Original Article

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

Su Yon Jung1,2,3, Herbert Yu4, Xianglong Tan5, Matteo Pellegrini6

1Translational Sciences Section, School of Nursing, University of California, Los Angeles, CA 90095, USA; 2Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA; 3Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA; 4Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA; 5Department of Biological Chemistry, University of California, Los Angeles, CA 90095, USA; 6Department 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 (B4GALNT4cg10321339, SV2Bcg18144285) had the strongest effect on CRC risk. Results from our DMR scans contained MIR663cg06007966, 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,
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 μU/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: μU/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 endpoint 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, \(CD8^+\)CD28CD45RA- 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 CpGs, 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 \(C\) (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 (B4GALNT4cg10321339 and SV2Bcg18144285) 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 (MIR663cg06007966 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

<table>
<thead>
<tr>
<th>Chr</th>
<th>CpG site§</th>
<th>Position</th>
<th>Age adjusted</th>
<th>BMI &amp; age adjusted</th>
<th>DM, IR, BMI &amp; age adjusted</th>
<th>CpG context</th>
<th>Gene</th>
<th>Gene region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>HR (95% CI)</td>
<td>P</td>
<td>HR (95% CI)</td>
<td>P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>cg14057946¥</td>
<td>713985</td>
<td>1.43 (1.25, 1.64)</td>
<td>&lt; 1E-007</td>
<td>1.43 (1.25, 1.64)</td>
<td>&lt; 1E-007</td>
<td>1.42 (1.23, 1.64)</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>chr1</td>
<td>cg04231937¥</td>
<td>714526</td>
<td>1.56 (1.33, 1.82)</td>
<td>&lt; 1E-007</td>
<td>1.56 (1.33, 1.82)</td>
<td>&lt; 1E-007</td>
<td>1.56 (1.33, 1.83)</td>
<td>&lt; 1E-007</td>
</tr>
<tr>
<td>chr1</td>
<td>cg04496485¥</td>
<td>714565</td>
<td>1.39 (1.24, 1.57)</td>
<td>&lt; 1E-007</td>
<td>1.39 (1.23, 1.56)</td>
<td>&lt; 1E-007</td>
<td>1.38 (1.23, 1.56)</td>
<td>&lt; 1E-007</td>
</tr>
<tr>
<td>chr1</td>
<td>cg02014020</td>
<td>1115461</td>
<td>0.67 (0.57, 0.79)</td>
<td>1.00E-06</td>
<td>0.67 (0.58, 0.79)</td>
<td>1.00E-06</td>
<td>0.67 (0.57, 0.79)</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>chr6</td>
<td>cg07572131*,¥</td>
<td>31430791</td>
<td>1.65 (1.37, 1.98)</td>
<td>&lt; 1E-007</td>
<td>1.64 (1.35, 1.98)</td>
<td>&lt; 1E-007</td>
<td>1.58 (1.29, 1.94)</td>
<td>9.00E-06</td>
</tr>
<tr>
<td>chr6</td>
<td>cg14498116¥</td>
<td>148823862</td>
<td>1.42 (1.25, 1.61)</td>
<td>&lt; 1E-007</td>
<td>1.43 (1.26, 1.62)</td>
<td>&lt; 1E-007</td>
<td>1.43 (1.26, 1.63)</td>
<td>&lt; 1E-007</td>
</tr>
<tr>
<td>chr7</td>
<td>cg04958124¥</td>
<td>111303252</td>
<td>1.50 (1.27, 1.77)</td>
<td>2.00E-06</td>
<td>1.49 (1.26, 1.77)</td>
<td>3.00E-06</td>
<td>1.47 (1.24, 1.75)</td>
<td>1.30E-05</td>
</tr>
<tr>
<td>chr10</td>
<td>cg18072629</td>
<td>8092036</td>
<td>1.30 (1.17, 1.44)</td>
<td>1.00E-06</td>
<td>1.30 (1.17, 1.44)</td>
<td>1.00E-06</td>
<td>1.30 (1.17, 1.44)</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>chr11</td>
<td>cg10321339</td>
<td>369810</td>
<td>1.54 (1.30, 1.83)</td>
<td>1.00E-06</td>
<td>1.55 (1.30, 1.84)</td>
<td>1.00E-06</td>
<td>1.54 (1.29, 1.84)</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>chr12</td>
<td>cg12704462</td>
<td>120151527</td>
<td>1.47 (1.26, 1.72)</td>
<td>1.00E-06</td>
<td>1.48 (1.27, 1.72)</td>
<td>1.00E-06</td>
<td>1.46 (1.24, 1.72)</td>
<td>5.00E-06</td>
</tr>
<tr>
<td>chr15</td>
<td>cg11823654¥</td>
<td>91643026</td>
<td>1.47 (1.26, 1.72)</td>
<td>1.00E-06</td>
<td>1.48 (1.27, 1.72)</td>
<td>1.00E-06</td>
<td>1.46 (1.24, 1.72)</td>
<td>5.00E-06</td>
</tr>
<tr>
<td>chr16</td>
<td>cg04872027</td>
<td>29149757</td>
<td>0.73 (0.64, 0.83)</td>
<td>1.00E-06</td>
<td>0.72 (0.64, 0.82)</td>
<td>1.00E-06</td>
<td>0.73 (0.64, 0.83)</td>
<td>3.00E-06</td>
</tr>
<tr>
<td>chr20</td>
<td>cg06007966</td>
<td>26188971</td>
<td>1.94 (1.49, 2.52)</td>
<td>1.00E-06</td>
<td>1.94 (1.49, 2.53)</td>
<td>1.00E-06</td>
<td>1.95 (1.48, 2.56)</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>chr20</td>
<td>cg17379901</td>
<td>61754940</td>
<td>0.62 (0.51, 0.74)</td>
<td>&lt; 1E-007</td>
<td>0.61 (0.51, 0.74)</td>
<td>&lt; 1E-007</td>
<td>0.62 (0.51, 0.75)</td>
<td>1.00E-06</td>
</tr>
</tbody>
</table>

