Supplementary Information

Epigenetic insights into colorectal cancer: comprehensive genome-wide DNA methylation profiling of 294 patients in Korea

Soobok Joe[#,1], Jinyong Kim[#,2], Jin-Young Lee[#,3], Jongbum Jeon[1], Iksu Byeon[1], Sae-Won Han[2,4], Seung-Bum Ryoo[5], Kyu Joo Park[5], Sang-Hyun Song[4], Sang-Hyun Song[4,] Sheehyun Cho[3], Hyeran Shim[3], Hoang Bao Khanh Chu[3], Jisun Kang[3], Hong Seok Lee[3], DongWoo Kim[6], Young-Joon Kim[3,7,*], Tae-You Kim[2,4,8,9,*] & Seon-Young Kim[1,*]

[1]Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, [2]Department of Internal Medicine, Seoul National University Hospital, Seoul National University College of Medicine, Seoul 03080, [3]Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, [4]Cancer Research Institute, Seoul National University College of Medicine, Seoul 03080, [5]Department of Surgery, Seoul National University Hospital, Seoul National University College of Medicine, Seoul 03080, [6]Cellgentek, Cheongju 28161, [7]LepiDyne Co., Ltd., Seoul 04779, [8]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul 08826, [9]IMBdx, Inc., Seoul 08506, Korea

*Corresponding authors. Seon-Young Kim, Tel: +82-42-879-8500; Fax: +82-42-879-8519; E-mail: kimsy@kribb.re.kr; Tae-You Kim, Tel: +82-2-2072-3943; Fax: +82-2-762-9662; E-mail: kimty@snu.ac.kr; Young-Joon Kim, Tel: +82-2-2123-2628; Fax: +82-2-363-4083; E-mail: yjkim@yonsei.ac.kr

[#]These authors contributed equally to this work.

SUPPLEMENTARY MATERIAL

Supplementary Note 1. Supplementary materials and methods

Study design and patients

Patients diagnosed with stage II, III, and IV CRC were enrolled from September 2017 to 2022 at the Seoul National University Hospital. Patients with stage II and III colon cancer underwent surgery, followed by adjuvant chemotherapy. Patients with stage IV colon cancer underwent surgery with preoperative or postoperative palliative chemotherapy. Patients with rectal cancer receive preoperative or postoperative concurrent chemoradiation therapy (CCRT), surgery, and adjuvant chemotherapy. Tissue samples were collected from the tumor and adjacent normal tissues at the time of operation and kept fresh and frozen. Peripheral blood samples were collected before and after each treatment (i.e., surgery, chemotherapy, and CCRT). The methylation dataset comprised 443 arrays, including 299 tumors and 144 matched normal samples. All samples were analyzed using an EPIC array and subjected to preprocessing and quality control. After filtering for mismatched clinical sex information, 294 tumor and 143 normal (142 matched) samples were analyzed using the EPIC array and subjected to preprocessing and quality control for downstream analysis.

Generation of EPIC array dataset from genomic DNA

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 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.

Normalization, batch correction, and probe filtering

To construct a reliable CRC methylome profile, the EPIC array dataset was processed using the *minfi* pipeline (1). Initially, we extracted the raw intensities of 865,859 probes from the green and red channels of raw .IDAT files. After raw intensity extraction, SWAN (2) was used to correct for technical differences between type I and II probes. Batch effects were addressed using the surrogate variable analysis tool in conjunction with the *combat* approach (3, 4). To handle additional batch corrections, we used the sentrix ID information of the array because the dataset consisted of one set before 2042203330001 and another set after 2042203330001 sentrix IDs. During this process, we manually removed 1047 probes based on the Illumina EPIC array manual version 1.05B. Subsequently, additional probes, such as sex chromosomes (19,575), known SNP locus (161,412), failed detection (2,123) probes, and low-variable 83,655 probes having beta range less than 0.1, were filtered out. Ultimately, methylation beta values of 616,162 probes from 294 tumor and 143 normal samples (142 matched) were used for downstream analysis.

Identification of CIMP groups from EPIC array data

For identification of CIMP groups, we used previously identified 258 CIMP marker genes (12) for the initial CIMP probe set consisting of 4,327 regional CpG island probes. We then filtered out 2,397 low-variable probes (i.e., selected probes with standard deviation 〉 0.15). Overall, we clustered CRC samples based on the methylation level of 1,930 probes. We divided the CRC samples into five clusters using 100 iterations of the K-means approach (28). In each iteration, we minimized

the local distance, based on the squared Euclidean distance, to finally identify the five optimized clusters. According to sorted grand mean methylation levels of each cluster, we finally identified the top-two clusters (mean beta values of 0.58 and 0.43) as CIMP-H, the third and fourth clusters as CIMP-L (mean beta values of 0.30 and 0.29), and the last cluster as non-CMIP (mean beta value of 0.19).

