Received: 16 October 2018 | Revised: 26 December 2018 | Accepted: 28 December 2018

DOI: 10.1111/cas.13937

ORIGINAL ARTICLE

WILEY Cancer Science

Differential regulation of CpG island methylation within divergent and unidirectional promoters in colorectal cancer

Shinichi Namba[1] | Kazuhito Sato[2] | Shinya Kojima[1] | Toshihide Ueno[1] | Yoko Yamamoto[2] | Yosuke Tanaka[1] | Satoshi Inoue[1] | Genta Nagae[3] | Hisae Iinuma[4] | Shoichi Hazama[5] | Soichiro Ishihara[2] | Hiroyuki Aburatani[3] | Hiroyuki Mano[1] | Masahito Kawazu[1]

[1]Division of Cellular Signaling, National Cancer Center Research Institute, Tokyo, Japan

[2]Department of Surgical Oncology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[3]Genome Science Division, Research Center for Advanced Science and Technologies, The University of Tokyo, Tokyo, Japan

[4]Department of Surgery, Teikyo University School of Medicine, Tokyo, Japan

[5]Department of Gastroenterological, Breast and Endocrine Surgery, Yamaguchi University Graduate School of Medicine, Yamaguchi, Japan

Correspondence
Masahito Kawazu, Division of Cellular Signaling, National Cancer Center Research Institute, Tokyo, Japan.
Email: mkawz-tky@umin.ac.jp

Funding information
Japan Agency for Medical Research and Development, Grant/Award Number: JP17am0001001 and JP17cm0106502; Japan Society for the Promotion of Science, Grant/Award Number: 16K07143 and 17J00386

The silencing of tumor suppressor genes by promoter CpG island (CGI) methylation is an important cause of oncogenesis. Silencing of *MLH1* and *BRCA1*, two examples of oncogenic events, results from promoter CGI methylation. Interestingly, both *MLH1* and *BRCA1* have a divergent promoter, from which another gene on the opposite strand is also transcribed. Although studies have shown that divergent transcription is an important factor in transcriptional regulation, little is known about its implication in aberrant promoter methylation in cancer. In this study, we analyzed the methylation status of CGI in divergent promoters using a recently enriched transcriptome database. We measured the extent of CGI methylation in 119 colorectal cancer (CRC) clinical samples (65 microsatellite instability high [MSI-H] CRC with CGI methylator phenotype, 28 MSI-H CRC without CGI methylator phenotype and 26 microsatellite stable CRC) and 21 normal colorectal tissues using Infinium MethylationEPIC BeadChip. We found that CGI within divergent promoters are less frequently methylated than CGI within unidirectional promoters in normal cells. In the genome of CRC cells, CGI within unidirectional promoters are more vulnerable to aberrant methylation than CGI within divergent promoters. In addition, we identified three DNA sequence motifs that correlate with methylated CGI. We also showed that methylated CGI are associated with genes whose expression is low in normal cells. Thus, we here provide fundamental observations regarding the methylation of divergent promoters that are essential for the understanding of carcinogenesis and development of cancer prevention strategies.

KEYWORDS
colorectal neoplasms, CpG Islands, DNA methylation, microsatellite instability, oligonucleotide array sequence analysis

1 | INTRODUCTION

Silencing of tumor suppressor genes by promoter CpG island (CGI) methylation is one of the major driver events that play important roles during tumor initiation and/or progression. For example, previous studies showed that silencing of *MLH1* and *BRCA1* causes microsatellite instability high colorectal cancer (MSI-H CRC) and triple negative breast cancer with homologous recombination deficiency, respectively.[1,2] In addition, silencing of *CDKN2A* is prevalent among several cancers including gastric adenocarcinoma, lung squamous cell carcinoma and esophageal adenocarcinoma.[3-5] The mechanistic basis of aberrant CGI methylation in cancer is largely unknown with only a few clues: a previous study demonstrated that mutations in *IDH1/2* and *TET2* can cause aberrant CGI methylation in a hematological malignancy,[6] and *Fusobacterium* colonization is associated with aberrant CGI methylation in CRC.[7] Further, there appears to be a specificity for targeted methylation of certain genes in disease. Among the genes associated with Lynch syndrome that encode proteins involved in DNA mismatch repair, such as *MSH2*, *MSH6* and *PMS2*, it is intriguing that only *MLH1* is prone to silencing with promoter CGI methylation. *BRCA1* but not *BRCA2* is silenced by promoter CGI methylation in breast cancer with defective homologous recombination.[8,9] This specificity of the genes affected by aberrant CGI methylation is an important issue to be addressed in clarifying the mechanism underlying promoter methylation in disease.

The completion of human genome sequencing revealed that more than 10% of human genes are associated with divergent promoters.[10] Owing to the advent of next-generation sequencing, tens of thousands of long non-coding RNA (lncRNA) have been identified in the past decade.[11] A substantial part of lncRNA are reported to be transcribed from promoters of protein-coding genes in the opposite direction,[12] despite gene promoters being intrinsically directional.[13,14] Although some lncRNA act in trans, recent investigations have demonstrated that lncRNA mainly regulate neighbor genes in cis by various manners: as scaffolds, as sRNA sponges or by transcription itself. Accumulating evidence has shown that transcriptional regulation is in part mediated by divergent transcription.[15-17] While the expression of several pairs of genes was inversely correlated with the methylation of CGI within the corresponding divergent promoters in cancer cell lines,[18] little is known about the implication of divergent transcription in the aberrant promoter CGI methylation seen in cancer. With the growing catalogue of lncRNA, it has become increasingly important to evaluate CGI within divergent promoters (div-CGI) associated with protein-coding genes and newly discovered lncRNA.