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+CD28CD45RA T cell, naive 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

<table>
<thead>
<tr>
<th>Chr</th>
<th>CpG site§</th>
<th>Position</th>
<th>GSE 48684</th>
<th>GSE199057</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR¶ (95% CI)</td>
<td>P</td>
</tr>
<tr>
<td>chr1</td>
<td>cg14057946¥</td>
<td>713985</td>
<td>1.46 (0.89, 2.87)</td>
<td>0.211</td>
</tr>
<tr>
<td>chr1</td>
<td>cg04231937¥</td>
<td>714526</td>
<td>1.06 (0.71, 1.68)</td>
<td>0.768</td>
</tr>
<tr>
<td>chr1</td>
<td>cg04496485¥</td>
<td>714565</td>
<td>1.30 (0.82, 2.48)</td>
<td>0.344</td>
</tr>
<tr>
<td>chr1</td>
<td>cg02014020</td>
<td>1115461</td>
<td>1.47 (0.98, 2.30)</td>
<td>0.076</td>
</tr>
<tr>
<td>chr6</td>
<td>cg07572131*¥</td>
<td>31430791</td>
<td>0.66 (0.42, 1.00)</td>
<td>0.057</td>
</tr>
<tr>
<td>chr6</td>
<td>cg25410010</td>
<td>41554543</td>
<td>1.06 (0.69, 1.57)</td>
<td>0.793</td>
</tr>
<tr>
<td>chr6</td>
<td>cg06498809¥</td>
<td>111303174</td>
<td>0.83 (0.55, 1.24)</td>
<td>0.362</td>
</tr>
<tr>
<td>chr6</td>
<td>cg00020352¥</td>
<td>111303252</td>
<td>0.61 (0.39, 0.93)</td>
<td>0.025</td>
</tr>
<tr>
<td>chr6</td>
<td>cg10924212¥</td>
<td>111303363</td>
<td>1.40 (0.92, 2.20)</td>
<td>0.126</td>
</tr>
<tr>
<td>chr6</td>
<td>cg14498116¥</td>
<td>111303482</td>
<td>0.32 (0.15, 0.60)</td>
<td>0.001</td>
</tr>
<tr>
<td>chr7</td>
<td>cg04958124¥</td>
<td>148823862</td>
<td>2.37 (1.43, 4.37)</td>
<td>0.002</td>
</tr>
<tr>
<td>chr10</td>
<td>cg18072629</td>
<td>8092036</td>
<td>3.00 (1.42, 10.25)</td>
<td>0.025</td>
</tr>
<tr>
<td>chr10</td>
<td>cg05970116</td>
<td>75351076</td>
<td>0.48 (0.29, 0.75)</td>
<td>0.002</td>
</tr>
<tr>
<td>chr11</td>
<td>cg10321339</td>
<td>369810</td>
<td>6.73 (2.29, 40.88)</td>
<td>0.008</td>
</tr>
<tr>
<td>chr12</td>
<td>cg12704462</td>
<td>120151527</td>
<td>0.84 (0.47, 1.28)</td>
<td>0.477</td>
</tr>
<tr>
<td>chr15</td>
<td>cg11823654¥</td>
<td>65117936</td>
<td>0.39 (0.23, 0.63)</td>
<td>2.45E-04</td>
</tr>
<tr>
<td>chr15</td>
<td>cg18144285</td>
<td>91643026</td>
<td>199.31 (5.83, 69923.50)</td>
<td>0.023</td>
</tr>
<tr>
<td>chr16</td>
<td>cg04872027</td>
<td>29149757</td>
<td>0.50 (0.25, 0.86)</td>
<td>0.025</td>
</tr>
<tr>
<td>chr20</td>
<td>cg06007966</td>
<td>26188971</td>
<td>4.20 (2.34, 8.59)</td>
<td>1.20E-05</td>
</tr>
<tr>
<td>chr20</td>
<td>cg17375901</td>
<td>61754940</td>
<td>0.07 (0.02, 0.22)</td>
<td>6.70E-05</td>
</tr>
</tbody>
</table>