Statistical analysis

Using *minfi's dmpfinder* function (1, 5), we identified DMPs between tumor and normal samples and classified them as hyper- or hypomethylated at the probe level. DMPs that were hypermethylated probes were identified when the mean difference between the tumor and normal tissues was greater than 0.15 with a q-value $\langle$ 0.0001. Hypomethylated probes were similarly identified, with a mean difference of less than -0.15. To compare the abundance of DMPs in the genic and CpG island regions, we calculated the OR of enrichment for each DMP group based on the genomic annotations of each probe. Fisher's exact test and t-test were used to compare the significance of CIMP cluster proportions for each tumor characteristic. For pair-difference analysis, we used 142 matched samples and selected the top 10,000 probes that had high standard deviations. One hundred forty-three samples were clustered into three groups using the K-means approach, and we only calculated proportions of CIMP groups for those samples. We used the Kolmogorov - Smirnov test to compare the cumulative distribution, as shown in Supplementary Fig. 7 (6). To estimate gene methylation levels, we used the mean beta values of the probes annotated as promoter-like regions (TSS1500, TSS200, 5' UTR, first exon). All statistical analyses were conducted using MATLAB2022a and the R software (v4.4).

Supplementary Table 1. List of total hypermethylated probes.

Supplementary_table_1_2.xlsx

Supplementary Table 2. List of total hypomethylated probes.

Supplementary_table_1_2.xlsx

Supplementary Table 3. Number of hypermethylated probes in the genic region.

Genic region	Hypermethylated probes	Total EPIC probes	Odds ratio* (p-value)
TSS1500	1529	80580	1.43 ($p < 0.0001$)
TSS200	1542	37964	3.38 ($p < 0.0001$)
5' UTR	1786	69168	2.07 ($p < 0.0001$)
1st exon	1185	22890	4.26 ($p < 0.0001$)
Body	2704	270424	0.57 ($p < 0.0001$)
3' UTR	163	18821	0.60 ($p < 0.0001$)
Genic DMPs	6933	424870	1.79 ($p < 0.0001$)
Total DMPs	8691	616162	

*The odds ratio was calculated using the odds of each genic region position (e.g., TSS1500 and TS200) that was given hypermethylated probes. The numbers within parentheses represent p-values from Fisher's exact test for each genic region.

Abbreviations: UTR, untranslated region; DMP, differentially methylated position

Supplementary Table 4. Number of hypomethylated probes in the genic region.

Genic region	Hypomethylated probes	Total EPIC probes	Odds ratio* (p-value)
TSS1500	2751	80580	0.63 ($p < 0.0001$)
TSS200	656	37964	0.31 ($p < 0.0001$)
5' UTR	2253	69168	0.60 ($p < 0.0001$)
1st exon	386	22890	0.31 ($p < 0.0001$)
Body	11239	270424	0.70 ($p < 0.0001$)
3' UTR	548	18821	0.55 ($p < 0.0001$)
Genic DMPs	16145	424870	0.46 ($p < 0.0001$)
Total DMPs	31312	616162	

*The odds ratio was calculated using the odds of each genic region position (e.g., TSS1500 and TS200) that was given hypomethylated probes. All p-values were obtained from Fisher's exact test.

Abbreviations: UTR, untranslated region; DMP, differentially methylated position

Supplementary Table 5. Number of hypermethylated probes in the CpG island region.

CpG island region	Hypermethylated probes	Total EPIC probes	Odds ratio* (p-value)
S shelf	94	23144	0.28 ($p < 0.0001$)
S shore	674	50025	0.95 ($p = 0.22$)
Island	5856	80705	14.70 ($p < 0.0001$)
N shore	1010	58973	1.25 ($p < 0.0001$)
N shelf	97	25008	0.26 ($p < 0.0001$)
Open sea	960	378307	0.08 ($p < 0.0001$)
Total DMPs	8691	616162	

*The odds ratio was calculated using the odds of each genic region position (e.g., island, shore, and shelf) that was given hypermethylated probes. All p-values were obtained from Fisher's exact test.

