

Statistical Methods for Epigenetic Data

Ya Wang

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Public Health
in the Mailman School of Public Health

COLUMBIA UNIVERSITY

2019

©2019

Ya Wang

All Rights Reserved

ABSTRACT

Statistical Methods for Epigenetic Data

Ya Wang

DNA methylation plays a crucial role in human health, especially cancer. Traditional DNA methylation analysis aims to identify CpGs/genes with differential methylation (DM) between experimental groups. Differential variability (DV) was recently observed that contributes to cancer heterogeneity and was also shown to be essential in detecting early DNA methylation alterations, notably epigenetic field defects. Moreover, studies have demonstrated that environmental factors may modify the effect of DNA methylation on health outcomes, or vice versa. Therefore, this dissertation seeks to develop new statistical methods for epigenetic data focusing on DV and interactions when efficient analytical tools are lacking. First, as neighboring CpG sites are usually highly correlated, we introduced a new method to detect differentially methylated regions (DMRs) that uses combined DM and DV signals between diseased and non-diseased groups. Next, using both DM and DV signals, we considered the problem of identifying epigenetic field defects, when CpG-site-level DM and DV signals are minimal and hard to be detected by existing methods. We proposed a weighted epigenetic distance-based method that accumulates CpG-site-level DM and DV signals in a gene. Here DV signals were captured by a pseudo-data matrix constructed using centered quadratic methylation measures. CpG-site-level association signal annotations were introduced as weights in distance calculations to up-weight signal CpGs and down-weight noise CpGs to further boost the study power. Lastly, we extended the weighted epigenetic distance-based method to incorporate DNA methylation by environment interactions in the detection of overall association between DNA methylation and health outcomes. A pseudo-data matrix was constructed with cross-product terms between DNA methylation and environmental factors that is able to capture their interactions. The superior performance of the proposed methods were shown through intensive simulation

studies and real data applications to multiple DNA methylation data.

Table of Contents

List of Figures	v
List of Tables	xiv
1 Introduction	1
1.1 Overview	1
1.2 DNA methylation	3
1.3 Statistical methods in DNA methylation studies	4
1.4 Epigenetic field defects	5
1.5 Distance-based method	5
1.6 Interactions between DNA methylation and Environmental factors	6
2 Accounting for Differential Variability in Detecting Differentially Methylated Regions	8
2.1 Introduction	8
2.2 Methods	12
2.3 Comparison methods	15
2.4 Simulation study	15
2.4.1 Simulation setup	15
2.4.2 Adaption to case-control designs	16
2.5 Results	16
2.5.1 Simulation results	16
2.5.2 Real data application	17

2.5.3	TCGA BRCA data	18
2.5.4	Replication analysis with GEO BRCA data	22
2.5.5	Identification of epigenetic field defects in the GEO BRCA data	24
2.6	Discussion	29
3	Detection of Epigenetic Field Defects Using a Weighted Epigenetic Distance-Based Method	31
3.1	Introduction	31
3.2	Materials and methods	33
3.2.1	Comparison methods	36
3.2.2	Simulation study	36
3.2.3	Simulation setup	37
3.3	Results	39
3.3.1	Simulation results	39
3.3.2	Real data application	44
3.3.3	Discovery analysis using the GEO BRCA data	44
3.3.4	Validation of the identified epigenetic field defects in the GEO BRCA data	51
3.3.5	Replication analysis using an independent data of normal tissues	51
3.4	Discussion	53
4	A Powerful and Flexible Weighted Distance-Based Method Incorporating Interactions Between DNA Methylation and Environmental Factors on Health Outcomes	55
4.1	Introduction	55
4.2	Methods	57
4.2.1	The proposed method	57
4.2.2	Comparison methods	59
4.3	Simulation studies	60
4.3.1	Simulation setup	60
4.3.2	Simulation results	62

4.4	Real data applications	65
4.4.1	CCCEH birth cohorts	65
4.4.2	Neurodevelopment outcomes	65
4.4.3	DNA methylation data processing	66
4.4.4	Risk of PAH, DNA methylation and their interactions on ADHD . .	66
4.4.5	Risk of PAH, DNA methylation and their interactions on MDI . . .	68
4.5	Discussion	71
5	Conclusions	73
I	Appendices	77
A	Appendix to Accounting for Differential Variability in Detecting Differentially Methylated Regions	78
A.1	Investigation of the distance limits to define clusters	79
A.2	Simulation studies for case-control designs	81
A.2.1	Simulation setup	81
A.2.2	Simulation results	82
A.3	Real data application	84
B	Appendix to Detection of Epigenetic Field Defects Using a Weighted Epigenetic Distance-Based Method	93
B.1	Additional Simulation Studies	94
B.1.1	Effects of gene sizes in Type I errors	94
B.1.2	Values of shape parameters in simulations	95
B.1.3	Simulation settings with one gene considering correlations among CpGs	96
B.2	Real data application	99
B.2.1	Discovery analysis	99
B.2.2	Validation analysis	109
B.2.3	Replication analysis	110

C	Appendix to A Powerful and Flexible Weighted Distance-Based Method Incorporating Interactions Between DNA Methylation and Environmen- tal Factors on Health Outcomes	116
C.1	Additional Simulation Studies	117
C.1.1	Effects of gene sizes in Type I errors	117
C.1.2	Simulation settings with different types of signals	118
C.1.3	Simulation settings with fixed number of signal items coming from different number of signal CpGs	119
C.2	Real data applications	121
C.2.1	DNA methylation data processing	121
C.2.2	Risk of PAH, DNA methylation and their interactions on ADHD . .	122
C.2.3	Risk of PAH, DNA methylation and their interactions on MDI . . .	129
II	Bibliography	133
	Bibliography	134

List of Figures

2.1	The pipeline of the proposed new DMR detection algorithm.	12
2.2	ROC curves from simulation studies where 10 true DMRs have different region sizes ranging from 3 to 15 CpG sites with (A) mean signals only; (B) variance signals only; and (C) both mean and variance signals. DMRs were defined as regions with minimum region size of $L \geq 3$ CpG sites.	18
2.3	Top two ranked DMRs uniquely identified by the new method in the TCGA BRCA data. DMR #1 (top row) and #2 (bottom row) are located on chromosomes 19 and 5. Vertical dashed lines define boundaries of DMRs. Left column shows the combined signal scores of the sites in the DMRs before (circles) and after (curves) smoothing, in which horizontal dotted lines define the threshold k that defines a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the DMRs comparing tumor and normal-adjacent tissues.	21
2.4	Two examples of overlapping DMRs among all DMRs identified by the new method in both the TCGA and the GEO BRCA data from the tumor versus normal-adjacent comparison. Plotted are the combined site-level signal scores in the DMRs before and after smoothing for the TCGA BRCA data (circles, solid curves) and the GEO BRCA data (crosses, dashed curves). Vertical lines define boundaries of DMRs, and horizontal lines define the threshold k that defines a candidate region.	23

2.5	<p>(A) The epigenetic field defects, i.e. the two DMRs identified in the GEO BRCA normal-adjacent versus normal comparison (crosses, dotted curves) together with the overlapping DMRs identified in the GEO BRCA tumor versus normal-adjacent comparison (circles, dashed curves), and the GEO BRCA tumor versus normal comparison (triangles, solid curves). Vertical lines define boundaries of DMRs, and horizontal lines define the threshold k that defines a candidate region. (B) Heat maps of the original DNA methylation measures of the sites from the epigenetic field defects, i.e. the two DMRs identified in the GEO BRCA normal-adjacent versus normal comparison. Green is for 50 normal tissues from age-matched cancer-free women. Blue is for 42 normal-adjacent tissues, and red is for 42 tumor tissues. (C) Two CpG sites selected of the 58 sites from the epigenetic field defects, i.e. the two DMRs. Plotted are the original DNA methylation measures of normal tissues from 50 age-matched cancer-free women (crosses), and normal-adjacent tissues (circles) and tumor tissues (triangles) from 42 BRCA patients. The three horizontal lines represent mean methylation levels of the three groups. $\lambda_i(\text{NN})$ is the site-level scaling parameter from the normal-adjacent versus normal comparison, and $\lambda_i(\text{TN})$ is that from the tumor versus normal-adjacent comparison. The three outlier samples were marked using solid circles (normal-adjacent tissues) and solid triangles (matching tumor tissues).</p>	27
3.1	<p>Power results for simulation settings with one gene. The signal gene has one signal CpG and increasing number of total CpGs, i.e., decreasing signal-to-noise ratios from 1:0, 1:24 to 1:49 (panel A for mean signals only, panel B for variance signals only), or with a fixed total number of CpGs 50 and increasing signal-to-noise ratios from 1:49, 3:47, to 5:45 (panel C for mean signals only, panel D for variance signals only).</p>	40
3.2	<p>Power results for simulation settings with 10 genes. We set each gene to have 25 CpGs and only one gene to have signals. The signal gene has 1 signal CpG and 24 noise CpGs, with signal CpG having mean signal only (panel A), variance signal only (panel B), and mean and variance signals with different sizes of mean signals (panels C and D).</p>	42

3.3	Power results for simulation settings with outlier samples. We set to have 10%, 15% and 20% outlier samples and two different signal-to-noise ratios 5:45 and 10:40 . . .	43
3.4	Manhattan plots with results from $\mathbf{D}^{w-DM-DV}$ and EWAS ^{min-P} . The solid horizontal line is the 0.0005 gene-level P -value threshold. The dashed horizontal line is the Bonferroni adjusted 0.05 significance level (0.05/19 271 genes = 0.0000026 adjusted gene-level P -value threshold). Genes annotated with stars are those identified by both methods at the 0.0005 gene-level P -value threshold.	45
3.5	(A) Heatmaps of DNA methylation measures of CpGs in the <i>CFTR</i> and <i>PLS1</i> genes. The CpGs underlined are the top 4 P -value ranked CpGs in the <i>CFTR</i> gene and the top 2 P -value ranked CpGs in the <i>PLS1</i> gene. (B) DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of the top 4 P -value ranked CpGs in the <i>CFTR</i> gene and the top 2 P -value ranked CpGs in the <i>PLS1</i> gene. Pm and Pv are P -values from CpG site-level mean and variance tests adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.	49
4.1	Power results for simulation settings with main signals only, interaction signals only and both main and interaction signals when there are 30 CpGs in a gene.	64
4.2	Boxplot of DNA methylation measures of the 13 CpGs in gene <i>CYP2E1</i> stratified by PAH and ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene <i>CYP2E1</i>) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$	68
4.3	Boxplots of DNA methylation measures of the 4 CpGs in gene <i>C8orf80</i> stratified by PAH and MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene <i>C8orf80</i>) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$	70

A.1	Histogram of the difference in the combined signal scores for neighboring CpG sites with the choice of difference distance limits.	80
A.2	ROC curves from simulation studies when 10 true DMRs have different region sizes varying from 3 to 15 CpG sites with: (A) mean signals only; (B) variance signals only; and (C) both mean and variance signals. DMRs were defined as regions with minimum region size $L \geq 3$ CpG sites.	83
A.3	DMR #1 (top row) and #2 (bottom row) located on chromosomes 10 and 8 (out of 170 DMRs) that were identified uniquely by the new method in the TCGA KIRC data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal-adjacent tissues.	88
A.4	DMR #1 (top row) and #2 (bottom row) located on chromosomes 1 and 10 (out of 89 DMRs) that were identified uniquely by the new method in the GEO BRCA tumor vs. normal-adjacent data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal-adjacent tissues. There are 3 gene, <i>SGCE</i> , <i>PEG10</i> and <i>PHOX2B</i> in these 2 DMRs. <i>SGCE</i> was reported to be associated with colorectal cancer (Ortega et al., 2010), <i>PEG10</i> was reported to be associated with hepatocellular carcinoma (Ip et al., 2007), and <i>PHOX2B</i> was reported to be associated with neuroblastoma (De Pontual et al., 2007)	89
A.5	Distributions of genome-wide site-level scale parameter λ_i in the GEO BRCA data. From left to right shows distribution of λ_i in (1) normal-adjacent vs. normal, (2) tumor vs. normal-adjacent, and (3) tumor vs. normal comparisons.	90

A.6	DMR #1 (top row) and #2 (bottom row) located on chromosomes 10 and 12 that were identified uniquely by the new method in the GEO BRCA normal-adjacent vs. normal data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between normal-adjacent and normal tissues.	91
A.7	DMR #1 (top row) and #2 (bottom row) located on chromosomes 6 (out of 15 DMRs) that were identified uniquely by the new method in the GEO BRCA tumor vs. normal data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal tissues.	92
B.1	Power results for simulation settings with one gene considering AR(1) correlation among neighboring CpGs with correlation coefficient $\rho=0.5$. The signal gene has one signal CpG and increasing number of total CpGs, i.e., decreasing signal-to-noise ratios from 1:0, 1:24 to 1:49 (panel A for mean signals only, panel B for variance signals only), or with a fixed total number of CpGs 50 and increasing signal-to-noise ratios from 1:49, 3:47, to 5:45 (panel C for mean signals only, panel D for variance signals only).	98
B.2	Heatmaps of original DNA methylation measures of the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues for 14 genes uniquely identified by $D^{w-DM-DV}$	101
B.3	Heatmaps of original DNA methylation measures of the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues for 7 genes uniquely identified by $EWAS^{min-P}$	102
B.4	Heatmaps of original DNA methylation measures of the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues for 7 genes identified by both $D^{w-DM-DV}$ and $EWAS^{min-P}$	103

B.5	DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of 16 CpGs in the <i>CFTR</i> gene that was uniquely identified by $D^{w-DM-DV}$, but ranked the last using $EWAS^{min-P}$ among all uniquely identified genes. Pm and Pv are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.	104
B.6	DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of 16 CpGs in the <i>PLS1</i> gene that was uniquely identified by $EWAS^{min-P}$, but ranked the last using $D^{w-DM-DV}$ among all uniquely identified genes. Pm and Pv are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.	105
B.7	Weighted distance matrices for genes <i>CFTR</i> and <i>PLS1</i>	106
B.8	DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of 13 CpGs in the <i>TMC4</i> gene that was identified by both $D^{w-DM-DV}$ and $EWAS^{min-P}$ and ranked on #1 and #2, respectively. Pm and Pv are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.	107
B.9	The selection probability of identifying a gene out of all genes of the same size. . .	108
B.10	$-\log_{10}(p\text{-value})$ from CpG site-level t -tests in (1) normal-adjacent versus normal comparison and (2) unmatched tumor versus normal-adjacent comparison in the GEO BRCA data for 21 genes identified by $D^{w-DM-DV}$	109
B.11	$-\log_{10}(P\text{-value})$ from CpG site-level t -tests in (1) normal-adjacent versus normal comparison and (2) unmatched tumor versus normal-adjacent comparison in the replication analysis for 7 replicated genes identified by $D^{w-DM-DV}$	111

