ColonCancer_Meth
Human_AdenoPolyp_Meth
Human_ColonicMucosa_Meth
Human_ColonNormal_Meth
large intestine WGBS Roadmap
large intestine WGBS Roadmap
sigmoid colon WGBS Roadmap
sigmoid colon WGBS Roadmap
100
0
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V43
ARHGAP20
RefSeq genes from NCBI
RefSeq Curated
Illumina 850k EPIC Methylation Array
cg08290076
cg21406100
cg03813168
cg04417028
cg24474723
cg03475182
cg26210187
cg26578621
cg18221133
cg24575245
cg24800426
cg03532274
cg12767055
cg10725952
cg04282419
cg04354564
cg09634887
cg23632529
cg08012199
cg07117674
cg03110368
cg04974587
cg14652031
cg26135012
cg03927037
cg14868764

F


Scale
chr11:
1 kb
hg38
47,336,500
47,337,000
47,337,500
47,338,000
47,338,500
47,339,000
Cancer
Adenoma
Normal
ColonCancer
ColonCancer
AdenoPolyp
ColonicMucosa
ColonNormal
ENCFF891TLY
ENCFF932VUM
ENCFF945PUR
ENCFF213FPF
Human_ColonCancer_Meth
Human_ColonCancer_Meth
Human_AdenoPolyp_Meth
Human_ColonicMucosa_Meth
Human_ColonNormal_Meth
large intestine WGBS Roadmap
large intestine WGBS Roadmap
sigmoid colon WGBS Roadmap
sigmoid colon WGBS Roadmap
100
0
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V43 (1 items filtered out)
MYBPC3
RefSeq genes from NCBI
RefSeq Curated
Illumina 850k EPIC Methylation Array
cg14642259
cg19126426
cg14590637
cg05649391
cg16711986
cg22697045
cg09140593
cg03778108
cg13200125

G



Supplementary Figure 13. WGBS-based Methylation levels of regions around selected CRC diagnosis markers. Visualization of WGBS-based methylation levels of regions around the marker using UCSC genome browser (A) cg09528825, (B) cg09623400, (C) cg16601494, (D) cg23383871, (E) cg26578621, (F) cg05649391, and (G) cg08737189. WGBS-based methylation levels of colorectal cancer (red, the 1st and 2nd tracks), adenoma (light brown, the 3rd track) and normal (black, the 4th - 9th tracks) are shown. The 10th and 11th tracks show gene annotation from GENCODE and RefSeq respectively. The 12th track shows the location of probes from Illumina 850K EPIC methylation array. Blue and yellow background highlight the methylation marker and the region around it, respectively.

Supplementary Table 1. Clinicopathological characteristics of 228 patients with colorectal cancer (CRC).

Clinicopathological characteristics	Total[a] (N = 228)	CIMP[b] group			*p* value
		CIMP-H 47 (20.6%)	CIMP-L 116 (50.9%)	non-CIMP 65 (28.5%)	
Sex					NS
Female	89 (40.5%)	21 (47.7%)	43 (37.7%)	25 (40.3%)	
Male	131 (59.5%)	23 (52.3%)	71 (62.3%)	37 (59.7%)	
Age (years)	64.4 ± 12.7	66.0 ± 13.3	64.7 ± 12.9	62.9 ± 11.8	NS
Location[c]					p 〈 0.05
Right-sided	50 (22.8%)	15 (34.1%)	31 (27.2%)	4 (6.6%)	
Left-sided	114 (52.1%)	19 (43.2%)	54 (47.4%)	41 (67.2%)	
Rectum	55 (25.1%)	10 (22.7%)	29 (25.4%)	16 (26.2%)	
MSI[d]					p 〈 0.05
MSI-H	15 (6.8%)	6 (13.6%)	6 (5.3%)	3 (4.8%)	
MSI-L	19 (8.6%)	7 (15.9%)	10 (8.8%)	2 (3.2%)	
MSS	186 (84.5%)	31 (70.5%)	98 (86.0%)	57 (91.9%)	
KRAS[e]					NS
MT	44 (20.0%)	12 (27.3%)	24 (21.1%)	8 (12.9%)	
WT	176 (80.0%)	32 (72.7%)	90 (78.9%)	54 (87.1%)	
Stage					NS
I	5 (2.3%)	1 (2.3%)	4 (3.5%)	0 (0%)	
II	86 (39.1%)	18 (40.9%)	39 (34.2%)	29 (46.8%)	
III	92 (41.8%)	18 (40.9%)	51 (44.7%)	23 (37.1%)	
IV	37 (16.8%)	7 (15.9%)	20 (17.5%)	10 (16.1%)	
T stage					NS
II	9 (4.1%)	1 (2.3%)	7 (6.1%)	1 (1.6%)	
III	167 (75.9%)	30 (68.2%)	91 (79.8%)	46 (74.2%)	