There are two major subtypes of CRC: the microsatellite stable subtype (MSS) and the MSI-H subtype. While MSI-H CRC harbor a large number of nucleotide substitutions caused by defective DNA mismatch repair machinery, MSS CRC are characterized by chromosomal instability.[19] MSI-H CRC, in which *MLH1* is functionally defective owing to silencing or mutation, include a subset called CGI methylator phenotype (CIMP). CIMP is found in various types of cancers, and was reported to be a clinically distinct subset in CRC.[20] Using the recently enriched transcriptome database and our recently published data from genome-wide methylation analysis of MSI-H CRC,[21] here we analyzed the methylation status of div-CGI and found that div-CGI were less methylated compared with CGI within unidirectional promoters (uni-CGI) in normal colon cells as well as in CRC cells. These results provide important insights to understand the aberrant promoter CGI methylation in cancer cells.

2 | MATERIALS AND METHODS

2.1 | Clinical specimens

The data used in this study were obtained and described in the previous report.[21] Patients with CRC gave written informed consent prior to their participation in the study. This project was approved by the institutional ethics committees of the University of Tokyo (The Human Genome, Gene Analysis Research Ethics Committee; G10063 and G3546), Teikyo University (#14-197) and Yamaguchi University (H17-83).

2.2 | Genome-wide DNA methylation analysis

Genome-wide DNA methylation analysis was performed with the Infinium Human MethylationEPIC BeadChip (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. We excluded probes that had single nucleotide polymorphism in ±5 bp. While *M*-value and β-value have been used as a general index of DNA methylation, we chose *M*-value because it was reported to have a higher detection power of methylation.[22] The extent of methylation was first examined by β-value, which was then put into the logit-like function to obtain the correlating *M*-value (slightly modified from reference [22]).

$$M = \log_2 \frac{\beta + 2^{-25}}{1 - \beta}$$

The reason for this modification was to convert the probe with a β-value of 0. The *M*-value was calculated for the probe with the smallest β-value above 0, which turned out to be approximately −21; therefore, 2^{-25} was added in the *M*-value calculation equation.

The location of the probe was calculated using the Liftover tool, from Hg19 to Hg38. The mean *M*-value of the probes on the island was used for the calculation of *M*-value as a CGI unit. The location of CGI was obtained using the UCSC Table Browser[23] on 15 November 2017. CGI with the number of valid probes under four were excluded from analysis.

2.3 | Methylation of promoter CpG islands

We defined promoter CGI as CGI that are located 0-500 bp upstream of transcript start sites (TSS). As described in the "Results" section, CpG islands were considered to be methylated when the *M*-value was over −1.6; when the *M*-value was below −1.6, the CpG island was considered as unmethylated.

2.4 | Phenotype-specific methylated CpG islands

The CGI that were specifically methylated in non-CIMP MSI-H or CIMP MSI-H CRC were identified with the *F*-test using Minfi package.[24] Non-CIMP MSI-H-specific CGI were defined as those that fulfill both of the following conditions: methylated in non-CIMP MSI-H CRC (median *M*-value, >−1.6) and unmethylated in normal samples (median *M*-value, ≤−1.6). CIMP MSI-H-specific CGI were limited to those that are methylated in CIMP MSI-H CRC and unmethylated in non-CIMP MSI-H CRC and normal samples.

2.5 | Forward genes

In each gene pair, genes with greater FPKM were calculated per sample. Data were obtained for all samples that underwent RNA-seq, and genes with a larger number of samples that had greater FPKM were defined as the forward gene.

2.6 | Motif analysis

MEME[25] was used to find methylated group-specific motifs. The setting of MEME was in discriminative mode, number of motifs was set as five and others were set as default in meme-suite.org/tools/meme-chip. Sequences 0-500 bp upstream of average forward TSS positions were scanned for motifs using FIMO[26] with default setting in meme-suite.org/tools/meme-chip. Motif matching was limited to those with $q < 0.05$. We used the motifs detected in 0-500 bp upstream of TSS of forward genes with CGI that are specifically methylated in any of three cell types: normal cells, non-CIMP MSI-H CRC cells or CIMP MSI-H CRC cells.

2.7 | Logistic regression

For logistic regression, we adopted generalized linear regression, and for stratified sampling, we used train data and test data in the ratio of 7:3. Existence of the motifs and the bidirectionality (C/C pairs, C/L pairs, L/L pairs, unidirectional [protein coding] and unidirectional [lncRNA] promoters) were used as explanatory variables, and the response variable was whether correspondent CpG islands were methylated in any of the three cell types: normal cells, non-CIMP MSI-H CRC or CIMP MSI-H CRC. For the existence of motifs, we did not differentiate whether there was more than one matched motif or not. We used two-sided DeLong's test for comparison of area under the receiver-operating characteristic curve.

2.8 | Statistics

Comparisons of the distribution of categorical variables in different groups were performed using χ^2-test. False discovery rate (FDR) was obtained using Benjamini-Hochberg method with some modification. Statistical analysis was performed using the computing environment R.

2.9 | GTEx

The median transcript per million (TPM) of sigmoid colon cells was used to examine the expression of normal colon cells. These data were downloaded from the GTEx portal (https://www.gtexportal.prg)[20] on 15 June 2018. The name of the file is GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.

2.10 | Data accessibility

Raw sequencing data were deposited in the Japanese Genotype-Phenotype Archive (http://trace.ddbj.nig.ac.jp/jga) under accession number JGAS00000000113 (National Bioscience Database Center no. hum0094).

3 | RESULTS

3.1 | Definition of divergent promoters

We first selected pairs of genes with head to head (HtH) orientation among 18 730 protein-coding genes and 29 413 lncRNA obtained from the Gencode database (Gencode_v27 all comprehensive gene annotation; Chr Filter, Autosome Only; Biotype Filter, Coding or LncRNA). For every transcript on the plus strand, the transcript encoded on the opposite strand and whose TSS was nearest to and upstream of the transcript was identified as a partner transcript with HtH orientation (Figure 1A). Given that most genes yield more than one transcript variant, other variants of the partner genes (whose TSS fulfilled these conditions) were also determined as partner transcripts. Partner transcripts were also identified for every transcript on the minus strand. A set of pairs of transcripts with HtH orientation ("HtH transcript pairs") was determined by combining the identified pairs and excluding duplicates. Further, HtH transcript pairs composed of an identical pair of genes were combined to determine a set of pairs of genes with HtH orientation. Distances between the paired genes were defined as the average of distances between the paired transcripts. Because the peak of transcript density, which were analyzed by RIKEN's CAGE-seq, accorded with the annotations of TSS for bidirectional promoters from the UCSC Human Known Genes database,[27] the variation of the TSS obtained from public databases seemed to have little influence on subsequent analyses.

