Am J Cancer Res 2024;14(5):2253-2271
www.ajcr.us /ISSN:2156-6976/ajcr0155540

Original Article

Novel DNA methylation-based epigenetic signatures in colorectal cancer from peripheral blood leukocytes

Su Yon Jung[1,2,3], Herbert Yu[4], Xianglong Tan[5], Matteo Pellegrini[6]

[1]Translational Sciences Section, School of Nursing, University of California, Los Angeles, CA 90095, USA; [2]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA; [3]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA; [4]Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA; [5]Department of Biological Chemistry, University of California, Los Angeles, CA 90095, USA; [6]Department of Molecular, Cell and Developmental Biology, Life Sciences Division, University of California, Los Angeles, CA 90095, USA

Received January 20, 2024; Accepted April 21, 2024; Epub May 15, 2024; Published May 30, 2024

Abstract: Colorectal cancer (CRC) is a multifactorial disease characterized by accumulation of multiple genetic and epigenetic alterations, transforming colonic epithelial cells into adenocarcinomas. Alteration of DNA methylation (DNAm) is a promising biomarker for predicting cancer risk and prognosis, but its role in CRC tumorigenesis is inconclusive. Notably, few DNAm studies have used pre-diagnostic peripheral blood (PB) DNA, causing difficulty in postulating the underlying biologic mechanism of CRC initiation. We conducted epigenome-wide association (EWA) scans in postmenopausal women from Women's Health Initiative (WHI) with their pre-diagnostic DNAm in PB leukocytes (PBLs) to prospectively evaluate CRC development. Our site-specific DNAm analyses across the genome adjusted for DNAm-age, leukocyte heterogeneities, as well as body mass index, diabetes, and insulin resistance. We validated 20 top EWA-CpGs in 2 independent CRC tissue datasets. Also, we detected differentially methylated regions (DMRs) associated with CRC, further mapped to transcriptomic profile, and finally conducted a Gene Set Enrichment Analysis. We detected multiple novel CpGs validated across WHI and tissue datasets. In particular, 2 CpGs (*B4GALNT4*cg10321339, *SV2B*cg18144285) had the strongest effect on CRC risk. Results from our DMR scans contained *MIR663*cg06007966, which was also validated in EWA analyses. Also, we detected 1 methylome region in *PEG10* of Chr7 shared across datasets. Our findings reflect both novel and well-established epigenomic and transcriptomic sites in CRC, warranting further functional validations. Our study contributes to better understanding of the complex interrelated mechanisms on the methylome underlying CRC tumorigenesis and suggests novel preventive DNAm-targets in PBLs for detecting at-risk individuals for CRC development.

Keywords: Colorectal cancer tumorigenesis, epigenetic signatures, pre-diagnostic DNA methylation, transcriptomics, postmenopausal women

Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed malignancy and the second leading cause of cancer death in both sexes worldwide [1] and the third leading cause of cancer mortality in women in the U.S. [2]. CRC is a multifactorial disease characterized by environmental/behavioral factors and long-term genetic/epigenetic alterations and their interplay, transforming colonic epithelial cells into adenocarcinomas [3, 4]. Multiple behavioral factors have long been identified as CRC determinants [5-7], but the full extent of CRC's genetic background remains incomplete, although it is equally critical for capturing the biologic mechanisms of CRC carcinogenesis. Identifying genetic variations has been challenging because CRC tumorigenesis is a complex process influenced by environmental/lifestyle factors that must be accounted for when exploring the genomic architecture of CRC development. Also, despite hundreds of mutations found in relation to CRC genomes, only a small set of functionally essential genes are proposed as driver mutations, which are insufficient in carcinogenesis [8, 9]. Epigenetic mechanisms may address these issues. In fact,

https://doi.org/10.62347/MXWJ1398

differential epigenetic patterns in CRC reflect interactions between environmental exposures and genetic influences, conferring cellular plasticity with specific cellular states, regulating gene expression, and consequently affecting cancer development [10-15]. Further, the epigenome influences accumulation of DNA mutations, controls important tumor cell phenotypes [16], and finally exerts a driving effect on CRC risk in the combined analyses of epigenetic alterations and genetic mutations [17], suggesting its crucial role during CRC development.

In particular, DNA methylation (DNAm) is a well-characterized major epigenetic modification that involves mitotically heritable and reversible attachment of methyl groups at the 5' carbon of cytosine in CpG dinucleotides (CpGs) [18, 19]. DNAm alteration has received growing attention as a promising biomarker for predicting CRC prognosis [20, 21], response to treatment [22], and early detection [23], since it presents high clinical sensitivity and dynamic changes by environmental cues [24] and occurs much in advance of the consequent changes in gene expression [25] and in clinical diagnoses of cancers, including CRC [26, 27]. Also, specific DNAm modification was detected in pre-cancerous "normal" colonic tissues that modify cancer risk [28-30], suggesting its occurrence at an earlier stage than carcinoma formation, thus playing a crucial role in CRC tumor initiation.

In general, global DNA hypomethylation is observed in CRC cells/tissues accompanied by local hypermethylation at regulatory regions such as promoters, leading to silencing of tumor-suppressor genes [31-33]. But the role of epigenetic mechanisms in CRC tumorigenesis is inconclusive, shown as a lack of consensus on DNAm markers in epigenetic studies; this is mainly owing to studies on heterogeneous populations in sex, age, and race/ethnicity, a lack of validation in independent samples, use of different biospecimens (e.g., tissues, blood, stool), and focus mainly on DNAm in promoter/CpG-island (i.e., CpG-rich) regions, although CpG-depleted regions, a large proportion of methylated positions, have potential effects on cancer [34-37]. In addition, few epigenome-wide studies [38-40] have examined non-invasive DNAm extracted from peripheral blood (PB) in CRC. Although DNAm is tissue specific, the correlations between PB and tissue are gene specific [26, 41, 42]. Given that obtaining intestinal tissues from healthy individuals is difficult, DNAm in PB is the most promising non-invasive risk marker for early identification of a population at high risk for CRC development [43]. Of note, most DNAm studies in PB for CRC have used post-diagnostic PB DNA, likely reflecting DNAm status as an early hematologic response to the presence of CRC cells, causing difficulty in postulating the underlying biologic mechanism of cancer initiation.

Our study addresses these critical gaps. We examined postmenopausal women, who are the most vulnerable to CRC (about 90% of CRC cases occur in individuals 50 years and older [2, 44]), focusing on white women. We conducted an epigenome-wide DNAm study by covering a majority of CpG-depleted regions or gene bodies in PB leukocytes (PBLs), reflecting the pre-diagnostic DNAm state (i.e., before CRC development). We validated our PBL-based findings in 2 independent CRC tissue cohorts by comparing the DNAm status between CRC tissues and normal colon tissues adjacent or obtained from non-tumor bearing patients. Finally, we mapped our findings to transcriptome profiles for investigating the cross-talk between DNAm and gene expression in CRC tissues. We hypothesized that the 2 CRC tissue cohorts are not exactly the same in DNAm status because the DNAm of normal tissues adjacent to CRC tissues differs from that of normal tissues derived from non-CRC patients [5, 45], but they are more similar to each other than DNAm from PBLs, reflecting somatic-specific epigenetic signatures. Ultimately, overlapping CpGs across the 3 cohorts may indicate long-term non-invasive surrogate markers in tissues, reflecting multiple CRC tumorigenic mechanisms in this population.

Materials and methods

Study population

For our epigenome-wide association (EWA) analysis in the discovery phase, we used data from the Women's Health Initiative (WHI) cohort, a large prospective study of postmenopausal women, ages 50-79 years at enrollment between 1993 and 1998 from 40 U.S. clinical

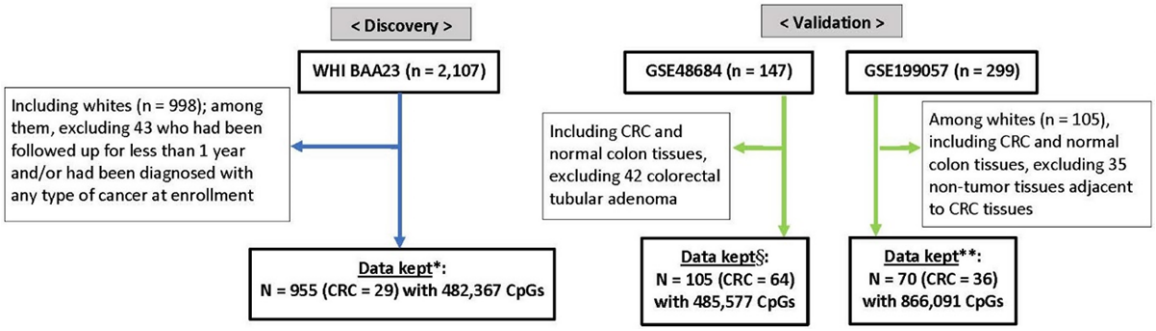


Figure 1. Diagram of EWA and CRC study populations from the WHI and GEO datasets. CpGs, CpG dinucleotides; CRC, colorectal cancer; EWA, epigenome-wide association; GEO, Gene Expression Omnibus; WHI, Women's Health Initiative. Note: *Data include 1 individual with missing data from covariates such as waist and hip circumferences, which was ultimately imputed on the basis of the na.roughfix method [111]. §Data include 64 CRC tissues from CRC patients and 41 normal colon tissues adjacent to CRC tissue from CRC patients. **Data include 36 CRC tissues from CRC patients and 34 normal colon tissues from non-CRC patients.

centers [46]. From the Database for Genotypes and Phenotypes (dbGaP) genetic repository, we included genome-wide DNAm data measured in PBLs available in a WHI ancillary study (AS), BAA23, by repurposing data that originally focused on the integrative genomics for heart disease and related phenotypes [47]. Because racial/ethnic variations exist in CRC-related DNAm [48, 49], we included in this study only self-reported non-Hispanic white women, a majority of the AS population: of 2,107 total, 998 whites, 600 African Americans, and 509 Asians/others. Among the 998 white women, 955 who had been followed for at least 1 year and not been diagnosed with any cancer at enrollment were included (**Figure 1**). They had been followed through March 6, 2021, with a 17-year mean follow-up, resulting in 29 who developed primary colorectal adenocarcinoma.

For our validation study, we used 2 independent methylation datasets from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) databases, which included 64 CRC and 41 normal adjacent tissues (accession number GSE48684 [50]) and 36 CRC tissues from CRC patients and 34 normal colon tissues from non-CRC patients (GSE199057 [45]), after excluding 42 colorectal tubular adenoma (GSE48684) and 35 non-tumor tissues adjacent to CRC tissues (GSE199057) (**Figure 1**). The institutional review boards of each WHI clinical center and the University of California, Los Angeles, approved this study.

Data collection, laboratory method, and CRC outcome

For the WHI participants, self-administered questionnaires had been provided at enrollment to collect demographic information such as age and race and comorbidities, e.g., ever having been treated for diabetes (DM). Anthropometric measurements, including height and weight, had been assessed by trained staff at screening. Blood samples from WHI women after fasting 8 hours or longer had been collected at enrollment by phlebotomists and assayed for glucose and insulin concentrations, using the hexokinase method on a Hitachi 747 analyzer (Boehringer Mannheim Diagnostics, Indianapolis, IN) for glucose and via radioimmunoassay (Linco Research, Inc., St. Louis, MO) or automated ES300 method (Boehringer Mannheim Diagnostics) for insulin. The results from the 2 methods for insulin measurement were comparable at insulin concentrations < 60 μIU/ml, and the intra-class correlation coefficient with repeatedly measured insulin was 0.7 [51]. Homeostatic model assessment-insulin resistance (HOMA-IR), as a surrogate of IR, was estimated as glucose (unit: mg/dl) × insulin (unit: μIU/ml)/405 [52] and was included in the analysis as a covariate.

Primary CRC development among the WHI participants was initially ascertained from their self-report of a new cancer diagnosis, determined by a committee of physicians on the basis of a review of the patients' medical records and pathology/cytology reports, and

finally coded into the WHI database according to the National Cancer Institute's Surveillance, Epidemiology, and End-Results guidelines [53]. The time from enrollment until CRC development, censoring, or study end-point was measured as the number of days and then converted into years.

CRC tissue sample-based data from the GEO databases included patients' sex, race, and age. For this study purpose, data from Caucasians only were analyzed.

Epigenome-wide DNAm array

Genome-wide DNAm array of the WHI participants was conducted with their extracted PBL DNA, via Illumina 450 BeadChip. DNAm quality control procedures excluded poor-performing CpGs with $P > 0.01$ in > 10% of the samples. Data were beta-mixture quantile (BMIQ)-normalized [54] and batch-adjusted via random intercept for plate and chip and a fixed effect for row [47]. Leukocyte heterogeneities were estimated (Table S1) and adjusted for in the analysis using Houseman's method [55] (for $CD4^+$ T cells, natural killer cells, monocytes, and granulocytes) and Horvath's method [56] (for plasma blasts, $CD8^+CD28^-CD45RA^-$ T cells, and naïve CD8 T cells).

In the GSE48684 cohort, a tissue-derived genome-wide DNAm assay was performed with the Illumina Infinium HM450K array, and unreliable probes were removed if $P > 0.05$. Using the R *minfi* package, data were normalized via Illumina background level corrections, color adjustment, and subset quantile within array normalization. CpGs were further filtered out if they contained single-nucleotide polymorphisms (SNPs) and were chromosome (Chr) X-associated. The ComBat algorithm was applied to correct for batch effects across all array runs [50]. In the GSE199057 cohort, Illumina EPIC array was performed, followed by data normalization via background correction based on normal-exponential out-of-band (Noob) [57] using *minfi*. SNP-associated and cross-reactive CpGs were excluded, and poor-quality CpGs with missing ≥ 20% of samples were also excluded. Batch effects were corrected using Bland Altman methods for replicate samples [45].

DNAm levels (β values) from Illumina 450K and EPIC array were calculated as the ratio of intensities between the methylated and unmethylated probes, ranging from 0 (completely unmethylated) to 1 (completely methylated) [58]. Also, epigenetic ages (DNAm-predicted ages) were estimated using the Horvath clock [56] in the WHI and GSE199057 cohorts, where relevant data were available.

Statistical analysis

DNAm levels were standardized across samples in each cohort, resulting in 482,367 CpGs in the WHI, 485,577 CpGs in the GSE48684, and 866,091 CpGs in the GSE199057 included in our analysis; the effect size from the analysis reflected a 1 standard-deviation (SD) increase in DNAm on CRC risk.

For the DNAm site-specific analysis across the genome with CRC development in the WHI data, we conducted multiple Cox proportional hazards regression, with an assumption test met via a Schoenfeld residual plot and rho, adjusting for DNAm-predicted age, biologic age, and leukocyte heterogeneity, as well as body mass index (BMI), DM, and IR levels as key confounding factors [7, 14, 15, 59, 60] in associations between DNAm probes and CRC. With 20 top CpGs detected at the genome-wide level, we next performed logit regression for CRC outcomes in each GEO dataset by adjusting for sex and, in the GSE199057 only, DNAm-predicted and biologic ages. Two-sided $P <$ 1E-007 (discovery) and < 2.5E-03 (= 0.05/20 top CpGs) (validation), after Bonferroni correction, were considered statistically significant.

Differences in DNAm levels of the modeled CpGs by CRC risk in each cohort and those of CpGs among the 3 cohorts of the CRC patients were tested using unpaired 2-sample t and 1-way ANOVA tests, respectively. If β values were skewed or had outliers, Mann-Whitney/Wilcoxon's ranked-sum and Kruskal-Wallis tests were used as appropriate.

In addition to individual GpGs, we detected differentially methylated regions (DMRs) associated with CRC, using R *DMRcate* package on the basis of kernel smoothing of the differential methylation signal, with 1,000 lambda (Gaussian kernel bandwidth) and 2 C (scaling factor for bandwidth) as recommended, so that half a kilobase represents 1 SD of support [61, 62]. This method is superior to others (e.g., Bumphunter and Probe Lasso), removing the bias

from irregularly spaced methylation sites and filtering probes possibly confounded by SNPs and cross-hybridization [61, 62].