IV	44 (20.0%)	13 (29.5%)	16 (14.0%)	15 (24.2%)	
N stage					NS
N0	92 (41.8%)	19 (43.2%)	44 (38.6%)	29 (46.8%)	
N1	71 (32.3%)	11 (25.0%)	39 (34.2%)	21 (33.9%)	
N2	57 (25.9%)	14 (31.8%)	31 (27.2%)	12 (19.4%)	
M stage					NS
M0	183 (83.2%)	37 (84.1%)	94 (82.5%)	52 (83.9%)	
M1	37 (16.8%)	7 (15.9%)	20 (17.5%)	10 (16.1%)	
Differentiation[f]					NS
WD	23 (10.6%)	4 (9.3%)	13 (11.6%)	6 (9.7%)	
MD	179 (82.5%)	33 (76.7%)	94 (83.9%)	52 (83.9%)	
PD	9 (4.1%)	2 (4.7%)	4 (3.6%)	3 (4.8%)	
Mucinous	6 (2.8%)	4 (9.3%)	1 (0.9%)	1 (1.6%)	
Lymphatic invasion[g]					NS
Yes	64(29.1%)	15(34.1%)	34(29.8%)	15(24.2%)	
No	156(70.9%)	29(65.9%)	80(70.2%)	47(75.8%)	
Venous invasion					NS
Yes	67(30.5%)	14(31.8%)	34(29.8%)	19(30.6%)	
No	153(69.5%)	30(68.2%)	80(70.2%)	43(69.4%)	
Perineural invasion					NS
Yes	105(47.7%)	18(40.9%)	62(54.4%)	25(40.3%)	
No	115(52.3%)	26(59.1%)	52(45.6%)	37(59.7%)	

[a]For each clinical category, we excluded patients without relevant information when we calculated the percentages in the table. For example, 220 patients (89 females and 131 males) were considered as 100% for sex category (40.5% and 59.5% for female and male, respectively). The relevant information of age, MSI, KRAS, NRAS, stage, and TNM-stages categories was missing for eight patients. Differentiation category was missing for 11 patients. When we classified the tumor locations into "left", "right", and "rectum" groups, we excluded one synchronous tumor sample (i.e., multiple presence of both left and right

location) and samples of the aforementioned eight patients from the groups. [b]CIMP: 5'-C-phosphate-G-3' island methylator phenotype. [c]Right-sided: ascending, cecum, hepatic flexure, transverse; Left-sided: descending, rectosigmoid, sigmoid, splenic flexure; [d]MSI: microsatellite instability; MSS: microsatellite stability; MSI-H: high microsatellite instability; MSI-L: low microsatellite instability; [e]MT: mutation; WT: wild type; [f]WD: well-differentiated; MD: moderately differentiated; PD: poorly differentiated; [g]Yes: Presence of cancer cells in lymph vessels or in blood vessels or surrounding nerves; No: Absence of cancer cells in lymph vessels or in blood vessels or surrounding nerves. p values represent the significance of association test with CIMP groups, ANOVA for continuous characteristic (age), Chi-square test for categorical characteristics (Sex, MSI, KRAS, N stage, M stage, lymphatic, venous, perineural invasion types), and Fisher's exact test for categorical characteristics with fewer than 5 samples in specific subtypes (AJCC stage, T stage, Differentiation). NS denotes not significant.