A total of 4387 pairs of protein-coding genes (C/C pairs) were identified, whereas 7445 pairs of protein-coding gene-lncRNA (C/L pairs) and 2949 pairs of lncRNA pairs (L/L pairs) were identified (Figure 1B). From the biphasic distribution of the distances, it was assumed that a subset of paired genes was placed closer to each other than expected by random distribution (Figure 1C). Thus, we defined the genomic regions between the TSS of paired genes with less than 1000 bp distance as divergent promoters in this study, in accordance with the previous report.[10] We next selected genes not containing any HtH transcripts and defined their promoters as unidirectional for comparison.

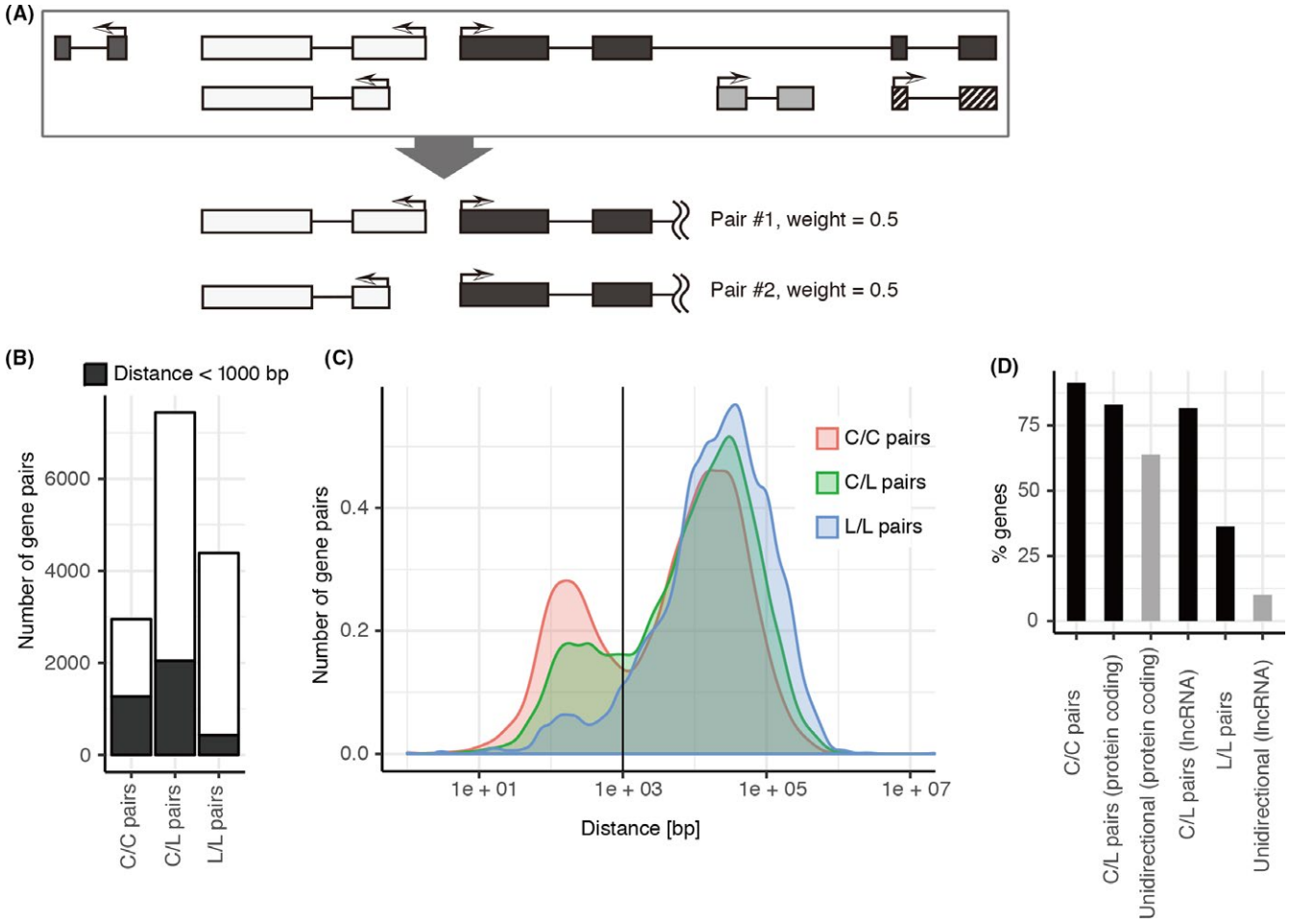

FIGURE 1 Definition and positional characteristics of divergent promoters. A, A schematic of head to head (HtH) oriented transcripts. An example of an HtH pair (white on left, black on right) is shown. The neighboring transcript (black striped box) is not selected when there is a transcript derived from other genes between the striped and white boxed transcript (boxes shown in gray). B, Bar plots show the number of gene pairs. Gene pairs with the distance of less than 1000 bp are shown in black. C/C, protein-coding genes (4387 pairs); C/L, protein-coding gene long non-coding RNA (lncRNA) (7445 pairs); L/L, lncRNA pairs (2949 pairs). C, The distribution of distances between the transcription start site (TSS) of paired genes. The vertical line indicates 1000 bp. D, Bar plots show the fraction of genes containing CpG islands

3.2 | CGI within the divergent promoters

The proportions of divergent and unidirectional promoters containing CGI were calculated (Figure 1D). CGI were more frequently observed in divergent promoters with C/C or C/L pairs than in unidirectional promoters with protein-coding genes (FDR < 2.2e-16, FDR = 1.7e-13, respectively). L/L pairs were also more likely to have CGI in their promoters than unidirectional lncRNA genes (FDR = 0.012). This observation was in accordance with those from a previous report.[10] To exclude effects caused by the difference of CGI proportion, we only included genes whose promoters contained CGI for further analysis.