- B.12 DNA methylation measures of 18 normal tissues from the replication data (GSE67919), 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors from the discovery data (GSE69914) of 16 CpGs in the *CFTR* gene that was uniquely identified in the discovery analysis and replicated by $D^{w-DM-DV}$. Pm1 and Pv1 are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene from the discovery analysis, and Pm2 and Pv2 are those from the replication analysis. Highlighted are the minimum adjusted DM and DV P -value across all P -values in the gene in each comparison. The four horizontal lines represent mean methylation levels of the four groups of tissues. . . . 113
- B.13 DNA methylation measures of 18 normal tissues from the replication data (GSE67919), 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors from the discovery data (GSE69914) of 7 CpGs in the *CXCL6* gene which was identified in the discovery analysis and replicated by both $D^{w-DM-DV}$ and $EWAS^{min-P}$. Pm1 and Pv1 are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene from the discovery analysis, and Pm2 and Pv2 are those from the replication analysis. Highlighted are the minimum adjusted DM and DV P -value across all P -values in the gene in each comparison. The four horizontal lines represent mean methylation levels of the four groups of tissues. 114
- B.14 DNA methylation measures of 18 normal tissues from the replication data (GSE67919), 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors from the discovery data (GSE69914) of 16 CpGs in the *PLS1* gene that was uniquely identified in the discovery analysis and replicated by $EWAS^{min-P}$. Pm1 and Pv1 are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene from the discovery analysis, and Pm2 and Pv2 are those from the replication analysis. Highlighted are the minimum adjusted DM and DV P -value across all P -values in the gene in each comparison. The four horizontal lines represent mean methylation levels of the four groups of tissues. . . . 115

C.1	Power results for simulation settings with main signals only, interaction signals only and both main and interaction signals when there are (A) 20 CpGs, (B) 30 CpGs, and (C) 40 CpGs in a gene.	118
C.2	Power results for simulation settings where there are 2 main signal items and 2 interaction signal items coming from 2, 3 and 4 signal CpGs, respectively when there are 30 CpGs in a gene.	120
C.3	Boxplot of DNA methylation measures of the 9 CpGs in gene <i>LOC84931</i> stratified by ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene <i>LOC84931</i>) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$.122	122
C.4	Boxplot of DNA methylation measures of the 4 CpGs in gene <i>HIST1H2BJ</i> stratified by ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene <i>HIST1H2BJ</i>) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$.123	123
C.5	Boxplot of DNA methylation measures of the 7 CpGs in gene <i>FAM35A</i> stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene <i>FAM35A</i>) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$	130
C.6	Boxplot of DNA methylation measures of the 3 CpGs in gene <i>DIRC1</i> stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene <i>DIRC1</i>) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$	131

C.7 Boxplot of DNA methylation measures of the 5 CpGs in gene *THSD1P* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *THSD1P*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$ 132

List of Tables

2.1	Significant DMRs identified in the TCGA BRCA data	19
2.2	Eleven genes identified in the top 10 ranked DMRs in TCGA BRCA data .	20
2.3	Significant DMRs identified in the GEO BRCA data	25
3.1	Type I error rates	39
3.2	Twenty one genes identified by $\mathbf{D}^{w-DM-DV}$ at the 0.0005 gene-level P -value threshold using the GEO BRCA Data	47
3.3	Fourteen genes identified by $\text{EWAS}^{\text{min-}P}$ at the 0.0005 gene-level P -value threshold using the GEO BRCA Data	48
4.1	Simulation settings with different types of signals	62
4.2	Type I error rates	63
4.3	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 10 genes by the proposed method $\mathbf{D}^{\text{w-main-int}}$ at the 0.005 gene-level P -value threshold	67
4.4	Application examining prenatal PAH, DNA methylation and their interactions on child MDI at age 3 identified 7 genes by the proposed method at the 0.005 gene-level P -value threshold	69
A.1	Significant DMRs Identified in the TCGA KIRC Data (160 matched pairs)	85
A.2	7 Cancer-Related Genes Identified in the Top 10 Ranked DMRs in TCGA KIRC Data ^a	86
A.3	11 Cancer-Related Genes Identified in the Top 10 Ranked DMRs in GEO BRCA Data (Tumor vs. Normal-adjacent) ^a	86

A.4	Significant DMRs Identified in the GEO BRCA Data (Tumor vs. Normal) .	87
B.1	Type I error rates in simulation settings with multiple genes of different sizes	94
B.2	Values of a_1 and b_1 for signal CpGs	95
B.3	Type I error rates in simulation settings with AR(1) correlation among neighboring CpGs with $\rho = 0.5$	97
B.4	11 genes identified by \mathbf{D}^{w-DM} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data	99
B.5	9 genes identified by \mathbf{D}^{w-DV} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data	99
B.6	2 significant genes identified by \mathbf{D}^{DM-DV} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data	100
B.7	6 genes identified by \mathbf{D}^{DM} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data	100
B.8	4 significant genes identified by \mathbf{D}^{DV} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data	100
B.9	Summary number of genes identified by comparison methods in both the discovery analysis and replication analysis at the 0.0005 gene-level P -value threshold	110
C.1	Type I error rates in simulation settings with multiple genes of different sizes	117
C.2	Simulation settings with 4 signal items and the same signal composition (2 main and 2 interaction signals) but from 2~4 signal CpGs	119
C.3	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 11 genes by \mathbf{D}^{w-main} at the 0.005 gene-level P -value threshold	124
C.4	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 14 genes by \mathbf{D}^{w-int} at the 0.005 gene-level P -value threshold	125

C.5	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 5 genes by $\mathbf{D}^{\text{main-int}}$ at the 0.005 gene-level P -value threshold	125
C.6	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 8 genes by \mathbf{D}^{main} at the 0.005 gene-level P -value threshold	126
C.7	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 11 genes by \mathbf{D}^{int} at the 0.005 gene-level P -value threshold	126
C.8	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 4 genes by L^S at the 0.005 gene-level P -value threshold	127
C.9	Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 4 genes by L^M at the 0.005 gene-level P -value threshold	127
C.10	Summary number of genes identified at the 0.005 gene-level P -value threshold and replicated at the 0.1 gene-level P -value threshold in the application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3	128
C.11	Summary number of genes identified at the 0.005 gene-level P -value threshold and replicated at the 0.1 gene-level P -value threshold in the application examining prenatal PAH, DNA methylation and their interactions on child MDI at age 3	129

Acknowledgments

I would like to express my sincere gratitude to my thesis adviser Dr. Shuang Wang, who has been an ideal advisor, for offering enormous effort, help and encouragement through my doctoral study. She provided inspiring advice and insightful advice throughout my dissertation research. Without her tremendous guidance and support, this dissertation would not be possible.

I am grateful to my dissertation committee members, Dr. Min Qian, Dr. Julie Herbstman, Dr. Yuanjia Wang and Dr. Iuliana Ionita-Laza for offering valuable thoughts and suggestions to help me improve my thesis. I feel absolutely privileged to have these supportive and inspiring mentors. Dr. Qian provided numerous constructive comments and advice on my thesis work and collaborated on two chapters of my thesis work.

Special thanks to Dr. Frederica Perera and Dr. Julie Herbstman for sponsoring my research assistantship in the past five years. We collaborated on multiple incredible projects that concern the impact of environmental exposures to the neurodevelopment of newborns and childrens at their early ages.

Last but not least, I would like to thank my family for their endless love and unconditional support in every moment of my life. I am also very grateful to all my friends for their incredible support.

Chapter 1

Introduction

1.1 Overview

DNA methylation plays a crucial role in human health, especially cancer. A typical analysis of DNA methylation involves identification of differentially methylated (DM) CpG sites or genes between different experimental groups, when traditional DM analysis aims to identify CpGs/genes with significant changes in mean methylation levels between different experimental groups, such as diseased and non-diseased samples. Differential variability (DV) was recently observed that contributes to cancer heterogeneity and was also shown to be essential in detecting DNA methylation alterations happening early in carcinogenesis, namely epigenetic field defects. In analysis of DV, the aim is to identify CpGs/genes with a significant change in variance of methylation levels between two experimental groups. Moreover, studies have demonstrated that environmental factors may modify the effect of DNA methylation on health outcomes, or vice versa. Therefore, this dissertation seeks to develop new statistical methods for epigenetic data focusing on DV and interactions when efficient analytical tools are lacking.

First, as neighboring CpG sites are usually highly correlated, we introduced a new method to detect differentially methylated regions (DMRs) that uses combined DM and DV signals between diseased and non-diseased groups. This work found that the proposed DMR method is much more powerful than existing methods that use either DM or DV signals when both signals exist. Applications to DNA methylation data of breast invasive carcinoma

(BRCA) and kidney renal clear cell carcinoma (KIRC) from The Cancer Genome Atlas (TCGA) and another DNA methylation data of BRCA from Gene Expression Omnibus (GEO) suggest that the proposed DMR method identified new cancer-related DMRs that were missed by methods considering only one type of signals. The two BRCA datasets allowed us for a replication analysis which suggests that the identified DMRs based on DV signals are reproducible.

Next, using both DM and DV signals, we considered the problem of identifying epigenetic field defects, when both DM and DV signals are minimal on CpG site levels and won't be detected by existing methods. We proposed a weighted epigenetic distance-based method accumulating both CpG-site-level DM and DV signals in a gene. Here DV signals were captured by a pseudo-data matrix constructed using centered quadratic methylation measures. CpG-site-level weights were introduced in distance calculations to up-weight signal CpGs and down-weight noise CpGs to further boost the study power. This work found that the proposed weighted epigenetic distance-based method has much greater power than non-weighted versions and site-level epigenome-wide association studies (EWAS). The application to the same GEO BRCA methylation data comparing normal-adjacent tissues to tumors of breast cancer patients (as a surrogate to pre-cancer tissues) to normal tissues of independent age-matched cancer-free women identified novel epigenetic field defects that were missed by comparison methods. Majority of the epigenetic field defects identified were previously reported to be associated with breast cancer and were confirmed the progression to breast cancer with some of them being further replicated.

Lastly, we extended the weighted epigenetic distance-based method to incorporate DNA methylation by environment interactions in the detection of overall association between DNA methylation and health outcomes. In this weighted epigenetic distance-based method, a pseudo-data matrix was constructed with cross-product terms between DNA methylation and environmental factors capturing their interactions. Weights were similarly considered to up-weight signals and down-weight noises in distance calculations. We demonstrated the superior performance of the proposed method over comparison methods. Applications to the data from the Mothers and Newborns birth cohort of the Columbia Center for Children's Environmental Health (CCCEH) identified associations between Attention Deficit Hyper-

activity Disorder and Mental Development Index at age 3 and several epigenetic genes due to interaction effects between DNA methylation and prenatal air polycyclic aromatic hydrocarbons (PAH) exposure when some of the genes identified were further replicated in the CCCEH replication data.

1.2 DNA methylation

DNA methylation is an epigenetic modification when methyl groups are added to the 5th position of cytosine within the CpG dinucleotide (Bird, 2002). It plays an essential role in gene expression (Baylin et al., 2001; Fahrner et al., 2002; Jones, 2012; Phillips, 2008) and cancer (Das and Singal, 2004; Ehrlich, 2002; Esteller and Herman, 2002; Kulis and Esteller, 2010). Studies have found that abnormal DNA methylation processes are related to many different types of cancers (Ruike et al., 2010; Teschendorff et al., 2009; Lasseigne et al., 2014; Hinoue et al., 2012). In general, two kinds of aberrant methylation are observed. One is local hyper-methylation which usually occurs in the promoter-related CpG island, and it often leads to the silence of downstream tumor suppressor genes (Koukoura et al., 2014; Baylin, 2005; Curradi et al., 2002; Herman and Baylin, 2003; Robertson, 2005). The other is global hypo-methylation that usually leads to instability of chromosomes (Robertson, 2005; Eden et al., 2003; Feinberg and Tycko, 2004; Jaenisch and Bird, 2003). In addition to cancers, research also found that aberrant DNA methylation is related to a range of other human diseases (Feinberg, 2007; Jager et al., 2014; Lund et al., 2004; Mill and Petronis, 2007, 2008; Mill et al., 2008; Nestler, 2014; Schanen, 2006), such as Alzheimer's disease (Jager et al., 2014), major depressive disorder (Mill and Petronis, 2007), drug addiction (Nestler, 2014), etc.

There are different methods to quantify DNA methylation, among which Bisulfite microarray and sequencing are two widely used technologies. This includes popular array technologies including the Illumina Infinium HumanMethylation 27K, 450K and 850K EPIC BeadChips with methylation β -values measuring the proportion of methylated intensities out of total intensities. Popular sequencing technologies include whole-genome bisulfite sequencing, reduced representation bisulfite sequencing and Agilent SureSelect Human

Methyl-Seq (Methyl-seq), which generate either ratio of methylated intensities versus total coverage at each CpG site or number of methylated or unmethylated cytosine.

1.3 Statistical methods in DNA methylation studies

Numerous methods were already developed to identify differentially methylated loci (DML) based on differences in mean methylation levels (DM, mean signals) between two experimental groups. Standard EWAS that focus on mean signals perform CpG site-level tests to identify differentially methylated CpGs between two experimental groups using standard tests such as a t -test, a regression-based test or its regularized versions (Tusher et al., 2001; Smyth, 2004; Wettenhall and Smyth, 2004), or a non-parametric Wilcoxon rank sum test.

Studies have also found that epigenetic instability of important genomic regions may lead to increased methylation variability in cancer, which also contribute to cancer heterogeneity (Phipson and Oshlack, 2014; Hansen et al., 2011a; Feinberg and Irizarry, 2010; Gervin et al., 2011; Jaffe et al., 2012a). Methods were developed to identify differential variability (DV, variance signals), i.e., CpGs sites with significant differences in variance of methylation levels between two experimental groups, using standard tests such as an F -test (Hansen et al., 2011b; Ho et al., 2008), the Bartlett's test or its regularized versions (Teschendorff et al., 2016a,b), or an empirical Bayes extension of the Levene's test (Phipson and Oshlack, 2014). Methods were also developed to identify CpGs with mean and variance combined signals at CpG site-level (Ahn and Wang, 2013; Chen et al., 2014; Ruan et al., 2016; Sun et al., 2017).