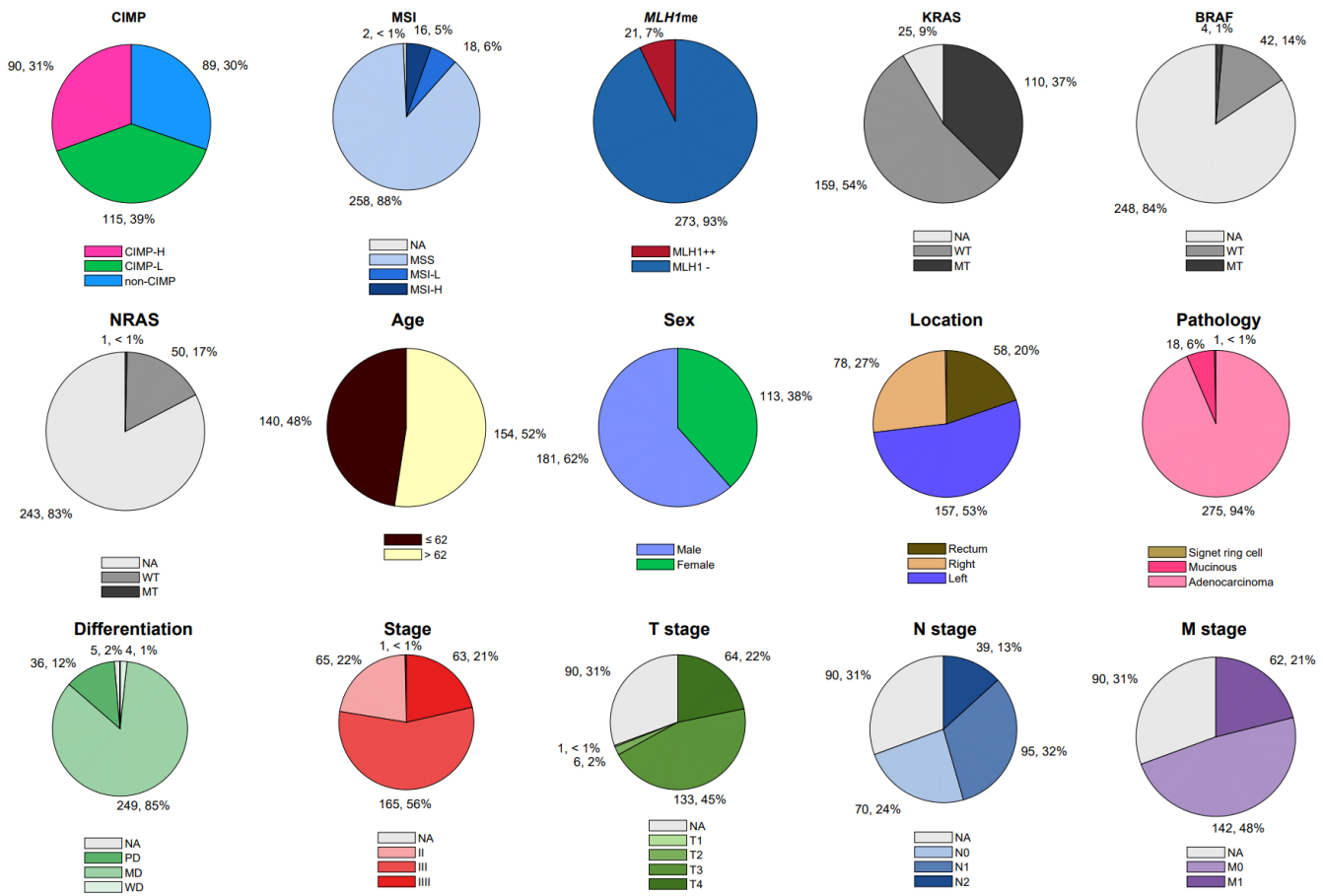
Abbreviation: DMP, differentially methylated position

Supplementary Table 6. Number of hypomethylated probes in the CpG island region.

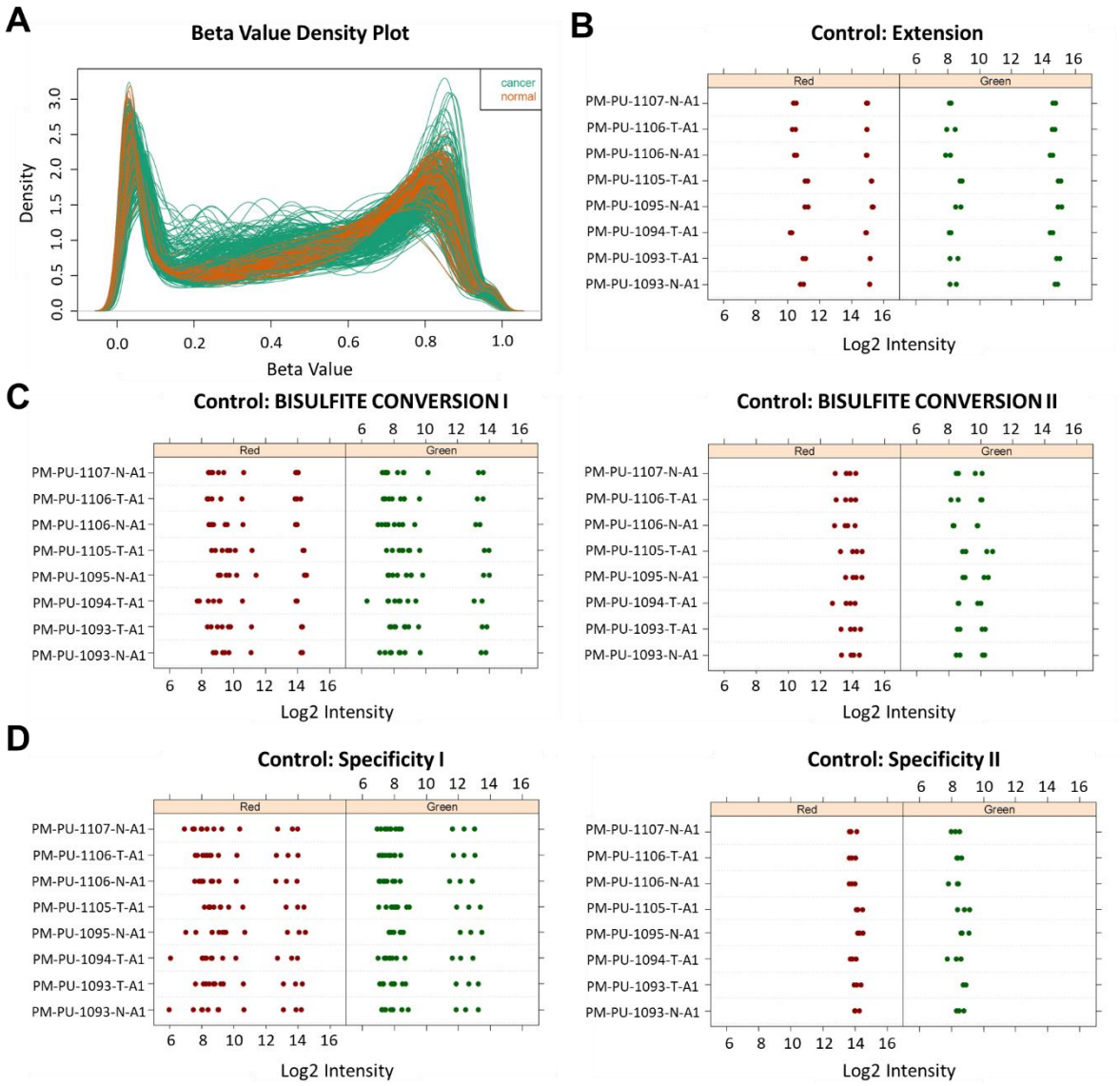
CpG island region	Hypomethylated probes	Total EPIC probes	Odds ratio* (p-value)
S shelf	824	23144	0.68 ($p < 0.0001$)
S shore	967	50025	0.35 ($p < 0.0001$)
Island	274	80705	0.06 ($p < 0.0001$)
N shore	1222	58973	0.37 ($p < 0.0001$)
N shelf	894	25008	0.68 ($p < 0.0001$)
Open sea	27131	378307	4.32 ($p < 0.0001$)
Total DMPs	31312	616162	

