SUPPLEMENTARY INFORMATION

Deciphering the DNA methylation landscape of colorectal cancer in a Korean cohort

Seok-Byung Lim[1,#], Soobok Joe[2,#], Hyo-Ju Kim[3,#], Jong Lyul Lee[1], In Ja Park[1], Yong Sik Yoon[1], Chan Wook Kim[1], Jong-Hwan Kim[2], Sangok Kim[2], Jin-Young Lee[3], Hyeran Shim[3], Hoang Bao Khanh Chu[3], Sheehyun Cho[3], Jisun Kang[3], Si-Cho Kim[3], Hong Seok Lee[3], Young-Joon Kim[3,4,*], Seon-Young Kim[2,*] & Chang Sik Yu[1,*]

[1]Division of Colon and Rectal Surgery, Department of Surgery, Asan Medical Center, College of Medicine, Ulsan University, Seoul 05505, [2]Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, [3]Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, [4]LepiDyne Co., Ltd., Seoul 04779, Korea

*Corresponding authors. Chang Sik Yu, Tel: +82-2-3010-3494; Fax: +82-2-3010-6701; E-mail: csyu@amc.seoul.kr; Seon-Young Kim, Tel: +82-42-879-8500; Fax: +82-42-879-8519; E-mail: kimsy@kribb.re.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 CONTENTS

Supplementary Results

Sample preprocessing for DNA methylome profile construction

Based on *minfi* pipeline, we first assessed the EPIC array quality by examining the overall beta-value distribution and control strip plots, including bisulfite conversion efficiency, extension quality, and specificity (**Supplementary Figure 1**). We then employed a Subset-quantile Within Array Normalization (SWAN) to correct for technical differences within each array. Next, we addressed the known batch effects specific to each EPIC array batch type by removing 1,050 batch-related probes. For the downstream analysis, we filtered out additional methylation probes, including sex chromosomes (19,164 probes), known SNPs (161,620 probes), and poorly detected sites (1, 990 probes). Subsequently, we calculated the maximum difference range of each probe's beta-value for all samples and excluded 90,405 probes with a maximum beta range of less than 0.1. In total, 610,674 probe methylation beta-values from 172 tumor and matched 128 normal samples were used for downstream analysis. During this process, we compared the beta-value distributions between raw and processed probes using principal component (PC) analysis and confirmed sex- and batch-related biases in the raw beta-values (**Supplementary Figure 2A** and **3A**). After normalization and filtering, we obtained high-quality harmonized data, effectively eliminating technical noise and sex-based biases, as demonstrated by the PC plots in **Supplementary Figures 2B** and **3B.**

Comaprison of methylation patterns in tumors to TCGA COAD and READ data

We next conducted a comparative analysis between the previously published methylome of colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) from The Cancer Genome Atlas (TCGA) and our Korean CRC profile. By using 300,708 overlapped probes, a strong correlation was observed between the Korean CRC profile and TCGA COAD and READ cohorts. This correlation was evident when we observed the mean methylation differences between tumor and normal samples for overlapping probes, as illustrated in **Supplementary Figure 5A**. Identified DMPs in our Korean CRC profile reaffirmed the consistency of results with the methylation levels of tumor and normal samples in overlapping TCGA probes (number of overlapped hypermethylated probes = 4,166, hypomethylated probes = 4,245), as depicted in **Supplementary Figure 5B**. We further extended our investigation to encompass a set of 15 established CRC diagnostic markers. For these known CRC markers, we calculated the differences in promoter methylation between tumor and normal samples and compared them across TCGA cohort and the Korean CRC profile. Of these 15 CRC marker genes, 13 displayed significant differences between tumor and normal samples within our Korean cohort, while in TCGA cohort, all 15 genes showed significant differences, as evidenced in **Supplementary Figure 5C**.