As DNA methylation levels of neighboring CpGs are strongly correlated, especially when consecutive CpGs in a genomic region are associated with health outcomes, statistical methods to detect differentially methylated regions (DMRs) were also developed. Existing DMR detection methods can be generally grouped into three types, to detect site-level signals first and then group adjacent loci into regions using ad hoc grouping rules (Hansen et al., 2012; Jaffe et al., 2012b; Butcher and Beck, 2015; Jühling et al., 2016; Hesse et al., 2015; Wen et al., 2016); or to define regions first and then test the significance of the defined regions (Ayyala et al., 2015; Sofer et al., 2013; Yip et al., 2014; Mayo et al., 2014); or to

use hidden Markov model that assumes three latent methylation states: hyper-methylation, hypo-methylation and no differential methylation, and then group adjacent sites with the same state into a region (Saito and Mituyama, 2015; Saito et al., 2014). However, existing DMR detection methods all focus on mean signals only.

1.4 Epigenetic field defects

Epigenetic field defects are notably DNA methylation alterations that usually occur in pre-cancer tissues and are crucial in cancer research because of its potential usage in early cancer detection (Teschendorff et al., 2016a,b). Current studies detecting epigenetic field defects usually compare normal tissues of healthy individuals to normal tissues adjacent to tumors (normal-adjacent tissues) of cancer patients as a surrogate of pre-cancer tissues that are difficult to collect.

Studies have successfully identified epigenetic field defects in breast cancer by comparing normal-adjacent tissues of breast cancer patients to normal tissues from healthy individuals. Teschendorff et al. identified epigenetic field defects in breast cancer based on DV signals with methylation site-level analyses (Teschendorff et al., 2016a).

In chapter 2, we developed a new DMR detection method that combines both DM and DV signals and successfully applied the developed method in identifying epigenetic field defects in breast cancer. Similarly as what was observed in Teschendorff et al. (Teschendorff et al., 2016a), epigenetic field defects were found to be mainly driven by increased variation in methylation due to several outlier normal-adjacent tissue samples.

1.5 Distance-based method

Distance-based method was originally developed in the field of ecology (McArdle and Anderson, 2001; Anderson, 2001) and had been proven to be powerful in genetic and gene expression studies (Zapala and Schork, 2006; Wessel and Schork, 2006; Han and Pan, 2010). A fundamental step is to construct a distance matrix to characterize the dissimilarities between pairs of individual samples in a study. It always has a dimension of $N \times N$ with N being the sample size regardless the added dimensionality from additional information

collected on the samples under the study.

Since epigenetic field defects are often characterized by increased variation in DNA methylation measures due to a few outlier normal-adjacent tissue samples, standard EWAS that perform CpG site-level tests are usually underpowered due to small mean differences as well as stringent multiple comparisons adjustment in a gene or a genetic region level. The common practice is to conduct site-level tests and select the site with minimum P -value in the region studied. These methods will have low power when site-level effects are minimal. On the other hand, distance-based methods accumulate site-level signals across all CpGs in a gene or a genetic region in calculating gene-level distances between pairs of samples thus boost the overall association power. This makes the distance-based methods the ideal methods to detect epigenetic field defects. In addition, distance-based methods are flexible and can be applied to a CpG site, a gene, a pathway, or an entire genome.

In chapter 3, we developed a weighted epigenetic distance-based method with a pseudo-data matrix constructed with centered quadratic methylation measures that is able to capture DV signals. By combining the original data matrix with the pseudo-data matrix, we are able to accumulate weak CpG-site-level DM and DV signals in a gene. CpG-site-level association signal annotations were introduced as weights in distance calculations to up-weight signal CpGs and down-weight noise CpGs to further boost the study power.

1.6 Interactions between DNA methylation and Environmental factors

Studies have demonstrated that DNA methylation may modify the risk of environmental factors on health outcomes. To identify interaction effects between DNA methylation and environmental factors on health outcomes has always been a great interest to researchers. Conventional methods such as EWAS usually focus on examining DM at CpG level or gene level combining multiple CpGs and/or finding environmental effects on health outcomes. Due to high dimensionality and low study power, current studies usually ignore the interaction between DNA methylation and environmental factors.

In chapter 4, we extended the weighted epigenetic distance-based method to incorpo-

rate DNA methylation by environment interactions. We introduced a pseudo-data matrix constructed with cross-product terms between DNA methylation and environmental factors that is able to capture their interaction signals. By combining the original data matrix with the pseudo-data matrix for interaction effects, we are able to identify both main and interaction signals. Weights were similarly considered to up-weight signals and down-weight noises in distance calculations.

Chapter 2

Accounting for Differential Variability in Detecting Differentially Methylated Regions

2.1 Introduction

DNA methylation plays an important role in gene expression (Baylin et al., 2001; Fahrner et al., 2002; Jones, 2012; Phillips, 2008) and cancer (Das and Singal, 2004; Ehrlich, 2002; Esteller and Herman, 2002; Kulis and Esteller, 2010). Two types of aberrant DNA methylation in cancer have been frequently observed, local hyper-methylation in some promoter-related CpG islands that often leads to silencing of downstream tumor suppressor genes (Koukoura et al., 2014; Baylin, 2005; Curradi et al., 2002; Herman and Baylin, 2003; Robertson, 2005), and global hypo-methylation that usually cause instability of chromosomes (Robertson, 2005; Eden et al., 2003; Feinberg and Tycko, 2004; Jaenisch and Bird, 2003). Studies have found that abnormal DNA methylation processes are related to many cancer types (Ruike et al., 2010; Teschendorff et al., 2009; Lasseigne et al., 2014; Hinoue et al., 2012) and a range of other human diseases (Feinberg, 2007; Jager et al., 2014; Lund et al., 2004; Mill and Petronis, 2007, 2008; Mill et al., 2008; Nestler, 2014; Schanen, 2006). Studies have also found that epigenetic instability of important genomic regions may lead to increased

methylation variability in cancer, which also contribute to cancer heterogeneity (Phipson and Oshlack, 2014; Hansen et al., 2011a; Feinberg and Irizarry, 2010; Gervin et al., 2011; Jaffe et al., 2012a). A study examining DNA methylation profiles of 1,505 CpG sites of both normal tissues and tumorigenic tissues observed that there is little variation in the DNA methylation patterns of these normal tissues but greater methylation heterogeneity among tumors (Fernandez et al., 2012). Studies have successfully identified epigenetic field defects in breast cancer based on differential variability (DV) (Teschendorff et al., 2016a). Epigenetic field defects are notably DNA methylation alterations that usually occur in pre-cancer tissues (Slaughter et al., 1953) and are crucial in cancer research because of its potential usage in early cancer detection (Teschendorff et al., 2016a,b).

Bisulfite microarray and sequencing are two widely used technologies to quantify DNA methylation. Popular array technologies include Illumina Infinium HumanMethylation 27K, 450K and 850K EPIC BeadChips, which produce methylation β -values measuring the proportion of methylated intensities out of total intensities. Popular sequencing technologies include whole-genome bisulfite sequencing, reduced representation bisulfite sequencing and Agilent SureSelect Human Methyl-Seq (Methyl-seq), which generate either ratio of methylated intensities versus total coverage at each CpG site or number of methylated or unmethylated cytosine.

Methods to identify differentially methylated loci (DML) based on differences in mean methylation levels between two groups are well-studied (Akalın et al., 2012; Chen et al., 2013; Huang et al., 2013; Shen et al., 2012; Sun and Wang, 2012, 2013). As DNA methylation levels of neighboring CpG sites are strongly correlated (Eckhardt et al., 2006; Irizarry et al., 2008) when genomic regions with consecutive CpG sites are associated with cancers (Hansen et al., 2011a; Irizarry et al., 2009; Lister et al., 2009), methods to detect differentially methylated regions (DMRs) were also developed. However, existing DMR detection methods all focus on mean signals only, and can be generally grouped into three types, to detect site-level signals first and then group adjacent loci into regions using ad hoc grouping rules (Hansen et al., 2012; Jaffe et al., 2012b; Butcher and Beck, 2015; Jühling et al., 2016; Hesse et al., 2015; Wen et al., 2016); or to define regions first and then test the significance of the defined regions (Ayyala et al., 2015; Sofer et al., 2013; Yip et al., 2014; Mayo et al.,

2014); or to use hidden Markov model that assumes three latent methylation states: hypermethylation, hypo-methylation and no differential methylation, and then group adjacent sites with the same state into a region (Saito and Mituyama, 2015; Saito et al., 2014). For array data, for example, bumpHunter (Jaffe et al., 2012b) uses surrogate variable analysis to account for potential batch effects, smoothes site-level signals within a predefined window and defines regions, as adjacent CpG sites with smoothed signals exceed a user-defined threshold. DMRcate (Peters et al., 2015) uses a tunable Gaussian kernel to smooth site-level differential methylation signals within a given window, then uses the method of Satterthwaite (Satterthwaite, 1946) to model the smoothed signals and group neighboring false discovery rate (FDR)-corrected significant CpG sites into regions. Probe Lasso (Butcher and Beck, 2015) uses a flexible window based on probe density to gather neighboring significant signals to define DMR boundaries. For Bisulfite sequencing data, for example, metilene (Jühling et al., 2016) uses a binary segmentation algorithm to identify candidate DMRs and then use a two-dimensional Kolmogorov-Smirnov (KS) test to assess the significance of candidate DMRs. MethylKit (Akalin et al., 2012) applies logistic regression to predefined regions after normalizing the read coverage across samples. Specific Methylation Analysis and Report Tool (SMART) (Liu et al., 2015) is an entropy-based framework that first calculates Tukey biweight to quantify methylation specificity at each CpG site, then uses specificity state, Euclidean distance-based methylation similarity, entropy-based methylation similarity and minimum distance requirement to indicate whether methylation patterns of two neighboring CpG sites are similar and then determines DMRs. All of these existing DMR detection methods use mean signals only. In addition to differential methylation, which refers to the difference between mean methylation measures between experimental groups, methods to identify DV, that is, experimental groups differ in terms of methylation variances, were also developed (Phipson and Oshlack, 2014; Teschendorff et al., 2016a; Ahn and Wang, 2013; Chen et al., 2014; Teschendorff et al., 2014; Teschendorff and Widschwendter, 2012; Ruan et al., 2016). We recently developed NEpiC and pETM methods where NEpiC is a network-based method that combines both mean and variance signals with a much improved power in searching for differentially methylated subnetworks using the protein-protein interaction network (Ruan et al., 2016); pETM is a penalized Exponen-

tial Tilt Model that detects both methylation mean and variance signals at CpG site level with the network-based regularization considering correlations among nearby CpGs (Sun et al., 2017). The study that identified epigenetic field defects in breast cancer through comparing DNA methylation levels in normal tissues adjacent to tumors (normal-adjacent) as a surrogate of pre-cancer tissues with those in normal tissues from healthy individuals would have erroneously concluded that there are no significant epigenetic field defects in breast cancer had the authors used a statistical method based on differential methylation only (Teschendorff et al., 2016a). The authors also observed increased variation in the normal-adjacent tissues driven by a relatively small number of outlier samples exhibiting much-different methylation values from the rest of the normal-adjacent samples (Teschendorff et al., 2016a), when conventional methods focused on mean signals are not able to detect such epigenetic alterations. On region levels, a new method that incorporates DV is needed, especially in detecting epigenetic field defects.

Here, we developed a new DMR detection method that uses combined signal from differential methylation and DV. Simulation studies showed the great performance of the new method. We further demonstrated the performance of the new method through applications to 450K DNA methylation data of tumor and normal-adjacent tissues of breast invasive carcinoma (BRCA) and kidney renal clear cell carcinoma (KIRC) from The Cancer Genome Atlas (TCGA) project, where some cancer-related genes were missed by the DMR detection methods that use only mean signals or variance signals. By applying the new method to an independent 450K DNA methylation data of BRCA tumor and normal-adjacent tissues from Gene Expression Omnibus (GEO), we concluded that DMRs detected using variance signals are reproducible. Further applications to the GEO DNA methylation data comparing normal-adjacent tissues from breast cancer patients and normal tissues from age-matched cancer-free women and comparing tumor tissues from breast cancer patients to normal tissues from age-matched cancer-free women not only identified epigenetic field defects in breast cancer but also confirmed that the epigenetic field defects are enriched in the progression to breast cancer (Teschendorff et al., 2016a). Importantly, the epigenetic field defects were only identified by the developed new DMR detection method that uses mean and variance combined signals.

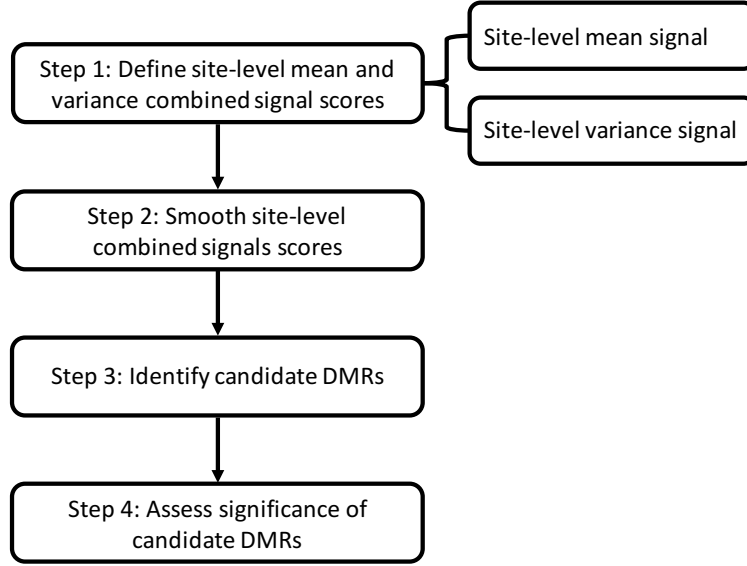


Figure 2.1: The pipeline of the proposed new DMR detection algorithm.

2.2 Methods

As matched case-control study designs with tumor and normal-adjacent tissues are widely used in DNA methylation studies of cancer; here, we focused on studies with a matched case-control design. The proposed new DMR detection method can be easily adapted to other types of designs. There are four steps in the new method (Figure 2.1): (1) define site-level mean and variance combined signal scores; (2) smooth site-level combined signal scores; (3) identify candidate DMRs; and (4) assess significance of candidate DMRs.