*The odds ratio was calculated using the odds of each genic region position (e.g., island, shore, and shelf) that was given hypomethylated probes. The numbers within parentheses represent p-values from Fisher's exact test for each genic region.

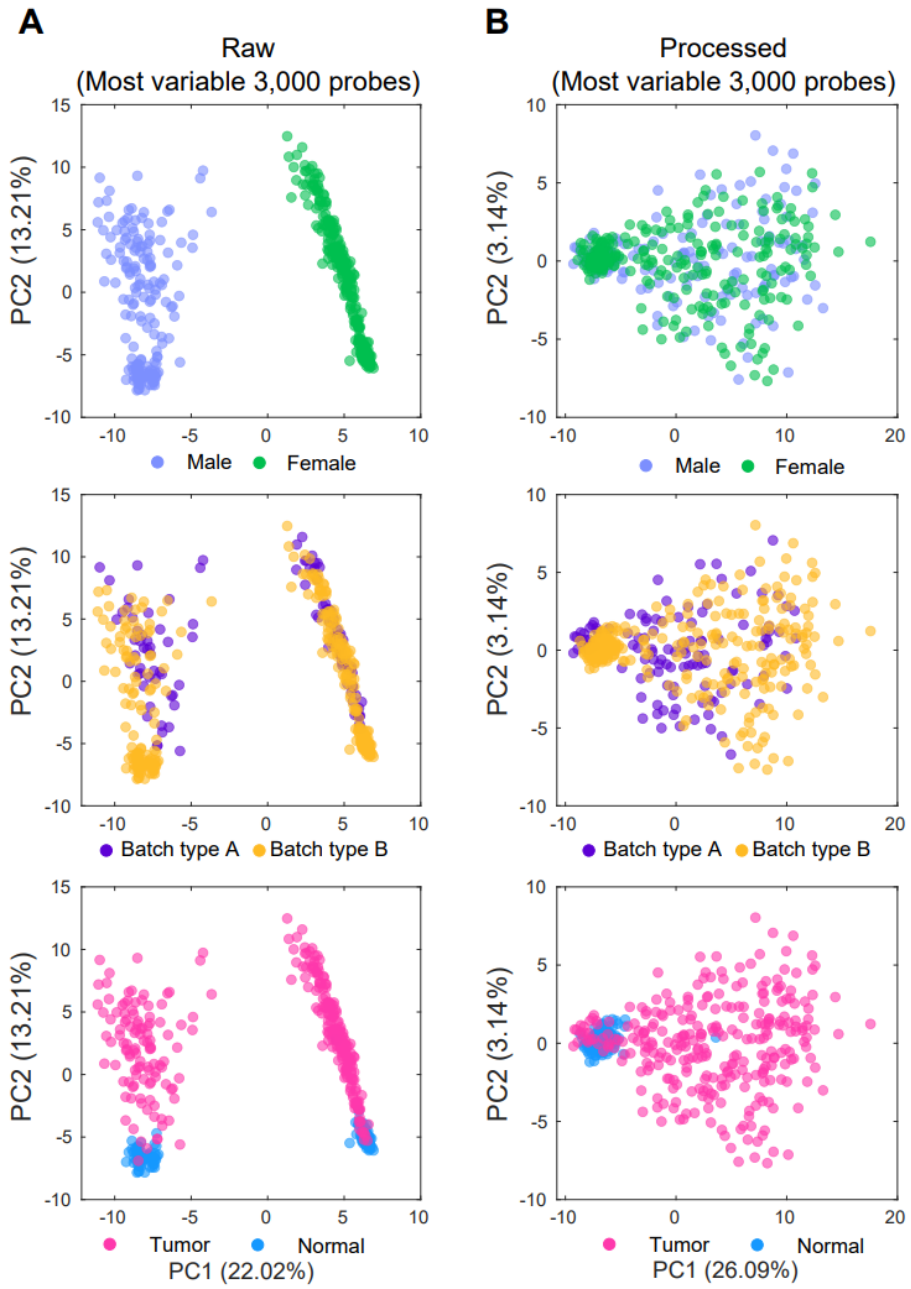
Abbreviation: DMP, differentially methylated position