Supplementary Methods

DNA extraction and EPIC array-based methylation assay

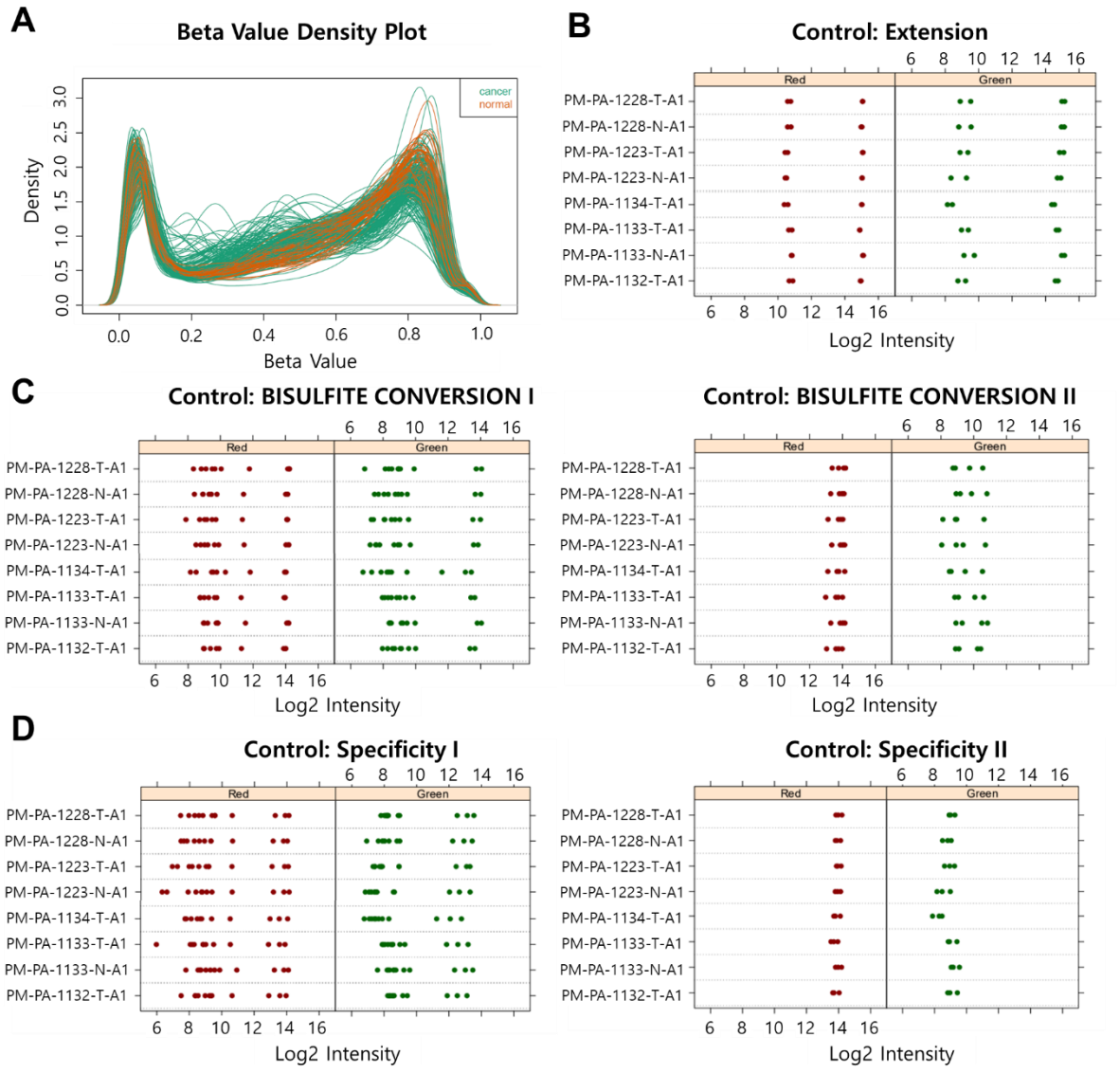
Genomic DNA was isolated from tumors and adjacent normal tissues using the PureLink Genomic DNA Mini Kit (Invitrogen, Waltham, MA, USA). The isolated genomic DNA was quantified using a NanoDrop® (ND-2000, Waltham, MA, USA) and gel electrophoresis (1% agarose gel, 100 V, 30 min). Intact genomic DNA was diluted to 50 ng/μl based on Quant-iT Picogreen (Invitrogen, Waltham, MA, USA) quantitation and subjected to bisulfite conversion using EZ DNA Methylation Kit (ZymoResearch, USA). Subsequently, the converted genomic DNA was amplified up to 1,000-fold using whole-genome amplification and hybridized to the Infinium MethylationEPIC BeadChip (V1; WG-317-1001, Illumina, San Diego, CA, USA), according to the standard Illumina protocol. After completing the single-base extension in the SD, the BeadChip was imaged using the iScan™ system (SY-101-1001, Illumina, San Diego, CA, USA) to yield raw data in IDAT format.