3.3 | Methylation status of CGI within divergent and unidirectional promoters

During the integrative genomic analysis of MSI-H CRC tumor samples, we measured the methylation of CGI using the Infinium MethylationEPIC Kit in 119 clinical specimens of CRC (65 MSI-H CIMP CRC, 28 MSI-H non-CIMP CRC and 26 MSS CRC) and 21 normal colorectal samples. *M*-values were calculated as an index for DNA methylation. The *M*-values of all promoter CGI across the analyzed samples showed a biphasic distribution (Figure 2A). As the intersection of the peaks was approximately −1.6 in *M*-value, we defined CGI with *M*-values higher than −1.6 as methylated.

The *M*-values of each CGI in the normal samples are represented as box plots in Figure 2B,C. The box plots are arranged in order of the median. div-CGI with C/C or C/L pairs contained significantly less methylated CGI than uni-CGI with protein-coding genes (Figure 2B,D, FDR < 2.2e-16, FDR < 2.2e-16, respectively). Similar differences were observed between div-CGI with L/L pairs and uni-CGI with lncRNA (Figures 2C,D). As shown in the box plots, uni-CGI consisted of much more CGI whose *M*-values were distributed across or above the value of −1.6, indicating the frequent methylation. In contrast, div-CGI with C/C or C/L pairs contained more CGI whose *M*-values distributed below −1.6, indicating stable unmethylation.

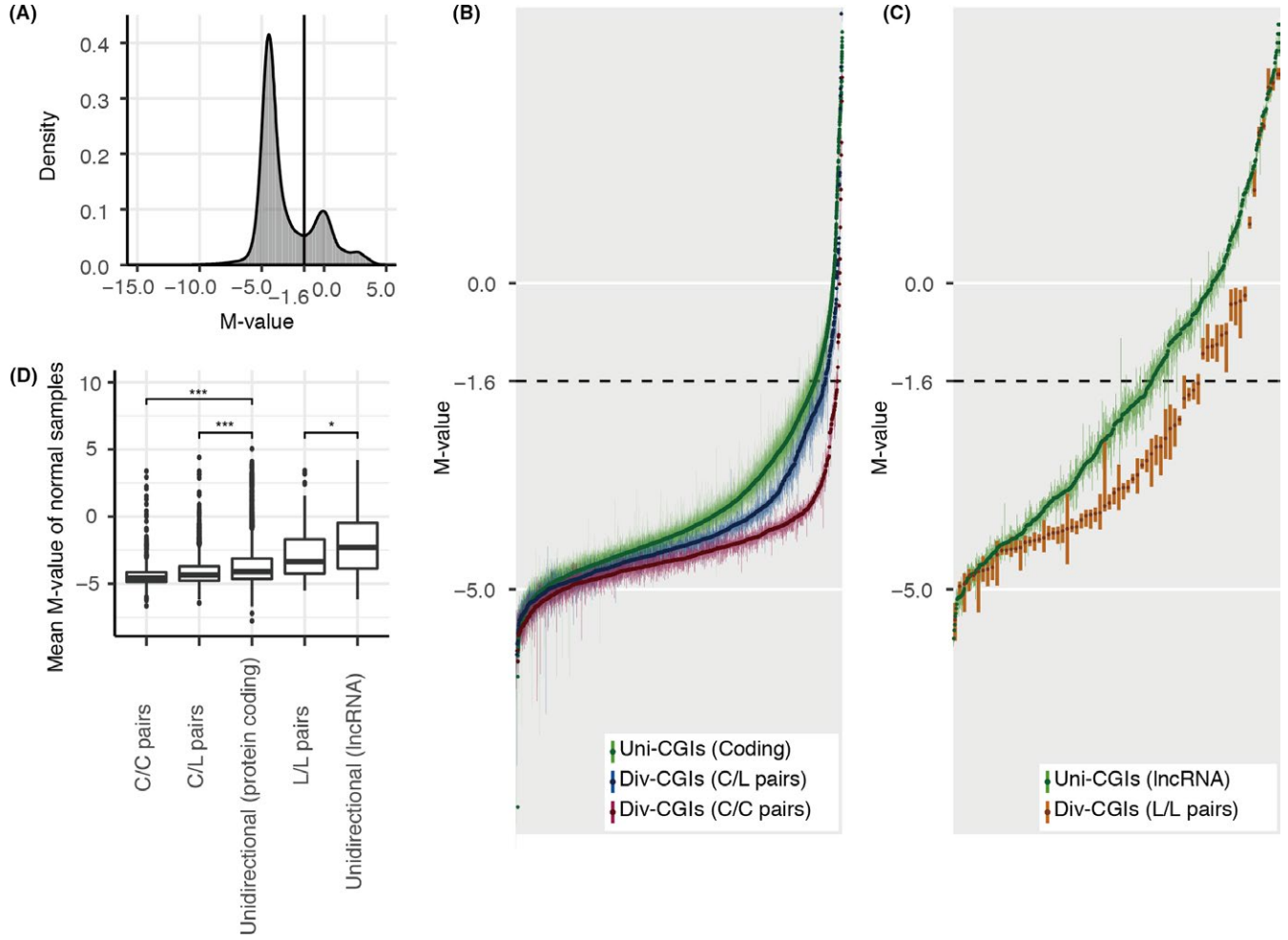

FIGURE 2 Methylation of CpG islands (CGI) within divergent promoters in normal cells. A, Distribution of *M*-values of all promoter CGI in all samples. The estimated *M*-value dividing the two major peaks is −1.6 (indicated by vertical line). B,C, The plots show box plots representing the *M*-values of CGI associated with protein-coding genes (B) and those associated with long non-coding RNA (C). Both data are from normal samples. Box plots are arranged in the order of *M*-values; the dark points are the medians. div-CGI, CGI within divergent promoters; uni-CGI, CGI within unidirectional promoters; C/C pair, pair of protein-coding genes; C/L pair, pair of protein-coding gene and long non-coding RNA; L/L pair, pair of long non-coding RNA. D, The box plots represent the mean *M*-value of CGI associated with divergent and unidirectional promoters in normal samples. *False discovery rate (FDR) < 0.05, ***FDR < 0.001

This observation suggested that the methylation of promoter CGI is regulated differently according to the positional relationship with other genes, implying the existence of underlying regulatory mechanisms that protect div-CGI from methylation.