Step 1: Define site-level mean and variance combined signal scores

We define mean and variance combined signal score S_i for CpG site i as follows:

$$S_i = \frac{|T_{mi}|}{T_{mi}} (\lambda m_i + (1 - \lambda) v_i) \quad (2.1)$$

where $m_i = \Phi^{-1}(1 - p_{mi})$ and $v_i = \Phi^{-1}(1 - p_{vi})$. Here, Φ is standard normal quantile function, and p_{mi} and p_{vi} are P -values from the two-sided paired t -test testing if the mean methylation measures are the same between tumor and normal-adjacent tissues and from the one-sided Pitman-Morgan test testing if the variance of the methylation measures in

tumor tissues is greater than that in normal-adjacent tissues at CpG site i (Morgan, 1939; Pitman, 1939), respectively. We set mean and variance signal scores that are smaller than zero (i.e. sites with $p_{mi} > 0.5$ and $p_{vi} > 0.5$) as zero and remove sites whose mean and variance signal scores are both zero. Here, T_{mi} is the test statistic from a paired t -test, where $\frac{|T_{mi}|}{T_{mi}}$ adds a sign to the combined signal score to indicate whether CpG site i is hyper-methylated (positive sign) or hypo-methylated (negative sign). Similarly, as in our previous work (Ruan et al., 2016), because of potential different scales of the site-level mean and variance signal scores m_i and v_i , we weight the two scores by λ and $1 - \lambda$, respectively, to balance the contribution of mean and variance signals to the combined score. We first define the site-level scaling parameter:

$$\lambda_i = \frac{v_i}{m_i + v_i} \quad (2.2)$$

At CpG site i , we then average across all sites from the whole genome to obtain the overall scaling parameter λ .

Step 2: Smooth site-level combined signal scores

As methylation levels of CpG sites within 1,000 base pairs (bps) are considered highly correlated (Eckhardt et al., 2006; Sofer et al., 2013), we assign any two neighboring CpG sites into the same cluster if the genomic distance between them is $< 1,000$ bps. We examined how neighboring CpG sites with different distance limits differ in the combined signal scores and summarized results in the Supplementary Data (Section A.1 Investigation of the distance limits to define clusters). We then smooth site-level combined signal scores within a defined cluster using the running median method with a window size of minimum of W sites. The running median method was chosen over the moving average (Wu et al., 2015) method because of its robustness to outliers. It was chosen over regression-based smoothing methods (Hansen et al., 2012; Jaffe et al., 2012b), which have been shown to have similar performance as the moving averaging method (Wu et al., 2015) because of its computational efficiency. After smoothing, we denote the smoothed combined signal score for CpG site i as \tilde{S}_i .

Step 3: Identify candidate DMRs

A candidate DMR is defined to be a region having at least L consecutive CpG sites of the same sign with $|\tilde{S}_i| > k$, where L is a predefined number, and k is a predefined threshold, e.g. the region size to be $L \geq 3$ CpG sites and the threshold to be $k = 99^{\text{th}}$ percentile of genome-wide $|\tilde{S}_i|$. Similar criteria of k (Hansen et al., 2012; Jaffe et al., 2012b; Wu et al., 2015; Hebestreit et al., 2013) and L (Hansen et al., 2012; Wu et al., 2015) were used in other DMR detection methods.

Step 4: Assess significance of candidate DMRs

We use permutation procedures to assess the significance of candidate DMRs, where we use the following measure to evaluate the strength of evidence for the j^{th} candidate DMR R_j : $A_j = \sum_{i \in R_j} |\tilde{S}_i|$. To assess the significance of the candidate DMR R_j via a permutation procedure under the global null hypothesis adjusting for multiple comparisons, we first shuffle tumor and normal-adjacent status within all pairs and then apply Steps 1-3 to the permuted data set. For the g^{th} permutation that generates n_g regions, we have the evidence of strength for each region as follows: $A_{\text{perm}_g, t}, t = 1, \dots, n_g$. We repeat the permutation procedure 1,000 times, and the empirical P -value of the candidate DMR R_j is calculated as:

$$P_j = \frac{\sum_{g=1}^{1000} \sum_{t=1}^{n_g} I(A_{\text{perm}_g, t} > A_j)}{\sum_{g=1}^{1000} n_g} \quad (2.3)$$

To account for multiple comparisons, we calculate the family-wise error rate (FWER) for the candidate DMR R_j as the proportion of permutations with $\max_{t \in [1, n_g]} (A_{\text{perm}_g, t}) > A_j$. The candidate DMR R_j is then considered to be significant if its $\text{FWER} \leq 0.05$.

The new method outputs a table of candidate DMRs with detailed information of each candidate DMR R_j : (1) chromosome location, (2) genomic locations of the first and last CpG sites, (3) strength of evidence A_j , (4) number of CpG sites, (5) unadjusted P -value P_j and (6) FWER. Users could also output intermediate results such as mean signal scores and variance signal scores of CpG sites before smoothing as an option.

2.3 Comparison methods

We compared the performance of the new method that combines mean and variance signals with those of the DMR detection methods that (1) consider mean signals only including the adapted bump hunting algorithm using two-sided paired t -test, DMRcate, Probe Lasso and the adapted new method with the test statistic from Wilcoxon signed-rank test as the nonparametric version of the mean signals, (2) variance signals only which is the adapted bump hunting algorithm using one-sided Pitman-Morgan test and (3) both mean and variance signals (the adapted new method with test statistic from KS test).

2.4 Simulation study

We conducted simulation studies to evaluate type I errors and the performance of the new method. We define type I errors as the proportions of simulations identified any significant DMRs when data are generated with no DMRs. We use receiver operating characteristic (ROC) curves to evaluate the performance of the new method where we define true positive as significant DMRs with any CpG sites that are in the true DMRs and false positive as significant DMRs with no CpG sites from the true DMRs.

2.4.1 Simulation setup

To simulate methylation measures for tumor and normal-adjacent tissues, we considered 1:1 matched study design with one tumor sample ($Y = 1$) and one normal-adjacent sample ($Y = 0$) on the matching variable Z . Given Y and Z , we assume logit2 transformed methylation measures (Du et al., 2010) X follows a conditional scaled normal distribution:

$$\begin{aligned} X|Y = 1, Z = z &\sim \sqrt{z}N(\mu, \Delta^T\Sigma\Delta) \\ X|Y = 0, Z = z &\sim \sqrt{z}N(0, \Sigma) \end{aligned}$$

where the matching variable $Z \sim Beta(a, b)$ and Σ is a variance-covariance matrix considering correlations among CpG sites within a predefined cluster. The mean vector $\mu = (\mu_1, \dots, \mu_h)^T$ and diagonal matrix $\Delta = \text{diag}(\sqrt{\delta_1}, \dots, \sqrt{\delta_h})$ control the mean and variance signals in a cluster of h consecutive sites. Here, we assume an $AR(1)$ correlation with correlation coefficient ρ , i.e. $\Sigma_{mn} = \sigma \times \rho^{|m-n|}$. We set $\rho = 0.5$ and $Z \sim Beta(1, 1)$ in simulation

studies based on our previous experience (Sun and Wang, 2013). In each simulation, we generated X of 10,000 CpG sites from 100 tumor and normal-adjacent pairs, where the genomic locations of these 10,000 sites are the first 10,000 sites of Chromosome 1 on the Illumina 450K array.

To evaluate type I errors, we set $\mu = 0$, $\sigma = 0.3$, where σ was estimated using methylation measures of the normal-adjacent tissues of the TCGA BRCA data. To evaluate the performance of the new method, we simulated 10 true DMRs with different sizes, varying from 3 to 15 CpG sites, and we considered scenarios when each CpG site in the true DMRs has (1) mean signals only, (2) variance signals only and (3) both mean and variance signals. For all other null CpG sites, we set $\mu = 0$ and $\sigma = 0.3$. For each simulation scenario, we conducted 1,000 simulations. In all simulation studies and real data applications, we defined the region size to be $L \geq 3$ CpG sites.

2.4.2 Adaption to case-control designs

We adapted the proposed new DMR detection method for case-control designs, which can adjust for relevant covariates. More specifically, we fit a linear regression model on logit2 transformed methylation β -values, M -values, adjusting for known confounders such as age and gender, and cell composition if necessary, and work on residuals in all subsequent steps. We conducted simulation studies parallel as for matched case-control designs to evaluate the type I errors and the performance. The simulation setup and results are summarized in the Supplementary Data (Section A.2 Simulation studies for case-control designs).

2.5 Results

2.5.1 Simulation results

Type I errors are all well controlled at the 0.05 significance level with values 0.055, 0.046, 0.050, 0.041 and 0.041 for the new method, DMR methods based on paired t -test, Pitman-Morgan test, Wilcoxon signed-rank test and KS test, respectively, while that for DMRcate and Probe Lasso are much more conservative with values 0.015 and 0, respectively.

For the ROC curve results (Figure 2.2), when the significance threshold was set from 0

to 0.05, we notice that when the true DMRs are set to have sites with mean signals only, the new method performs slightly inferior to paired t -test and similarly to KS test, and much better than the Wilcoxon signed-rank test, while Pitman-Morgan test that considers variance signals only could not detect any true DMRs. On the other hand, DMRcate appears to perform better than the new method with higher true-positive rates and zero false-positive rates. This is because DMRcate uses Stouffer transformation (Stouffer et al., 1949) of the limma-derived FDRs for individual CpG sites constituting a DMR to assess the overall significance of the DMR, which in general is much smaller than the P -values by the new method assessing significance of candidate DMRs via 1,000 permutations. We also noticed that DMRcate may not be able to identify regions with small effect sizes comparing with t -test (with true-positive rates up to around 6 of the 10 regions with signals, while true-positive rates for t -test could be up to around 8); Probe Lasso also has small false-positive rates, but the true-positive rates are smaller than that of the new method, and it also uses Stouffer's method to combine weighted individual P -values, and thus also leads to much smaller P -values for DMRs compared with the new method. Similarly, when the true DMRs are set to have sites with variance signals only, the new method performs slightly inferior to Pitman-Morgan test that considers variance signals only while all other five comparison methods could not detect any true DMRs. When the true DMRs are set to have sites with both mean and variance signals, the new method performs much better than all of the six comparison methods.

The type I errors and ROC curve results of the adapted algorithm are summarized in the Supplementary Data (Section A.2 Simulation studies for case-control designs).

2.5.2 Real data application

We used two data sets, TCGA BRCA data (tumor and normal-adjacent pairs) and GEO BRCA data (tumor and normal-adjacent pairs, normal controls from age-matched cancer-free women), to demonstrate the performance of the new method from three aspects: (1) identification of DMRs associated with tumor and normal-adjacent status (TCGA BRCA tumor versus normal-adjacent comparison; and GEO BRCA tumor versus normal-adjacent comparison); (2) replication with two independent BRCA data (TCGA BRCA tumor versus

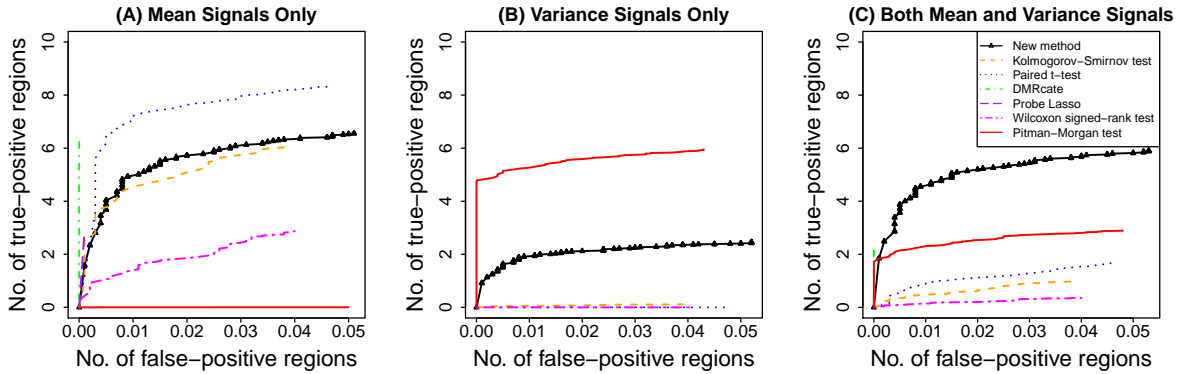


Figure 2.2: ROC curves from simulation studies where 10 true DMRs have different region sizes ranging from 3 to 15 CpG sites with (A) mean signals only; (B) variance signals only; and (C) both mean and variance signals. DMRs were defined as regions with minimum region size of $L \geq 3$ CpG sites.

normal-adjacent comparison; and GEO BRCA tumor versus normal-adjacent comparison); (3) identification of epigenetic field defects (GEO age-matched cancer-free versus normal-adjacent comparison); (4) enrichment of epigenetic alterations from age-matched cancer-free to normal-adjacent to tumor tissues (GEO age-matched cancer-free versus normal-adjacent comparison, GEO BRCA tumor versus normal-adjacent comparison and GEO age-matched cancer-free versus tumor comparison).

2.5.3 TCGA BRCA data

We applied the new method and the six comparison methods to the TCGA BRCA 450K DNA methylation data of tumor and normal-adjacent tissues. The original data have DNA methylation measures on 485,577 CpG sites for 96 tumor and normal-adjacent pairs. We conducted standard quality control steps where we removed sites on sex chromosomes and sites overlap with known single-nucleotide polymorphisms (SNPs). We also required at least 95% CpG coverage per sample and 70% sample coverage per CpG sites. We ended up with 326,105 CpG sites for 90 matched tumor and normal-adjacent pairs. We then corrected for the type II probe bias using the ‘watermelon’ package (Pidsley et al., 2013).

We found that DMRs identified by the Wilcoxon signed-rank test and KS test are larger than others (both in terms of number of sites and bps) in general, while those by the mean-

Table 2.1: Significant DMRs identified in the TCGA BRCA data

DMRs ($L^a \geq 3$)	New method	Wilcoxon		paired t -test	DMRcate	Probe Lasso	Pitman-Morgan test	KS test ^b
		signed-rank test	test					
Total number of DMRs (total number of DMR-covered CpG sites)	986 (18,654)	135 (4,295)	1,473 (21,777)	20,657 (133,410)	7,190 (36,936)	610 (13,208)	720 (16,047)	
Mean (SD) number of CpG sites per DMR	19 (8)	32 (8)	15 (8)	6 (5)	5 (5)	22 (10)	22 (11)	
Mean (SD) number of base pairs per DMR	3,373 (1,989)	5,341 (2,387)	2,669 (1,726)	1,185 (1,064)	748 (1,148)	3,804 (2,361)	4,144 (2,492)	
Number of overlapping DMRs ^c	-	129	806	986	642	533	588	

^a L : minimum region size, i.e., minimum number of CpG sites.

^bKolmogorov-Smirnov test.

^cNumber of overlapping DMRs: a DMR identified by the new method is considered to overlap with DMRs identified by each comparison method if there is any overlap.