With the top genome-wide CpGs in the WHI discovery and those significant at the validation level in the GEO datasets, we finally conducted a Gene Set Enrichment Analysis (GSEA) using WebGestalt [63]. All statistical analyses were performed with R through UCLA's Hoffman2 high-performance computing cluster.

Transcriptomics analysis

Using R *TCGAbiolinks* package, we retrieved data from The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) and TCGA-Rectum Adenocarcinoma (READ) projects, integrating 701 RNA-sequence (Seqs) samples comprising 1 metastatic, 2 recurrent, and 647 primary tumor tissues plus 51 normal adjacent tissues. Raw count normalization between the cancer and normal groups, followed by differential expression analysis, was conducted via *DESeq2* and *org.Hs.eg.db* package. Further, we calculated z-scores for each modeled gene and performed Uniform Manifold Approximation and Projection (UMAP) and heatmap analyses, producing graphic visualizations.

Results

Site-specific CpG analysis across genome for CRC outcomes

With 482,367 CpGs in the WHI discovery, our genome-wide DNAm scan detected the 20 top CpGs differentially methylated by CRC development (**Table 1**). The hazard ratios were consistent across the analyses accounted for age only; age plus BMI; and age plus BMI, DM, and IR levels. In the validations with 2 GEO datasets (**Table 2**), the effect sizes and directions of the top 20 CpGs were in general similar between datasets, reflecting somatic-specific DNAm profiles. Of the top 20 genome-wide CpGs, 11 were also significant at the validation level in either or both of the GEO cohorts. Six of the 11 CpGs presented similar risk magnitudes between the WHI and either/both GEO cohorts and in each dataset, the area under the receiver operating characteristic curve has been reported (Figure S1): cg04958124, cg10321339, cg12704462, cg18144285, cg06007966, and cg17375901. In particular, 2 CpGs (*B4GALNT4*cg10321339 and *SV2B*cg18144285) had the strongest effect on CRC risk (32 and 22 times greater risk, respectively, each with a 1-SD increase in DNAm) in the GSE199057. Also, 2 other CpGs (*MIR663*cg06007966 and cg17375901) were validated in both GEO datasets; both are located in Chr20 with the same direction and similar risk magnitudes in the WHI discovery and both validation GEO datasets, but having more profound effects in the validations. Of interest, 1 CpG (cg05970116 in Chr10) had genome-wide significance in the discovery and both validation datasets, presenting different directions: a positive association of its 1-SD increase in DNAm with CRC development in the WHI, but an inverse association with CRC tissues in both GEO datasets.

We compared the DNAm levels of the top 20 CpGs by CRC status across Chr, CpG context, enhancer and/or promoter, and gene region within the WHI (**Figure 2**) and each GEO dataset (Figure S2). The mean levels of DNAm differed in Chr1, 6, 7, 10, and 15 in the WHI, where DNAm levels were higher in those with CRC development than in those without. Similarly, hypermethylation in CRC tissues was observed in Chr7 in GSE48684, but more substantial differences in Chr11 were found in the GEO datasets. Chr12 presented hypomethylation in CRC across all 3 cohorts, shown more profoundly in the WHI, and an apparent difference in DNAm mean level by CRC status was observed in GEO199057. Whereas CpG islands and S-Shores were hypermethylated in the WHI women with CRC, N-Shores were hypermethylated in both CRC GEO datasets. In the WHI, both enhancer and promoter were hypermethylated in CRC patients, but the opposite direction was observed in GSE199057, where promoter was hypomethylated in CRC tissues. In both the WHI and the GEO datasets, 5' untranslated regions (5'UTR) were hypermethylated in CRC patients and tissues.

Further, we compared DNAm levels of the top 20 CpGs within the CRC patients across the 3 cohorts in terms of Chr, CpGs, CpG context, and gene region (Figure S3), showing consistent patterns in **Figures 2** and S2. We also compared among CRC patients the DNAm levels of 3 individual CpGs that were genome-wide significant at the validation level in both GEO datasets (**Figure 3**). Except for cg05970116, which

Table 1. WHIBAA23 dataset: differentially DNA-methylated top 20 CpGs genome-wide associated with CRC risk

Chr	CpG site§	Position	Age adjusted		BMI & age adjusted		DM, IR, BMI & age adjusted		CpG context	Gene	Gene region
			HR¶ (95% CI)	*P*	HR¶ (95% CI)	*P*	HR¶ (95% CI)	*P*			
chr1	cg14057946¥	713985	1.43 (1.25, 1.64)	< 1E-007	1.43 (1.25, 1.64)	< 1E-007	1.42 (1.23, 1.64)	2.00E-06	Island		Intergenic
chr1	cg04231937¥	714526	1.56 (1.33, 1.82)	< 1E-007	1.56 (1.33, 1.82)	< 1E-007	1.56 (1.33, 1.83)	< 1E-007	Island		Intergenic
chr1	cg04496485¥	714565	1.39 (1.24, 1.57)	< 1E-007	1.39 (1.23, 1.56)	< 1E-007	1.38 (1.23, 1.56)	< 1E-007	S Shore		Intergenic
chr1	cg02014020	1115461	0.67 (0.57, 0.79)	1.00E-06	0.67 (0.58, 0.79)	1.00E-06	0.67 (0.57, 0.79)	2.00E-06	N Shore	*TTLL10*	Body
chr6	cg07572131*,¥	31430791	1.65 (1.37, 1.98)	< 1E-007	1.64 (1.35, 1.98)	< 1E-007	1.58 (1.29, 1.94)	9.00E-06	OpenSea	*HCP5*	TSS200
chr6	cg25410010	41554543	0.59 (0.49, 0.72)	< 1E-007	0.59 (0.49, 0.72)	< 1E-007	0.57 (0.47, 0.70)	< 1E-007	OpenSea	*FOXP4*	Body
chr6	cg06498809¥	111303174	1.45 (1.26, 1.68)	< 1E-007	1.45 (1.25, 1.67)	1.00E-06	1.45 (1.25, 1.68)	1.00E-06	Island	*RPF2*	TSS200
chr6	cg00020352¥	111303252	1.50 (1.27, 1.77)	2.00E-06	1.49 (1.26, 1.77)	3.00E-06	1.47 (1.24, 1.75)	1.30E-05	Island	*RPF2*	TSS200
chr6	cg10920427¥	111303363	1.33 (1.19, 1.49)	1.00E-06	1.32 (1.18, 1.48)	1.00E-06	1.33 (1.18, 1.48)	1.00E-06	Island	*RPF2*	Body
chr6	cg14498116¥	111303482	1.39 (1.22, 1.58)	< 1E-007	1.38 (1.22, 1.57)	1.00E-06	1.38 (1.21, 1.57)	1.00E-06	Island	*RPF2*	Body
chr7	cg04958124¥	148823862	1.42 (1.25, 1.61)	< 1E-007	1.43 (1.26, 1.62)	< 1E-007	1.43 (1.26, 1.63)	< 1E-007	Island	*ZNF398*	5'UTR
chr10	cg18072629	8092036	1.30 (1.17, 1.44)	1.00E-06	1.30 (1.17, 1.44)	1.00E-06	1.30 (1.17, 1.44)	2.00E-06	Island		Intergenic
chr10	**cg05970116**	**75351076**	**1.48 (1.28, 1.71)**	**< 1E-007**	**1.50 (1.29, 1.75)**	**< 1E-007**	**1.51 (1.29, 1.76)**	**< 1E-007**	OpenSea		Intergenic
chr11	cg10321339	369810	1.54 (1.30, 1.83)	1.00E-06	1.55 (1.30, 1.84)	1.00E-06	1.54 (1.29, 1.84)	1.00E-06	OpenSea	*B4GALNT4*	1st Exon
chr12	cg12704462	120151527	0.74 (0.66, 0.83)	1.00E-06	0.74 (0.66, 0.84)	1.00E-06	0.75 (0.66, 0.84)	2.00E-06	OpenSea	*MIR1178*	Body
chr15	cg11823654¥	65117936	1.65 (1.35, 2.03)	1.00E-06	1.65 (1.35, 2.02)	1.00E-06	1.68 (1.37, 2.06)	1.00E-06	S Shore	*PIF1*	TSS200
chr15	cg18144285	91643026	1.47 (1.26, 1.72)	1.00E-06	1.48 (1.27, 1.72)	1.00E-06	1.46 (1.24, 1.72)	5.00E-06	Island	*SV2B*	TSS200
chr16	cg04872027	29149757	0.73 (0.64, 0.83)	1.00E-06	0.72 (0.64, 0.82)	1.00E-06	0.73 (0.64, 0.83)	3.00E-06	N Shelf		Intergenic
chr20	**cg06007966**	**26188971**	**1.94 (1.49, 2.52)**	**1.00E-06**	**1.94 (1.49, 2.53)**	**1.00E-06**	**1.95 (1.48, 2.56)**	**2.00E-06**	Island	*MIR663*	TSS200
chr20	**cg17375901**	**61754940**	**0.62 (0.51, 0.74)**	**< 1E-007**	**0.61 (0.51, 0.74)**	**< 1E-007**	**0.62 (0.51, 0.75)**	**1.00E-06**	Island		Intergenic

BMI, body mass index; Chr, chromosome; CI, confidence interval; CpG, CpG dinucleotide; CRC, colorectal cancer; DM, ever having been treated for diabetes mellitus; HR, hazard ratio; IR, insulin resistance; TSS200, 0-200 bp upstream of transcription start site; UTR, untranslated region; WHI, Women's Health Initiative. CpGs in bold face are among those statistically significant, shared ones across WHIBAA23, GSE48684, and GSE199057. §Annotation used R v.0.6.0.*IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays.* ¶HR adjusted by leukocyte heterogeneities ($CD8^+CD28^-CD45RA^-$ T cell, naïve CD8 T cell, plasma blast, $CD4^+$ T cell, natural killer cell, monocyte, and granulocyte) plus DNA methylation-predicted age. ¥Promoter associated. *Enhancer associated.

Table 2. GSE datasets: differentially DNA-methylated top 20 CpGs identified from WHIBAA23 in association with CRC risk

Chr	CpG site§	Position	GSE 48684		GSE199057		CpG context	Gene	Gene region
			OR¶ (95% CI)	*P*	OR£ (95% CI)	*P*			
chr1	cg14057946¥	713985	1.46 (0.89, 2.87)	0.211	0.97 (0.59, 1.58)	0.907	Island		Intergenic
chr1	cg04231937¥	714526	1.06 (0.71, 1.68)	0.768	1.40 (0.83, 3.03)	0.273	Island		Intergenic
chr1	cg04496485¥	714565	1.30 (0.82, 2.48)	0.344	1.48 (0.88, 2.84)	0.167	S Shore		Intergenic
chr1	cg02014020	1115461	1.47 (0.98, 2.30)	0.076	3.55 (1.88, 7.69)	3.52E-04	N Shore	*TTLL10*	Body
chr6	cg07572131*,¥	31430791	0.66 (0.42, 1.00)	0.057	1.04 (0.62, 1.74)	0.875	OpenSea	*HCP5*	TSS200
chr6	cg25410010	41554543	1.06 (0.69, 1.57)	0.793			OpenSea	*FOXP4*	Body
chr6	cg06498809¥	111303174	0.83 (0.55, 1.24)	0.362	0.27 (0.11, 0.55)	0.001	Island	*RPF2*	TSS200
chr6	cg00020352¥	111303252	0.61 (0.39, 0.93)	0.025			Island	*RPF2*	TSS200
chr6	cg10920427¥	111303363	1.40 (0.92, 2.20)	0.126	1.55 (0.94, 2.66)	0.096	Island	*RPF2*	Body
chr6	cg14498116¥	111303482	0.32 (0.15, 0.60)	0.001	0.61 (0.33, 1.03)	0.085	Island	*RPF2*	Body
chr7	cg04958124¥	148823862	2.37 (1.43, 4.37)	0.002	1.60 (0.97, 2.85)	0.084	Island	*ZNF398*	5'UTR
chr10	cg18072629	8092036	3.00 (1.42, 10.25)	0.025			Island		Intergenic
chr10	**cg05970116**	**75351076**	**0.48 (0.29, 0.75)**	**0.002**	**0.13 (0.04, 0.31)**	**2.80E-05**	OpenSea		Intergenic
chr11	cg10321339	369810	6.73 (2.29, 40.88)	0.008	32.38 (4.94, 568.21)	0.003	OpenSea	*B4GALNT4*	1st Exon
chr12	cg12704462	120151527	0.84 (0.47, 1.28)	0.477	0.19 (0.06, 0.46)	0.001	OpenSea	*MIR1178*	Body
chr15	cg11823654¥	65117936	0.39 (0.23, 0.63)	2.45E-04	0.65 (0.36, 1.08)	0.109	S Shore	*PIF1*	TSS200
chr15	cg18144285	91643026	199.31 (5.83, 59923.50)	0.023	22.60 (4.28, 246.53)	0.003	Island	*SV2B*	TSS200
chr16	cg04872027	29149757	0.50 (0.25, 0.86)	0.025	0.48 (0.24, 0.83)	0.018	N Shelf		Intergenic
chr20	**cg06007966**	**26188971**	**4.20 (2.34, 8.59)**	**1.20E-05**	**8.05 (3.44, 25.24)**	**2.80E-05**	Island	*MIR663*	TSS200
chr20	**cg17375901**	**61754940**	**0.07 (0.02, 0.22)**	**6.70E-05**	**0.03 (0.003, 0.12)**	**1.23E-04**	Island		Intergenic

Chr, chromosome; CI, confidence interval; CpG, CpG dinucleotide; CRC, colorectal cancer; OR, odds ratio; TSS200, 0-200 bp upstream of transcription start site; UTR, untranslated region; WHI, Women's Health Initiative. CpGs in bold face are among those statistically significant, shared ones across WHIBAA23, GSE48684, and GSE199057. §Annotation used R v.0.6.0.*IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays.* ¶OR adjusted by sex. £OR adjusted by sex plus age and DNA methylation-predicted age. ¥Promoter associated. *Enhancer associated.

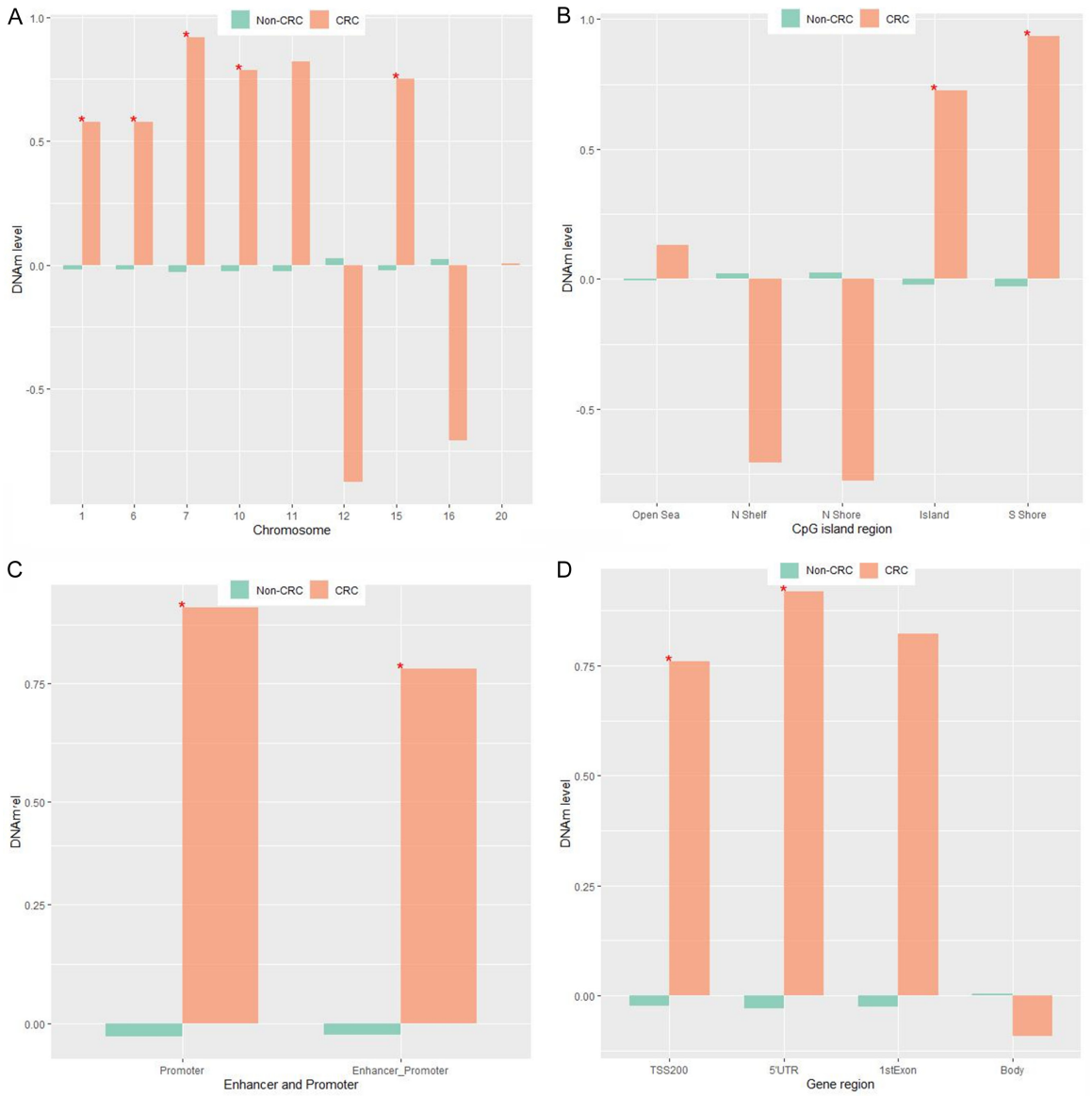


Figure 2. WHIBAA23: Bar plots for mean difference in DNAm levels of top 20 genome-wide CpGs stratified by CRC status. A. By chromosome. B. By CpG context. C. By enhancer and/or promoter. D. By gene region. CpG, CpG dinucleotide; CRC, colorectal cancer; DNAm, DNA methylation; TSS200, 0-200 bp upstream of transcription start site; UTR, untranslated region; WHI, Women's Health Initiative. Note: *Statistical significance after multiple comparison correction.

presented a different direction for CRC risk between the WHI and GEO datasets, 2 other CpGs (*MIR663*cg06007966 and cg17375901) had similar DNAm levels across the cohorts, suggesting DNAm parallelisms between PBLs and tissues in CRC patients.