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.
DNAm in CRC from PBLs

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 (MIR663cg06007966 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, PIF1cg11823654, RPF2cg14498116, and ZNF398cg04958124 in the GSE48984, and TTL10cg02014020, SV2Bcg18144285, B4GALNT4cg10321339, RPF2cg06498809, MIR1178cg12704462, 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

<table>
<thead>
<tr>
<th># of DMR</th>
<th>Seqnames</th>
<th>Start</th>
<th>End</th>
<th>Width</th>
<th>Fisher</th>
<th>DMR: No. of CpGs</th>
<th>Overlapping CpGs†</th>
<th>DMR: genes</th>
<th>Overlapping genes¥</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;GSE48684&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>chr10</td>
<td>8088801</td>
<td>8103673</td>
<td>14873</td>
<td>4.18E-215</td>
<td>90</td>
<td>cg18072629</td>
<td>RP11-379F12.4, GATA3, GATA3-AS1, RP11-379F12.3</td>
<td></td>
</tr>
<tr>
<td>2782</td>
<td>chr11</td>
<td>368351</td>
<td>369875</td>
<td>1525</td>
<td>2.56E-30</td>
<td>27</td>
<td>cg10321339</td>
<td>B4GALNT4</td>
<td></td>
</tr>
<tr>
<td>2910</td>
<td>chr15</td>
<td>91642470</td>
<td>91643742</td>
<td>1273</td>
<td>2.58E-29</td>
<td>15</td>
<td>cg18144285</td>
<td>SV2B</td>
<td></td>
</tr>
<tr>
<td>2924</td>
<td>chr15</td>
<td>65115218</td>
<td>65119016</td>
<td>3799</td>
<td>3.34E-29</td>
<td>13</td>
<td>cg11823654§</td>
<td>PIF1§</td>
<td></td>
</tr>
<tr>
<td>2931</td>
<td>chr20</td>
<td>26188639</td>
<td>26189240</td>
<td>602</td>
<td>3.72E-29</td>
<td>7</td>
<td>cg06007966§,</td>
<td>MIR663§*,</td>
<td></td>
</tr>
<tr>
<td>6682</td>
<td>chr1</td>
<td>1113501</td>
<td>1115920</td>
<td>2420</td>
<td>8.03E-13</td>
<td>16</td>
<td>cg02014020</td>
<td>TTL10, TTL10-AS1</td>
<td></td>
</tr>
<tr>
<td>8302</td>
<td>chr6</td>
<td>111302792</td>
<td>111303792</td>
<td>1064</td>
<td>9.06E-10</td>
<td>13</td>
<td>cg00020352,</td>
<td>RPF2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg06498809,</td>
<td>MIR663§*,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg10920427,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg14498116§</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ZNF398, ZNF425,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RN7SL521P</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ZNF398§</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17740</td>
<td>chr7</td>
<td>148822673</td>
<td>148823965</td>
<td>1293</td>
<td>0.013</td>
<td>13</td>
<td>cg04958124§</td>
<td>ZNF398, ZNF425,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RN7SL521P</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ZNF398§</td>
<td></td>
</tr>
<tr>
<td>20166</td>
<td>chr12</td>
<td>120151527</td>
<td>120152127</td>
<td>601</td>
<td>0.090</td>
<td>7</td>
<td>cg12704462</td>
<td>CIT, MIR1178</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MIR1178</td>
<td></td>
</tr>
<tr>
<td>&lt;GSE199057&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1979</td>
<td>chr20</td>
<td>26188639</td>
<td>26190354</td>
<td>1716</td>
<td>4.78E-56</td>
<td>10</td>
<td>cg06007966§,</td>
<td>MIR663§*,</td>
<td></td>
</tr>
<tr>
<td>4442</td>
<td>chr1</td>
<td>1113624</td>
<td>1115920</td>
<td>2297</td>
<td>8.25E-36</td>
<td>17</td>
<td>cg02014020§</td>
<td>TTL10, TTL10-AS1</td>
<td></td>
</tr>
<tr>
<td>4640</td>
<td>chr15</td>
<td>91641719</td>
<td>91643742</td>
<td>2024</td>
<td>4.98E-35</td>
<td>18</td>
<td>cg18144285§</td>