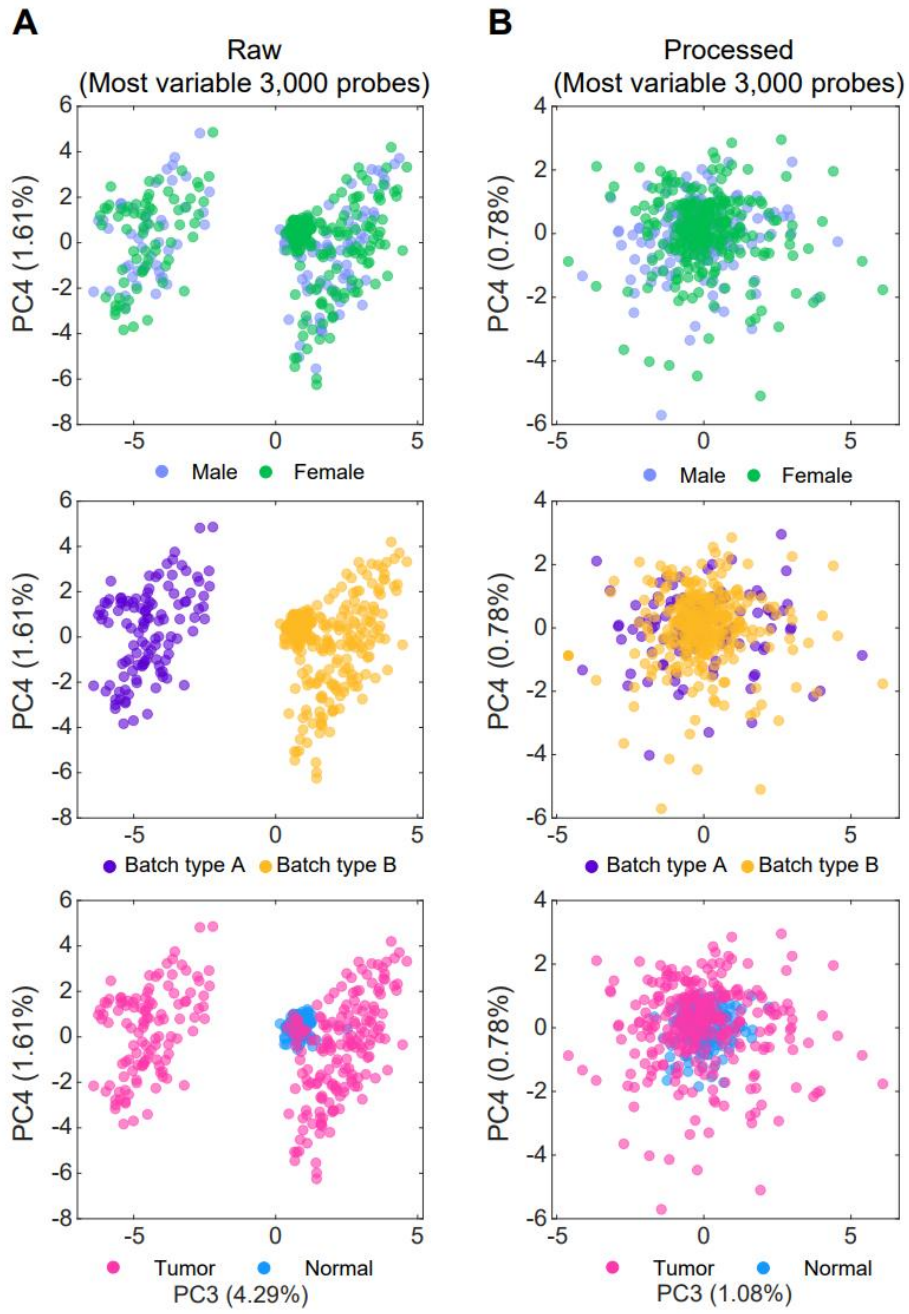
Supplementary Figure 1. Distribution of the characteristics of 294 patients with colorectal cancer. Twelve clinical characteristics, including the CIMP group and *MLH1* methylation statuses, are shown. Abbreviation: CIMP, CpG island methylator phenotype