DMR scans for CRC

Our DMR analyses showed distinct patterns between PBL- and tissue-based databases. In particular, both GEO datasets detected similar DMRs, showing that 4 of each top 5 DMRs (Figure S4) and > 70 of each top 100 DMRs (Table S2) overlapped. Also, the combined results of our EWA and DMR analyses in each GEO contained multiple CpGs overlapping between the top 20 CpGs and the CpGs detected from DMR scans (**Table 3**). Of them, *PIF1*cg11823654, *RPF2*cg14498116, and *ZNF398*cg04958124 in the GSE48984, and *TTLL10*cg02014020, *SV2B*cg18144285, *B4GALNT4*cg10321339, *RPF2*cg06498809, *MIR1178*cg12704462, and cg05970116 in the

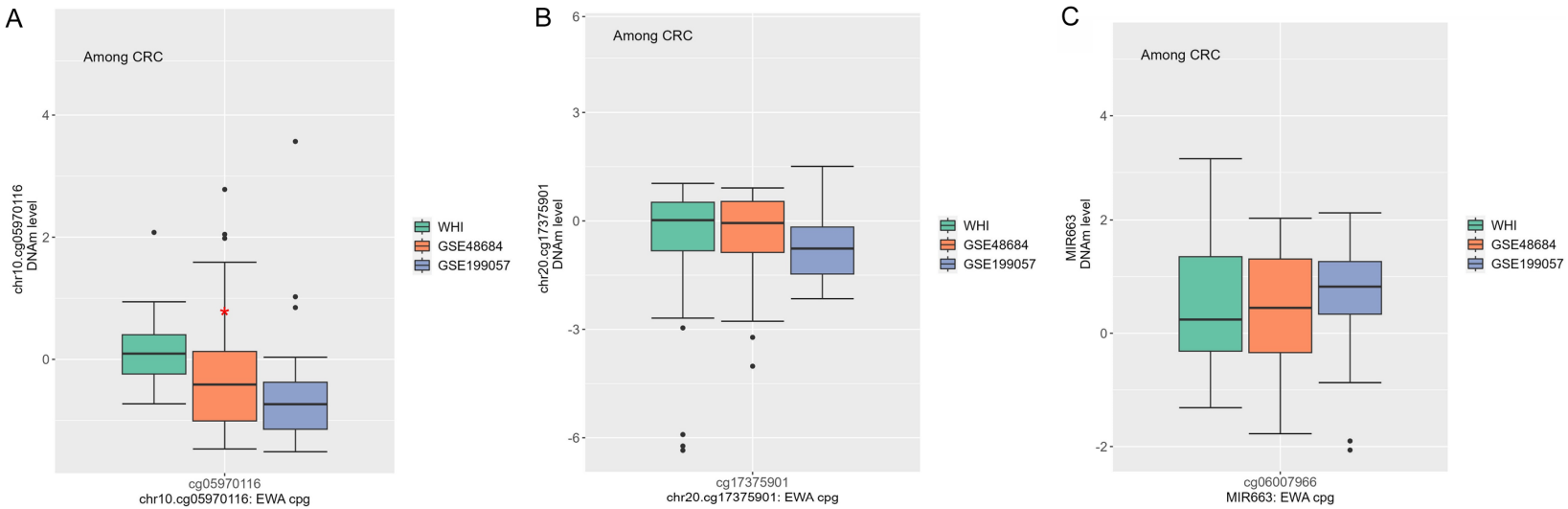


Figure 3. WHIBAA23, GSE48684, and GSE199057 in CRC patients (peripheral leukocytes for WHIBAA23 and CRC tissues for GSEs): Comparisons among the 3 studies for DNAm levels of 3 individual CpGs that are statistically significant and shared across the studies. A. Chr10, cg05970116. B. Chr20, cg17375901. C. Chr20, cg06007966, *MIR663*. Chr, chromosome; CpG, CpG dinucleotide; CRC, colorectal cancer; DNAm, DNA methylation; EWA, epigenome-wide association; WHI, Women's Health Initiative. Note: *Statistical significance after multiple comparison correction.

Table 3. Combined results from EWA and DMR analyses in GSE datasets

# of DMR	Seqnames	Start	End	Width	Fisher	DMR: No. of CpGs	Overlapping CpGs¶	DMR: genes	Overlapping genes¥
<GSE48684>									
55	chr10	8088801	8103673	14873	4.18E-215	90	cg18072629	*RP11-379F12.4, GATA3, GATA3-AS1, RP11-379F12.3*	
2782	chr11	368351	369875	1525	2.56E-30	27	cg10321339	*B4GALNT4*	*B4GALNT4*
2910	chr15	91642470	91643742	1273	2.58E-29	15	cg18144285	*SV2B*	*SV2B*
2924	chr15	65115218	65119016	3799	3.34E-29	13	cg11823654§	*PIF1*	*PIF1§*
2931	chr20	26188639	26189240	602	3.72E-29	7	cg06007966§,*	*MIR663A*	*MIR663§,**
6682	chr1	1113501	1115920	2420	8.03E-13	16	cg02014020	*TTLL10, TTLL10-AS1*	*TTLL10*
8302	chr6	111302729	111303792	1064	9.06E-10	13	cg00020352, cg06498809, cg10920427, cg14498116§	*RPF2*	*RPF2§*
17740	chr7	148822673	148823965	1293	0.013	13	cg04958124§	*ZNF398, ZNF425, RN7SL521P*	*ZNF398§*
20166	chr12	120151527	120152127	601	0.090	7	cg12704462	*CIT, MIR1178*	*MIR1178*
<GSE199057>									
1979	chr20	26188639	26190354	1716	4.78E-56	10	cg06007966§,*	*MIR663A*	*MIR663§,**
4442	chr1	1113624	1115920	2297	8.25E-36	17	cg02014020§	*TTLL10, TTLL10-AS1*	*TTLL10§*
4640	chr15	91641719	91643742	2024	4.98E-35	18	cg18144285§	*SV2B*	*SV2B§*
5895	chr15	65116255	65119016	2762	1.46E-30	16	cg11823654	*PIF1*	*PIF1*
13206	chr11	368351	369875	1525	1.88E-16	25	cg10321339§	*B4GALNT4*	*B4GALNT4§*
14389	chr6	111301798	111303792	1995	4.06E-15	17	cg06498809§, cg10920427, cg14498116	*RPF2*	*RPF2§*
18934	chr12	120151527	120152127	601	5.19E-11	5	cg12704462§	*CIT, MIR1178*	*MIR1178§*
21755	chr10	75351076	75351888	813	5.84E-09	5	cg05970116§	*USP54*	
33351	chr7	148822673	148823862	1190	0.079	13	cg04958124	*ZNF398, ZNF425, RN7SL521P*	*ZNF398*

Among top 20 genome-wide CpGs, overlapping CpGs¶ and nearby genes¥. Chr, chromosome; CpG, CpG dinucleotide; DMR, differentially methylated region; EWA, epigenome-wide association. §CpGs and nearby genes that overlap between EWA and DMR analyses in each GSE, which are statistically significant at the validation level. *CpG and nearby gene that are statistically significant at the validation level and overlapping across the 2 GSE datasets.

Table 4. Differentially methylated regions (DMRs) overlapping across WHIBAA23, GSE48684, and GSE199057

Seqnames	Start	End	Width	Score	No. of CpGs	CpGs§,¥	Position	CpG context	Overlapping Gene	Gene region
Chromosome 7	94286086	94286267	182	3	10	cg24885794	94286086	Island	PEG10	5'UTR
						cg26997085	94286110	Island	*PEG10*	5'UTR
						cg22331138	94286131	Island	*PEG10*	5'UTR
						cg16492735	94286208	Island	*PEG10*	5'UTR
						cg09512080	94286219	Island	*PEG10*	5'UTR
						cg00906934	94286232	Island	*PEG10*	5'UTR
						cg26503018	94286243	Island	*PEG10*	5'UTR
						cg27120649	94286261	Island	*PEG10*	5'UTR
						cg21771834	94286263	Island	*PEG10*	5'UTR
						cg27001184	94286267	Island	*PEG10*	5'UTR

UTR, untranslated region; WHI, Women's Health Initiative. Note: The score of 3 indicates that the 3 datasets have overlapping DMRs, and the 10 CpGs are not genome-wide site-specific CpGs. §Annotation used R v.0.6.0.*IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays*. ¥All 10 CpGs are promoter associated.

GSE199057 were significant at the validation level. Of note, *MIR663*cg06007966, which was validated as positively associated with CRC in both PBL- and tissue-based databases, was also detected as an overlapped probe in the DMR scans of both GEO datasets.

A different pattern was observed in the DMR analysis for the WHI cohort, demonstrating no overlapping genome-wide CpGs in the DMRs. Moreover, the DMR shared across all 3 cohorts was only 1 region in *PEG10* of Chr7, and 10 CpGs detected in this DMR did not overlap with any genome-wide CpGs (**Table 4**). We further estimated the effect sizes of DNAm for these individual CpGs, displaying a consistently increased risk of CRC across the cohorts (Table S3).

Transcriptomic profile and GSEA

Among 8 genome-wide genes from the EWA analysis plus 1 additional gene overlapped across the DMRs of the 3 cohorts, 7 passed the FDR < 0.05 (**Figure 4A**, **4B**). In particular, *B4GALNT4* and *PIF1*, whose related CpGs were hypermethylated (in both PBL- and tissue-based CRC) and hypomethylated (in CRC tissues), respectively, showed the strongest upregulation of mRNA-Seqs in CRC tissues (**Figure 4C**, **4D**). In contrast, *SV2B* presented the strongest downregulation of mRNA-Seqs in CRC, where associated CpGs in our analyses of the WHI and GEOs showed hypermethylation in CRC (**Figure 4E**). Further, *FOXP4*, *RPF2*, and *TTLL10* were upregulated in CRC tissues with relevant-CpGs' hypomethylation in our CRC cohorts, whereas *ZNF398* displayed weak upregulation with hypermethylation of associated-CpGs in CRC (Figure S5).

Finally, with genome-wide CpGs from our EWA scan, we performed multiple analyses of GSEA gene ontology (GO) with biologic process, cellular component, and molecular functions, pathways with *KEGG* and *Reactome*, and diseases via *DisGeNET* and *GLAD4U* databases (Table S4). GO with biologic and molecular functions identified DNA/RNA biosynthetic processes, telomeres' organism/DNA binding, p53-mediated signal transduction, and catalytic/transferase activity on glycosyl groups. Gene-enrichment pathways were involved in extracellular matrix (ECM)-receptor interaction, which plays an important role in regulating cell behavior, communicating cell proliferation and migration, implicating a key role in CRC development [64, 65]. *Reactome* pathways and diseases were involved in neurotransmitter transport, infection, and neoplasms.

Discussion

To our knowledge, this study is the first genome-wide scan in postmenopausal women, the population most vulnerable to CRC, with pre-diagnostic DNAm in PBLs to prospectively evaluate CRC development in both CpG site-specific and regionally differentiated methylation fashions. We further validated in CRC tissue-level datasets and finally, mapped to transcriptome profiles. As hypothesized, the DNAm levels and

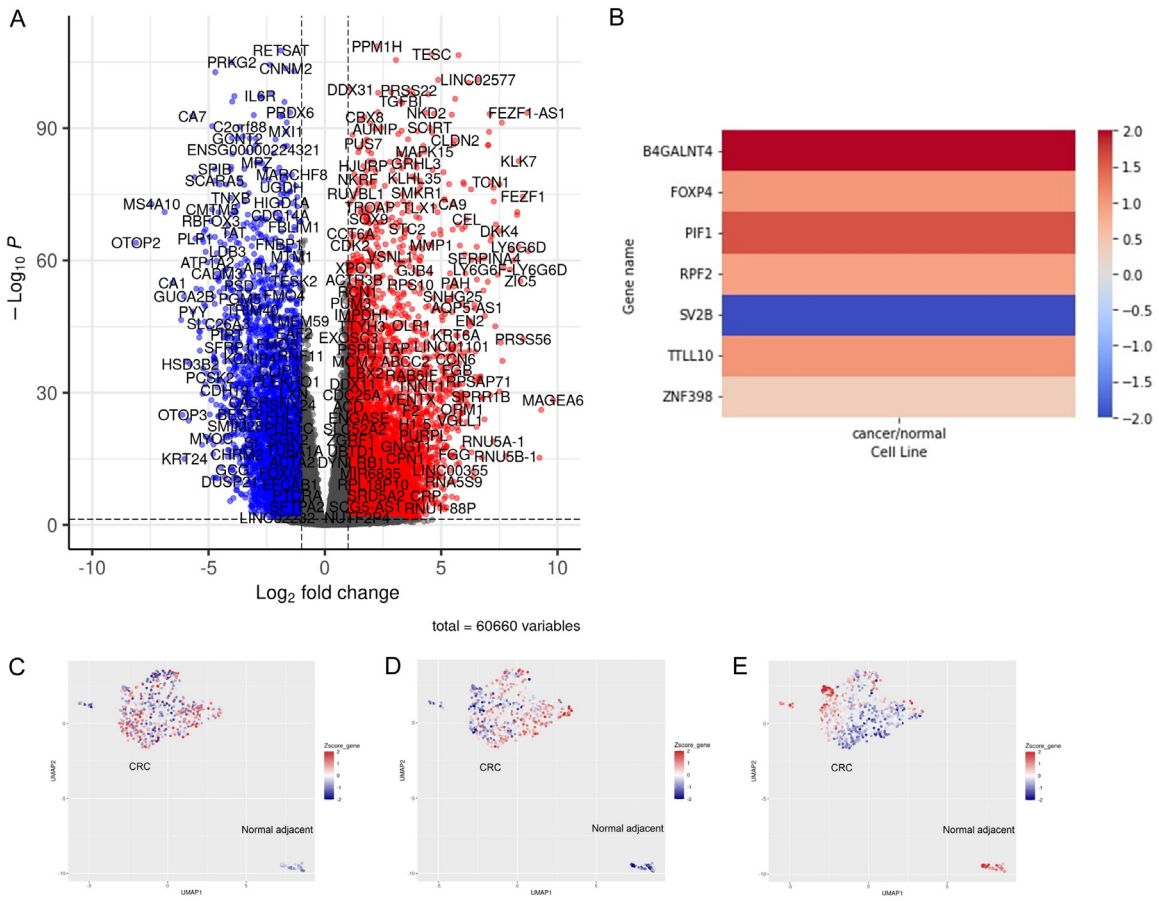


Figure 4. mRNA-sequences mapping to genes in TCGA COAD and READ datasets. A. Volcano plot: Differentially expressed genes between CRC and normal adjacent tissues. B. Heat plot: Log_2 fold changes in modeled genes. C. UMAP plot: *B4GALNT4*. D. UMAP plot: *PIF1*. E. UMAP plot: *SV2B*. CRC, colorectal cancer; UMAP, Uniform Manifold Approximation and Projection.

risk magnitudes of detected CpGs as well as DMR patterns between 2 independent tissue datasets were similar, indicating somatic-level epigenetic signatures. Also, several genome-wide CpGs in genes overlapped across PBL- and tissue-based datasets, suggesting DNAm parallelisms between PBLs and tissues in a site-/gene-specific manner; these validated DNAm probes may have further implications as the best long-term surrogate markers in non-invasive tissues, reflecting multiple interconnected CRC tumorigenesis mechanisms.