<td>SV2B§</td>
<td></td>
</tr>
<tr>
<td>5895</td>
<td>chr15</td>
<td>65116255</td>
<td>65119016</td>
<td>2762</td>
<td>1.46E-30</td>
<td>16</td>
<td>cg11823654</td>
<td>PIF1§</td>
<td></td>
</tr>
<tr>
<td>13206</td>
<td>chr11</td>
<td>368351</td>
<td>369875</td>
<td>1525</td>
<td>1.88E-16</td>
<td>25</td>
<td>cg10321339§</td>
<td>B4GALNT4§</td>
<td></td>
</tr>
<tr>
<td>14389</td>
<td>chr6</td>
<td>111301798</td>
<td>111303792</td>
<td>1995</td>
<td>4.06E-15</td>
<td>17</td>
<td>cg06498809§,</td>
<td>RPF2§</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg10920427,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg14498116§</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18934</td>
<td>chr12</td>
<td>120151527</td>
<td>120152127</td>
<td>601</td>
<td>5.19E-11</td>
<td>5</td>
<td>cg12704462§</td>
<td>CIT, MIR1178</td>
<td></td>
</tr>
<tr>
<td>21755</td>
<td>chr10</td>
<td>75351076</td>
<td>75351888</td>
<td>813</td>
<td>5.84E-09</td>
<td>5</td>
<td>cg05970116§</td>
<td>USP54</td>
<td></td>
</tr>
<tr>
<td>33351</td>
<td>chr7</td>
<td>148822673</td>
<td>148823862</td>
<td>1190</td>
<td>0.079</td>
<td>13</td>
<td>cg04958124</td>
<td>ZNF398, ZNF425,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RN7SL521P</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ZNF398§</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Seqnames</th>
<th>Start</th>
<th>End</th>
<th>Width</th>
<th>Score</th>
<th>No. of CpGs</th>
<th>Cpgs§,¥ Position</th>
<th>Cpg context</th>
<th>Overlapping Gene</th>
<th>Gene region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr7</td>
<td>94286086</td>
<td>94286267</td>
<td>182</td>
<td>3</td>
<td>10</td>
<td>cg24885794 94286086</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg26997085 94286110</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg22331138 94286131</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg16492735 94286208</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg09512080 94286219</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg00906934 94286232</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg26503018 94286243</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg27120649 94286261</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg21771834 94286263</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cg27001184 94286267</td>
<td>Island</td>
<td>PEG10</td>
<td>5'UTR</td>
</tr>
</tbody>
</table>

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, MIR663cg06007966, 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
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-
DNAm in CRC from PBLs

Synes. 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


DNAm in CRC from PBLs


DNAm in CRC from PBLs


Molecular characterization of prostatic small-cell neuroendocrine carcinoma. Prostate 2003; 55: 55-64.


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

<table>
<thead>
<tr>
<th></th>
<th>Houseman’s method</th>
<th>Horvath’s method</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4(^+) T</td>
<td>Mean (SD)</td>
<td>CD8(^+)CD28CD45RA T Naive CD8 T</td>
</tr>
<tr>
<td>Natural killer cell</td>
<td>22.40 (7.76)</td>
<td>1.80 (0.20)</td>
</tr>
<tr>
<td>Monocyte</td>
<td>1.00 (5.32)</td>
<td>11.13 (3.76)</td>
</tr>
<tr>
<td>Granulocyte</td>
<td>8.15 (2.94)</td>
<td>196.95 (44.88)</td>
</tr>
<tr>
<td>Plasma blast</td>
<td>52.03 (12.03)</td>
<td></td>
</tr>
</tbody>
</table>