3.4 | Methylation status of CGI within divergent and unidirectional promoters in CRC cells

Next, we compared *M*-values of promoter CGI between CRC subtypes (non-CIMP MSI-H vs normal samples is shown in Figures 3A,B, and S1A-E; CIMP MSI-H vs non-CIMP MSI-H is shown in Figures 3C,D, and S1F-J; MSS vs non-CIMP MSI-H is shown in Figure S1K-O). The box plots are arranged in the order of median *M*-values of the latter populations. The methylation profiles of MSS and non-CIMP MSI-H subtypes were concordant on the whole and were supposed to be general methylation profiles of CRC without CIMP (Figure S1K-O). Intriguingly, methylation profiles of div-CGI with C/C pairs were similar among MSS, non-CIMP MSI-H, and CIMP MSI-H CRC (Figure 3A,C). In contrast, there was a large number of uni-CGI that were methylated in non-CIMP MSI-H CRC but not in normal tissues (Figure 3B) and those that were methylated in CIMP MSI-H CRC but not in non-CIMP MSI-H CRC (Figure 3D).

The proportion of CGI that are specifically methylated in non-CIMP MSI-H CRC or in CIMP MSI-H CRC is presented along with that of CGI that are methylated in normal cells in Figure 3E-G. As in normal cells, uni-CGI were more frequently methylated in CRC cells than div-CGI: 0.96% and 2.5% of div-CGI with C/C pairs were specifically methylated in non-CIMP MSI and CIMP MSI-H CRC, respectively, while 5.7% and 10.4% of uni-CGI with protein-coding genes were specifically methylated in non-CIMP MSI-H and CIMP MSI-H CRC, respectively. This observation indicated the possible existence of a methylation-defense mechanism that is effective in div-CGI.