In detail, DNAm of cg10321339 in the first exon of *B4GALNT4* was strongly associated with increased risk for CRC development in both PBL and tissue datasets, and also, the strongest upregulated expression of *B4GALNT4* was observed in CRC tissues. *B4GALNT4*, encoding an enzyme β-1,4-N-acetylgalactosaminyltransferase 4, is involved in LacdiNAc group synthesis, which is important in embryonic development and disease progression [66, 67]. It has been associated with progression of cancers, with decreased expression in breast carcinomas (BC) [68] and esophageal squamous cell carcinomas [69]. *B4GALNT3* has also been studied in CRC cells, demonstrating overexpression [70], similar to our transcriptomics finding. Our detected genome-wide CpG and overexpression of these *B4GALNT* gene groups in CRC tissues are novel findings, deserving further validation and functional studies.

The DNAm of cg18144285 in the CpG island within 200 bp upstream of the transcription start site (TSS200) of *SV2B* displayed the second strongest effect on increased risk for CRC, with strong down-regulation of *SV2B* in CRC tis-

sues. Synaptic vesicle glycoprotein 2B (SV2B) is essential to the synaptic machinery in neural and endocrine cells [71, 72] and is overexpressed in prostate small-cell neuroendocrine carcinoma [73] and glioblastoma [74]. Of note, our GSEA-GO analysis in CRC detected the ECM-receptor interaction pathways, which play an important role in modulating cancer-cell behaviors [64, 65], with *SV2B* as a key driver; this is consistent with previous findings in gastric cancer [75], which identified *SV2B* as a strong indicator of ECM-receptor interactions. However, the role of *SV2B* in tumors is still inconclusive.

Some other genes we detected at genome-wide significance are also involved in cancer development and progression. For instance, *ZNF398*, encoding zinc finger protein 398, enables transcription of *TGF-β* downstream pluripotency/epithelial characters in human pluripotent stem cells [76] and, as an oncogene, was upregulated in tumor tissues [77, 78]. *PIF1*, conserving non-processive 5'-to-3' DNA helicase, has a functional role in tumor cell viability during replication stress and inhibits apoptosis, which is essential in the early stage of tumorigenesis [79, 80]; it is also overexpressed in lung cancer [81], BC [82], and neuroblastoma [83]. However, these genes' biologic function and detected DNAm probes' potential involvement in CRC tumorigenesis remain elusive, calling for functional/mechanical studies on the methylome of these genes in CRC.

Of noteworthy, cg06007966 in the CpG island at TSS200 of *MIR663* was validated genome-wide across PBLs and 2 tissue datasets. MicroRNAs (miRNAs) are short non-coding RNAs that control gene expression by targeting mRNAs to promote either translation regression or RNA degradation [84, 85]. Aberrant miRNAs have been found in human cancers, correlated with tumorigenesis and progression. In particular, miR-663 has a strong binding affinity to *AATF* (an anti-apoptotic gene) mRNA, thus, promoting apoptosis in cancer cells, known as "apopto-miR" [86]. The miR-663 is regulated epigenetically; in particular, the CpG island promoter region of miR-663 is hypermethylated, showing decreased expression [85, 87-90], resulting in tumor cell growth, invasion, and metastasis in multiple cancer cells [85, 87-89, 91-95], including CRC [96, 97]. Our CpG in miR-663 was hypermethylated in CRC, presuming downregulation, consistent with those previous study findings. In contrast, miR-663 is also considered an "onco-miR" in several cancer cells with different target genes and downstream signaling involved in carcinogenesis and cancer growth [84, 90, 98-102]. Overall, the role of miR-663 and its abnormal expression regulated by the methylome is little known in CRC, warranting functional validation studies.

Finally, our DMR analyses detected 1 region shared across PBL and tissue cohorts in *PEG10* of Chr7 with 10 related CpGs, although these probes were not significant genome wide in our analysis. *PEG10* is considered an oncogene, a proliferation-positive, paternally expressed imprinted gene, overexpressed in cancer cells/tissues [103-108]. CRC tissues also showed overexpression of *PEG10* through which a long non-coding RNA sponges miR-574 [109]. Interestingly, *PEG10* was the only gene differentially expressed in a study [110] comparing gene expression between early- and late-onset (≥ age 65 years) CRC, in which its overexpression was found only in the early-onset group; this supports our finding of the 10 CpGs in *PEG10* that were hypermethylated in CRC (i.e., a negative effect on gene expression) in our postmenopausal women.

Our analysis of GSE48684 did not include DNAm-age prediction and tumor purity owing to a lack of data availability. Our transcriptome profile did not analyze miRNAs, as the data contained mRNA-Seqs only; this deserves future functional/mechanical laboratory studies of miRNAs for biologic implications in CRC. Also, data from the methylome for our EWA analyses and from the transcriptomics for gene expression were not paired; thus, our findings should be interpreted with caution. The two GEO tissue datasets have different tissue sources - tissues from CRC patients compared with their normal adjacent tissues and tissues from CRC patients compared with those from non-CRC patients - supporting that our validation studies reflect complex pathways underlying CRC. However, few DNAm probes from the GEO databases demonstrated an extreme risk magnitude, a replication study with a larger dataset is warranted. In addition, because we repurposed data from the WHI AS, samples

analyzed for our study may not fully reflect the source population, resulting in limited statistical power, and our study findings should not be generalized to populations other than white postmenopausal women.

In summary, we found multiple site-specific CpGs and differentially methylated regions across PBL- and tissue-level data at genome-wide significance for CRC development which had been prospectively evaluated. Some are novel, but others are well-established in CRC, warranting epigenetic and functional validation. Our study contributes to elucidating the complex interrelated mechanisms on the methylome underlying CRC tumorigenesis and suggests novel preventive DNAm-targets in PBLs for capturing individuals at high risk for CRC development.

Acknowledgements

We thank Michael Carey in the UCLA Department of Biological Chemistry, for valuable discussions and support during transcriptomics analyses and in preparation of this manuscript. Part of the data for this project was provided by the WHI program, which is funded by the National Heart, Lung, and Blood Institute, the National Institutes of Health, and the U.S. Department of Health and Human Services through 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, and 75N92021D00005. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession (phs000200.v11.p3). This study was supported by the NINR (K01NR017852) and the NIGMS (R01GM074701).

Written informed consent was obtained from the participants at the source.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Su Yon Jung, Translational Sciences Section, School of Nursing, University of California, 700 Tiverton Avenue, 3-264 Factor Building, Los Angeles, CA 90095, USA. Tel: 310-825-2840; Fax: 310-267-0413; E-mail: sjung@sonnet.ucla.edu

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A and Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71: 209-249.

[2] American Cancer Society. Cancer facts and figures 2023. American Cancer Society, Inc. 2023. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf.

[3] Sakai E, Nakajima A and Kaneda A. Accumulation of aberrant DNA methylation during colorectal cancer development. World J Gastroenterol 2014; 20: 978-87.

[4] Fatemi N, Tierling S, Es HA, Varkiani M, Mojarad EN, Aghdaei HA, Walter J and Totonchi M. DNA methylation biomarkers in colorectal cancer: clinical applications for precision medicine. Int J Cancer 2022; 151: 2068-2081.

[5] Wang T, Maden SK, Luebeck GE, Li CI, Newcomb PA, Ulrich CM, Joo JE, Buchanan DD, Milne RL, Southey MC, Carter KT, Willbanks AR, Luo Y, Yu M and Grady WM. Dysfunctional epigenetic aging of the normal colon and colorectal cancer risk. Clin Epigenetics 2020; 12: 5.

[6] Wang Y, Zhang M, Hu X, Qin W, Wu H and Wei M. Colon cancer-specific diagnostic and prognostic biomarkers based on genome-wide abnormal DNA methylation. Aging (Albany NY) 2020; 12: 22626-22655.

[7] Johnson CM, Wei C, Ensor JE, Smolenski DJ, Amos CI, Levin B and Berry DA. Meta-analyses of colorectal cancer risk factors. Cancer Causes Control 2013; 24: 1207-22.

[8] Prasetyanti PR and Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. Mol Cancer 2017; 16: 41.

[9] Mendiratta G, Ke E, Aziz M, Liarakos D, Tong M and Stites EC. Cancer gene mutation frequencies for the U.S. population. Nat Commun 2021; 12: 5961.

[10] Li M, Zhu C, Xue Y, Miao C, He R, Li W, Zhang B, Yu W, Huang X, Lv M, Xu Y and Huang Q. A DNA methylation signature for the prediction of tumour recurrence in stage II colorectal cancer. Br J Cancer 2023; 128: 1681-1689.

[11] Tiffon C. The impact of nutrition and environmental epigenetics on human health and disease. Int J Mol Sci 2018; 19: 3425.

[12] Boughanem H, Izquierdo AG, Hernandez-Alonso P, Arranz-Salas I, Casanueva FF, Tinahones FJ, Crujeiras AB and Macias-Gonzalez M. An epigenetic signature is associated with serum

25-hydroxyvitamin D in colorectal cancer tumors. Mol Nutr Food Res 2021; 65: e2100125.
[13] Khayami R, Goltzman D, Rabbani SA and Kerachian MA. Epigenomic effects of vitamin D in colorectal cancer. Epigenomics 2022; 14: 1213-1228.
[14] Crujeiras AB, Morcillo S, Diaz-Lagares A, Sandoval J, Castellano-Castillo D, Torres E, Hervas D, Moran S, Esteller M, Macias-Gonzalez M, Casanueva FF and Tinahones FJ. Identification of an episignature of human colorectal cancer associated with obesity by genome-wide DNA methylation analysis. Int J Obes (Lond) 2019; 43: 176-188.
[15] Dong L, Ma L, Ma GH and Ren H. Genome-wide analysis reveals DNA methylation alterations in obesity associated with high risk of colorectal cancer. Sci Rep 2019; 9: 5100.
[16] Heide T, Househam J, Cresswell GD, Spiteri I, Lynn C, Mossner M, Kimberley C, Fernandez-Mateos J, Chen B, Zapata L, James C, Barozzi I, Chkhaidze K, Nichol D, Gunasri V, Berner A, Schmidt M, Lakatos E, Baker AM, Costa H, Mitchinson M, Piazza R, Jansen M, Caravagna G, Ramazzotti D, Shibata D, Bridgewater J, Rodriguez-Justo M, Magnani L, Graham TA and Sottoriva A. The co-evolution of the genome and epigenome in colorectal cancer. Nature 2022; 611: 733-743.
[17] Barfield R, Huyghe JR, Lemire M, Dong X, Su YR, Brezina S, Buchanan DD, Figueiredo JC, Gallinger S, Giannakis M, Gsur A, Gunter MJ, Hampel H, Harrison TA, Hopper JL, Hudson TJ, Li CI, Moreno V, Newcomb PA, Pai RK, Pharoah PDP, Phipps AI, Qu C, Steinfelder RS, Sun W, Win AK, Zaidi SH, Campbell PT, Peters U and Hsu L. Genetic regulation of DNA methylation yields novel discoveries in GWAS of colorectal cancer. Cancer Epidemiol Biomarkers Prev 2022; 31: 1068-1076.
[18] Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, Hedman AK, Sandling JK, Li LA, Irvin MR, Zhi D, Deloukas P, Liang L, Liu C, Bressler J, Spector TD, North K, Li Y, Absher DM, Levy D, Arnett DK, Fornage M, Pankow JS and Boerwinkle E. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. Hum Mol Genet 2015; 24: 4464-79.
[19] Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H, Waldenberger M, Peters A, Erdmann J, Hengstenberg C, Cambien F, Goodall AH, Ouwehand WH, Schunkert H, Thompson JR, Spector TD, Gieger C, Tregouet DA, Deloukas P and Samani NJ. DNA methylation and body-mass index: a genome-wide analysis. Lancet 2014; 383: 1990-8.
[20] Yang X, Wen X, Guo Q, Zhang Y, Liang Z, Wu Q, Li Z, Ruan W, Ye Z, Wang H, Chen Z, Fan JB, Lan P, Liu H and Wu X. Predicting disease-free survival in colorectal cancer by circulating tumor DNA methylation markers. Clin Epigenetics 2022; 14: 160.
[21] Liu Z, Georgakopoulos-Soares I, Ahituv N and Wong KC. Risk scoring based on DNA methylation-driven related DEGs for colorectal cancer prognosis with systematic insights. Life Sci 2023; 316: 121413.
[22] Ouchi K, Takahashi S, Yamada Y, Tsuji S, Tatsuno K, Takahashi H, Takahashi N, Takahashi M, Shimodaira H, Aburatani H and Ishioka C. DNA methylation status as a biomarker of anti-epidermal growth factor receptor treatment for metastatic colorectal cancer. Cancer Sci 2015; 106: 1722-9.
[23] Shen Y, Wang D, Yuan T, Fang H, Zhu C, Qin J, Xu X, Zhang C, Liu J, Zhang Y, Wen Z, Tang J and Wang Z. Novel DNA methylation biomarkers in stool and blood for early detection of colorectal cancer and precancerous lesions. Clin Epigenetics 2023; 15: 26.
[24] Gao D, Herman JG and Guo M. The clinical value of aberrant epigenetic changes of DNA damage repair genes in human cancer. Oncotarget 2016; 7: 37331-37346.
[25] Zitt M, Zitt M and Muller HM. DNA methylation in colorectal cancer--impact on screening and therapy monitoring modalities? Dis Markers 2007; 23: 51-71.
[26] Widschwendter M, Apostolidou S, Raum E, Rothenbacher D, Fiegl H, Menon U, Stegmaier C, Jacobs IJ and Brenner H. Epigenotyping in peripheral blood cell DNA and breast cancer risk: a proof of principle study. PLoS One 2008; 3: e2656.
[27] Kohonen-Corish MR, Sigglekow ND, Susanto J, Chapuis PH, Bokey EL, Dent OF, Chan C, Lin BP, Seng TJ, Laird PW, Young J, Leggett BA, Jass JR and Sutherland RL. Promoter methylation of the mutated in colorectal cancer gene is a frequent early event in colorectal cancer. Oncogene 2007; 26: 4435-41.
[28] Kaneda A and Feinberg AP. Loss of imprinting of IGF2: a common epigenetic modifier of intestinal tumor risk. Cancer Res 2005; 65: 11236-40.
[29] Baylin SB and Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer 2006; 6: 107-16.
[30] Feinberg AP, Ohlsson R and Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet 2006; 7: 21-33.
[31] Dor Y and Cedar H. Principles of DNA methylation and their implications for biology and medicine. Lancet 2018; 392: 777-786.