Figure S1. The area under the receiver operating characteristic curve (AUC) analysis with six CpGs (cg04958124, cg10321339, cg12704462, cg18144285, cg06007966, and cg17375901).
DNA methylation (DNAm) in colorectal cancer (CRC) from peripheral blood leukocytes (PBLs)

**<GSE48684>**

**A** By chromosome

**B** By CpG context

**C** By enhancer and/or promoter

**D** By gene region
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 DNA methylation (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.
DNAm in CRC from PBLs

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)
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

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

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Length</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr10</td>
<td>118030848</td>
<td>118034357</td>
<td>3510</td>
<td>2</td>
</tr>
<tr>
<td>Chr10</td>
<td>131756487</td>
<td>131772187</td>
<td>15701</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>2158555</td>
<td>2165961</td>
<td>7407</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>31817810</td>
<td>31823282</td>
<td>5473</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>31824327</td>
<td>31841980</td>
<td>17654</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>32447944</td>
<td>32452839</td>
<td>4896</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>32454216</td>
<td>32461240</td>
<td>7025</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>44324759</td>
<td>44333192</td>
<td>8434</td>
<td>2</td>
</tr>
<tr>
<td>Chr11</td>
<td>128553855</td>
<td>128566958</td>
<td>13104</td>
<td>2</td>
</tr>
<tr>
<td>Chr12</td>
<td>5018229</td>
<td>5021871</td>
<td>3643</td>
<td>2</td>
</tr>
<tr>
<td>Chr12</td>
<td>85304514</td>
<td>85307424</td>
<td>2911</td>
<td>2</td>
</tr>
<tr>
<td>Chr12</td>
<td>114840854</td>
<td>11487641</td>
<td>6788</td>
<td>2</td>
</tr>
<tr>
<td>Chr13</td>
<td>28491326</td>
<td>28499045</td>
<td>7720</td>
<td>2</td>
</tr>
<tr>
<td>Chr13</td>
<td>28500882</td>
<td>28503508</td>
<td>2627</td>
<td>2</td>
</tr>
<tr>
<td>Chr13</td>
<td>112707805</td>
<td>112717707</td>
<td>9903</td>
<td>2</td>
</tr>
<tr>
<td>Chr13</td>
<td>112758379</td>
<td>112763220</td>
<td>4842</td>
<td>2</td>
</tr>
<tr>
<td>Chr14</td>
<td>95233665</td>
<td>95240560</td>
<td>6896</td>
<td>2</td>
</tr>
<tr>
<td>Chr15</td>
<td>74419428</td>
<td>74429109</td>
<td>9682</td>
<td>2</td>
</tr>
<tr>
<td>Chr15</td>
<td>83951663</td>
<td>83954849</td>
<td>3187</td>
<td>2</td>
</tr>
<tr>
<td>Chr16</td>
<td>51183363</td>
<td>51190201</td>
<td>6839</td>
<td>2</td>
</tr>
<tr>
<td>Chr16</td>
<td>54965084</td>
<td>54974287</td>
<td>9204</td>
<td>2</td>
</tr>
<tr>
<td>Chr18</td>
<td>74960629</td>
<td>74963807</td>
<td>3179</td>
<td>2</td>
</tr>
<tr>
<td>Chr20</td>
<td>25061762</td>
<td>25065553</td>
<td>3792</td>
<td>2</td>
</tr>
<tr>
<td>Chr20</td>
<td>57424521</td>
<td>57431303</td>
<td>6783</td>
<td>2</td>
</tr>
<tr>
<td>Chr20</td>
<td>61806628</td>
<td>61810902</td>
<td>4275</td>
<td>2</td>
</tr>
<tr>
<td>Chr21</td>
<td>38076709</td>
<td>38083586</td>
<td>6878</td>
<td>2</td>
</tr>
</tbody>
</table>

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.