Preprocessing for normalization, batch correction, and probe filtering

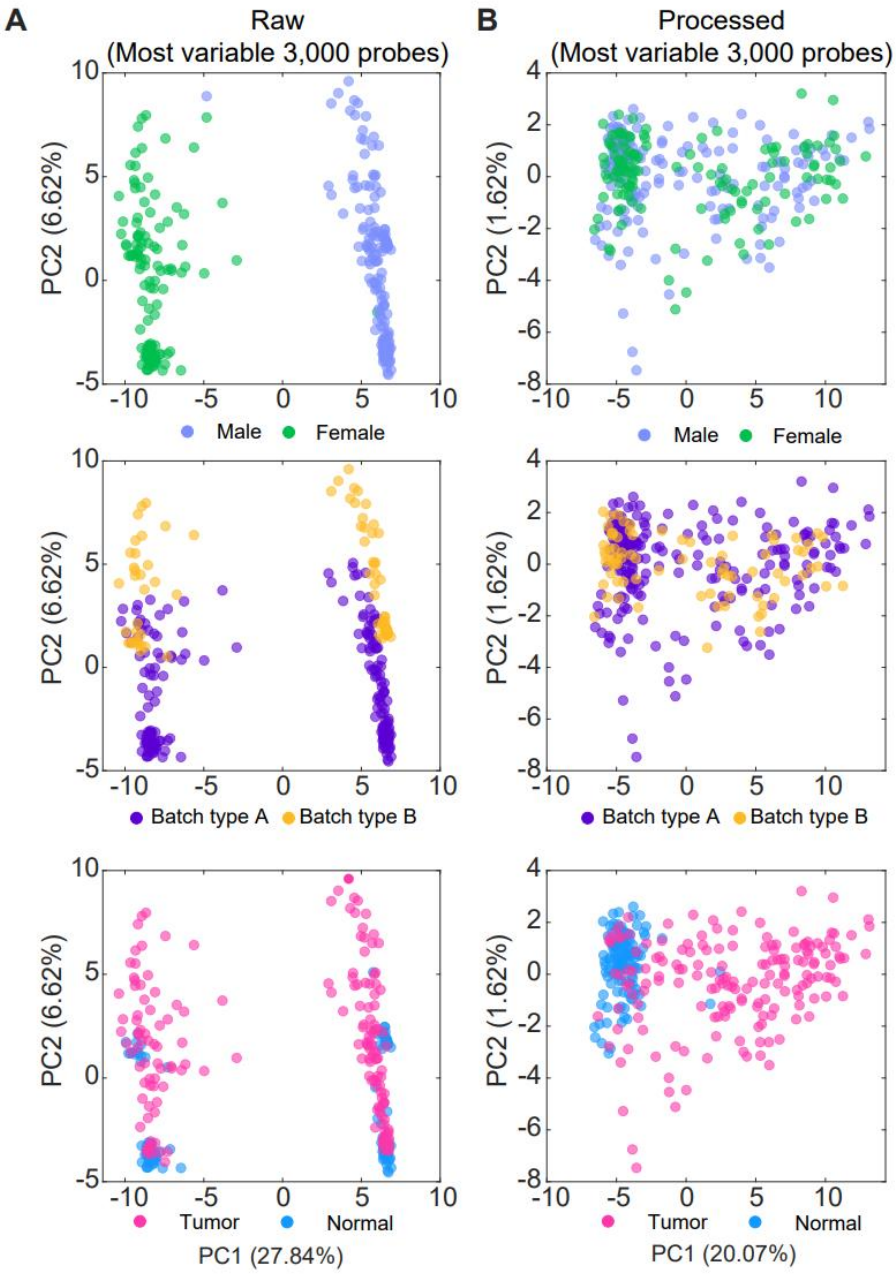
The EPIC array dataset was processed using the established *minfi* pipeline (1). Initially, raw intensities from 865,859 probes were extracted from the green and red channels of raw IDAT files. Following extraction, the SWAN (2) was applied to correct for technical differences between Type I and Type II probes. Batch effects were addressed using Surrogate Variable Analysis (SVA) (3) with the *combat* approach (4). To handle additional batch corrections, we used the sentrix ID information of the array; the dataset consisted of one set before and one set after the 2042203330001 and 2042203330001 sentrix IDs, respectively. During this process, we manually removed 1050 probes based on Illumina EPIC array manual version 1.05B. Subsequently, 254,135 probes were filtered out and sex chromosomes (19,164 probes), known locus of single nucleotide polymorphisms (SNPs) (161,620 probes), poorly detected sites (1,990 probes) with a detection significance of $p > 0.01$, and 90,405 probes with a maximum range < 0.1 were included. Consequently, 610,674 probe methylation beta-values, from 172 tumor and 128 normal samples, were used in downstream analysis.