[32] Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, Feinberg AP and Irizarry RA. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med 2014; 6: 61.
[33] Hong SN, Kim SJ, Kim ER, Chang DK and Kim YH. Epigenetic silencing of NDRG2 promotes colorectal cancer proliferation and invasion. J Gastroenterol Hepatol 2016; 31: 164-71.
[34] Muller D and Gyorffy B. DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer. Biochim Biophys Acta Rev Cancer 2022; 1877: 188722.
[35] Yu H, Wang X, Bai L, Tang G, Carter KT, Cui J, Huang P, Liang L, Ding Y, Cai M, Huang M, Liu H, Cao G, Gallinger S, Pai RK, Buchanan DD, Win AK, Newcomb PA, Wang J, Grady WM and Luo Y. DNA methylation profile in CpG-depleted regions uncovers a high-risk subtype of early-stage colorectal cancer. J Natl Cancer Inst 2023; 115: 52-61.
[36] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012; 13: 484-92.
[37] Galamb O, Kalmar A, Peterfia B, Csabai I, Bodor A, Ribli D, Krenacs T, Patai AV, Wichmann B, Bartak BK, Toth K, Valcz G, Spisak S, Tulassay Z and Molnar B. Aberrant DNA methylation of WNT pathway genes in the development and progression of CIMP-negative colorectal cancer. Epigenetics 2016; 11: 588-602.
[38] Onwuka JU, Li D, Liu Y, Huang H, Xu J, Liu Y, Zhang Y and Zhao Y. A panel of DNA methylation signature from peripheral blood may predict colorectal cancer susceptibility. BMC Cancer 2020; 20: 692.
[39] Liu Y, Wang Y, Hu F, Sun H, Zhang Z, Wang X, Luo X, Zhu L, Huang R, Li Y, Li G, Li X, Lin S, Wang F, Liu Y, Rong J, Yuan H and Zhao Y. Multiple gene-specific DNA methylation in blood leukocytes and colorectal cancer risk: a case-control study in China. Oncotarget 2017; 8: 61239-61252.
[40] Luo X, Huang R, Sun H, Liu Y, Bi H, Li J, Yu H, Sun J, Lin S, Cui B and Zhao Y. Methylation of a panel of genes in peripheral blood leukocytes is associated with colorectal cancer. Sci Rep 2016; 6: 29922.
[41] Conway K, Edmiston SN, Tse CK, Bryant C, Kuan PF, Hair BY, Parrish EA, May R and Swift-Scanlan T. Racial variation in breast tumor promoter methylation in the carolina breast cancer study. Cancer Epidemiol Biomarkers Prev 2015; 24: 921-30.
[42] Shah UJ, Xie W, Flyvbjerg A, Nolan JJ, Hojlund K, Walker M, Relton CL and Elliott HR; RISC consortium. Differential methylation of the type 2 diabetes susceptibility locus KCNQ1 is associated with insulin sensitivity and is predicted by CpG site specific genetic variation. Diabetes Res Clin Pract 2019; 148: 189-199.
[43] Li L, Choi JY, Lee KM, Sung H, Park SK, Oze I, Pan KF, You WC, Chen YX, Fang JY, Matsuo K, Kim WH, Yuasa Y and Kang D. DNA methylation in peripheral blood: a potential biomarker for cancer molecular epidemiology. J Epidemiol 2012; 22: 384-94.
[44] Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, Etzioni R, McKenna MT, Oeffinger KC, Shih YT, Walter LC, Andrews KS, Brawley OW, Brooks D, Fedewa SA, Manassaram-Baptiste D, Siegel RL, Wender RC and Smith RA. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA Cancer J Clin 2018; 68: 250-281.
[45] Ghosh J, Schultz BM, Chan J, Wultsch C, Singh R, Shureiqi I, Chow S, Doymaz A, Varriano S, Driscoll M, Muse J, Kleiman FE, Krampis K, Issa JJ and Sapienza C. Epigenome-wide study identifies epigenetic outliers in normal mucosa of patients with colorectal cancer. Cancer Prev Res (Phila) 2022; 15: 755-766.
[46] Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials 1998; 19: 61-109.
[47] Holliday KM, Gondalia R, Baldassari A, Justice AE, Stewart JD, Liao D, Yanosky JD, Jordahl KM, Bhatti P, Assimes TL, Pankow JS, Guan W, Fornage M, Bressler J, North KE, Conneely KN, Li Y, Hou L, Vokonas PS, Ward-Caviness CK, Wilson R, Wolf K, Waldenberger M, Cyrys J, Peters A, Boezen HM, Vonk JM, Sayols-Baixeras S, Lee M, Baccarelli AA and Whitsel EA. Gaseous air pollutants and DNA methylation in a methylome-wide association study of an ethnically and environmentally diverse population of U.S. adults. Environ Res 2022; 212: 113360.
[48] Devall MA, Sun X, Eaton S, Cooper GS, Willis JE, Weisenberger DJ, Casey G and Li L. A race-specific, DNA methylation analysis of aging in normal rectum: implications for the biology of aging and its relationship to rectal cancer. Cancers (Basel) 2022; 15: 45.
[49] Devall M, Sun X, Yuan F, Cooper GS, Willis J, Weisenberger DJ, Casey G and Li L. Racial disparities in epigenetic aging of the right vs left colon. J Natl Cancer Inst 2021; 113: 1779-1782.
[50] Luo Y, Wong CJ, Kaz AM, Dzieciatkowski S, Carter KT, Morris SM, Wang J, Willis JE, Makar KW, Ulrich CM, Lutterbaugh JD, Shrubsole MJ, Zheng W, Markowitz SD and Grady WM. Differences in DNA methylation signatures reveal multiple pathways of progression from adeno-

ma to colorectal cancer. Gastroenterology 2014; 147: 418-29, e8.
[51] Langer RD, White E, Lewis CE, Kotchen JM, Hendrix SL and Trevisan M. The Women's Health Initiative Observational Study: baseline characteristics of participants and reliability of baseline measures. Ann Epidemiol 2003; 13 Suppl: S107-21.
[52] Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF and Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985; 28: 412-9.
[53] National Cancer Institute. SEER program: comparative staging guide for cancer. 1993.
[54] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D and Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 2013; 29: 189-96.
[55] Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK and Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012; 13: 86.
[56] Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013; 14: R115.
[57] Triche TJ Jr, Weisenberger DJ, Van Den Berg D, Laird PW and Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 2013; 41: e90.
[58] Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavare S and Ritchie ME. Statistical issues in the analysis of Illumina data. BMC Bioinformatics 2008; 9: 85.
[59] Kabat GC, Kim MY, Peters U, Stefanick M, Hou L, Wactawski-Wende J, Messina C, Shikany JM and Rohan TE. A longitudinal study of the metabolic syndrome and risk of colorectal cancer in postmenopausal women. Eur J Cancer Prev 2012; 21: 326-32.
[60] Gunter MJ, Hoover DR, Yu H, Wassertheil-Smoller S, Rohan TE, Manson JE, Howard BV, Wylie-Rosett J, Anderson GL, Ho GY, Kaplan RC, Li J, Xue X, Harris TG, Burk RD and Strickler HD. Insulin, insulin-like growth factor-I, endogenous estradiol, and risk of colorectal cancer in postmenopausal women. Cancer Res 2008; 68: 329-37.
[61] Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, Clark SJ and Molloy PL. De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin 2015; 8: 6.
[62] Peters TJ, Buckley MJ, Chen Y, Smyth GK, Goodnow CC and Clark SJ. Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate. Nucleic Acids Res 2021; 49: e109.
[63] Liao Y, Wang J, Jaehnig EJ, Shi Z and Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res 2019; 47: W199-W205.
[64] Nersisyan S, Novosad V, Engibaryan N, Ushkaryov Y, Nikulin S and Tonevitsky A. ECM-receptor regulatory network and its prognostic role in colorectal cancer. Front Genet 2021; 12: 782699.
[65] Stankevicius V, Vasauskas G, Noreikiene R, Kuodyte K, Valius M and Suziedelis K. Extracellular matrix-dependent pathways in colorectal cancer cell lines reveal potential targets for anticancer therapies. Anticancer Res 2016; 36: 4559-67.
[66] Gotoh M, Sato T, Kiyohara K, Kameyama A, Kikuchi N, Kwon YD, Ishizuka Y, Iwai T, Nakanishi H and Narimatsu H. Molecular cloning and characterization of beta1,4-N-acetylgalactosaminyltransferases IV synthesizing N,N'-diacetyllactosediamine. FEBS Lett 2004; 562: 134-40.
[67] Fiete D, Beranek M and Baenziger JU. Molecular basis for protein-specific transfer of N-acetylgalactosamine to N-linked glycans by the glycosyltransferases beta1,4-N-acetylgalactosaminyl transferase 3 (beta4GalNAc-T3) and beta4GalNAc-T4. J Biol Chem 2012; 287: 29194-203.
[68] Hirano K, Matsuda A, Kuji R, Nakandakari S, Shirai T and Furukawa K. Enhanced expression of the beta4-N-acetylgalactosaminyltransferase 4 gene impairs tumor growth of human breast cancer cells. Biochem Biophys Res Commun 2015; 461: 80-5.
[69] Baba H, Kanda M, Sato Y, Sawaki K, Shimizu D, Koike M, Motoyama S, Kodera Y and Fujii T. Expression and malignant potential of B4GALNT4 in esophageal squamous cell carcinoma. Ann Surg Oncol 2020; 27: 3247-3256.
[70] Huang J, Liang JT, Huang HC, Shen TL, Chen HY, Lin NY, Che MI, Lin WC and Huang MC. Beta1,4-N-acetylgalactosaminyltransferase III enhances malignant phenotypes of colon cancer cells. Mol Cancer Res 2007; 5: 543-52.
[71] Lekholm E, Ceder MM, Forsberg EC, Schioth HB and Fredriksson R. Differentiation of two human neuroblastoma cell lines alters SV2 expression patterns. Cell Mol Biol Lett 2021; 26: 5.
[72] Q7L1I2 SV2B_HUMAN. UniProtKB, Global core biodata resource. Accessed July, 2023. https://www.uniprot.org/uniprotkb/Q7L1I2/entry#function.
[73] Clegg N, Ferguson C, True LD, Arnold H, Moorman A, Quinn JE, Vessella RL and Nelson PS.

Molecular characterization of prostatic small-cell neuroendocrine carcinoma. Prostate 2003; 55: 55-64.
[74] Zhang Y, Yang X, Zhu XL, Hao JQ, Bai H, Xiao YC, Wang ZZ, Hao CY and Duan HB. Bioinformatics analysis of potential core genes for glioblastoma. Biosci Rep 2020; 40: BSR20201625.
[75] Yang X, Chen L, Mao Y, Hu Z and He M. Progressive and prognostic performance of an extracellular matrix-receptor interaction signature in gastric cancer. Dis Markers 2020; 2020: 8816070.
[76] Zorzan I, Pellegrini M, Arboit M, Incarnato D, Maldotti M, Forcato M, Tagliazucchi GM, Carbognin E, Montagner M, Oliviero S and Martello G. The transcriptional regulator ZNF398 mediates pluripotency and epithelial character downstream of TGF-beta in human PSCs. Nat Commun 2020; 11: 2364.
[77] Li WF, Alfason L, Huang C, Tang Y, Qiu L, Miyagishi M, Wu SR and Kasim V. p52-ZER6: a determinant of tumor cell sensitivity to MDM2-p53 binding inhibitors. Acta Pharmacol Sin 2023; 44: 647-660.
[78] Huang C, Wu S, Li W, Herkilini A, Miyagishi M, Zhao H and Kasim V. Zinc-finger protein p52-ZER6 accelerates colorectal cancer cell proliferation and tumour progression through promoting p53 ubiquitination. EBioMedicine 2019; 48: 248-263.
[79] Gagou ME, Ganesh A, Phear G, Robinson D, Petermann E, Cox A and Meuth M. Human PIF1 helicase supports DNA replication and cell growth under oncogenic-stress. Oncotarget 2014; 5: 11381-98.
[80] Gagou ME, Ganesh A, Thompson R, Phear G, Sanders C and Meuth M. Suppression of apoptosis by PIF1 helicase in human tumor cells. Cancer Res 2011; 71: 4998-5008.
[81] Zhang T, Wu DM, Luo PW, Liu T, Han R, Deng SH, He M, Zhao YY and Xu Y. CircNEIL3 mediates pyroptosis to influence lung adenocarcinoma radiotherapy by upregulating PIF1 through miR-1184 inhibition. Cell Death Dis 2022; 13: 167.
[82] Chisholm KM, Aubert SD, Freese KP, Zakian VA, King MC and Welcsh PL. A genomewide screen for suppressors of Alu-mediated rearrangements reveals a role for PIF1. PLoS One 2012; 7: e30748.
[83] Chen B, Hua Z, Gong B, Tan X, Zhang S, Li Q, Chen Y, Zhang J and Li Z. Downregulation of PIF1, a potential new target of MYCN, induces apoptosis and inhibits cell migration in neuroblastoma cells. Life Sci 2020; 256: 117820.
[84] Dahiya N, Sherman-Baust CA, Wang TL, Davidson B, Shih IeM, Zhang Y, Wood W 3rd, Becker KG and Morin PJ. MicroRNA expression and identification of putative miRNA targets in ovarian cancer. PLoS One 2008; 3: e2436.
[85] Lehmann U, Hasemeier B, Christgen M, Muller M, Romermann D, Langer F and Kreipe H. Epigenetic inactivation of microRNA gene hsa-mir-9-1 in human breast cancer. J Pathol 2008; 214: 17-24.
[86] Benakanakere MR, Zhao J, Finoti L, Schattner R, Odabas-Yigit M and Kinane DF. MicroRNA-663 antagonizes apoptosis antagonizing transcription factor to induce apoptosis in epithelial cells. Apoptosis 2019; 24: 108-118.
[87] Yang Y, Wang LL, Wang HX, Guo ZK, Gao XF, Cen J, Li YH, Dou LP and Yu L. The epigenetically-regulated miR-663 targets H-ras in K-562 cells. FEBS J 2013; 280: 5109-17.
[88] Yanokura M, Banno K, Adachi M, Aoki D and Abe K. Genome-wide DNA methylation sequencing reveals miR-663a is a novel epimutation candidate in CIMP-high endometrial cancer. Int J Oncol 2017; 50: 1934-1946.
[89] Carden T, Singh B, Mooga V, Bajpai P and Singh KK. Epigenetic modification of miR-663 controls mitochondria-to-nucleus retrograde signaling and tumor progression. J Biol Chem 2017; 292: 20694-20706.
[90] Hu H, Li S, Cui X, Lv X, Jiao Y, Yu F, Yao H, Song E, Chen Y, Wang M and Lin L. The overexpression of hypomethylated miR-663 induces chemotherapy resistance in human breast cancer cells by targeting heparin sulfate proteoglycan 2 (HSPG2). J Biol Chem 2013; 288: 10973-85.
[91] Zhang C, Chen B, Jiao A, Li F, Sun N, Zhang G and Zhang J. miR-663a inhibits tumor growth and invasion by regulating TGF-beta1 in hepatocellular carcinoma. BMC Cancer 2018; 18: 1179.
[92] Zhang Z, Ao P, Han H, Zhang Q, Chen Y, Han J, Huang Q, Huang H and Zhuo D. LncRNA PLAC2 upregulates miR-663 to downregulate TGF-beta1 and suppress bladder cancer cell migration and invasion. BMC Urol 2020; 20: 94.
[93] Wang L, Lang B, Zhou Y, Ma J and Hu K. Upregulation of miR-663a inhibits the cancer stem cell-like properties of glioma via repressing the KDM2A-mediated TGF-beta/SMAD signaling pathway. Cell Cycle 2021; 20: 1935-1952.
[94] Wang Z, Zhang H, Zhang P, Dong W and He L. MicroRNA-663 suppresses cell invasion and migration by targeting transforming growth factor beta 1 in papillary thyroid carcinoma. Tumour Biol 2016; 37: 7633-44.
[95] Zhang Y, Xu X, Zhang M, Wang X, Bai X, Li H, Kan L, Zhou Y, Niu H and He P. MicroRNA-663a is downregulated in non-small cell lung cancer and inhibits proliferation and invasion by targeting JunD. BMC Cancer 2016; 16: 315.