<p>| WHIBAA23: All HRs were adjusted by leukocyte heterogeneities plus DNA methylation-predicted age. |
|-----------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>CpG</th>
<th>age.HR*</th>
<th>age.SE</th>
<th>age.P</th>
<th>bmi.HR**</th>
<th>bmi.SE</th>
<th>bmi.P</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg24885794</td>
<td>1.42</td>
<td>0.193</td>
<td>0.067</td>
<td>1.41</td>
<td>0.193</td>
<td>0.072</td>
</tr>
<tr>
<td>cg26997085</td>
<td>1.44</td>
<td>0.152</td>
<td>0.017</td>
<td>1.42</td>
<td>0.154</td>
<td>0.023</td>
</tr>
<tr>
<td>cg22331138</td>
<td>1.25</td>
<td>0.187</td>
<td>0.240</td>
<td>1.25</td>
<td>0.189</td>
<td>0.236</td>
</tr>
<tr>
<td>cg16492735</td>
<td>1.07</td>
<td>0.202</td>
<td>0.731</td>
<td>1.07</td>
<td>0.202</td>
<td>0.738</td>
</tr>
<tr>
<td>cg09512080</td>
<td>1.21</td>
<td>0.183</td>
<td>0.305</td>
<td>1.20</td>
<td>0.183</td>
<td>0.324</td>
</tr>
<tr>
<td>cg00906934</td>
<td>1.50</td>
<td>0.187</td>
<td>0.031</td>
<td>1.49</td>
<td>0.187</td>
<td>0.032</td>
</tr>
<tr>
<td>cg26503018</td>
<td>1.25</td>
<td>0.199</td>
<td>0.266</td>
<td>1.25</td>
<td>0.198</td>
<td>0.269</td>
</tr>
<tr>
<td>cg27120649</td>
<td>1.26</td>
<td>0.208</td>
<td>0.274</td>
<td>1.27</td>
<td>0.208</td>
<td>0.251</td>
</tr>
<tr>
<td>cg21771834</td>
<td>1.11</td>
<td>0.218</td>
<td>0.625</td>
<td>1.11</td>
<td>0.219</td>
<td>0.627</td>
</tr>
<tr>
<td>cg27001184</td>
<td>1.17</td>
<td>0.188</td>
<td>0.402</td>
<td>1.17</td>
<td>0.188</td>
<td>0.412</td>
</tr>
</tbody>
</table>

<p>| GSE48684: All ORs were adjusted by sex. |
|-----------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>CpG</th>
<th>OR§</th>
<th>SE</th>
<th>P</th>
<th>OR¥</th>
<th>SE</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg24885794</td>
<td>3.63</td>
<td>0.342</td>
<td>0.0002</td>
<td>3.63</td>
<td>0.342</td>
<td>0.0002</td>
</tr>
<tr>
<td>cg26997085</td>
<td>2.91</td>
<td>0.320</td>
<td>0.001</td>
<td>2.91</td>
<td>0.320</td>
<td>0.001</td>
</tr>
<tr>
<td>cg22331138</td>
<td>1.97</td>
<td>0.260</td>
<td>0.009</td>
<td>1.97</td>
<td>0.260</td>
<td>0.009</td>
</tr>
<tr>
<td>cg16492735</td>
<td>2.75</td>
<td>0.313</td>
<td>0.001</td>
<td>2.75</td>
<td>0.313</td>
<td>0.001</td>
</tr>
<tr>
<td>cg09512080</td>
<td>2.43</td>
<td>0.294</td>
<td>0.002</td>
<td>2.43</td>
<td>0.294</td>
<td>0.002</td>
</tr>
<tr>
<td>cg00906934</td>
<td>2.50</td>
<td>0.378</td>
<td>0.016</td>
<td>2.50</td>
<td>0.378</td>
<td>0.016</td>
</tr>
<tr>
<td>cg26503018</td>
<td>1.63</td>
<td>0.233</td>
<td>0.036</td>
<td>1.63</td>
<td>0.233</td>
<td>0.036</td>
</tr>
<tr>
<td>cg27120649</td>
<td>2.61</td>
<td>0.340</td>
<td>0.005</td>
<td>2.61</td>
<td>0.340</td>
<td>0.005</td>
</tr>
<tr>
<td>cg21771834</td>
<td>1.29</td>
<td>0.218</td>
<td>0.244</td>
<td>1.29</td>
<td>0.218</td>
<td>0.244</td>
</tr>
<tr>
<td>cg27001184</td>
<td>1.75</td>
<td>0.243</td>
<td>0.022</td>
<td>1.75</td>
<td>0.243</td>
<td>0.022</td>
</tr>
</tbody>
</table>