Supplementary Table 2. List of hyper- and hypomethylated positions.

A total of 7,824 hypermethylated (A) and 30,783 hypomethylated positions (B) are shown with the differences of the mean beta values between the tumor and normal tissues as well as the relevant annotations in terms of CpG-island-associated and genic regions.

See the attached excel file named “Supplementary Table 2.xlsx”

Supplementary Table 3. Distribution of hyper- and hypomethylated positions in the genic regions

Genic region	Number of annotated EPIC probes	Number of hyper-methylated DMPs	Odds ratio of hyper-methylated DMPs	Number of hypo-methylated DMPs	Odds ratio of hypo-methylated DMPs	p value of the odds ratio (hyper-methylated DMPs)	p value of the odds ratio (hypo-methylated DMPs)
TSS1500	79055	1344	1.3972	2672	0.6245	$< 1 \times 10^{-4}$	1
TSS200	36337	1420	3.5963	643	0.3243	$< 1 \times 10^{-4}$	1
UTR5	67792	1653	2.1671	2263	0.6209	$< 1 \times 10^{-4}$	1
F_Exon	22006	1012	4.1059	391	0.3313	$< 1 \times 10^{-4}$	1
BODY	267555	2561	0.6174	11115	0.7092	1	1
UTR3	18586	147	0.6052	550	0.5651	1	1
Total	609046	7824		30783			

Supplementary Table 4. Distribution of hyper- and hypomethylated positions in the CpG-island-associated regions

CpG Island	Number of annotated EPIC probes	Number of hyper-methylated DMPs	Odds ratio of hyper-methylated DMPs	Number of hypo-methylated DMPs	Odds ratio of hypo-methylated DMPs	p value of the odds ratio (hyper-methylated DMPs)	p value of the odds ratio (hypo-methylated DMPs)
S shelf	22990	96	0.3138	804	0.6722	1	1
S shore	49264	591	0.9276	945	0.3474	0.9644	1
Island	77220	5131	13.9850	286	0.0611	$< 1 \times 10^{-4}$	1
N shore	58149	857	1.1678	1173	0.3625	$< 1 \times 10^{-4}$	1
N shelf	24862	111	0.3352	860	0.6637	1	1
Open sea	376561	1038	0.0919	26715	4.2877	1	$< 1 \times 10^{-4}$
Total	609046	7824		30783			

Supplementary Table 5. List of 16 CIMP methylation markers

Probe ID	Chromosome	Locus (hg38)	CpG island	Gene name	Genic region	Mean Difference (CIMP-H vs CIMP-L)	Mean Difference (CIMP-L vs non-CIMP)
cg02455397	chr11	119422674-119422676	Island	THY1	5UTR	0.239	0.228
cg03853987	chr2	100417816-100417818	Island	CHST10	TSS200	0.275	0.231
cg05807690	chr2	100417820-100417822	Island	CHST10	TSS200	0.259	0.244
cg07922007	chr8	66962622-66962624	Island			0.239	0.234
cg09639725	chr10	133087789-133087791	Island	GPR123	TSS200	0.219	0.229
cg10502884	chr10	124092798-124092800	Island			0.299	0.213
cg15825786	chr10	133087792-133087794	Island	GPR123	TSS200	0.222	0.210
cg16288399	chr11	119422666-119422668	Island	THY1;USP2-AS1	5UTR;Body	0.224	0.227
cg18255353	chr4	153791269-153791271	Island			0.245	0.223
cg19082230	chr4	182448600-182448602	Island	ODZ3	Body	0.235	0.214
cg20577765	chr2	100417832-100417834	Island	CHST10	TSS200	0.287	0.228
cg20680720	chr19	36916313-36916315	Island	ZNF568;ZNF829	TSS200;TSS200	0.211	0.213
cg24292703	chr14	56797920-56797922	Island			0.207	0.221
cg26747293	chr5	38258567-38258569	Island	EGFLAM;EGFLAM	1stExon;5UTR	0.227	0.204
cg27515369	chr3	141051756-141051758	Island	SPSB4	TSS200	0.215	0.227
cg27591450	chr17	77528921-77528923	Island			0.240	0.205