[96] Yu S, Xie H, Zhang J, Wang D, Song Y, Zhang S, Zheng S and Wang J. MicroRNA-663 suppresses the proliferation and invasion of colorectal cancer cells by directly targeting FSCN1. Mol Med Rep 2017; 16: 9707-9714.
[97] Tian W, Du Y, Ma Y, Zhang B, Gu L, Zhou J and Deng D. miR663a-TTC22V1 axis inhibits colon cancer metastasis. Oncol Rep 2019; 41: 1718-1728.
[98] Zhao S, Xiong W and Xu K. MiR-663a, regulated by lncRNA GAS5, contributes to osteosarcoma development through targeting MYL9. Hum Exp Toxicol 2020; 39: 1607-1618.
[99] Zhou L, Pan X, Li Z, Chen P, Quan J, Lin C, Lai Y, Xu J, Xu W, Guan X, Li H, Gui Y and Lai Y. Oncogenic miR-663a is associated with cellular function and poor prognosis in renal cell carcinoma. Biomed Pharmacother 2018; 105: 1155-1163.
[100] Jiang FN, Liang YX, Wei W, Zou CY, Chen GX, Wan YP, Liu ZZ, Yang Y, Han ZD, Zhu JG and Zhong WD. Functional classification of prostate cancer-associated miRNAs through CRISPR/Cas9-mediated gene knockout. Mol Med Rep 2020; 22: 3777-3784.
[101] Fiori ME, Villanova L, Barbini C, De Angelis ML and De Maria R. miR-663 sustains NSCLC by inhibiting mitochondrial outer membrane permeabilization (MOMP) through PUMA/BBC3 and BTG2. Cell Death Dis 2018; 9: 49.
[102] Li S, Lu X, Zheng D, Chen W, Li Y and Li F. Methyltransferase-like 3 facilitates lung cancer progression by accelerating m6A methylation-mediated primary miR-663 processing and impeding SOCS6 expression. J Cancer Res Clin Oncol 2022; 148: 3485-3499.
[103] Deng X, Hu Y, Ding Q, Han R, Guo Q, Qin J, Li J, Xiao R, Tian S, Hu W, Zhang Q and Xiong J. PEG10 plays a crucial role in human lung cancer proliferation, progression, prognosis and metastasis. Oncol Rep 2014; 32: 2159-67.
[104] Li X, Xiao R, Tembo K, Hao L, Xiong M, Pan S, Yang X, Yuan W, Xiong J and Zhang Q. PEG10 promotes human breast cancer cell proliferation, migration and invasion. Int J Oncol 2016; 48: 1933-42.
[105] Peng YP, Zhu Y, Yin LD, Zhang JJ, Wei JS, Liu X, Liu XC, Gao WT, Jiang KR and Miao Y. PEG10 overexpression induced by E2F-1 promotes cell proliferation, migration, and invasion in pancreatic cancer. J Exp Clin Cancer Res 2017; 36: 30.
[106] Shapovalova M, Lee JK, Li Y, Vander Griend DJ, Coleman IM, Nelson PS, Dehm SM and LeBeau AM. PEG10 promoter-driven expression of reporter genes enables molecular imaging of lethal prostate cancer. Cancer Res 2019; 79: 5668-5680.
[107] Kawai Y, Imada K, Akamatsu S, Zhang F, Seiler R, Hayashi T, Leong J, Beraldi E, Saxena N, Kretschmer A, Oo HZ, Contreras-Sanz A, Matsuyama H, Lin D, Fazli L, Collins CC, Wyatt AW, Black PC and Gleave ME. Paternally expressed gene 10 (PEG10) promotes growth, invasion, and survival of bladder cancer. Mol Cancer Ther 2020; 19: 2210-2220.
[108] Zhang L, Wan Y, Zhang Z, Jiang Y, Gu Z, Ma X, Nie S, Yang J, Lang J, Cheng W and Zhu L. IGF2BP1 overexpression stabilizes PEG10 mRNA in an m6A-dependent manner and promotes endometrial cancer progression. Theranostics 2021; 11: 1100-1114.
[109] Yu Y, Xue W, Liu Z, Chen S, Wang J, Peng Q, Xu L, Liu X, Cui C and Fan JB. A novel DNA methylation marker to identify lymph node metastasis of colorectal cancer. Front Oncol 2022; 12: 1000823.
[110] Watson KM, Gardner IH, Byrne RM, Ruhl RR, Lanciault CP, Dewey EN, Anand S and Tsikitis VL. Differential expression of PEG10 contributes to aggressive disease in early versus late-onset colorectal cancer. Dis Colon Rectum 2020; 63: 1610-1620.
[111] Breiman L. Manual for setting up, using, and understanding random forest V4.0. http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf.

Table S1. Summary of leukocyte heterogeneities in the WHI discovery dataset

Houseman's method					Horvath's method	
$CD4^+$ T	Natural killer cell	Monocyte	Granulocyte	Plasma blast	$CD8^+CD28^-CD45RA^-$ T	Naïve CD8 T
Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
22.40 (7.76)	1.00 (5.32)	8.15 (2.94)	52.03 (12.03)	1.80 (0.20)	11.13 (3.76)	196.95 (44.88)

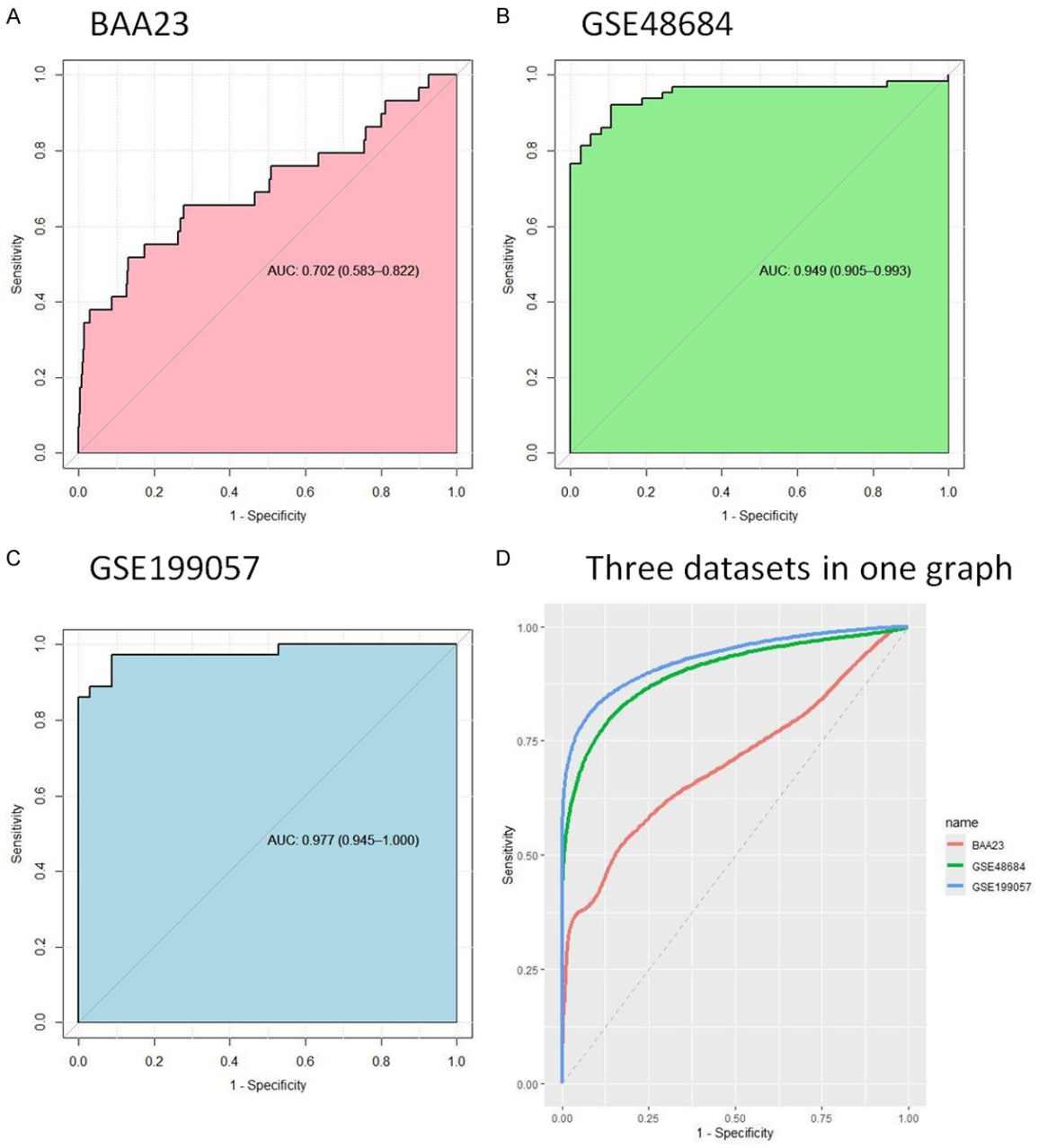
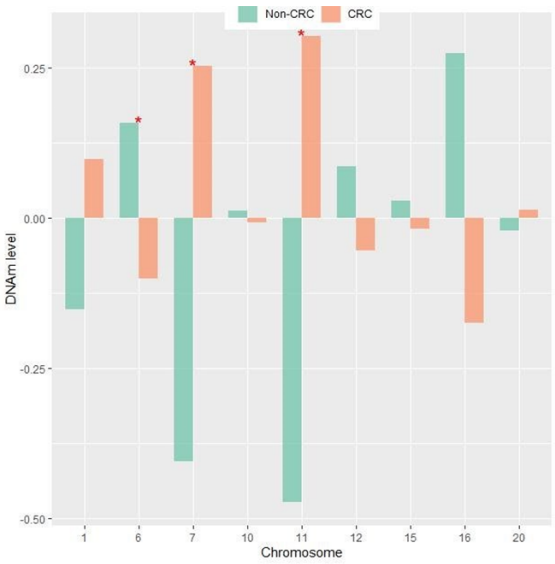


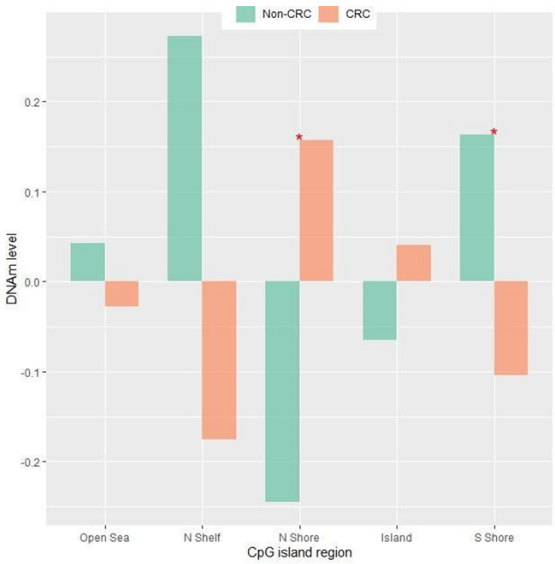
Figure S1. The area under the receiver operating characteristic curve (AUC) analysis with six CpGs (cg04958124, cg10321339, cg12704462, cg18144285, cg06007966, and cg17375901).

<GSE48684>

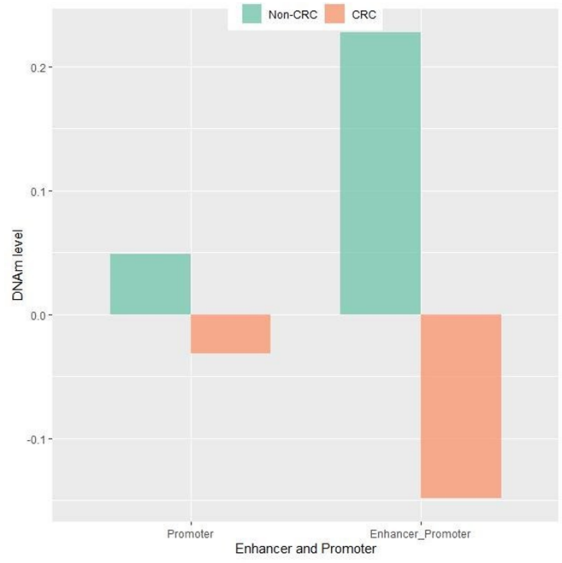
A By chromosome



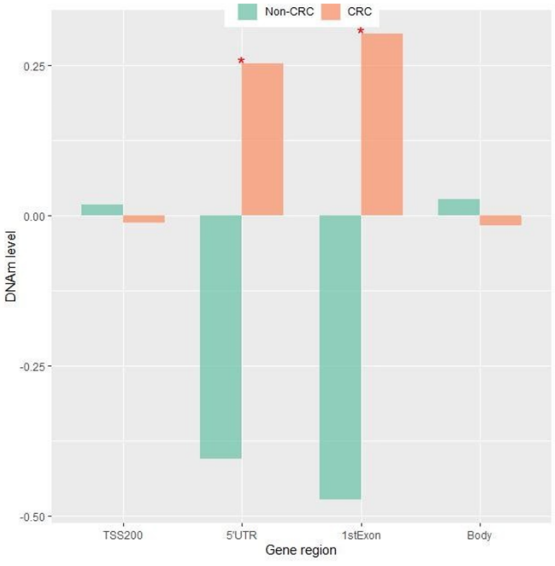
B By CpG context



C By enhancer and/or promoter



D By gene region



<GSE199057>

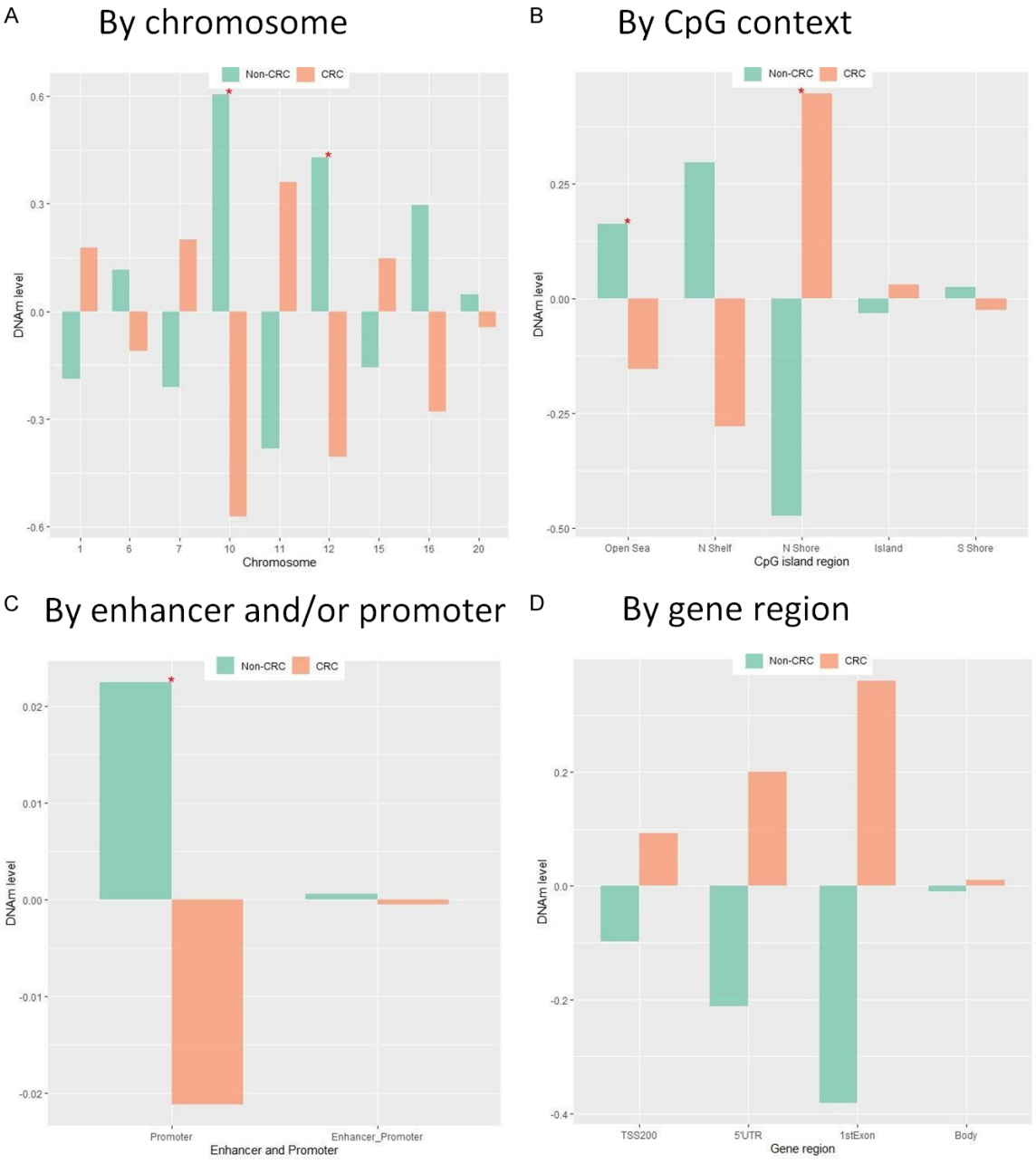


Figure S2. GSE datasets: Bar plots for mean difference in DNAm levels of top 20 genome-wide CpGs across chromosome, CpG context, enhancer and/or promoter, and gene region, stratified by CRC status. CpG, CpG dinucleotide; CRC, colorectal cancer; DNAm, DNA methylation; TSS200, 0-200 bp upstream of transcription start site; UTR, untranslated region. Note: *Statistical significance after multiple comparison correction.