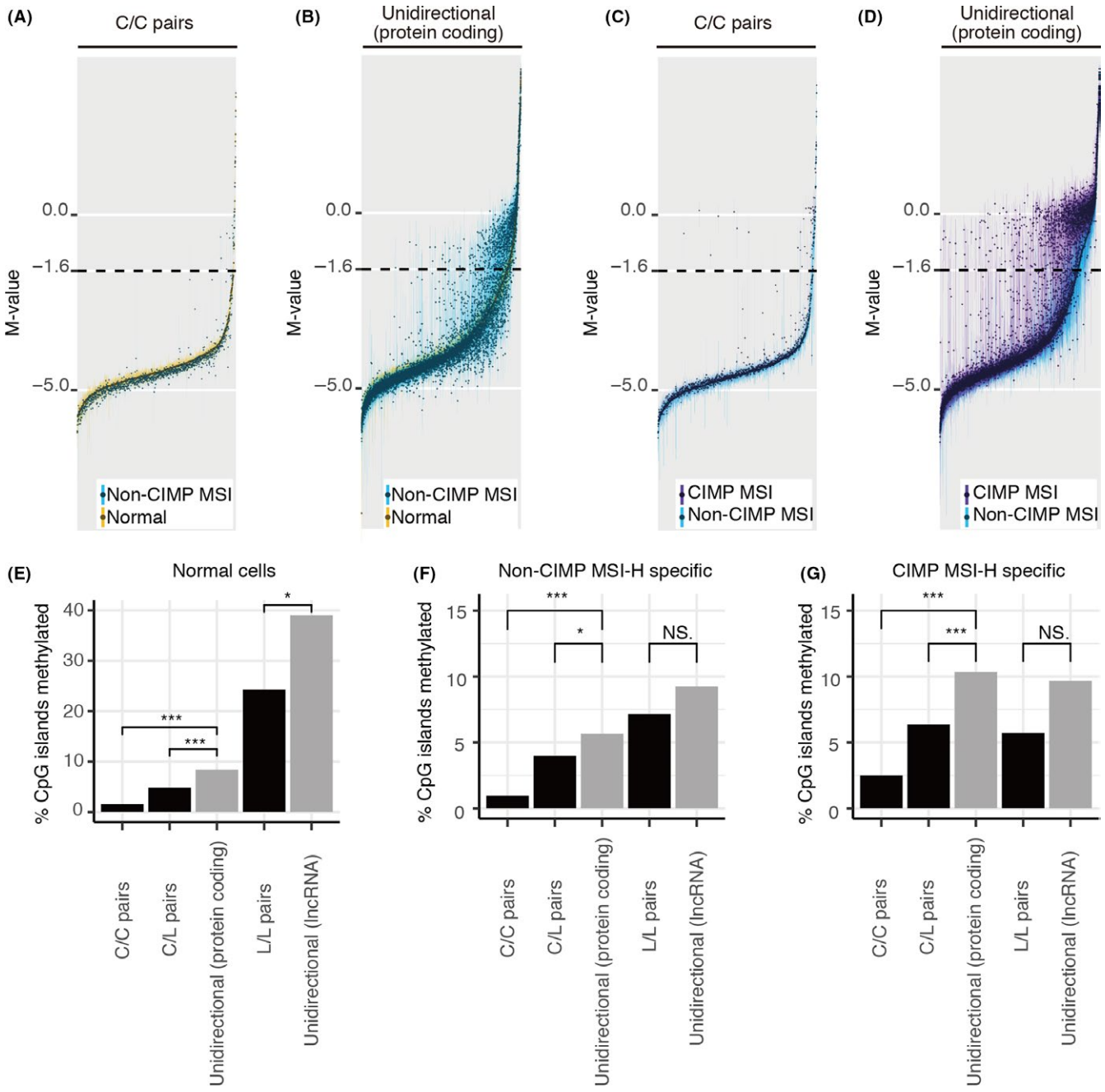

FIGURE 3 Comparison of CpG islands (CGI) methylation profiles in cancer cells. A, Box plots representing the *M*-value of CGI within divergent promoters (div-CGI) with pairs of protein-coding genes (C/C pairs). Yellow, normal; blue, non-CGI methylator phenotype (CIMP) microsatellite instability high colorectal cancer(MSI-H CRC). B, Box plots representing the *M*-value of CGI within unidirectional promoters (uni-CGI) with protein-coding genes. Yellow, normal; blue, non-CIMP MSI-H CRC. C, Box plots representing the *M*-value of div-CGI with C/C pairs. Blue, non-CIMP MSI-H CRC; purple, CIMP MSI-H CRC. D, Box plots representing the *M*-value of uni-CGI with protein-coding genes. Blue, non-CIMP MSI-H CRC; purple, CIMP MSI-H CRC. Points represent the median values. CGI are in the order of median values of the former samples. See Figure S1 for box plots of CGI in other groups. E, Bar plots show fraction of CGI methylated in normal cells. F,G, Bar plots show fraction of CGI specifically methylated in non-CIMP MSI-H CRC (E) or in CIMP MSI-H CRC (F). NS, not significant (False discovery rate [FDR] > 0.05); *FDR < 0.05; ***FDR < 0.001. C/L pair, pair of protein-coding gene and long non-coding RNA; L/L pair, pair of long non-coding RNA

3.5 | Methylation of CGI is not correlated with TSS density

To investigate if TSS density, which may be relatively high in bidirectional promoters, affects the methylation of CGI, we examined the number of genes whose TSS is within CGI shores (2000 bp upstream/downstream of CGI) (Figure S2A). By definition, CGI that contain a TSS of only one gene were limited to unidirectional CGI. In contrast, CGI that contain TSS of two genes consisted of all types of CGI. The majority of the CGI were shorter than 2000 bp

in length, and the distribution of the length was similar in all types (Figure S2B). Based on these results, we performed similar analyses, as shown in Figures 2 and 3, with a specific subset of CGI: CGI that contain TSS of two genes and are shorter than 2000 bp in length (3265/13 774). All the analyses had similar results as the analyses presented above, with a small difference; the C/C pair remained significant ($P < 0.001$) while the C/L pair and L/L pair were not significant ($P > 0.05$) (Figure S2C-F).

Next, we examined the correlation between TSS density and CGI methylation on a broader scale. We plotted the *M*-value of CGI against the number of genes with a TSS within 10 000 bp upstream or 100 000 bp downstream of CGI for all samples in this study (Figure S2G). Excluding CGI with only one gene (because these consisted only of unidirectional CGI), the absolute value of Spearman's correlation coefficient was calculated, which was smaller than 0.2 in all cases. These data suggested that TSS density did not substantially affect CGI methylation.

3.6 | Characterization of methylated promoter CGI

We first illustrated an overview of the promoter CGI (Figures 4A,B and S3). Interestingly, when CGI were plotted according to the increment of methylation levels among the populations (normal, non-CIMP MSI-H and CIMP MSI-H), the distribution of CGI was clearly divided between those methylated in any of the populations and those never methylated in these populations (Figure 4A).

To more closely examine the difference between methylated and unmethylated CGI, motif analysis was conducted using MEME.[25] Among the CGI methylated in any of the populations, we found motifs that did not exist in unmethylated CGI. The sequence between position −500 and 0 of the TSS were used for the analysis; in case of div-CGI, a susceptible TSS sequence with higher expression was chosen. As a result, we identified three common types of sequence motifs seen in every methylated group (Figure 4C). These were robust in various methods of sequence analysis (Figure S4). Of note, CCG and CGG repeats were reportedly seen near the TSS of coding genes that were paired with non-coding RNA.[28] These motifs also existed in *MLH1* and *BRCA1* at 500 bp upstream of the TSS (Motif 2 in *MLH1* and motifs 2 and 3 in *BRCA1*). *MLH1*-relevant CGI were specifically methylated in CIMP MSI-H CRC (Figure S5).

There was a weak correlation between the existence of these motifs and the bidirectionality (Spearman's rho was 0.08 with motif 1, 0.23 with motif 2, and 0.12 with motif 3). We performed logistic regression analysis under several conditions to examine the differential effect of the bidirectionality and the methylation-specific motifs on DNA methylation. Logistic regression with the bidirectionality