Supplementary Table 6. Eleven hypermethylated gene markers associated with six cancer-related pathways

Genes	Promoter-like* probes	GOBP/KEGG pathways (reported in this study)					
		WNT signaling pathway	TGF-beta signaling pathway	BMP signaling pathway	Cell adhesion	Regulation of angiogenesis	cAMP signaling pathway
SFRP1 (15)	cg04255616 cg10406295 cg21517947	O		O	O	O	
SFRP2 (15)	cg00082664 cg03202804 cg05164933 cg05961809 cg06549216 cg10942078 cg11354906 cg14063488 cg14330641 cg23121156 cg23207990 cg23292160 cg25645268 cg25775322	O		O	O		
SOX17 (15)	cg04672706 cg15186181 cg24891539 cg26059468	O					
WIF1 (15)	cg03509412 cg19427610 cg26733786	O					
SMAD1 (18)	cg16071998		O	O		O	
SMAD2 (18)	cg26130023		O	O			
CDH13	cg05374412				O		

(19)							
TMEFF2 (20]	cg01808545 cg02288301 cg06008912 cg06856528 cg09237843 cg18107367 cg18221862 cg24899822				O		
ADAMTS1 (21-23)	cg00472814 cg12282100 cg15621322 cg24262066				O	O	
ADCY1 (24)	cg07651242 cg07960450 cg24676071						O
ADCY4 (25)	cg05031016 cg12265829 cg23179456						O

*Promoter-like represents the regions which were annotated with TSS200, TSS1500, 5'UTR, and first exon.

Supplementary Table 7. Comparison of tumor stages between The Cancer Genome Atlas (TCGA) colorectal cancer dataset and the proposed Korean patients with colorectal cancer.

Data source	Stage			
	I	II	III	IV
Korea_CRC	5	86	92	37
TCGA_CRC	56	103	69	44

Supplementary Table 8. List of 21 diagnostic methylation markers

Probe ID	Chromosome	Locus (hg38)	CpG island	Gene name	Genic region	Mean Difference (Korean CRC) Tumor - Normal	Mean Difference (TCGA) Tumor - Normal
cg01425188	chr8	28621759-28621761	N_Shore			0.196	0.186
cg05391255	chr12	68931181-68931183	N_Shelf	CPM	Body	0.178	0.091
cg09528825	chr16	28063066-28063068	Island	GSG1L	Body	0.330	0.485
cg09623400	chr20	24469716-24469718	Island	TMEM90B	5UTR, 1stExon	0.186	0.416
cg16601494	chr1	1540356-1540358	N_Shore	C1orf70	5UTR, 1stExon	0.423	0.557
cg21427213	chr3	188155186-188155188	S_Shore	LPP	5UTR	0.173	0.215
cg22226904	chr1	27492548-27492550	S_Shelf			0.162	0.172
cg23383871	chr20	49318449-49318451	Island			0.228	0.327
cg26578621	chr11	110712653-110712655	Island	ARHGAP20	5UTR, 1stExon	0.235	0.358
cg27026192	chr16	57803047-57803049	Island	KIFC3	TSS1500	0.270	0.348
cg00228984	chr20	1491237-1491239	OpenSea	SIRPB2	Body	-0.167	-0.285
cg04605287	chr1	54487812-54487814	N_Shore			-0.187	-0.167
cg05649391	chr11	47337563-47337565	N_Shore	MYBPC3	Body	-0.298	-0.367
cg06418131	chr6	32055872-32055874	OpenSea	TNXB	Body	-0.179	-0.149
cg06825878	chr7	75843221-75843223	OpenSea			-0.207	-0.389
cg08224563	chr16	20904982-20904984	S_Shelf	LYRM1	5UTR	-0.191	-0.143
cg08737189	chr7	131538657-131538659	OpenSea	PODXL	Body	-0.267	-0.286
cg12523691	chr4	168796529-168796531	OpenSea	PALLD	Body	-0.182	-0.226
cg15554966	chr19	53679931-53679933	OpenSea	MIR519E	TSS200	-0.160	-0.195
cg26314722	chr1	234731552-234731554	OpenSea			-0.176	-0.284
cg27450744	chr8	142564299-142564301	Island			-0.190	-0.244