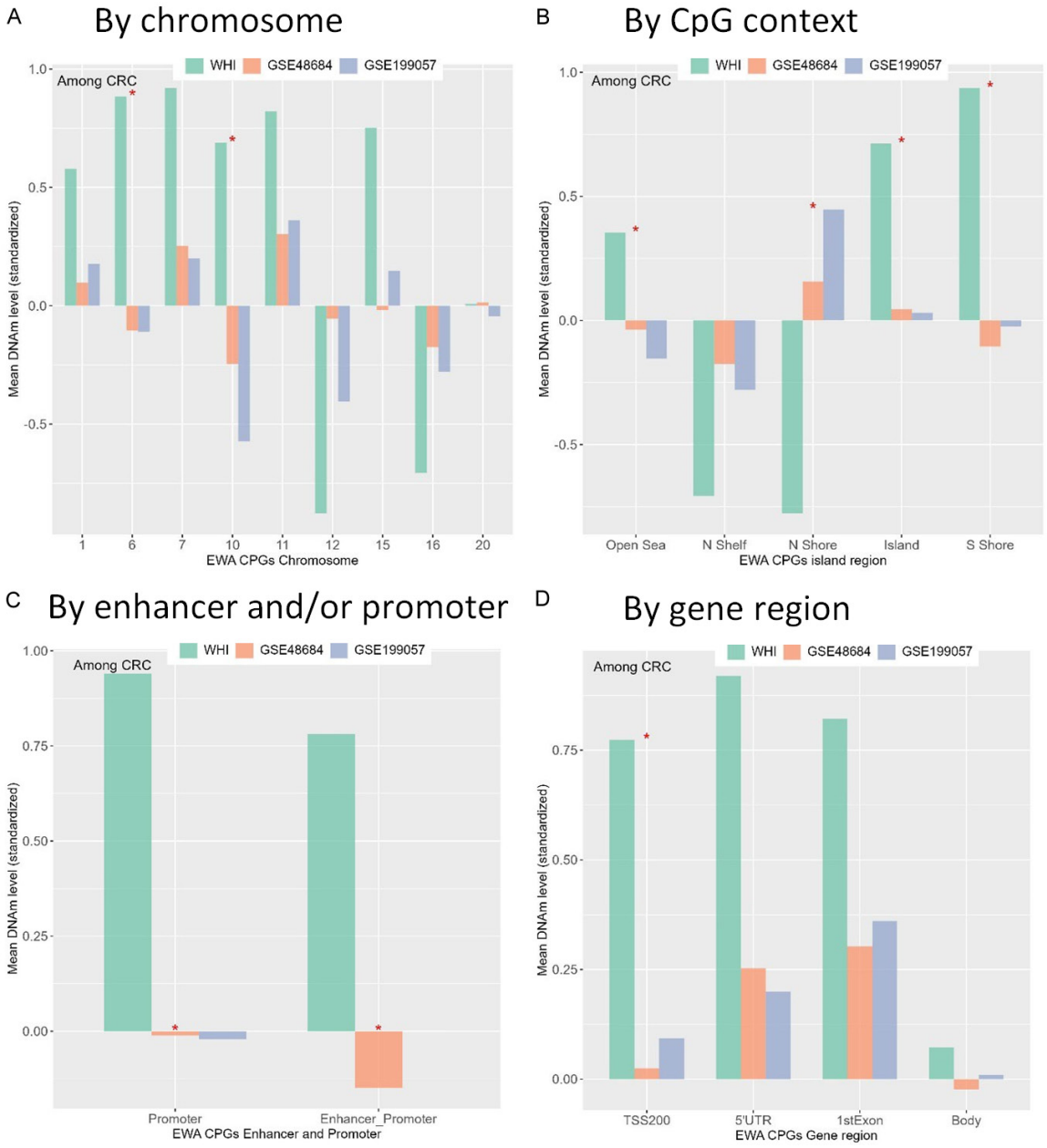
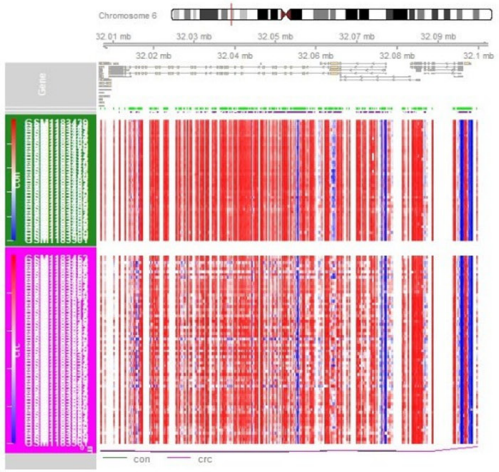
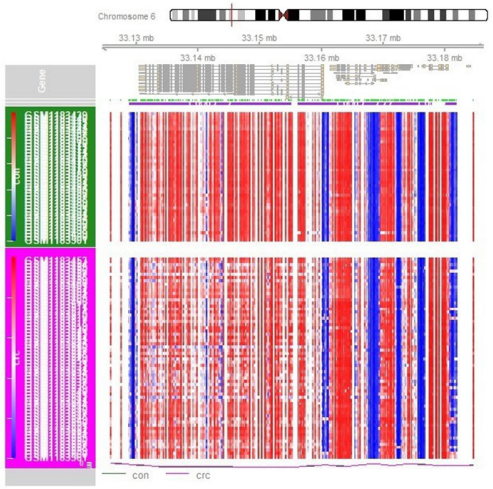


Figure S3. WHIBAA23, GSE48684, and GSE199057 in CRC patients (peripheral leukocytes for WHIBAA23 and CRC tissues for GSEs): Comparisons among the 3 studies for mean differences in DNAm levels of top 20 genome-wide CpGs across chromosome, CpG context, enhancer and/or promoter, and gene region. CpG, CpG dinucleotide; CRC, colorectal cancer; DNAm, DNA methylation; EWA, epigenome-wide association; TSS200, 0-200 bp upstream of transcription start site; UTR, untranslated region; WHI, Women's Health Initiative. Note: *Statistical significance after multiple comparison correction.

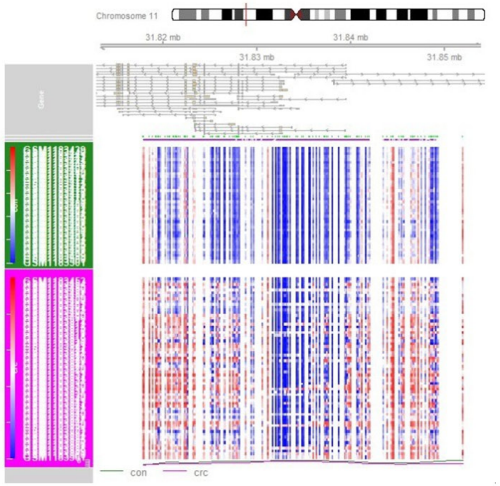
A1 DMR1 (Chr6: 32036449-32059605; TNXB, RNA5SP206)



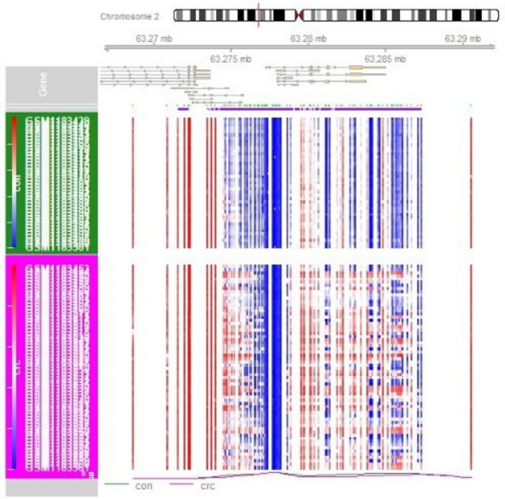
B1 DMR2 (Chr6: 33128825-33155135; COL11A2)



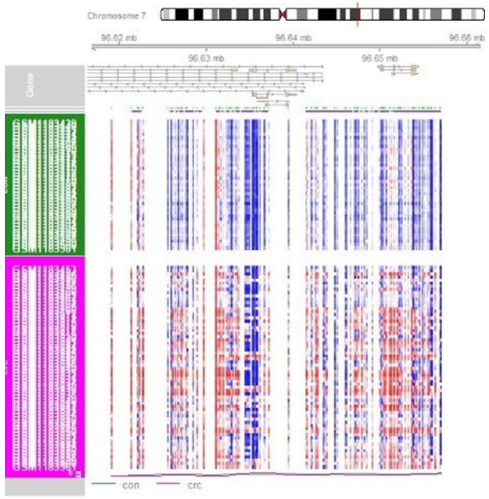
C1 DMR3 (Chr11: 31817810-31841980; RCN1, PAX6)



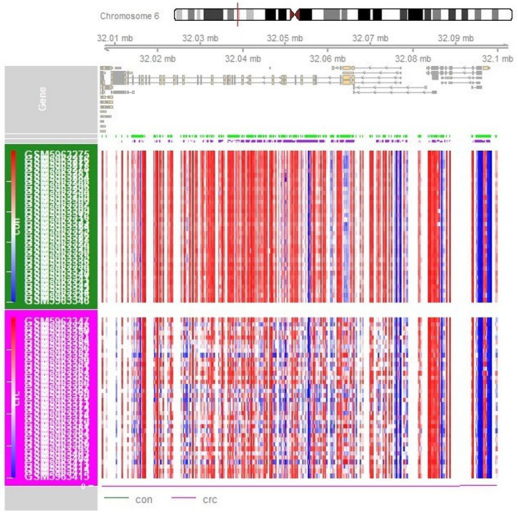
D1 DMR4 (chr2: 63273436-63287288; EHBP1, OTX1, AC009501.4)



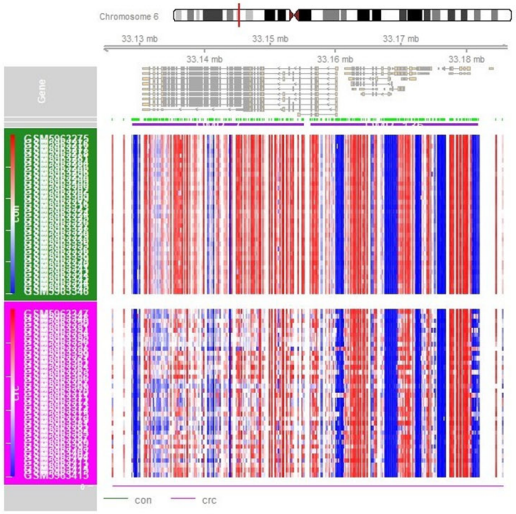
E1 DMR5 (chr7: 96641456-96657023; DLX6-AS1, DLX5)



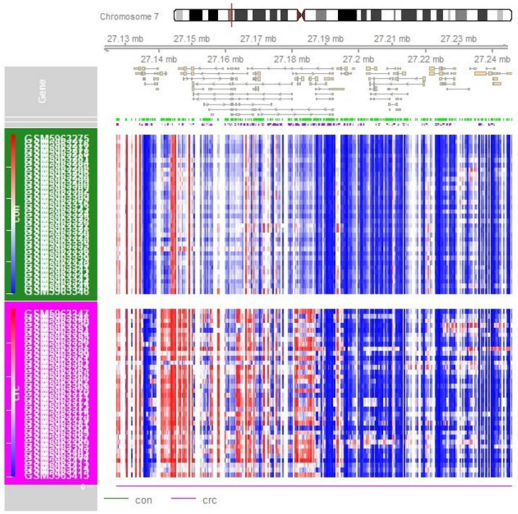
A2 DMR1 (chr6: 32036449-32059605; TNXB, RNA5SP206)



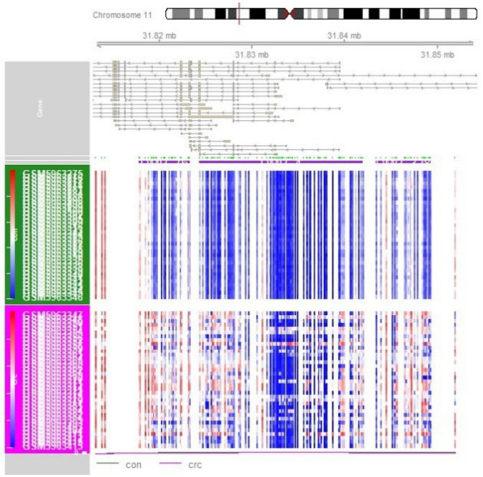
B2 DMR2 (chr6: 33128903-33155135; COL11A2)



C2 DMR3 (chr7: 27178861-27198374; HOXA-AS3, RP1-170O19.21, HOXA3, RP1-170O19.22, HOXA5, HOXA6, RP1-170O19.23, HOXA7)



D2 DMR4 (chr11: 31824327-31841980; RCN1, PAX6)



E2 DMR5 (chr2: 63273436-63287686; EHBP1, OTX1, AC009501.4)

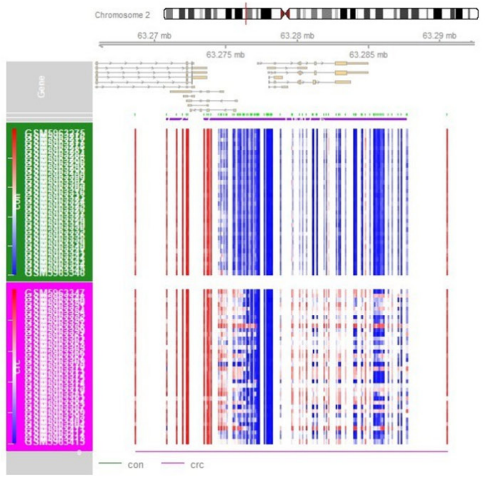


Figure S4. Top 5 differentially methylated regions (DMRs) in each GSE dataset. (A1-E1) for GSE48684 and (A2-E2) for GSE199057. A vertical bar on the chromosome schematic locates plotted region. The first track shows the CpG island context; the second track (yellow) shows the gene context i.e., the location of the DMR in the genome, the position of any genes that are nearby; the third track (light green) shows the base pair positions of the CpGs. Heatmap shows the methylation levels of the individual samples by colorectal cancer (CRC) status. The last smooth line graph shows the mean methylation levels for the samples grouped by CRC status.