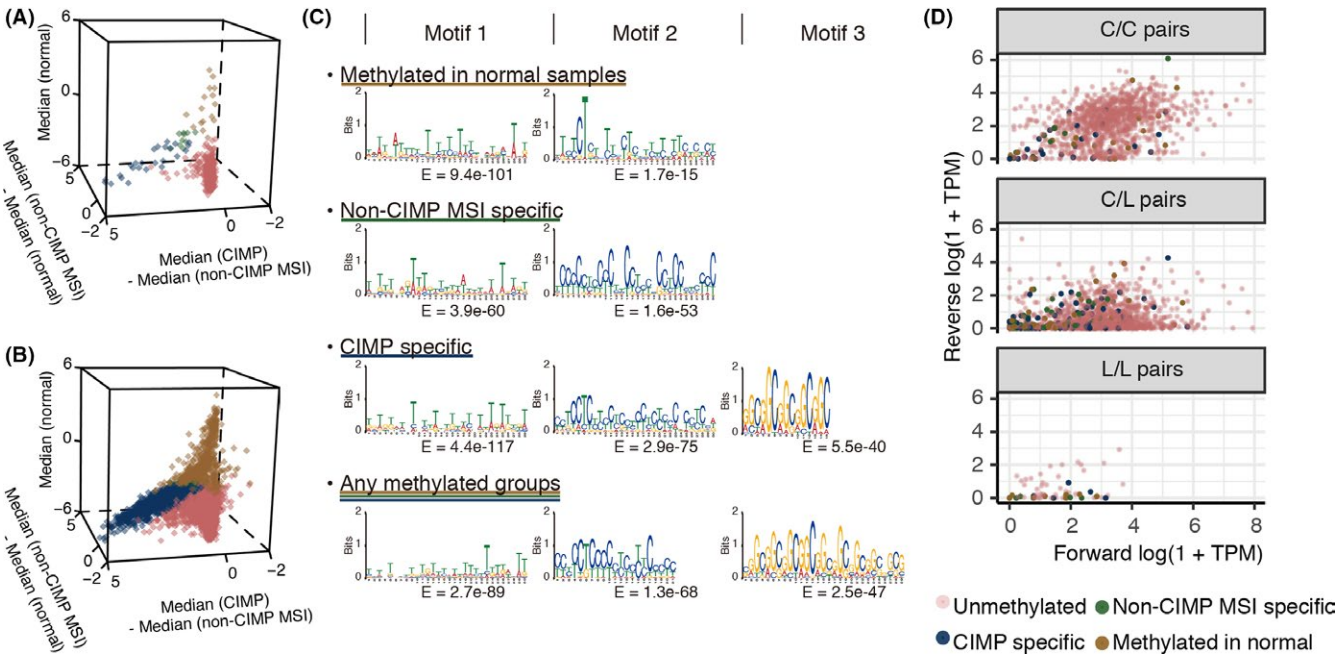

FIGURE 4 Characterization of methylated promoter CpG islands (CGI). A,B, *X*-axis indicates difference in the median of *M*-value between non-CGI methylator phenotype (CIMP) microsatellite instability high colorectal cancer(MSI-H) samples and normal samples. *Y*-axis indicates difference in the median of *M*-value between CIMP MSI-H samples and non-CIMP MSI-H samples. *Z*-axis indicates the median *M*-value in normal samples. A, CGI within divergent promoters with pairs of protein-coding genes (C/C pairs). B, CGI within unidirectional promoters of protein-coding genes. The plots of CGI in other groups are shown in Figure S2. C, De novo motifs found using MEME are shown. Promoter CGI specifically methylated in respective groups were compared with promoter CGI in unmethylated groups. The sequence 0-500 bp upstream of the forward gene transcription start site was used as query. Motifs found using other query sequences are shown in Figure S3. D, Expressions of genes in normal cells are shown in association with the methylation status of related CGI color coded. Expression data of normal sigmoid colon cells were obtained from the GTEx project. *X*-axis indicates log-transformed expression (1+ transcript per million [TPM]) of forward genes. *Y*-axis indicates log-transformed expression (1+ TPM) of reverse genes. C/L pair, pair of protein-coding genes and long non-coding RNA; L/L pair, pair of long non-coding RNA

predicted methylation of CGI in any of the three cell types was better than that with the methylation-specific motifs (Table S1). The predictive potential of logistic regression incorporating both the bidirectionality and the motifs did not significantly differ compared with that with bidirectionality only (Table S1), suggesting that the effect of the bidirectionality on CGI methylation may be larger than that of the identified motifs.

In addition, we used the TPM data of normal sigmoid colon cells from GTEx[20] to look into the expression of both ends of genes in div-CGI (Figure 4D). We found that expressions of both genes in normal cells were low when the associated CGI were methylated in any of the three cell types. In contrast, methylation of the associated CGI was not observed in any of the cell types when expression of either end of the gene was high in normal cells. This observation was consistent among C/C, C/L and L/L pairs, suggesting that low expression of both genes may be necessary for div-CGI to be methylated in normal cells or in CRC cells.

4 | DISCUSSION

In this study, we analyzed the genome-wide methylation status in association with the structural configuration of genes. We report several observations that are essential for the understanding of the epigenetic regulation of transcription and aberrant transcriptional regulation in cancer. The data presented in this report provide important clues for the understanding of carcinogenesis and the development of strategies for cancer prevention.

First, we demonstrated that div-CGI are less frequently methylated compared with uni-CGI in normal cells and cancer cells, as well as cancer cells with CIMP. Given div-CGI are resistant to methylation, silencing of *MLH1* and *BRCA1* appear to be deviated from normal status. Because the unmethylated status of div-CGI was observed genome-wide, we speculate that transcription itself, rather than the functions of individual RNA, is what contributes to the maintenance of the unmethylated status of div-CGI. Under this assumption, the observation that the number of methylated div-CGI is less in C/C pairs than C/L pairs is consistent with the fact that coding genes are generally transcribed at higher rates than lncRNA. Given that de novo methylation of CGI requires recruitment of several proteins including DNMT1,[29] the transcription of neighbor genes may inhibit these proteins from combining with DNA strands. Further studies are required to elucidate the mechanistic bases of methylation resistance of div-CGI.

Second, we showed that paired genes are more likely to be located adjacent to CGI. Previous reports have established that house-keeping genes and cell type-specific genes (including nervous system-specific genes) enrich CGI in their promoter regions.[30] These data suggested that divergent structures prevent important genes, including tumor suppressor genes like *MLH1* and *BRCA1*, from promoter CGI methylation. In terms of evolution, methylated CGI have been shown to influence genetic variation; in comparing the human and primate genomes, CGI are conserved with low mutation rates where CGI are hypomethylated in the germ line.[31,32] If CGI within divergent promoters are also hypomethylated in the germ line, it may be reasonable that CGI within divergent promoters are conserved over generations.

Third, we identified three DNA sequence motifs that were associated with CGI methylation. These motifs may be the target of molecules that regulate the methylation of CGI. Future studies should investigate the molecules that recognize and bind to these motifs. It is of particular interest that the *MLH1* promoter, in which CGI were methylated in MSI-H CRC despite its bidirectionality, contains the motif associated with CGI methylation. Although much remains to be revealed for the precise understanding of the regulation of DNA methylation, elucidation of the mechanistic basis of aberrant CGI methylation through the analysis of such molecules would enable the prevention of cancers that are caused by silencing of tumor suppressor genes by methylation.

Regarding *MLH1*-silenced CRC, Fang et al reported that oncoprotein BRAF(V600E) induced phosphorylation of MAFG and MACH1, resulting in the promoter methylation of *MLH1* by DNMT1.[33] However, our genome-wide analysis did not detect any motifs similar to a MAF recognition element. This suggests that *MLH1* silencing may occur from a different mechanism and have a different time course from other CGI.