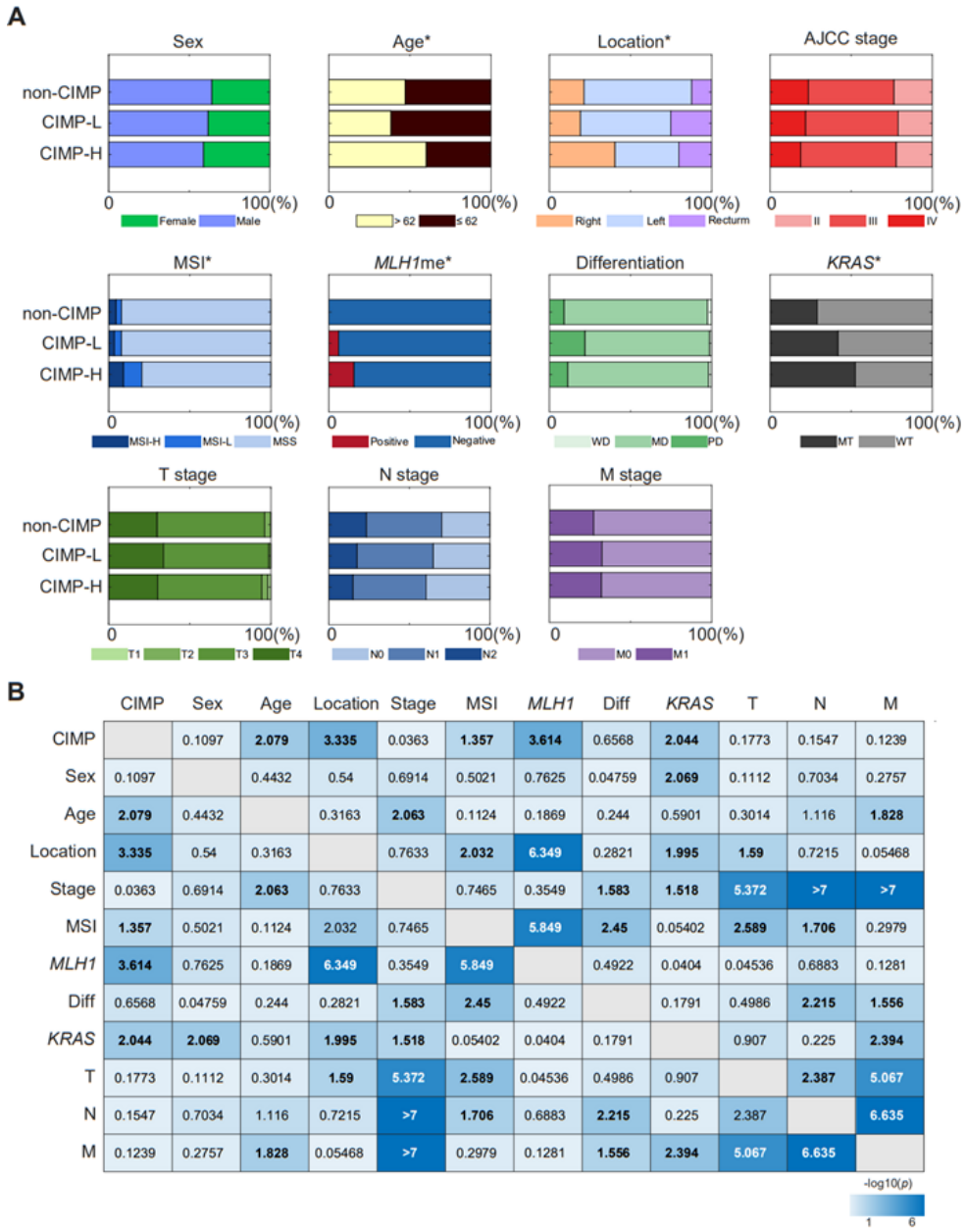
Supplementary Figure 2. Density plot of methylation beta values and control strip plots. (A) Density plot of methylation beta values from individual samples (orange: normal; green: tumor). (B) Control strip plots of extension efficiency, (C) bisulfite conversion efficiency, and (D) specificity.