Table S2. Among top 100 differentially methylated regions (DMRs) selected from each GSE dataset (GSE48684 and GSE199057), DMRs overlapping across the GSE datasets

Seqnames	Start	End	Width	Score
Chr1	25253237	25259034	5798	2
Chr1	50879560	50893984	14425	2
Chr1	119526060	119532925	6866	2
Chr1	217306700	217314284	7585	2
Chr1	221057236	221070193	12958	2
Chr2	45155201	45163188	7988	2
Chr2	63273436	63287288	13853	2
Chr2	119602212	119613877	11666	2
Chr2	223161771	223173061	11291	2
Chr3	62353312	62365402	12091	2
Chr3	128203414	128212476	9063	2
Chr3	137481938	137491164	9227	2
Chr3	147121892	147132559	10668	2
Chr4	4854459	4864902	10444	2
Chr4	96468962	96471143	2182	2
Chr4	154709441	154714852	5412	2
Chr4	174447847	174453287	5441	2
Chr5	1882188	1888033	5846	2
Chr5	37833969	37840839	6871	2
Chr5	134361983	134367394	5412	2
Chr5	170734312	170740937	6626	2
Chr6	29520527	29521803	1277	2
Chr6	30078080	30080782	2703	2
Chr6	32036449	32059605	23157	2
Chr6	32060681	32066582	5902	2
Chr6	32184296	32193235	8940	2
Chr6	33128903	33155135	26233	2
Chr6	84417445	84419360	1916	2
Chr6	108484512	108492769	8258	2
Chr6	133561224	133564578	3355	2
Chr6	152125861	152130332	4472	2
Chr7	1265197	1281585	16389	2
Chr7	19155785	19158954	3170	2
Chr7	27140797	27144854	4058	2
Chr7	27180888	27185512	4625	2
Chr7	49812836	49815938	3103	2
Chr7	94284258	94287242	2985	2
Chr7	96645989	96657023	11035	2
Chr7	130129946	130133110	3165	2
Chr8	25897201	25909599	12399	2
Chr8	69241923	69244553	2631	2
Chr8	70980488	70984917	4430	2
Chr8	72753268	72758701	5434	2
Chr8	97169621	97174382	4762	2
Chr8	145103393	145107857	4465	2
Chr10	7450112	7455714	5603	2

Chr10	118030848	118034357	3510	2
Chr10	131756487	131772187	15701	2
Chr11	2158555	2165961	7407	2
Chr11	31817810	31823282	5473	2
Chr11	31824327	31841980	17654	2
Chr11	32447944	32452839	4896	2
Chr11	32454216	32461240	7025	2
Chr11	44324759	44333192	8434	2
Chr11	128553855	128566958	13104	2
Chr12	5018229	5021871	3643	2
Chr12	85304514	85307424	2911	2
Chr12	114840854	114847641	6788	2
Chr13	28491326	28499045	7720	2
Chr13	28500882	28503508	2627	2
Chr13	112707805	112717707	9903	2
Chr13	112758379	112763220	4842	2
Chr14	95233665	95240560	6896	2
Chr15	74419428	74429109	9682	2
Chr15	83951663	83954849	3187	2
Chr16	51183363	51190201	6839	2
Chr16	54965084	54974287	9204	2
Chr18	74960629	74963807	3179	2
Chr20	25061762	25065553	3792	2
Chr20	57424521	57431303	6783	2
Chr20	61806628	61810902	4275	2
Chr21	38076709	38083586	6878	2

The score of 2 indicates that the 2 datasets have overlapping DMRs. Chr, chromosome.

Table S3. Effect size of 10 CpGs in the differentially methylated region (Chr7) which overlaps across the WHIBAA23 and GSE datasets

WHIBAA23: All HRs were adjusted by leukocyte heterogeneities plus DNA methylation-predicted age.									
CpG	age.HR*	age.SE	age.*P*	bmi.HR**	bmi.SE	bmi.*P*	DM.IR.HR¶	DM.IR.SE	DM.IR.*P*
cg24885794	1.42	0.193	0.067	1.41	0.193	0.072	1.45	0.196	0.058
cg26997085	1.44	0.152	0.017	1.42	0.154	0.023	1.45	0.158	0.019
cg22331138	1.25	0.187	0.240	1.25	0.189	0.236	1.33	0.193	0.134
cg16492735	1.07	0.202	0.731	1.07	0.202	0.738	1.07	0.210	0.750
cg09512080	1.21	0.183	0.305	1.20	0.183	0.324	1.21	0.191	0.326
cg00906934	1.50	0.187	0.031	1.49	0.187	0.032	1.45	0.197	0.060
cg26503018	1.25	0.199	0.266	1.25	0.198	0.269	1.20	0.203	0.359
cg27120649	1.26	0.208	0.274	1.27	0.208	0.251	1.29	0.216	0.241
cg21771834	1.11	0.218	0.625	1.11	0.219	0.627	1.19	0.219	0.438
cg27001184	1.17	0.188	0.402	1.17	0.188	0.412	1.18	0.195	0.405
GSE48684: All ORs were adjusted by sex.									
CpG	OR§	SE	*P*	OR¥	SE	*P*	OR£	SE	*P*
cg24885794	3.63	0.342	0.0002	3.63	0.342	0.0002	3.63	0.342	0.0002
cg26997085	2.91	0.320	0.001	2.91	0.320	0.001	2.91	0.320	0.001
cg22331138	1.97	0.260	0.009	1.97	0.260	0.009	1.97	0.260	0.009
cg16492735	2.75	0.313	0.001	2.75	0.313	0.001	2.75	0.313	0.001
cg09512080	2.43	0.294	0.002	2.43	0.294	0.002	2.43	0.294	0.002
cg00906934	2.50	0.378	0.016	2.50	0.378	0.016	2.50	0.378	0.016
cg26503018	1.63	0.233	0.036	1.63	0.233	0.036	1.63	0.233	0.036
cg27120649	2.61	0.340	0.005	2.61	0.340	0.005	2.61	0.340	0.005
cg21771834	1.29	0.218	0.244	1.29	0.218	0.244	1.29	0.218	0.244
cg27001184	1.75	0.243	0.022	1.75	0.243	0.022	1.75	0.243	0.022
GSE199057: All ORs were adjusted by sex plus age and DNA methylation-predicted age.									
CpG	OR§	SE	*P*	OR¥	SE	*P*	OR£	SE	*P*
cg24885794	2.99	0.445	0.014	5.73	0.538	0.001	5.08	0.455	0.0004
cg26997085	3.79	0.511	0.009	3.39	0.515	0.018	3.56	0.410	0.002
cg22331138	3.40	0.465	0.008	3.58	0.526	0.015	3.66	0.419	0.002
cg16492735	4.83	0.532	0.003	6.18	0.555	0.001	4.19	0.426	0.001
cg09512080	2.75	0.420	0.016	3.67	0.535	0.015	2.80	0.395	0.009
cg00906934	2.14	0.347	0.028	4.54	0.530	0.004	3.20	0.406	0.004
cg26503018	4.61	0.556	0.006	11.29	0.597	0.00005	5.40	0.463	0.0003
cg27120649	6.69	0.598	0.001	7.27	0.543	0.0003	4.22	0.429	0.001
cg21771834	5.65	0.620	0.005	4.47	0.529	0.005	5.72	0.484	0.0003
cg27001184	1.91	0.319	0.043	2.03	0.492	0.151	2.38	0.377	0.022

BMI, body mass index; Chr, chromosome; CpG, CpG dinucleotide; DM, ever having been treated for diabetes; HR, hazard ratio; IR, insulin resistance; OR, odds ratio; SE, standard error; WHI, Women's Health Initiative. *Age adjusted; **BMI and age adjusted; ¶DM, IR, BMI and age adjusted; §CpG as continuous variable; ¥CpG as categorical variable (binary using a median); £CpG as categorical variable (ternary using 1st and 3rd quartiles).

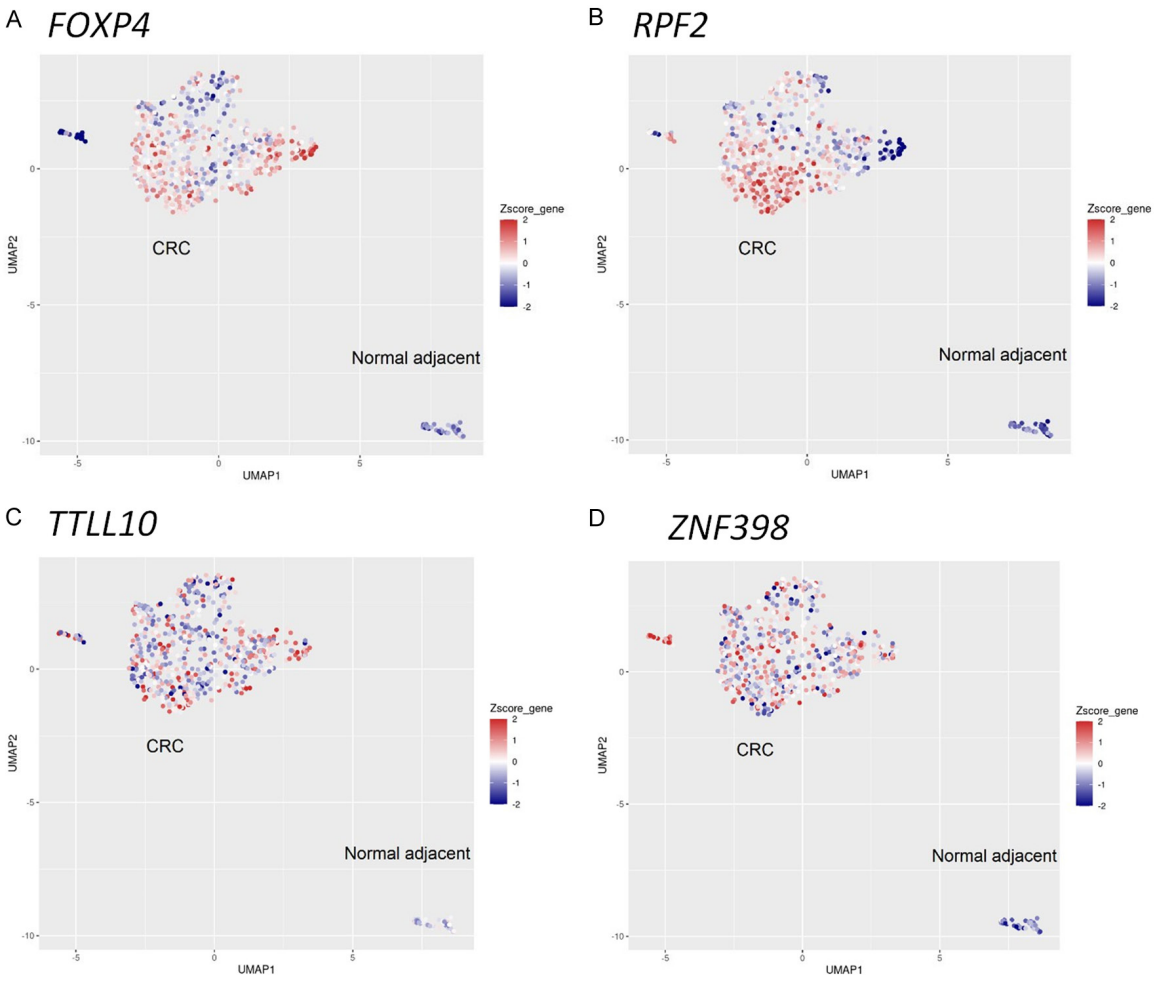


Figure S5. UMAP plots of mRNA-sequences mapping to modeled genes in TCGA COAD and READ datasets. CRC, colorectal cancer; UMAP, Uniform Manifold Approximation and Projection.

Table S4. Over-representation analysis (reference gene set: Agilent wholegenome.4x44.v2)

	BAA23*			GSE48684¶			GSE199057¥		
	Description	Enrichment Ratio	Key Driver	Description	Enrichment Ratio	Key Driver	Description	Enrichment Ratio	Key Driver
GO - biological process	Mitochondrial genome maintenance	168.27	PIF1	Mitochondrial genome maintenance	252.41	PIF1	Signal transduction by p53 class mediator	33.34	RPF2
	Telomere organization	33.90	PIF1	Telomere organization	50.85	PIF1	rRNA metabolic process	32.27	RPF2
	DNA biosynthetic process	26.77	PIF1	DNA biosynthetic process	40.16	PIF1	Ribonucleoprotein complex subunit organization	30.33	RPF2
	Signal transduction by p53 class mediator	22.22	RPF2	Signal transduction by p53 class mediator	33.34	RPF2	Protein localization to nucleus	27.72	RPF2
	rRNA metabolic process	21.51	RPF2	rRNA metabolic process	32.27	RPF2	Neurotransmitter transport	26.98	SV2B
	Ribonucleoprotein complex subunit organization	20.22	RPF2	Ribonucleoprotein complex subunit organization	30.33	RPF2	ncRNA processing	20.73	RPF2
	DNA conformation change	20.05	PIF1	DNA conformation change	30.07	PIF1	Ribonucleoprotein complex biogenesis	17.11	RPF2
	Protein localization to nucleus	18.48	RPF2	Protein localization to nucleus	27.72	RPF2			
	DNA replication	18.33	PIF1	DNA replication	27.50	PIF1			
	Negative regulation of transferase activity	18.19	PIF1	Negative regulation of transferase activity	27.29	PIF1			
GO - cellular component	Replication fork	47.27	PIF1	N/A			Golgi stack	37.38	B4GALNT4
	Golgi stack	24.92	B4GALNT4				Sperm part	26.21	SV2B
	Sperm part	17.47	SV2B				Transport vesicle	13.28	SV2B
	Chromosomal region	10.08	PIF1				Presynapse	10.35	SV2B
	Transport vesicle	8.86	SV2B						
	Presynapse	6.90	SV2B						
GO - molecular function	Catalytic activity, acting on a glycoprotein	170.38	B4GALNT4	Telomeric DNA binding rRNA binding	163.28 91.84	PIF1 RPF2	Catalytic activity, acting on a glycoprotein	255.57	B4GALNT4
	Telomeric DNA binding	108.85	PIF1	Helicase activity	40.82	PIF1	rRNA binding	91.84	RPF2
	rRNA binding	61.23	RPF2	Catalytic activity, acting on DNA	33.21	PIF1	Transferase activity, transferring glycosyl groups	22.27	B4GALNT4
	Helicase activity	27.21	PIF1	Magnesium ion binding	28.96	PIF1			
	Catalytic activity, acting on DNA	22.14	PIF1	Catalytic activity, acting on RNA	17.60	PIF1			
	Magnesium ion binding	19.30	PIF1	Enzyme inhibitor activity	17.09	PIF1			
	Transferase activity, transferring glycosyl groups	14.84	B4GALNT4	ATPase activity	13.80	PIF1			
	Catalytic activity, acting on RNA	11.73	PIF1						
	Enzyme inhibitor activity	11.39	PIF1						
	ATPase activity	9.20	PIF1						
Pathway - KEGG	ECM-receptor interaction	44.33	SV2B	N/A			ECM-receptor interaction	44.33	SV2B
	Metabolic pathways	2.77	B4GALNT4				Metabolic pathways	2.77	B4GALNT4

Pathway - Reactome§	Toxicity of botulinum toxin type D (bont/d)	1006.20	SV2B	N/A	Toxicity of botulinum toxin type D (BoNT/D)	1055.40	SV2B
	Toxicity of botulinum toxin type F (BoNT/F)	1006.20	SV2B		Toxicity of botulinum toxin type F (BoNT/F)	1055.40	SV2B
	Neurotoxicity of Clostridium toxins	503.10	SV2B		Neurotoxicity of Clostridium toxins	527.70	SV2B
	Uptake and actions of bacterial toxins	162.29	SV2B		Uptake and actions of bacterial toxins	170.23	SV2B
	Infectious disease	13.82	SV2B		Carboxyterminal post-translational modifications of tubulin	146.58	TTLL10
	Disease	5.00	SV2B				
	Generic transcription pathway	4.50	ZNF398		Infectious disease	13.81	SV2B
	RNA polymerase II transcription	4.07	ZNF398		Disease	5.01	SV2B
	Gene expression (transcription)	3.68	ZNF398		Post-translational protein modification	3.70	TTLL10
Disease - Disgenet	Kidney diseases	35.08	SV2B	N/A		N/A	
	Prostatic neoplasms	7.99	FOXP4				
Disease - GLAD4U	Cardiac output, low	227.16	FOXP4	N/A		N/A	
	Lymphocytic choriomeningitis	149.07	FOXP4				
	Epilepsy, temporal lobe	71.20	SV2B				
	Epilepsy	19.47	SV2B				
	Williams syndrome	18.14	ZNF398				
	Seizures	17.80	SV2B				
	Prostatic neoplasms	16.17	FOXP4				
	Nelson syndrome	10.37	ZNF398				

N/A, not available. *BAA23 genes: 20 top CpGs at the genome-wide significance. ¶GSE48684 genes: among top 20 CpGs, only CpGs significant at the validation level. ¥GSE199057 genes: among top 20 CpGs, only CpGs significant at the validation level. §GSE199057 of Pathway - Reactome: from Genome as Reference Gene Set.