Statistical analysis

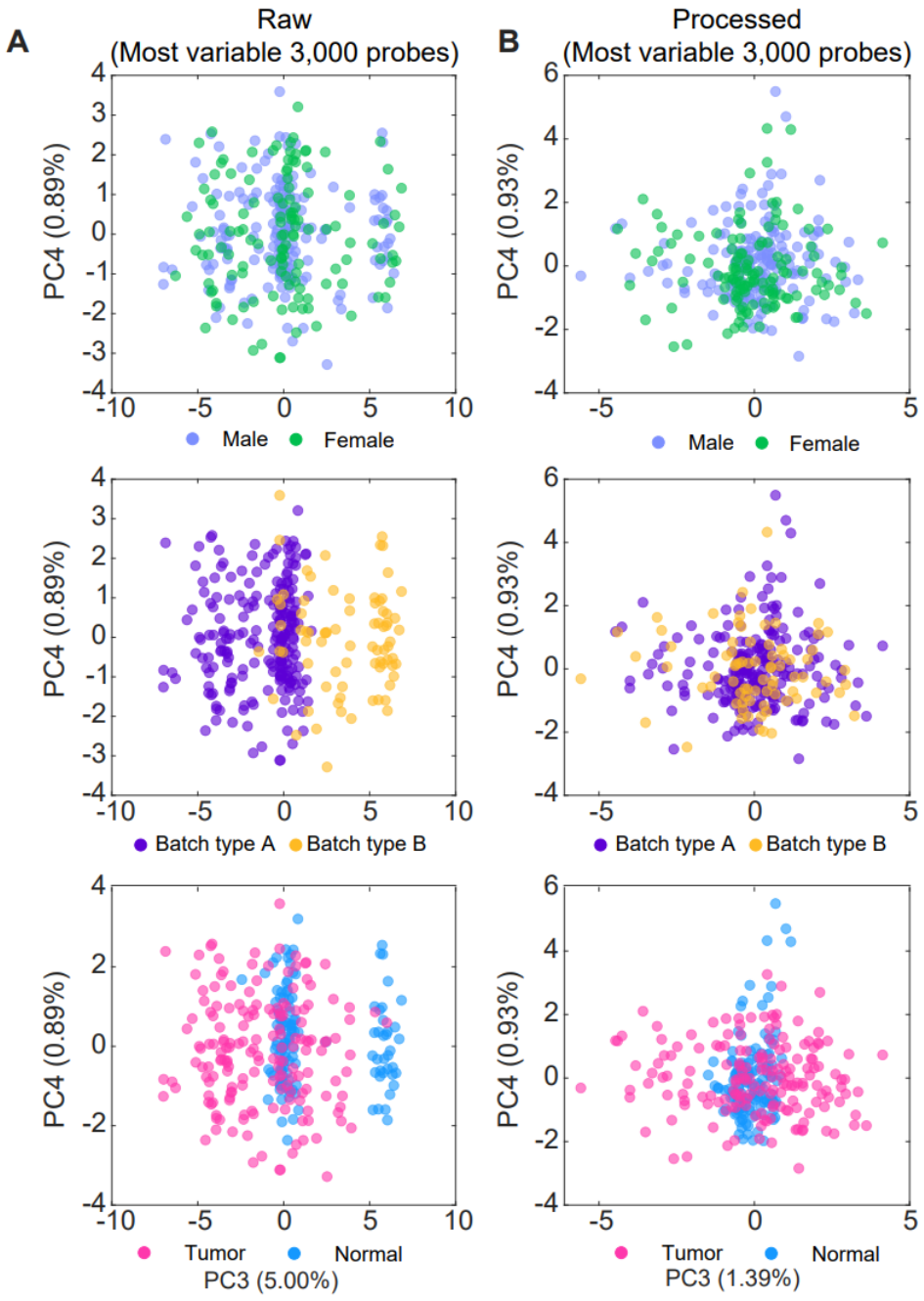
If the mean difference in beta-values between tumor and normal samples surpassed 0.15 and the q-values were below 0.000001, the DMP probes were classified as hyper- and hypo-methylated, using *minfi's dmpfinder* (5) function. We annotated the genomic regions using the EPIC array manual 1.05B (TSS1500:1500 to 200 base pairs upstream of the transcription start site [TSS]; TSS200:200 base pairs upstream of the TSS; Shore:2 kb from each end of the island; Shelf: 2 to 4 kb from the CpG island; Open sea: outside of CpG islands, shores, and shelves). We calculated the odds ratio (OR) of enrichment for each DMP group based on the genomic annotations. To compare with TCGA COAD and READ, probes were selected by using overlapped 450K and 850K arrays. For gene methylation levels, mean beta-values of the probes annotated as promoter-like regions (TSS1500, TSS200, 5'UTR, first exon) were used. Epigenetically *MLH1* silenced samples were classified in relation to *MLH1* methylation if they exhibited a methylation level less than 0.3 All statistical analyses were performed using MATLAB2022a and R software (v4.4).