<p>| GSE199057: All ORs were adjusted by sex plus age and DNA methylation-predicted age. |
|-----------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>CpG</th>
<th>OR§</th>
<th>SE</th>
<th>P</th>
<th>OR¥</th>
<th>SE</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg24885794</td>
<td>2.99</td>
<td>0.445</td>
<td>0.014</td>
<td>5.73</td>
<td>0.538</td>
<td>0.001</td>
</tr>
<tr>
<td>cg26997085</td>
<td>3.79</td>
<td>0.511</td>
<td>0.009</td>
<td>3.39</td>
<td>0.515</td>
<td>0.018</td>
</tr>
<tr>
<td>cg22331138</td>
<td>3.40</td>
<td>0.465</td>
<td>0.008</td>
<td>3.58</td>
<td>0.526</td>
<td>0.015</td>
</tr>
<tr>
<td>cg16492735</td>
<td>4.83</td>
<td>0.532</td>
<td>0.003</td>
<td>6.18</td>
<td>0.555</td>
<td>0.001</td>
</tr>
<tr>
<td>cg09512080</td>
<td>2.75</td>
<td>0.420</td>
<td>0.016</td>
<td>3.67</td>
<td>0.535</td>
<td>0.015</td>
</tr>
<tr>
<td>cg00906934</td>
<td>2.14</td>
<td>0.347</td>
<td>0.028</td>
<td>4.54</td>
<td>0.530</td>
<td>0.004</td>
</tr>
<tr>
<td>cg26503018</td>
<td>4.61</td>
<td>0.556</td>
<td>0.006</td>
<td>11.29</td>
<td>0.597</td>
<td>0.00005</td>
</tr>
<tr>
<td>cg27120649</td>
<td>6.69</td>
<td>0.598</td>
<td>0.001</td>
<td>7.27</td>
<td>0.543</td>
<td>0.0003</td>
</tr>
<tr>
<td>cg21771834</td>
<td>5.65</td>
<td>0.620</td>
<td>0.005</td>
<td>4.47</td>
<td>0.529</td>
<td>0.005</td>
</tr>
<tr>
<td>cg27001184</td>
<td>1.91</td>
<td>0.319</td>
<td>0.043</td>
<td>2.03</td>
<td>0.492</td>
<td>0.151</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Description</th>
<th>Enrichment Ratio</th>
<th>Key Driver</th>
<th>Description</th>
<th>Enrichment Ratio</th>
<th>Key Driver</th>
<th>Description</th>
<th>Enrichment Ratio</th>
<th>Key Driver</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO - biological process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial genome maintenance</td>
<td>168.27</td>
<td>PIF1</td>
<td>Mitochondrial genome maintenance</td>
<td>252.41</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telomere organization</td>
<td>33.90</td>
<td>PIF1</td>
<td>Telomere organization</td>
<td>50.85</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA biosynthetic process</td>
<td>26.77</td>
<td>PIF1</td>
<td>DNA biosynthetic process</td>
<td>40.16</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signal transduction by p53 class mediator</td>
<td>22.22</td>
<td>RPF2</td>
<td>Signal transduction by p53 class mediator</td>
<td>33.34</td>
<td>RPF2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rRNA metabolic process</td>
<td>21.51</td>
<td>RPF2</td>
<td>rRNA metabolic process</td>
<td>32.27</td>
<td>RPF2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribonucleoprotein complex subunit organization</td>
<td>20.22</td>
<td>RPF2</td>
<td>Ribonucleoprotein complex subunit organization</td>
<td>30.33</td>
<td>RPF2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA conformation change</td>
<td>20.05</td>
<td>PIF1</td>
<td>DNA conformation change</td>
<td>30.07</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein localization to nucleus</td>
<td>18.48</td>
<td>RPF2</td>
<td>Protein localization to nucleus</td>
<td>27.72</td>
<td>RPF2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA replication</td>
<td>18.33</td>
<td>PIF1</td>
<td>DNA replication</td>
<td>27.50</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative regulation of transferase activity</td>
<td>18.19</td>
<td>PIF1</td>
<td>Negative regulation of transferase activity</td>
<td>27.29</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO - cellular component</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication fork</td>
<td>47.27</td>
<td>PIF1</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golgi stack</td>
<td>24.92</td>
<td>B4GALNT4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sperm part</td>
<td>17.47</td>
<td>SV2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosomal region</td>
<td>10.08</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transport vesicle</td>
<td>8.86</td>
<td>SV2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presynapse</td>
<td>6.90</td>
<td>SV2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO - molecular function</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalytic activity, acting on a glycoprotein</td>
<td>170.38</td>