Table 2.2: Eleven genes identified in the top 10 ranked DMRs in TCGA BRCA data

Cancer	Gene
Breast cancer	<i>LBH</i> (Many and Brown, 2010)
Chordomas	<i>NPR3</i> (Alholle et al., 2015)
Clear cell renal cell carcinoma	<i>SMPD3</i> (Wang et al., 2015)
Gastric cancer	<i>FGF19</i> (Zhao et al., 2013)
Head and neck squamous cell carcinomas	<i>PHF21B</i> (Bertonha et al., 2015)
Hepatocellular carcinoma	<i>MYADM</i> (Song et al., 2013), <i>DBX2</i> (Zhang et al., 2013c)
Non-small cell lung cancer	<i>KCNC3</i> (Lokk et al., 2012)
Oral squamous cell carcinoma	<i>ATP8B2</i> (Yong-Deok et al., 2015)
Prostate cancer	<i>AQP10</i> (Raza and Jaiswal, 2013), <i>STEAP2</i> (Gomes et al., 2012)

only method are the smallest (paired t -test, DMRcate and Probe Lasso), and those by the new method and Pitman-Morgan test are in between (Table 2.1). On the CpG site level, 69.2, 13.1, 15.4 and 93.6% of sites in the DMRs that were identified by the mean-only methods: paired t -test, DMRcate, Probe Lasso and Wilcoxon signed-rank test were also identified by the new method; 83.6% of sites in the DMRs identified by the Pitman-Morgan test were also identified by the new method, and 77.4% of sites in the DMRs identified by the KS test were also identified by the new method. Further investigation reveals that DMRs identified uniquely by the paired t -test and Pitman-Morgan test were all defined by the new method but did not reach significance.

When comparing DMRs identified by the new method to those by the six comparison methods, the new method did not identify any unique DMRs. All DMRs identified by the new method overlap with those identified by DMRcate, where we define overlap if any CpG sites in a DMR identified by the new method are also in a DMR identified by DMRcate. When comparing DMRs identified by the new method to those by the five of the six comparison methods but not DMRcate, the new method uniquely identified 22 DMRs. Among these 22 DMRs, we further examined the top 10 DMRs ranked by the evidence of strength of each region. There are 11 genes in these 10 DMRs, and all were previously reported to be associated with cancer (Table 2.2). We plotted the top ranked #1

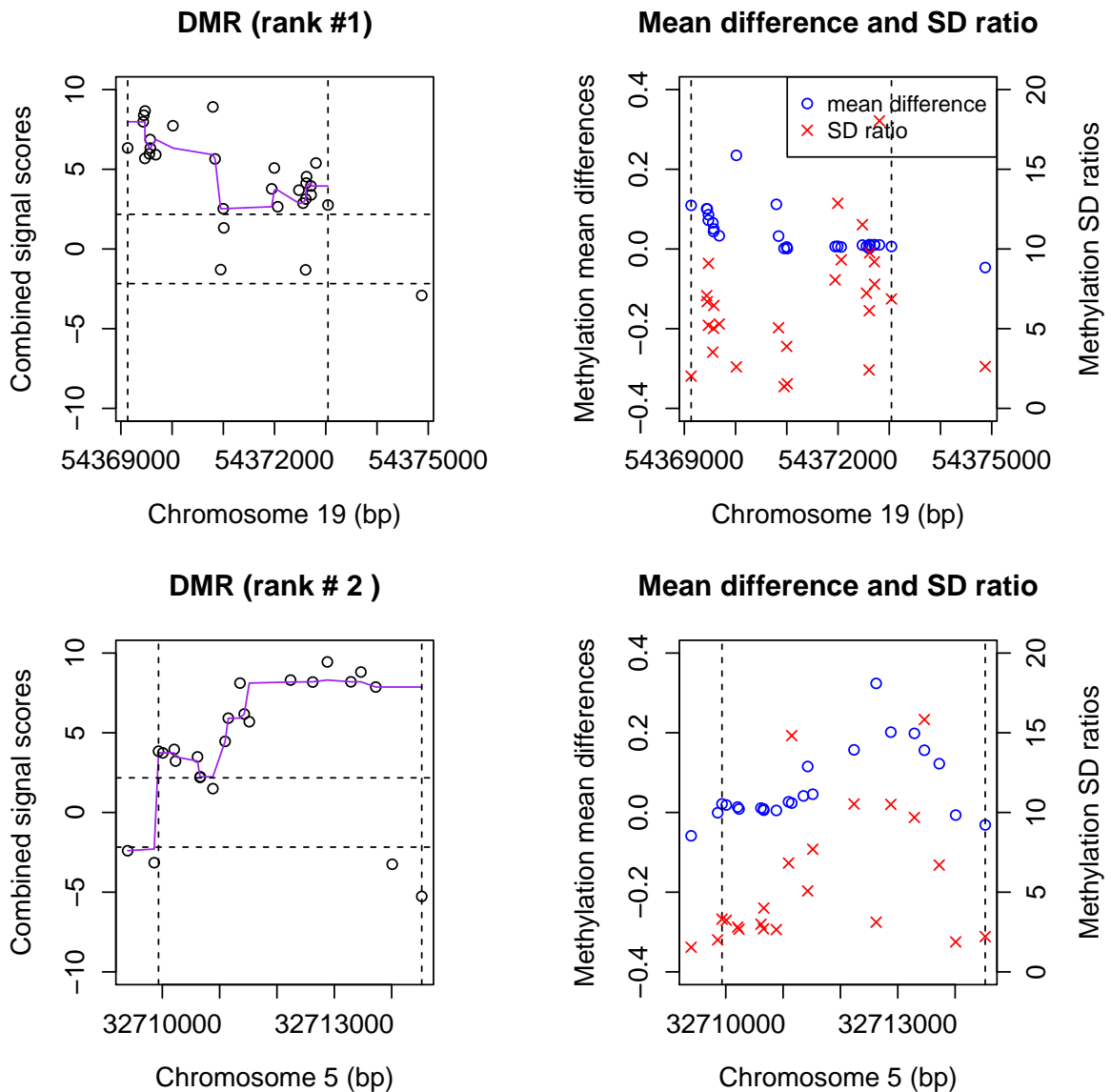


Figure 2.3: Top two ranked DMRs uniquely identified by the new method in the TCGA BRCA data. DMR #1 (top row) and #2 (bottom row) are located on chromosomes 19 and 5. Vertical dashed lines define boundaries of DMRs. Left column shows the combined signal scores of the sites in the DMRs before (circles) and after (curves) smoothing, in which horizontal dotted lines define the threshold k that defines a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the DMRs comparing tumor and normal-adjacent tissues.

and #2 DMRs out of the 22 uniquely identified DMRs for illustration (Figure 2.3), where both DMRs were hyper-methylated. In DMR #1, sites in the second half of the region do not have any mean differences between tumor and normal-adjacent tissues but have large variance differences. However, the variance signals are not strong enough to be detected by the variance-only method. For the sites in the first half of DMR #1, there are both mean and variance differences, but are not strong enough to be detected by most of the mean-only or variance-only methods.

We also applied the new method to the TCGA KIRC 450K DNA methylation data, and observed similar patterns as in the TCGA BRCA data. Results are included in the Supplementary Data (Supplementary Table A.1 and A.2 and Supplementary Figure A.3).

2.5.4 Replication analysis with GEO BRCA data

We performed a replication analysis using an independent DNA methylation data of BRCA tumor and normal-adjacent tissues from GEO (GSE69914) (Teschendorff et al., 2016a). The original GEO BRCA DNA methylation data have methylation measures on 385,184 CpG sites from 42 tumor and normal-adjacent pairs, 50 normal/benign controls from age-matched cancer-free women and 263 tumor tissues from independent breast cancer patients. We followed the same quality control steps as for the TCGA BRCA data and kept the same sets of CpG sites as in the TCGA BRCA data for comparison purpose. We ended up with 326,105 CpG sites from 42 tumor and normal-adjacent pairs.

We compared results from the TCGA BRCA data and the GEO BRCA data and found that 94.7, 94.4, 87.6, 87.2, 86.3, 80.2 and 95.4% of sites in the DMRs identified in the GEO BRCA tumor versus normal-adjacent comparison were also identified in the TCGA BRCA tumor versus normal-adjacent comparison by the new method, paired t -test, DMRcate, Probe Lasso, Wilcoxon signed-rank test, Pitman-Morgan test and KS test, respectively. We plotted two example overlapping DMRs (Figure 2.4). The first DMR is hypo-methylated and ranks #1 among all DMRs identified by the new method in both BRCA data sets. The second DMR is hyper-methylated and ranks #5 among all DMRs identified by the new method in the TCGA BRCA data and ranks #13 among all DMRs identified by the new method in the GEO BRCA data.

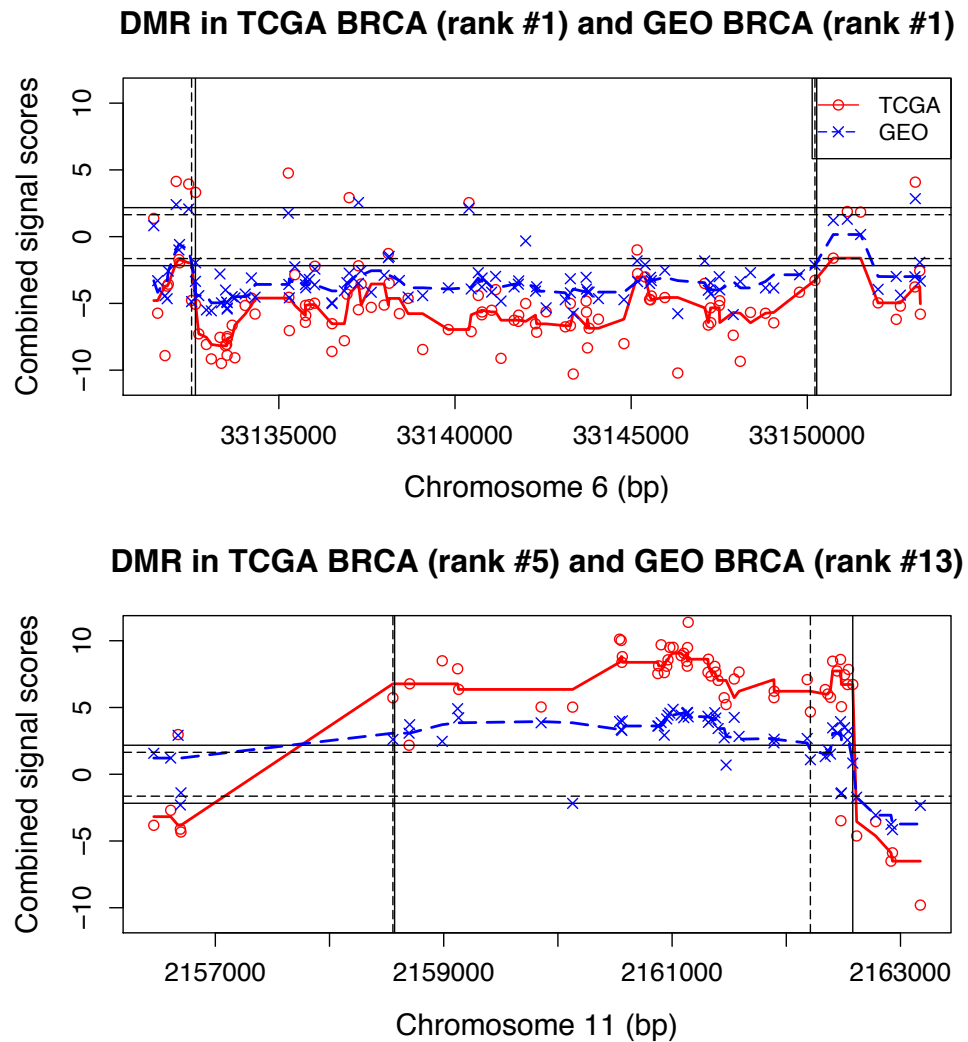


Figure 2.4: Two examples of overlapping DMRs among all DMRs identified by the new method in both the TCGA and the GEO BRCA data from the tumor versus normal-adjacent comparison. Plotted are the combined site-level signal scores in the DMRs before and after smoothing for the TCGA BRCA data (circles, solid curves) and the GEO BRCA data (crosses, dashed curves). Vertical lines define boundaries of DMRs, and horizontal lines define the threshold k that defines a candidate region.

Although sites in DMRs identified using the Pitman-Morgan test have the smallest replication rate (80.2%) as expected, it is still large enough to conclude that DMRs detected using variance signals are reproducible. Sites in DMRs identified in the two BRCA data sets using the new method, paired t -test and KS test can be almost perfectly reproduced with replication rates 94.7, 94.4 and 95.4%, respectively. This agrees with the general belief that DMR findings might be more reliable than DML findings, supporting the meaning of detecting DMRs.

Details of the DMRs identified by the new method and the six comparison methods in the GEO BRCA data are summarized in Table 2.3. We found that 67.5, 7.7, 10.7 and 86.8% of sites in the DMRs that were identified by the mean-only methods: paired t -test, DMRcate, Probe Lasso and Wilcoxon signed-rank test were also identified by the new method, 83.0% of sites in the DMRs identified by the Pitman-Morgan test were also identified by the new method and 81.9% of sites in the DMRs identified by KS test were also identified by the new method. Similarly as in the TCGA BRCA and KIRC data, DMRs identified uniquely by the paired t -test or Pitman-Morgan test were all defined by the new method but did not reach significance.

We similarly plotted the top #1 and #2 ranked DMRs in the GEO BRCA tumor versus normal-adjacent comparison (Supplementary Figure A.4) and further investigated genes in the top 10 DMRs (Supplementary Table A.3), where similar patterns were observed as in the TCGA BRCA and KIRC data.

2.5.5 Identification of epigenetic field defects in the GEO BRCA data

Teschendorff et al. (Teschendorff et al., 2016a) showed in their recent paper that the identification of early epigenetic alterations, commonly known as epigenetic field defects, through comparing DNA methylation measures of normal-adjacent tissues from breast cancer patients to normal tissues from age-matched cancer-free women is meaningful in the study of breast cancer development, and the differences are expected to be larger in comparisons between tumor and normal-adjacent tissues, and between tumor and normal tissues from cancer-free women. The original paper investigated the epigenetic field defects on the CpG site level. Here, we further investigated epigenetic field defects on the region level.