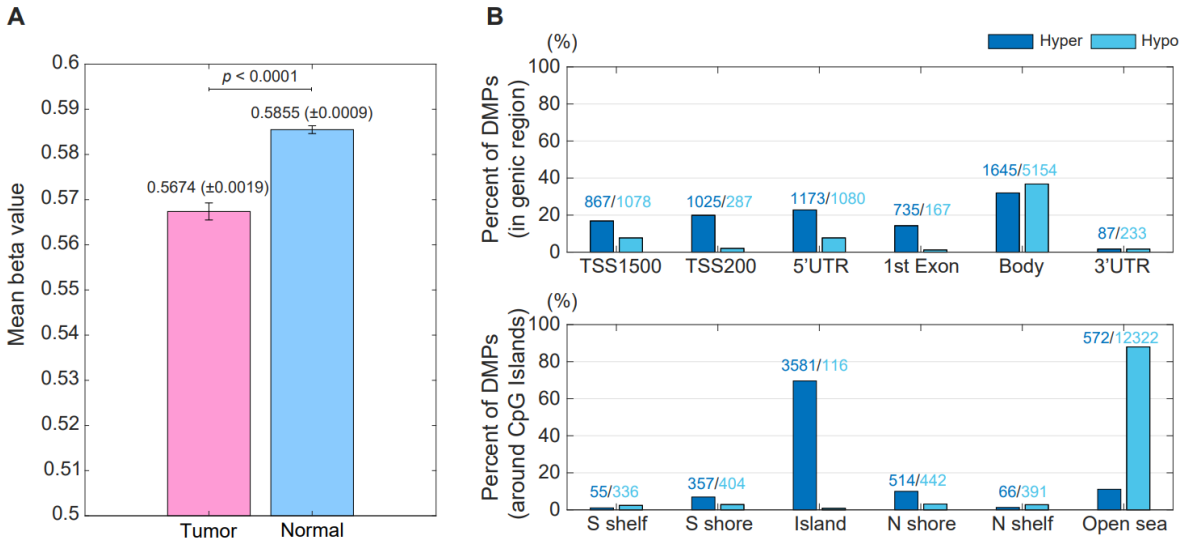
Supplementary Figure 1. Density plot of methylation beta-values and control strip plots. (A) Density plot of methylation beta-values from individual samples (Orange: Normal, Green: Tumor). Control strip plots of (B) Extension efficiency, (C) bisulfite conversion efficiency, and (D) Specificity.