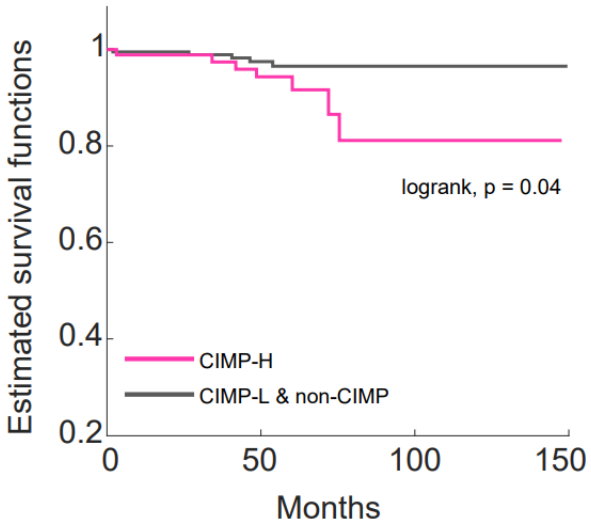
Supplementary Figure 3. Principal component (PC) plot. Among the total beta values of the probes, the top 3,000 variable beta values were used for the PC analysis plot. We tested (A) raw and (B) processed probes according to sex (top: male and female), batch number (middle: batch types), and tumor status (bottom: tumor and normal) with PC1 and PC2.



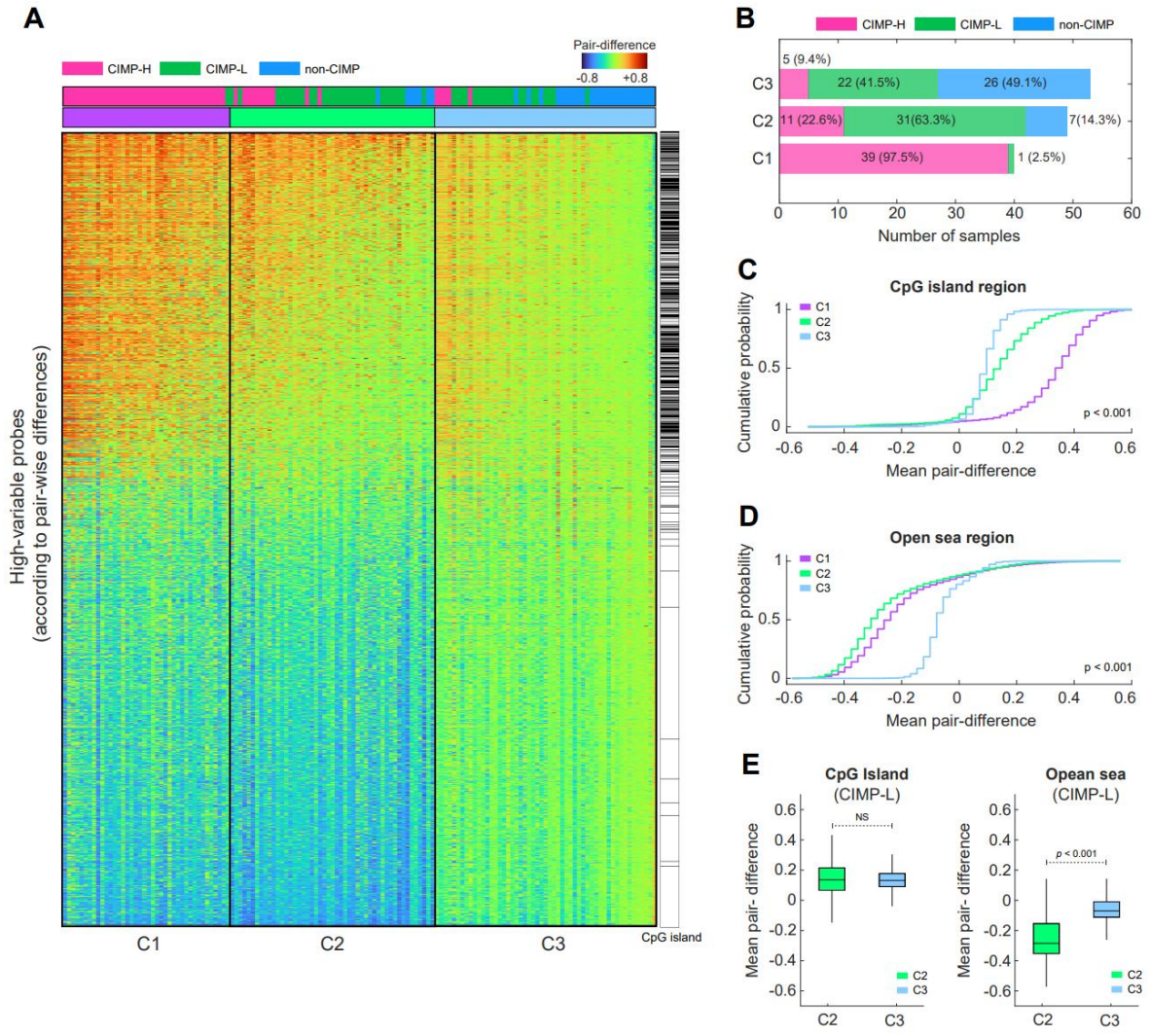
Supplementary Figure 4. Principal component (PC) plot. Among the total beta values of the probes, the top 3,000 variable beta values were used for the PC analysis plot. We tested (A) raw and (B) processed probes according to sex (top: male and female), batch number (middle: batch types), and tumor status (bottom: tumor and normal) with PC3 and PC4.