Supplementary References

1. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74
2. Hansen KD, Timp W, Bravo HC et al (2011) Increased methylation variation in epigenetic domains across cancer types. Nat Genet 43, 768-775
3. Berman BP, Weisenberger DJ, Aman JF et al (2011) Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet 44, 40-46
4. Song Q, Decato B, Hong EE et al (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One 8, e81148
5. Aryee MJ, Jaffe AE, Corrada-Bravo H et al (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363-1369
6. Maksimovic J, Gordon L and Oshlack A (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13, R44
7. Leek JT, Johnson WE, Parker HS, Jaffe AE and Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882-883
8. Ritchie ME, Phipson B, Wu D et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47
9. Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289-300
10. Huang da W, Sherman BT and Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57
11. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58, 267-288
12. McInnes T, Zou D, Rao DS et al (2017) Genome-wide methylation analysis identifies a core set of hypermethylated genes in CIMP-H colorectal cancer. BMC Cancer 17, 228
13. Sidak Z (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62, 626-633
14. Peto R and Peto J (1972) Asymptotically efficient rank invariant test

procedures. Journal of the Royal Statistical Society: Series A (General) 135, 185-198
15. Silva A-L, Dawson SN, Arends MJ et al (2014) Boosting Wnt activity during colorectal cancer progression through selective hypermethylation of Wnt signaling antagonists. BMC cancer 14, 1-10
16. Guo X and Wang XF (2009) Signaling cross-talk between TGF-beta/BMP and other pathways. Cell Res 19, 71-88
17. Huang F and Chen YG (2012) Regulation of TGF-beta receptor activity. Cell Biosci 2, 9
18. Ding N, Luo H, Zhang T, Peng T, Yao Y and He Y (2023) Correlation between SMADs and Colorectal Cancer Expression, Prognosis, and Immune Infiltrates.
19. Hibi K, Kodera Y, Ito K, Akiyama S and Nakao A (2004) Methylation pattern of CDH13 gene in digestive tract cancers. Br J Cancer 91, 1139-1142
20. Yan H, He J, Guan Q et al (2017) Identifying CpG sites with different differential methylation frequencies in colorectal cancer tissues based on individualized differential methylation analysis. Oncotarget 8, 47356-47364
21. Choi JE, Kim DS, Kim EJ et al (2008) Aberrant methylation of ADAMTS1 in non-small cell lung cancer. Cancer Genet Cytogenet 187, 80-84
22. Lind GE, Kleivi K, Meling GI et al (2006) ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis. Cell Oncol 28, 259-272
23. Cal S and Lopez-Otin C (2015) ADAMTS proteases and cancer. Matrix Biol 44-46, 77-85
24. Zhang Y, Yang J, Wang X and Li X (2021) GNG7 and ADCY1 as diagnostic and prognostic biomarkers for pancreatic adenocarcinoma through bioinformatic-based analyses. Sci Rep 11, 20441
25. Fan Y, Mu J, Huang M et al (2019) Epigenetic identification of ADCY4 as a biomarker for breast cancer: an integrated analysis of adenylate cyclases. Epigenomics 11, 1561-1579