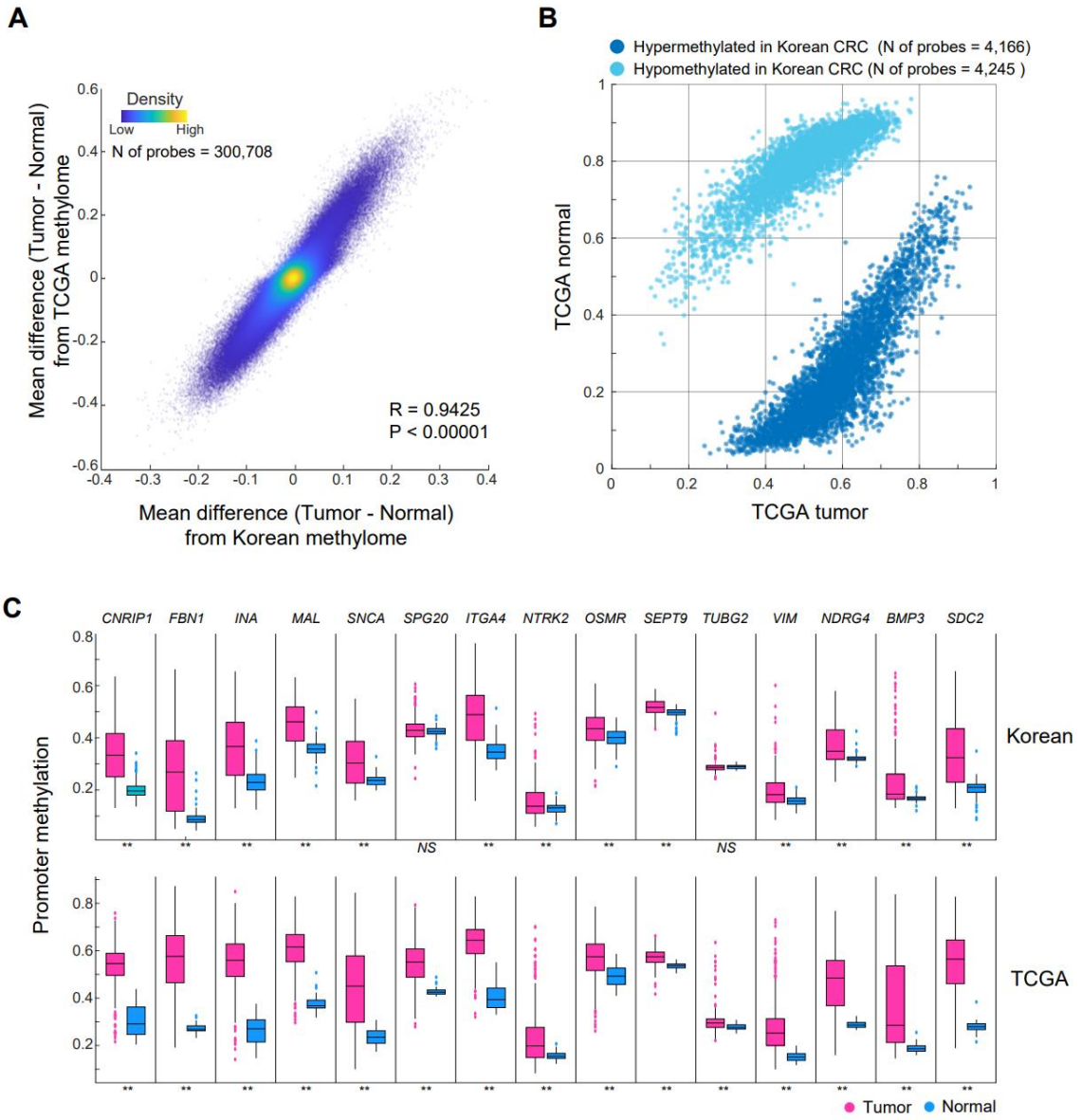
Supplementary Figure 2. Principal component (PC) plot according to PC1 and PC2. Among the total probe beta values, the top 3,000 variable beta values were used to generate the PCA plot. We tested raw (left) and processed (right) cg probes according to gender (top: Male and Female), batch number (middle: batch types), and tumor status (bottom: Tumor and Normal) with PC1 and PC2.