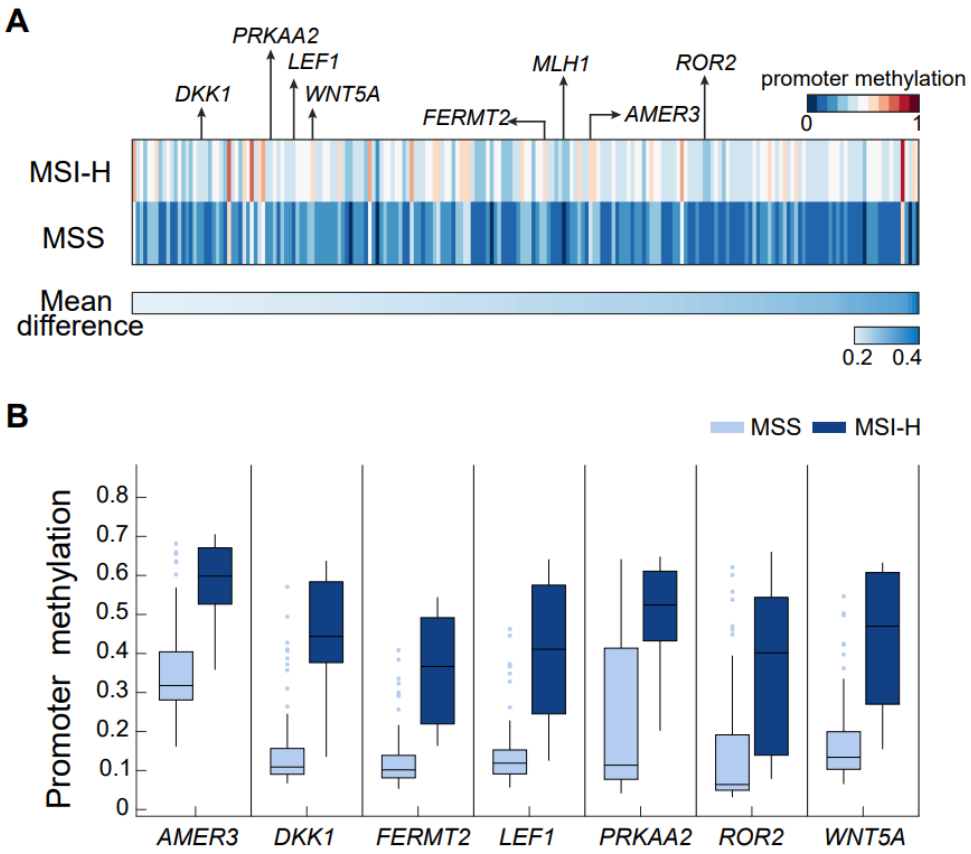
Supplementary Figure 5. Comparison of study participant characteristics with CIMP status. (A) Proportion of Korean patients with colorectal cancer (N = 294) in each clinical characteristic group according to CIMP status. Asterisks (*) represent significance based on chi-square tests corresponding to CIMP groups. (B) Heatmap for the correlation among 12 cancer characteristics, including CIMP and MLH1 methylation status. The bold numbers represent significant correlations based on chi-square tests (-log10 [p] when p<0.05). Diff, T, N, and M represent cancer differentiation and T, N, and M cancer stage, respectively.



Supplementary Figure 6. Comparison of overall survival between CIMP-H and others. Kaplan-Meier plot for CRC patient survival outcomes according to CIMP status. Red and gray represent CIMP-H and non CIMP-H. p represents the significance of the log-rank test between the two groups.



Supplementary Figure 7. Pair-wise difference analysis in 142 matched CRC samples. (A) Heatmap of methylation of 142 matched CRC samples based on pair-differences of the top 10k high-variable probes. The color bar on the top shows the CIMP groups, and the second color bar represents the sub-clusters that were designed using pair-differences of 10k of probes (purple, C1; green, C2; sky-blue, C3). The black-white bar to the right represents the CpG island (black) probes. Samples in each sub-cluster (i.e., C1, C2, and C2) were sorted by mean values of pair-differences in the CpG island region, and probes were sorted by mean values of pair-differences of sub-cluster C2. (B) Proportion of CIMP groups in each sub-cluster from the paired difference analysis. (C, D) Cumulative distribution of the pair-differences in the (C) CpG island and (D) open-sea regions corresponding to C1, C2, and C3. (E) Comparison of pair-differences between C2 and C3 sub-clusters in the CpG island and open-sea regions focusing on CIMP-L.



Supplementary Figure 8. 207 hypermethylated genes in MSI-H compared with MSS samples on CIMP-H. (A) Heatmap of promoter methylation of 207 genes. Each arrow-marked gene represents WNT-related genes except for *MLH1*. (B) Comparison of promoter methylation of seven WNT-related genes between MSS and MSI-H in CIMP-H groups.

REFERENCES

1. 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
2. Maksimovic J, Gordon L and Oshlack A (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13, R44
3. 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
4. Johnson WE, Li C and Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118-127
5. 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
6. Marsaglia G, Tsang WW and Wang J (2003) Evaluating Kolmogorov's Distribution. J Stat Softw 8, 1-4