Table 2.3: Significant DMRs identified in the GEO BRCA data

DMRs ($L^a \geq 3$)	New method	Wilcoxon signed-rank test	paired t -test	DMRcate	Probe Lasso	Pitman-Morgan test	KS test ^b
Total number of DMRs (total number of DMR-covered CpG sites)	382 (8,769)	126 (3,948)	490 (9,606)	15,609 (104,713)	4,505 (24,053)	22 (653)	22 (653)
Mean (SD) number of CpG sites per DMR	23 (8)	31 (8)	20 (8)	7 (5)	5 (5)	30 (7)	26 (8)
Mean (SD) number of base pairs per DMR	4,047 (2,130)	5,216 (2,279)	3,420 (1,880)	1,127 (999)	713 (1,097)	3,718 (2,457)	4,478 (2,034)
Number of overlapping DMRs ^c	-	114	284	382	262	19	175

^a L : minimum region size, i.e., minimum number of CpG sites.

^bKolmogorov-Smirnov test.

^cNumber of overlapping DMRs: a DMR identified by the new method is considered to overlap with DMRs identified by each comparison method if there is any overlap.

We kept the same 326,105 CpG sites as in the GEO BRCA tumor versus normal-adjacent comparison.

We first examined the distributions of the estimated genome-wide site-level scaling parameter λ_i from the three comparisons (Supplementary Figure A.5) (1) normal-adjacent tissues from breast cancer patients versus normal tissues from age-matched cancer-free women, (2) tumor tissues versus matched normal-adjacent tissues from breast cancer patients and (3) tumor tissues from breast cancer patients versus normal tissues from age-matched cancer-free women in the GEO BRCA data.

As defined in Equation 2.2, the site-level scaling parameter λ_i reflects the relative strength of the mean and variance signals at CpG site i . CpG sites with $\lambda_i = 0$ do not have any variance signals, CpG sites with $\lambda_i = 1$ do not have any mean signals and CpG sites with $0 < \lambda_i < 1$ have both mean and variance signals, within which sites with $\lambda_i > 0.5$ have stronger variance signals than mean signals. Supplementary Figure A.5 suggests that in the normal-adjacent versus normal comparison, there are much fewer sites with both mean and variance signals and a lot more sites with only variance signals comparing with the other two comparisons. The parameter λ that reflects the genome-wide relative signal strength is also the largest in the normal-adjacent versus normal comparison. This suggests that differential variation exists earlier in disease progression, which is consistent with the findings by Teschendorff et al. (Teschendorff et al., 2016a) that there is increased variability in DNA methylation within the normal-adjacent tissues comparing with normal breast tissue from age-matched cancer-free women.

We then examined the identified DMRs in the three comparisons using the GEO BRCA data (1) normal-adjacent versus normal, (2) tumor versus normal-adjacent and (3) tumor versus normal. In the normal-adjacent versus normal comparison that aims for epigenetic field defects, the new method identified two DMRs (Supplementary Figure A.6 shows the mean and variance signals of the two DMRs), both hyper-methylated, while all the six comparison methods identified none. Importantly, all 58 CpG sites covered by these two DMRs of epigenetic field defects are also in the DMRs identified in the tumor versus normal-adjacent, and tumor versus normal comparisons (results summarized in Supplementary Table A.4 and Supplementary Figure A.7). These 58 sites cover two genes, *NKX6-2* and

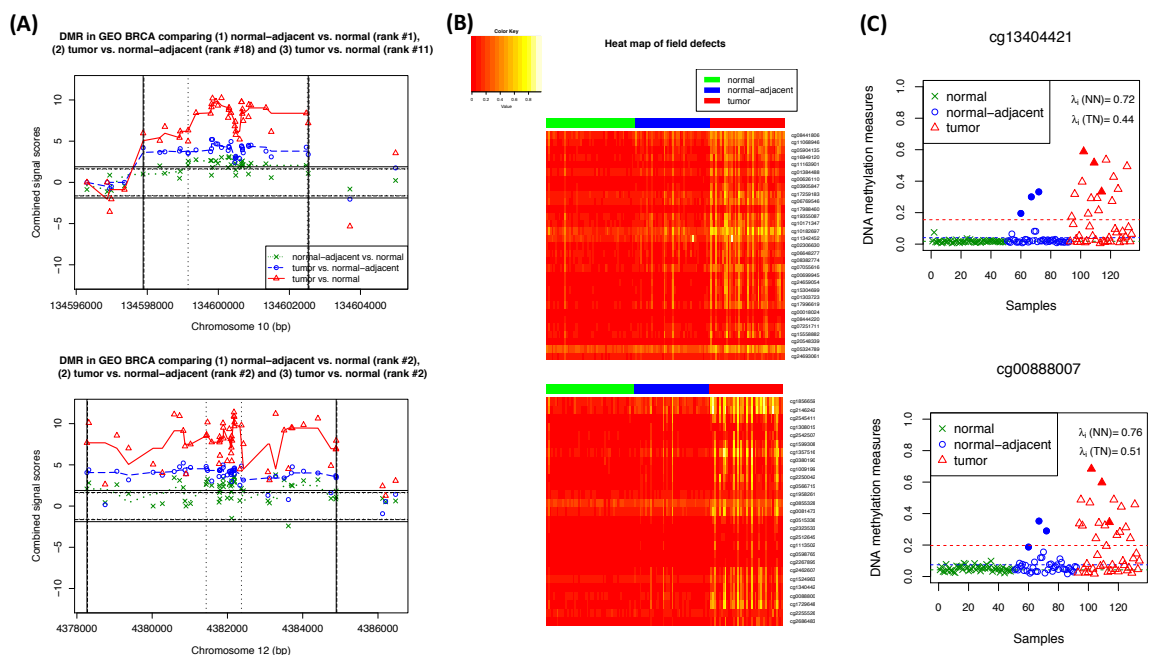


Figure 2.5: (A) The epigenetic field defects, i.e. the two DMRs identified in the GEO BRCA normal-adjacent versus normal comparison (crosses, dotted curves) together with the overlapping DMRs identified in the GEO BRCA tumor versus normal-adjacent comparison (circles, dashed curves), and the GEO BRCA tumor versus normal comparison (triangles, solid curves). Vertical lines define boundaries of DMRs, and horizontal lines define the threshold k that defines a candidate region. (B) Heat maps of the original DNA methylation measures of the sites from the epigenetic field defects, i.e. the two DMRs identified in the GEO BRCA normal-adjacent versus normal comparison. Green is for 50 normal tissues from age-matched cancer-free women. Blue is for 42 normal-adjacent tissues, and red is for 42 tumor tissues. (C) Two CpG sites selected of the 58 sites from the epigenetic field defects, i.e. the two DMRs. Plotted are the original DNA methylation measures of normal tissues from 50 age-matched cancer-free women (crosses), and normal-adjacent tissues (circles) and tumor tissues (triangles) from 42 BRCA patients. The three horizontal lines represent mean methylation levels of the three groups. $\lambda_i(NN)$ is the site-level scaling parameter from the normal-adjacent versus normal comparison, and $\lambda_i(TN)$ is that from the tumor versus normal-adjacent comparison. The three outlier samples were marked using solid circles (normal-adjacent tissues) and solid triangles (matching tumor tissues).

CCND2. Both were previously reported to be differentially methylated in breast cancer (Fackler et al., 2011; Sharma et al., 2007; Virmani et al., 2003). Moreover, for the two DMRs of epigenetic field defects, the #1 ranked DMR ranks #18 in the tumor versus normal-adjacent comparison and ranks #11 in the tumor versus normal comparison (Figure 2.5A); the #2 ranked DMR ranks #2 in the other two comparisons (Figure 2.5A). Sites in these two DMRs have larger combined signal scores in the tumor versus normal-adjacent, and tumor versus normal comparisons than those from the normal-adjacent versus normal comparison. This suggests that there exists epigenetic field defects earlier in disease progression, and the epigenetic field defects are enriched in the progression to breast cancer, confirming what is observed before on the CpG site level (Teschendorff et al., 2016a) using region-based method.

To further investigate whether the epigenetic field defects identified in the normal-adjacent versus normal comparison are because of a few outlier samples as Teschendorff et al. (Teschendorff et al., 2016b) noticed, and whether the notable DNA methylation alterations at the identified CpG sites are real but not technical artifact, we plotted two heat maps of sites in the two DMRs of epigenetic field defects (Figure 2.5B). It is clear that there is little variation in DNA methylation measures of normal tissues, and increased variation in those of normal-adjacent tissues because of three to five samples and much increased variability in those of tumor tissues. We further selected two sites with large variance signals ($\lambda_i = 0.72$ and 0.76) out of the 58 sites covered by the two DMRs and plotted their methylation measures (Figure 2.5C). It is clear that at these two sites, there is little variation in DNA methylation measures of normal tissues and increased variation in those of normal-adjacent tissues, mainly because of to three to five outlier samples, and there is no mean difference in DNA methylation measures between the normal and normal-adjacent tissues (Figure 2.5C). We also notice that the three outlier samples exhibit greater methylation deviations in tumor tissues than in normal-adjacent tissues, indicating an enriched methylation alteration with cancer progression (Figure 2.5C). We would like to emphasize that the two DMRs of epigenetic field defects in the normal-adjacent versus normal comparison were only identified by the new DMR detection method that uses mean and variance combined signals but were missed by all the other six comparison methods, which suggests

the great power achieved by the new DMR detection method.

2.6 Discussion

Here, we proposed a new DMR detection method that uses combined signals from differential methylation and DV. Simulation studies showed the correct type I error and the much improved power of the new method when true DMRs have sites with both mean and variance signals. Applications to the TCGA BRCA, TCGA KIRC and GEO BRCA DNA methylation data showed that the majority of genes in the uniquely identified DMRs by the new method were previously reported to be associated with cancers. Replication analysis results using two independent BRCA data sets suggest that DMRs detected with variance signals are reproducible.

Importantly, further application to the DNA methylation data of GEO BRCA normal-adjacent tissues from breast cancer patients and normal tissues from age-matched cancer-free women identified epigenetic field defects in two DMRs only by the new method, while the comparison mean-only and variance-only methods identified none. These two DMRs were also identified, and the methylation alterations were enriched in the comparisons of tumor versus normal-adjacent tissues and tumor versus normal tissues. The identified epigenetic field defects in these two DMRs could potentially be marks for breast cancer early detection with future investigations. Owing to the fact that the identified early DNA methylation alterations in breast cancer are characterized by increased variability because of a few ‘outlier’ samples when both mean and variance signals are weak and mean-only method and variance-only method could detect no differences, existing methods that focus on mean signals only or adapted methods that focus on variance signals only will be seriously underpowered. This shows the importance of using mean and variance combined signal, especially in identifying epigenetic field defects.

Although we did not consider correcting for differences in variances between batches, in the context of the data presented in this article, this is not an issue for the following two reasons (1) many previous studies (Teschendorff et al., 2016a; Teschendorff and Widschwendter, 2012; Teschendorff et al., 2012) have unequivocally demonstrated that most of

the differentially variable loci (DVL) are not batch or technical effects; (2) DVL are indeed generally characterized by fairly large changes in DNA methylation (>30% if not more) affecting a small number of samples, whereas batch effects generally involve smaller (10-15%) changes in DNA methylation, which affect most if not all the samples within a batch.

Furthermore, we did not adjust for cell-type composition in our analysis as Teschendorff et al. (Teschendorff et al., 2016a) have clearly demonstrated that DVL are not driven by changes in cell-type composition: (1) the DVL do not map to markers of adipose cells or immune cells, which are two main types of cell contaminants in breast tissue, (2) changes in cell-type composition between two phenotypes (e.g. normal versus normal-adjacent, or normal versus tumor) only involve relatively smaller changes in DNA methylation (10-15%). In contrast, DVL generally involve much larger changes in DNA methylation (>30%), which only affect a smaller number of samples. Their previous study also demonstrated that (3) the same DVL were found after adjustment for changes in cell-type composition. Put together, it is clear that most of the DVL are unrelated to cell-type composition changes, and that they instead mark pre-cancerous cells on route to becoming cancerous.

One thing we noticed in using methylation variance signals is, when methylation M -values are used, the mean and variance signals may not be completely separated. We have conducted some simulation studies in our previous work and found that if only mean signals are designed in the M -values, there will be both mean and variance signals in β -values after the transformation (Sun et al., 2017).

In summary, we proposed a new DMR detection method that uses mean and variance combined signals. Although we applied the new method to multiple cancer data sets, the method can be applied to other complex diseases. We focused on methylation array data in this work, but the new method is readily applied to sequencing data with sequencing data being preprocessed to methylation proportions.

Chapter 3

Detection of Epigenetic Field Defects Using a Weighted Epigenetic Distance-Based Method

3.1 Introduction

Identifying molecular alterations that happen early in carcinogenesis, known as field defects, is important for early cancer detection. One common approach is to compare normal tissue of healthy individuals to normal tissue adjacent to tumor (normal-adjacent tissue) of cancer patients as a surrogate of pre-cancer tissue that are difficult to collect. There have been studies in identifying epigenetic field defects (Katsurano et al., 2012; Bernstein et al., 2013; Teschendorff et al., 2016a,b), notably early DNA methylation alterations. DNA methylation is an epigenetic modification that has been shown to be crucial in gene expression (Baylin et al., 2001; Fahrner et al., 2002; Jones, 2012; Phillips, 2008) and cancers (Das and Singal, 2004; Ehrlich, 2002; Esteller and Herman, 2002; Kulis and Esteller, 2010). There are mainly two types of aberrant DNA methylation in cancers, local hyper-methylation in promoter-related CpGs that leads to the silencing of down-stream tumor suppressor genes (Koukoura et al., 2014; Baylin, 2005; Curradi et al., 2002; Herman and Baylin, 2003; Robertson, 2005), and global hypo-methylation that leads to chromosome instability (Robertson, 2005; Eden

et al., 2003; Feinberg and Tycko, 2004; Jaenisch and Bird, 2003). Studies have successfully identified epigenetic field defects in breast cancer by comparing normal-adjacent tissue of breast cancer patients to normal tissue from healthy individuals. Teschendorff et al. identified epigenetic field defects in breast cancer based on differential variability (DV), i.e. variance signals in DNA methylation (Teschendorff et al., 2016a), using methylation site-level analyses. Our previous work (Wang et al., 2017) identified epigenetic field defects in breast cancer based on both differential methylation (DM), i.e. mean signals, and DV, using methylation region-level analyses. In both studies, epigenetic field defects were found to be mainly driven by increased variation in methylation due to several outlier normal-adjacent tissue samples.