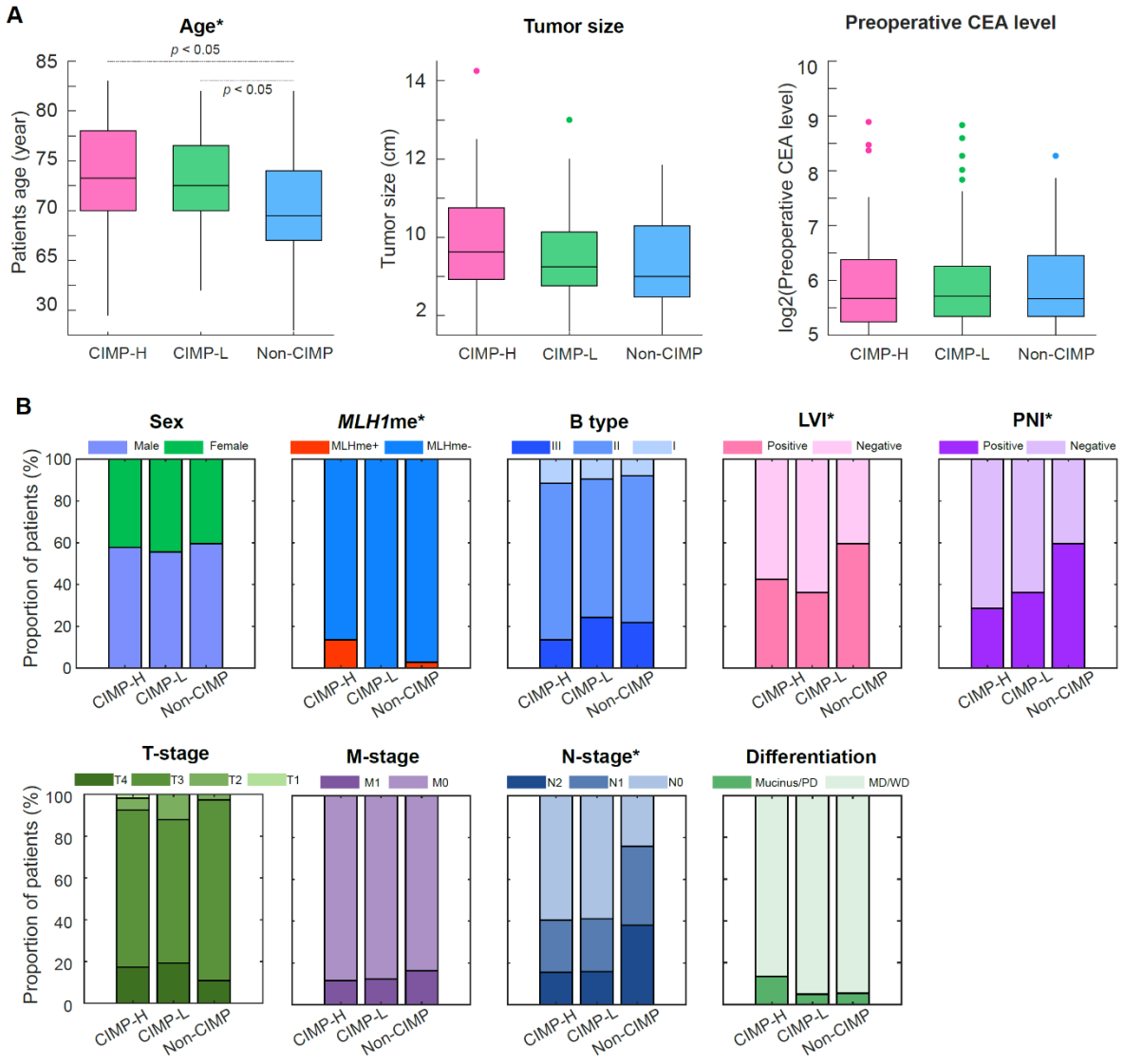
Supplementary Figure 3. Principal component (PC) plot according to PC3 and PC4. Among the total cg probe beta values, the top 3,000 variable beta values were used to generate the PCA plot. We tested raw (left) and processed (right) cg probes according to gender (top: Male and Female), batch number (middle: batch types), and tumor status (bottom: Tumor and Normal) with PC3 and PC4.