<td>B4GALNT4</td>
<td>Telomeric DNA binding rRNA binding</td>
<td>163.28</td>
<td>PIF1</td>
<td>Catalytic activity, acting on a glycoprotein</td>
<td>255.57</td>
<td>B4GALNT4</td>
</tr>
<tr>
<td>Telomeric DNA binding</td>
<td>108.85</td>
<td>PIF1</td>
<td>Helicase activity</td>
<td>40.82</td>
<td>PIF1</td>
<td>rRNA binding</td>
<td>91.84</td>
<td>RPF2</td>
</tr>
<tr>
<td>rRNA binding</td>
<td>61.23</td>
<td>RPF2</td>
<td>Catalytic activity, acting on DNA</td>
<td>33.21</td>
<td>PIF1</td>
<td>Transferase activity, transferring glycosyl groups</td>
<td>22.27</td>
<td>B4GALNT4</td>
</tr>
<tr>
<td>Helicase activity</td>
<td>27.21</td>
<td>PIF1</td>
<td>Magnesium ion binding</td>
<td>28.96</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalytic activity, acting on DNA</td>
<td>22.14</td>
<td>PIF1</td>
<td>Catalytic activity, acting on RNA</td>
<td>17.60</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnesium ion binding</td>
<td>19.30</td>
<td>PIF1</td>
<td>Enzyme inhibitor activity</td>
<td>17.09</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transferase activity, transferring glycosyl groups</td>
<td>14.84</td>
<td>B4GALNT4</td>
<td>ATPase activity</td>
<td>13.80</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalytic activity, acting on RNA</td>
<td>11.73</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enzyme inhibitor activity</td>
<td>11.39</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATPase activity</td>
<td>9.20</td>
<td>PIF1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathway - KEGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECM-receptor interaction</td>
<td>44.33</td>
<td>SV2B</td>
<td>N/A</td>
<td></td>
<td></td>
<td>ECM-receptor interaction</td>
<td>44.33</td>
<td>SV2B</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td>2.77</td>
<td>B4GALNT4</td>
<td></td>
<td></td>
<td></td>
<td>Metabolic pathways</td>
<td>2.77</td>
<td>B4GALNT4</td>
</tr>
<tr>
<td>Pathway - Reactome§</td>
<td>Toxicity of botulinum toxin type D (bont/d)</td>
<td>1006.20</td>
<td>SV2B</td>
<td>N/A</td>
<td>Toxicity of botulinum toxin type D (BoNT/D)</td>
<td>1055.40</td>
<td>SV2B</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------</td>
<td>---------</td>
<td>------</td>
<td>-----</td>
<td>--------------------------------------------</td>
<td>---------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>Pathway - Reactome§</td>
<td>Toxicity of botulinum toxin type F (BoNT/F)</td>
<td>1006.20</td>
<td>SV2B</td>
<td></td>
<td>Toxicity of botulinum toxin type F (BoNT/F)</td>
<td>1055.40</td>
<td>SV2B</td>
<td></td>
</tr>
<tr>
<td>Pathway - Reactome§</td>
<td>Neurotoxicity of Clostridium toxins</td>
<td>503.10</td>
<td>SV2B</td>
<td></td>
<td>Neurotoxicity of Clostridium toxins</td>
<td>527.70</td>
<td>SV2B</td>
<td></td>
</tr>
<tr>
<td>Pathway - Reactome§</td>
<td>Uptake and actions of bacterial toxins</td>
<td>162.29</td>
<td>SV2B</td>
<td></td>
<td>Uptake and actions of bacterial toxins</td>
<td>170.23</td>
<td>SV2B</td>
<td></td>
</tr>
<tr>
<td>Pathway - Reactome§</td>
<td>Infectious disease</td>
<td>13.82</td>
<td>SV2B</td>
<td></td>
<td>Carboxyterminal post-translational modifications of tubulin</td>
<td>146.58</td>
<td>TTLL10</td>
<td></td>
</tr>
<tr>
<td>Disease - Disgenet</td>
<td>Kidney diseases</td>
<td>35.08</td>
<td>SV2B</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Prostatic neoplasms</td>
<td>7.99</td>
<td>FOXP4</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Cardiac output, low</td>
<td>227.16</td>
<td>FOXP4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Lymphocytic choriomeningitis</td>
<td>149.07</td>
<td>FOXP4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Epilepsy, temporal lobe</td>
<td>71.20</td>
<td>SV2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Epilepsy</td>
<td>19.47</td>
<td>SV2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Williams syndrome</td>
<td>18.14</td>
<td>ZNF398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Seizures</td>
<td>17.80</td>
<td>SV2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Prostatic neoplasms</td>
<td>16.17</td>
<td>FOXP4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease - GLAD4U</td>
<td>Nelson syndrome</td>
<td>10.37</td>
<td>ZNF398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.