Due to the fact that CpG site-level signals for epigenetic field defects may be very small, existing methods based on differences (DM or DV or both) on CpG site-level may not have good power. Standard epigenome-wide association studies (EWAS) that focus on mean signals (EWAS-DM) perform CpG site-level tests to identify differentially methylated CpGs between two experimental groups using standard tests such as a t -test, a regression-based test or its regularized versions (Tusher et al., 2001; Smyth, 2004; Wettenhall and Smyth, 2004), or a non-parametric Wilcoxon rank sum test (Wilcoxon, 1945). EWAS that focus on variance signals (EWAS-DV) perform CpG site-level tests to identify differential variation CpGs between two experimental groups using standard tests such as the F -test (Hansen et al., 2011b; Ho et al., 2008), the Bartlett’s test or its regularized version (Teschendorff et al., 2016a,b), or an empirical Bayes extension of the Levene’s test (Phipson and Oshlack, 2014). The F -test and Bartlett’s test are sensitive to departures from normality which is usually the case for methylation data, while the Levene’s test is more robust to non-normality. On the other hand, distance-based methods that characterize (dis)similarity between pairwise samples across a gene, a genetic region, a pathway or an entire genome have been proven to be powerful in genetic and gene expression studies (Zapala and Schork, 2006; Wessel and Schork, 2006; McArdle and Anderson, 2001; Anderson, 2001; Han and Pan, 2010). While standard EWAS perform CpG site-level tests with stringent multiple comparisons adjustment, in a gene or a genetic region level, the common practice using non-distance-based methods is to select the minimum P -value out of all CpGs in that region.

These methods will not be powerful when site-level effects are very small. Alternatively, the distance-based methods accumulate any CpG site-level signals from a gene or a genetic region via the (dis)similarity matrix thus boost the overall association power, making them the ideal methods for detection of epigenetic field defects.

Here, we developed a weighted epigenetic distance-based method to identify epigenetic field defects at gene or genetic-region levels using both DM and DV signals. CpG site-level weights were incorporated in the calculation of (dis)similarity matrix to further boost signals and reduce noises. Specifically, we used original DNA methylation measures to examine DM and centered quadratic methylation measures to examine DV and considered site-level weights based on strengths of site-level DM and DV signals. Simulation studies showed much improved performance of the proposed weighted epigenetic distance-based method over several comparison methods including non-weighted versions and methods that use either DM or DV signals as well as standard EWAS methods. We further demonstrated the performance of the proposed method through an application to the 450K DNA methylation data of normal-adjacent tissue of breast invasive carcinoma (BRCA) patients and normal tissue from independent age-matched cancer-free women from Gene Expression Omnibus (GEO). The proposed method that accumulates weighted DM and DV signals identified genes with epigenetic field defects that were missed by standard EWAS methods and non-weighted distance-based methods. Many of these epigenetic field defects were previously reported to be associated with breast cancer. Further examination confirmed their enrichment in the progression to breast cancer and replicated some of these identified epigenetic field defects.

3.2 Materials and methods

Case-control designs using normal tissue from healthy individuals ($Y = 0$) and normal tissue adjacent to tumor from cancer patients ($Y = 1$) as a surrogate of pre-cancer tissue are widely used to identify epigenetic field defects in cancers. We therefore focused on case-control designs and illustrated and applied the proposed weighted epigenetic distance-based method on gene level. However, the proposed method can be easily adapted to other types

of design and on genetic region or genome levels. There are three steps in the proposed distance-based method: (i) to define gene-level weighted epigenetic distance matrix; (ii) to calculate pseudo- F statistic and (iii) to assess statistical significance using permutations.

Step 1: Define gene-level weighted epigenetic distance matrix

Define epigenetic distance matrix. For each gene, let \mathbf{X}^m be a $2N \times n$ matrix with original DNA methylation measures for N cases and N controls of n CpG sites in a gene, where element x_{ij}^m harbors DNA methylation measure of the j -th CpG site, $j = 1, \dots, n$ in the gene, for the i -th subject, $i = 1, \dots, N$. This \mathbf{X}^m matrix will be used to examine differential methylation (DM) capturing methylation mean signals. Let \mathbf{X}^v be a $2N \times n$ pseudo-data matrix of variability score capturing methylation variance signals, which will be used to examine differential variability (DV). The element $x_{ij}^v = (x_{ij}^m - \bar{x}_j^m)^2$ harbors centered quadratic methylation measure of the same j -th CpG site for the i -th subject. Here $\bar{x}_j^m = \frac{1}{N} \sum_{i=1}^N x_{ij}^m$ is the mean methylation measure of the j -th CpG site across N cases and N controls separately. The quadratic terms are centered to better capture variance signals. By using $\mathbf{X}^{mv} = [\mathbf{X}^m, \mathbf{X}^v]$, a $2N \times 2n$ matrix, we will be able to capture both methylation mean and methylation variance signals of the n CpG sites. Before constructing the epigenetic distance between any pair of subjects, we performed normalization on each column of \mathbf{X}^{mv} such that each column has mean zero and unit standard deviation.

We define the $2N \times 2N$ epigenetic distance matrix \mathbf{D}^{DM-DV} with element d_{st}^{DM-DV} that captures dissimilarities between any given pair of individuals s and t , $s, t = 1, \dots, 2N$ as

$$d_{st}^{DM-DV} = \sqrt{\sum_{j=1}^n \left\{ \frac{1}{2n} (x_{sj}^m - x_{tj}^m)^2 + \frac{1}{2n} (x_{sj}^v - x_{tj}^v)^2 \right\}} \quad (3.1)$$

Incorporate CpG site-level weights into epigenetic distance matrix. We construct CpG site-level weights aiming to up-weight signal CpGs (mean or variance) and to down-weight noise CpGs in calculating distances between pairs of subjects. Therefore, we define weights for mean and variance signals at CpG site j as follows:

$$w_j^m = \frac{-\log_{10}(p_j^m)}{\sum_{j=1}^n -\log_{10}(p_j^m)}, \quad w_j^v = \frac{-\log_{10}(p_j^v)}{\sum_{j=1}^n -\log_{10}(p_j^v)} \quad (3.2)$$

where p_j^m and p_j^v are the P -values from the two-sided two-sample t -test testing if the mean methylation measures are the same between cases and controls and from the one-sided Levene's test testing if the variance of the methylation measures in cases is greater than that in controls at CpG site $j, j = 1, \dots, n$ in a gene. Note that $\sum_{j=1}^n w_j^m = \sum_{j=1}^n w_j^v = 1$.

The corresponding $2N \times 2N$ weighted epigenetic distance matrix $\mathbf{D}^{w-DM-DV}$ with element $d_{st}^{w-DM-DV}$ that captures weighted dissimilarities between individuals s and $t, s, t = 1, \dots, 2N$ can be defined as

$$d_{st}^{w-DM-DV} = \sqrt{\sum_{j=1}^n \left\{ \frac{w_j^m}{2} (x_{sj}^m - x_{tj}^m)^2 + \frac{w_j^v}{2} (x_{sj}^v - x_{tj}^v)^2 \right\}} \quad (3.3)$$

Step 2: Calculate pseudo- F statistic

We apply distance-based regression originally developed in the field of ecology (McArdle and Anderson, 2001; Anderson, 2001) to test if DNA methylation measures in a gene is associated with the case-control status. Specifically, for each gene, we calculate a pseudo- F statistic based on the weighted epigenetic distance matrix $\mathbf{D}^{w-DM-DV}$ introduced above

$$F^{w-DM-DV} = \frac{tr(\mathbf{HGH})}{tr[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})]} \quad (3.4)$$

where $\mathbf{H} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ is a $2N \times 2N$ projection matrix, \mathbf{Y} is a $2N \times 1$ vector with case ($Y = 1$) and control ($Y = 0$) status, $\mathbf{G} = (\mathbf{I} - \frac{1}{2N}\mathbf{1}\mathbf{1}^T)\mathbf{A}(\mathbf{I} - \frac{1}{2N}\mathbf{1}\mathbf{1}^T)$ is the Gowers centered matrix, $\mathbf{A} = (a_{st}) = \left(-\frac{1}{2}(d_{st}^{w-DM-DV})^2 \right)$, $\mathbf{1}$ is a $2N$ -dimensional column vector with elements 1, and \mathbf{I} is a $2N \times 2N$ identity matrix. The pseudo- F statistic is used to evaluate the association between epigenetic distances of a gene with n CpG sites and the case/control status.

Step 3: Assess statistical significance using permutations

To assess significance of all G genes tested, we use permutation procedures, where we randomly shuffle cases ($Y = 1$) and controls ($Y = 0$) and repeat Steps 1-2 on the permuted data. In order to have more granular P -values, we pool pseudo- F statistics of all G genes from all permutations, as well as those from the observed data, to compute the empirical

P -value (Friedman et al., 2001). We repeat the permutation procedure 999 times, and calculate the empirical P -value for gene g , $g = 1, \dots, G$, as follows:

$$P_g^{\text{w-DM-DV}} = \frac{\sum_{g'=1}^G \{1 + \sum_{\text{perm}=1}^{999} I(F_{g',\text{perm}}^{\text{w-DM-DV}} \geq F_g^{\text{w-DM-DV}})\}}{G \times (1 + 999)} \quad (3.5)$$

In the real data application, we have $G=19,271$ genes, which helps to have high resolution gene-level empirical P -values.

To investigate if genes with different sizes, i.e., number of CpGs, will have different distributions for pseudo- F statistics under the null hypothesis, we conducted simulation studies to compare the type I error rates when the P -value for each gene is calculated based on pooled pseudo- F statistics of all G genes across all permutations (Supplementary Section B.1.1).

3.2.1 Comparison methods

We compare the performance of the proposed method $\mathbf{D}^{w-DM-DV}$ that considers site-level weights for mean and variance signals to that of several comparison methods, including the weighted distance-based methods that consider mean signals only \mathbf{D}^{w-DM} or variance signals only \mathbf{D}^{w-DV} , and distance-based methods without weights that consider both mean and variance signals \mathbf{D}^{DM-DV} , mean signals only \mathbf{D}^{DM} , variance signals only \mathbf{D}^{DV} , and standard EWAS methods on each CpG site with multiple comparisons adjustment of number of CpGs in a gene based on mean signals EWAS^{DM} or variance signals EWAS^{DV}.

3.2.2 Simulation study

We conducted simulation studies to evaluate type I error rate and power of the proposed method $\mathbf{D}^{w-DM-DV}$ and those of the comparison methods described above. Type I error rate is defined as the proportion of simulations with any significant genes when the data is generated under the null hypothesis of no genes are associated with case-control status. Power is defined as the proportion of simulations with any significant genes when the data is generated under the alternative hypothesis.

3.2.3 Simulation setup

We simulated methylation measures X of cases ($Y = 1$) and controls ($Y = 0$) at every CpG site in a gene from beta distributions:

$$X|Y = 0 \sim \text{Beta}(a_0, b_0)$$

$$X|Y = 1 \sim \text{Beta}(a_1, b_1)$$

where shape parameters a_0 and b_0 for samples in the control group were chosen based on estimates from the 50 normal tissue samples from cancer-free women in the GEO BRCA data (GSE69914), and shape parameters a_1 and b_1 for samples in the case group were chosen based on estimates from the 42 normal-adjacent tissues in the GEO BRCA data. More specifically, the average of the methylation means and standard deviations (SDs) of all CpG sites with gene information for the 50 normal tissue samples is 0.47 and 0.05, respectively. Therefore, we set $a_0 = 46.36$ and $b_0 = 52.28$ for noise CpGs such that the corresponding mean and SD of the beta distribution are 0.47 and 0.05, respectively. We generated methylation measures for 40 cases and 40 controls to mimic the size of the GEO BRCA study. We set $a_1 = a_0$ and $b_1 = b_0$ for all CpG sites in case and control groups to evaluate type I error rates. For power scenarios, we considered scenarios when signal CpGs have different mean or variance signals through varying shape parameters a_1 and b_1 . We conducted 1,000 simulations in each simulation setting.

3.2.3.1 Simulation settings with one gene

We first considered one gene with different number of CpGs with different signal-to-noise ratios of the CpGs. That is, the ratio between number of signal CpGs and number of noise CpGs in this gene ranges from 1:0, 1:24, 1:49, 3:47, to 5:45. We considered scenarios when signal CpGs have different mean or variance signals by varying shape parameters a_1 and b_1 such that the mean differences in methylation measures between cases and controls are 0.02, 0.04, 0.06, 0.08 and 0.1, and the ratios of SDs for cases and controls are 1.25, 1.50, 1.75, 2, 2.25 and 2.50, respectively. The values of a_1 , b_1 in those scenarios and the corresponding effect sizes are summarized in the Supplementary Table B.2. We consider a gene to be significant at the 0.05 significance level.

3.2.3.2 Simulation settings with 10 genes

We then considered 10 genes with one gene having signals when there are 25 CpGs in each of the 10 genes. In the signal gene, we set one CpG to have mean or/and variance signals with different effect sizes. Here we test for the global null and consider a simulation study to be significant if any gene is significant after Bonferroni adjustment for testing 10 genes. The empirical P -value for each gene is calculated using formula 3.5, where $G = 10$.

3.2.3.3 Simulation settings with outliers

Since epigenetic field defects are often characterized by increased variation in DNA methylation due to a few outlier normal-adjacent tissue samples (Teschendorff et al., 2016a; Wang et al., 2017), we considered simulation scenarios with outlier samples. Here, we only considered one gene with 50 CpGs for illustration purposes. We considered two signal-to-noise ratios in this gene to be either 5:45 or 10:40. We set 10%, 15% or 20% of cases to be outlier samples with DNA methylation alterations at some signal CpGs, while the rest cases have the same methylation measures as controls at those signal CpGs when different outlier samples could have DNA methylation alterations at different signal CpGs. For each signal CpG, we generated methylation measures X for cases from a mixture distribution $X = (1 - Z)X_1 + ZX_2$, and methylation measures for controls from $X_1 \sim \text{Beta}(a_0, b_0)$. Specifically, at each signal CpG, we randomly assigned 40 cases to be either outlier samples ($Z = 1$) or non-outlier samples ($Z = 0$) by $Z \sim \text{Bernoulli}(p)$, where p is the probability of any case being an outlier sample. We then generated methylation measures of outlier samples from $X_2 \sim \text{Beta}(a_2, b_2)$ and non-outlier samples from $X_1 \sim \text{Beta}(a_0, b_0)$.

3.2.3.4 Simulation settings with one gene considering correlations among CpGs

Since neighboring CpGs are known to be correlated, we considered simulation scenarios that assume an $AR(1)$ correlation among CpGs in a gene with a correlation coefficient 0.5. The detailed information for simulation setup for this scenario is summarized in the Supplementary File (Section B.1.3 Simulation settings with one gene considering correlations among CpGs).