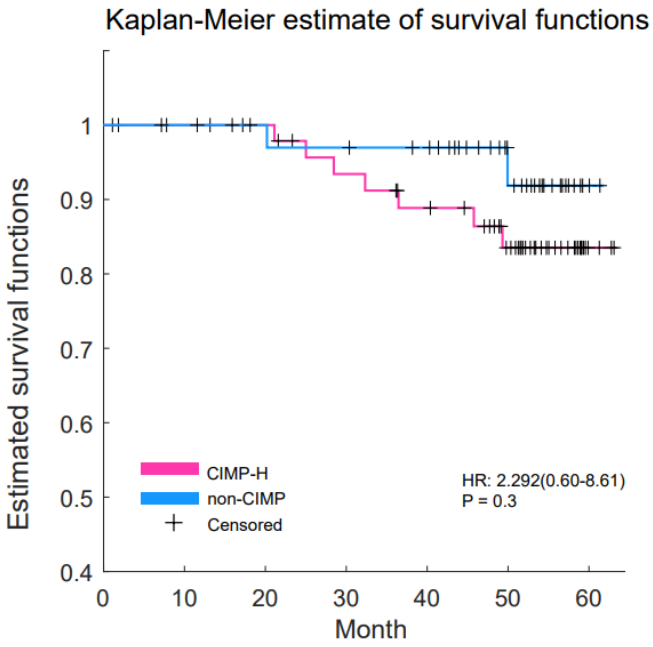
Supplementary Figure 4. Overall methylome levels between tumor and normal samples and DMP distribution based on genomic region. (A) Grand mean methylation levels in tumor and normal tissue samples. The numbers above each bar represent the mean beta values for all tumor and normal samples, with the standard errors seen in parentheses. (B) Distribution of DMPs in genic regions (upper graph) and around CpG islands (lower graph). Y-axis represents the percent of regional DMPs among total DMPs for each hyper- and hypo-methylated probe set, and the numbers above each bar indicate the count of hyper- and hypo-methylated probes in each respective region.



Supplementary Figure 5. Methylome comparison with TCGA. (A) Correlation of Korean CRC methylome with TCGA based on mean differences of overlapped probes (n=300,708) between tumor and normal tissue samples. Pearson's correlation coefficient “R” and statistical significance “P” are shown. (B) Comparison of mean methylation levels between tumor and normal samples in TCGA using the differentially methylated probes identified in this study, and overlapped with the probes from TCGA 450K methylome. The hypermethylated and hypomethylated probes are represented by sky-blue and blue dots, respectively. (C) Comparison of promoter methylation levels in Korean and TCGA CRC samples for the 15 known diagnostic CRC marker genes. Tumor and normal samples are denoted by magenta and blue bars, respectively.



Supplementary Figure 6. Representation of study participant characteristics with CIMP status. Proportion of Korean patients with colorectal cancer (n = 172) in each clinical characteristic group according to CIMP status. The cancer characteristics with *notation represent the significance of chi-square test or post-hoc test, corresponding to CIMP groups. LVI and PNI represent lymphovascular invasion and perineural invasion. CEA represents carcinoembryonic antigen. B type represents Bormann type.



Supplementary Figure 7. Comparisons of patient survival between CIMP-H and non-CIMP. Kaplan - Meier plot for survival outcomes of CRC patients according to CIMP status. Magenta and blue represent CIMP-H and non-CIMP. HR and *p* represent hazard ratio and the significance of log-rank test between the two 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