In conclusion, we demonstrated that div-CGI have a methylation-resistant nature in normal colorectal cells and are insusceptible to methylation in non-CIMP MSI-H or CIMP MSI-H CRC. Further investigation on the mechanistic basis of these observations may pave the way to the development of strategies for cancer prevention.

ACKNOWLEDGMENTS

We thank Ms Miki Tamura, Dr Manabu Soda and Dr Yoshihiro Yamashita for technical assistance. We are grateful to all the patients and families who contributed to this study. This study was supported in part by Grants-in-Aid for Scientific Research (KAKENHI, grant no. 17J00386 and 16K07143) from the Japan Society for the Promotion of Science (JSPS) and by grants from Leading Advanced Projects for Medical Innovation (LEAP, JP17am0001001) to H.M. and from the Project for Cancer Research and Therapeutic Evolution (P-CREATE, JP17 cm0106502) to M.K., M.S., H.A. and T.W. from the Japan Agency for Medical Research and Development. K.S. is a Research Fellow of JSPS.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Shoichi Hazama https://orcid.org/0000-0002-5239-8570
Hiroyuki Mano https://orcid.org/0000-0003-4645-0181

Masahito Kawazu https://orcid.org/0000-0003-4146-3629

REFERENCES

1. Herman JG, Umar A, Polyak K, et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci USA*. 1998;95:6870-6875.
2. Rice JC, Ozcelik H, Maxeiner P, et al. Methylation of the BRCA1 promoter is associated with decreased BRCA1 mRNA levels in clinical breast cancer specimens. *Carcinogenesis*. 2000;21:1761-1765.
3. Bass AJ, Thorsson V, Shmulevich I, et al. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*. 2014;513:202-209.
4. Hammerman PS, Lawrence MS, Voet D, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489:519-525.
5. Kim J, Bowlby R, Mungall AJ, et al. Integrated genomic characterization of oesophageal carcinoma. *Nature*. 2017;541:169-175.
6. Sasaki M, Knobbe CB, Munger JC, et al. IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics. *Nature*. 2012;488:656-659.
7. Tahara T, Yamamoto E, Suzuki H, et al. Fusobacterium in colonic flora and molecular features of colorectal carcinoma. *Cancer Res*. 2014;74:1311-1318.
8. Kawazu M, Kojima S, Ueno T, et al. Integrative analysis of genomic alterations in triple-negative breast cancer in association with homologous recombination deficiency. *PLoS Genet*. 2017;13:e1006853.
9. Polak P, Kim J, Braunstein LZ, et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet*. 2017;49:1476-1486.
10. Trinklein NT, Force Aldred S, Hartman SJ, et al. An abundance of bidirectional promoters in the human genome. *Genome Res*. 2004;14:62-66.
11. Kung JTY, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. *Genetics*. 2013;193:651-669.
12. Sigova AA, Mullen AC, Molinie B, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci USA*. 2013;110:2876-2881.
13. Almada AE, Wu X, Kriz AJ, et al. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature*. 2013;499:360-363.
14. Duttke SHC, Lacadie SA, Ibrahim MM, et al. Human promoters are intrinsically directional. *Mol Cell*. 2015;57:674-684.
15. Engreitz JM, Haines JE, Perez EM, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. 2016;539:452-455.
16. Yamamoto N, Agata K, Nakashima K, et al. Bidirectional promoters link cAMP signaling with irreversible differentiation through promoter-associated non-coding RNA (pancRNA) expression in PC12 cells. *Nucleic Acids Res*. 2016;44:5105-5122.
17. Joung J, Engreitz JM, Konermann S, et al. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature*. 2017;548:343-346.
18. Shu J, Jelinek J, Chang H, et al. Silencing of bidirectional promoters by DNA methylation in tumorigenesis. *Cancer Res*. 2006;66:5077-5084.
19. Trautmann K, Terdiman JP, French AJ, et al. Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clin Cancer Res*. 2006;12:6379-6385.
20. Weisenberger DJ, Liang G, Lenz HJ. DNA methylation aberrancies delineate clinically distinct subsets of colorectal cancer and provide novel targets for epigenetic therapies. *Oncogene*. 2018;37:566-577.
21. Sato K, Kawazu M, Yamamoto Y, et al. Fusion kinases identified by genomic analyses of sporadic microsatellite instability-high colorectal cancers. *Clin Cancer Res*. 2018;25:378-389. clincanres.1574.2018.
22. Du P, Zhang X, Huang C-C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11:587.
23. Karolchik D. The UCSC table browser data retrieval tool. *Nucleic Acids Res*. 2004;32:493D-496D.
24. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30:1363-1369.
25. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc Second Int Conf Intell Syst Mol Biol*. 1994;2:28-36.
26. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27:1017-1018.
27. Yang M, Elnitski L. Orthology-driven mapping of bidirectional promoters in human and mouse genomes. *BMC Bioinformatics*. 2014;15:S1.
28. Uesaka M, Nishimura O, Go Y, et al. Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC Genom*. 2014;15:35.
29. Akhavan-Niaki H, Samadani AA. DNA methylation and cancer development: molecular mechanism. *Cell Biochem Biophys*. 2013;67:501-513.
30. Orekhova AS, Rubtsov PM. Bidirectional promoters in the transcription of mammalian genomes. *Biochem*. 2013;78:335-341.
31. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA*. 2006;103:1412-1417.
32. Weber M, Hellmann I, Stadler MB, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*. 2007;39:457-466.
33. Fang M, Ou J, Hutchinson L, et al. The BRAF oncoprotein functions through the transcriptional repressor MAFG to mediate the CpG Island Methylator Phenotype. *Mol Cell*. 2014;55:904-915.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Namba S, Sato K, Kojima S, et al. Differential regulation of CpG island methylation within divergent and unidirectional promoters in colorectal cancer. *Cancer Sci*. 2019;110:1096–1104. https://doi.org/10.1111/cas.13937