Table 3.1: Type I error rates

Methods	1 gene			10 genes ^a
	1 CpG ^b	25 CpGs	50 CpGs	25 CpGs
$\mathbf{D}^{w-DM-DV}$	0.044	0.044	0.037	0.050
\mathbf{D}^{w-DM}	0.046	0.032	0.048	0.053
\mathbf{D}^{w-DV}	0.048	0.056	0.048	0.049
\mathbf{D}^{DM-DV}	0.044	0.052	0.045	0.054
\mathbf{D}^{DM}	0.046	0.043	0.041	0.057
\mathbf{D}^{DV}	0.044	0.052	0.054	0.045
EWAS ^{DM}	0.046	0.030	0.039	0.050
EWAS ^{DV}	0.044	0.047	0.040	0.037

^aType I error rates after Bonferroni adjustment for 10 genes.

^bNumber of CpG sites in a gene.

3.3 Results

3.3.1 Simulation results

3.3.1.1 Type I error rate

Type I error rates are well controlled at the 0.05 significance level in settings with one gene and 10 genes after Bonferroni adjustment for multiple comparisons (Table 3.1), respectively.

3.3.1.2 Power for simulation settings with one gene

Power results for simulation settings with one gene are summarized in Figure 3.1. When there are only mean signals at signal CpGs, \mathbf{D}^{w-DV} , \mathbf{D}^{DV} and EWAS^{DV} that consider variance signals only do not have any power as expected. When there is only one CpG in the gene, the non-weighted distance-based methods are the same as the weighted versions, as well as the EWAS method as expected. When there is one signal CpG and increasing number of noise CpGs in the gene, power of \mathbf{D}^{DM} decreases drastically while power of the weighted version \mathbf{D}^{w-DM} are well maintained. This suggests that incorporating weights to

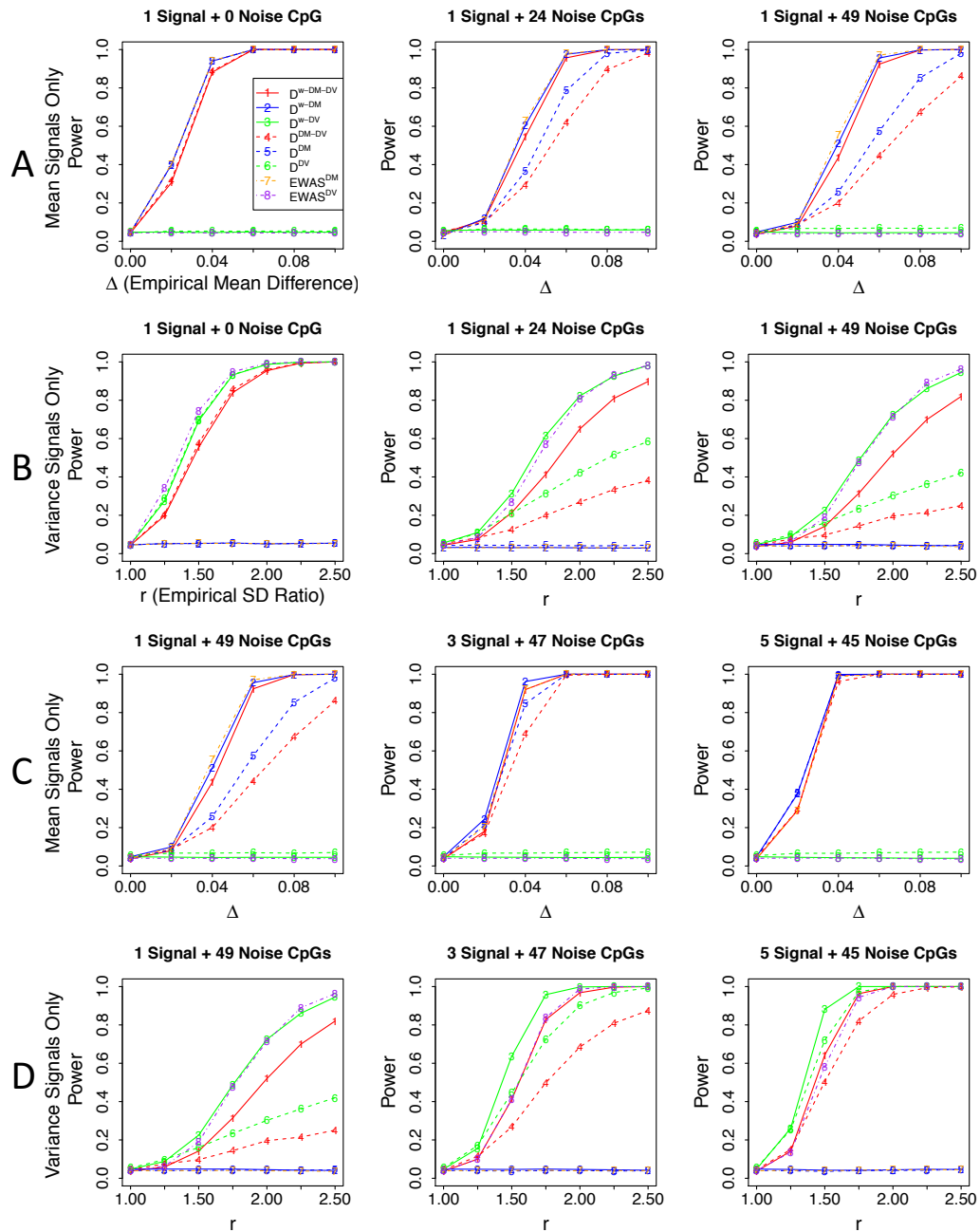


Figure 3.1: Power results for simulation settings with one gene. The signal gene has one signal CpG and increasing number of total CpGs, i.e., decreasing signal-to-noise ratios from 1:0, 1:24 to 1:49 (panel A for mean signals only, panel B for variance signals only), or with a fixed total number of CpGs 50 and increasing signal-to-noise ratios from 1:49, 3:47, to 5:45 (panel C for mean signals only, panel D for variance signals only).

CpGs indeed helps to up-weight signal CpGs and down-weight noise CpGs in constructing the distance matrix, thus improves the performance. When the size of a gene, i.e., number of CpGs in a gene, is fixed, among which when the number of signal CpGs increases, power of \mathbf{D}^{w-DM} increases much slower than that of \mathbf{D}^{DM} while \mathbf{D}^{w-DM} always has greater power than that of \mathbf{D}^{DM} . This implies that adding weights is most effective when a small percent of CpGs in a gene are signals. Similar power patterns are observed between weighted and non-weighted versions of the distance-based methods that consider both mean and variance signals, $\mathbf{D}^{w-DM-DV}$ and \mathbf{D}^{DM-DV} . We also notice that $\mathbf{D}^{w-DM-DV}$ is slightly less powerful than \mathbf{D}^{w-DM} because the overall mean signals are diluted by the inclusion of pseudo-sites for variance when there are only mean signals in the data. Moreover, \mathbf{D}^{w-DM} slightly outperform EWAS^{DM} when there are several signal CpGs. This is because the distance-based method has the advantage to accumulate weak signals and thus boost the overall power.

Similar power patterns are observed when signal CpGs are set to have variance signals only. \mathbf{D}^{w-DM} , \mathbf{D}^{DM} and EWAS^{DM} that consider mean signals only do not have any power, and the weighted distance-based methods outperform the non-weighted versions in the presence of noise CpGs, and \mathbf{D}^{w-DV} performs better than $\mathbf{D}^{w-DM-DV}$, and \mathbf{D}^{w-DV} outperforms EWAS^{DV} when there are several signal CpGs.

3.3.1.3 Power for simulation settings with 10 genes

Power results for simulation settings with 10 genes are summarized in Figure 3.2. When signal CpGs have either mean or variance signals, we observed similar patterns as in the simulation settings with one gene. When signal CpGs have non-negligible mean signals and variance signals ranging from weak to strong, $\mathbf{D}^{w-DM-DV}$ performs the best when variance signals are also weak to moderate as expected. This confirms that the potential area of usage for distance-based methods to be most effective is when there are weak signals that could be accumulated to boost the study power. When there are very strong signals at some sites, any methods will perform well. One observation that we need to point out is, powers of \mathbf{D}^{w-DM} , \mathbf{D}^{DM} and EWAS^{DM} that only consider mean signals actually decrease as variance signals increase when mean signals exist. This is due to the fact that we worked on the

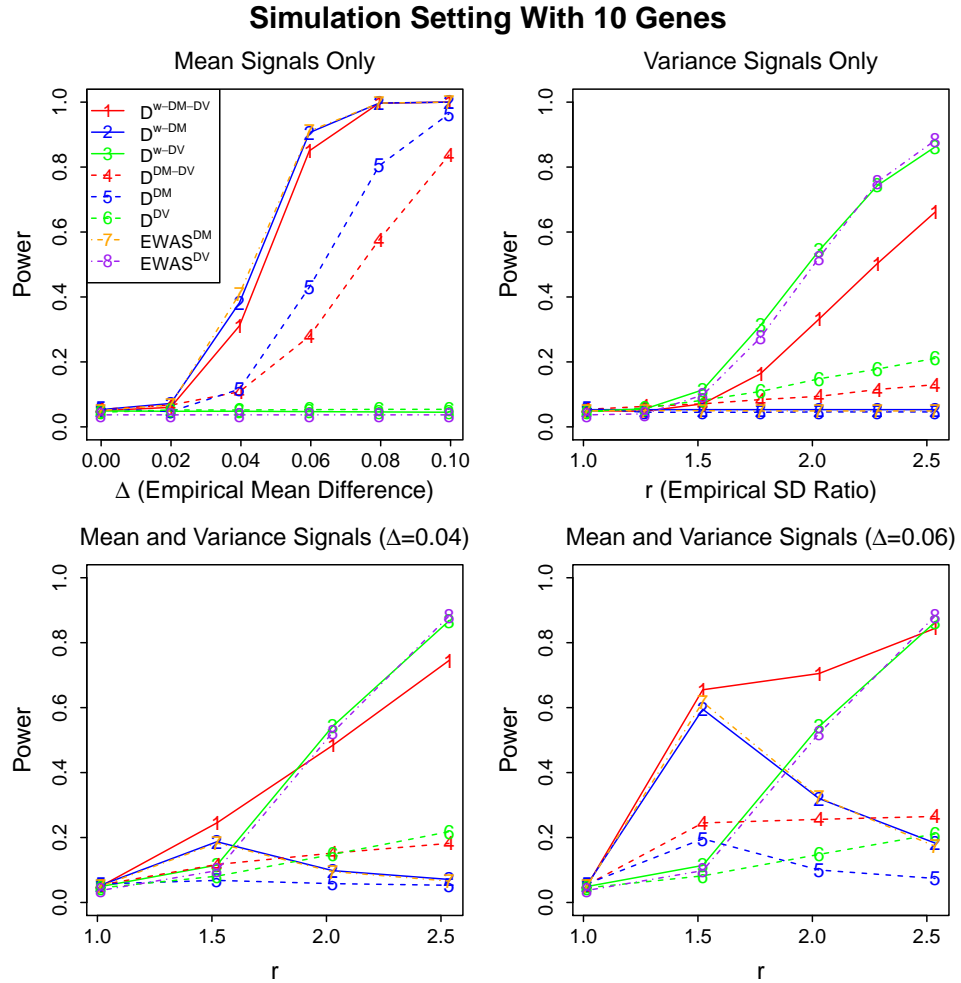


Figure 3.2: Power results for simulation settings with 10 genes. We set each gene to have 25 CpGs and only one gene to have signals. The signal gene has 1 signal CpG and 24 noise CpGs, with signal CpG having mean signal only (panel A), variance signal only (panel B), and mean and variance signals with different sizes of mean signals (panels C and D)

standardized data in $\mathbf{X}^{mv} = [\mathbf{X}^m, \mathbf{X}^v]$, and the effect sizes of mean signals (standardized mean difference) decrease as the effect sizes of variance signals (ratio of standard deviation for cases and controls) increase after standardization.

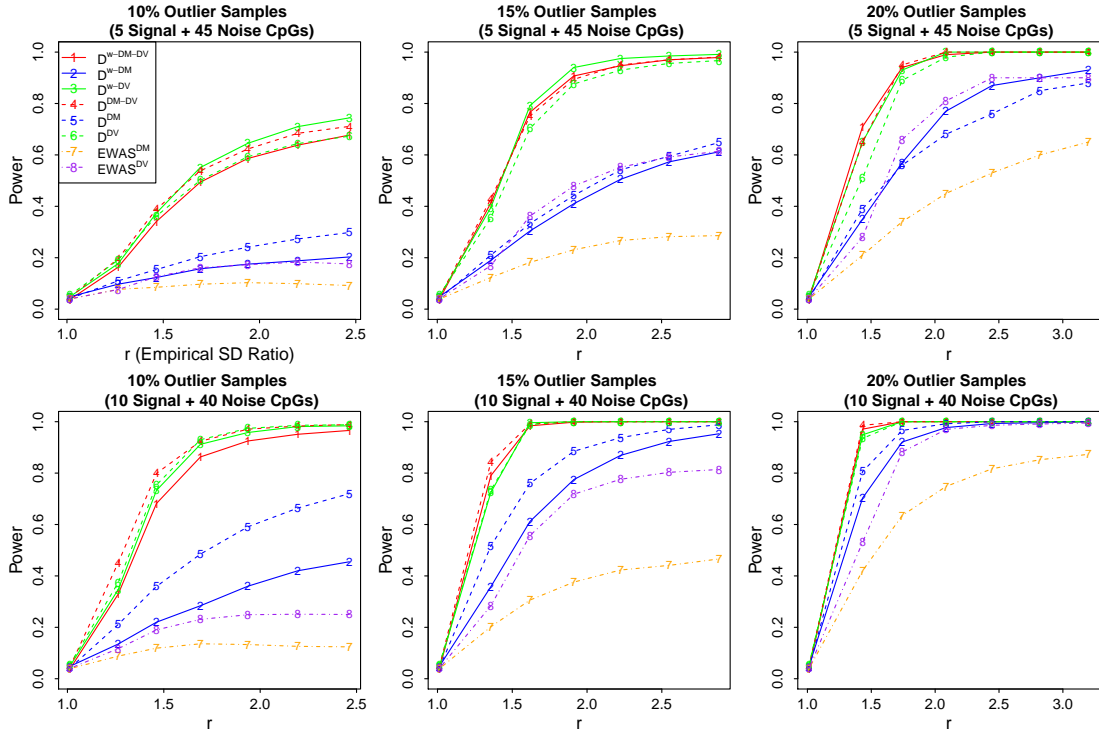


Figure 3.3: Power results for simulation settings with outlier samples. We set to have 10%, 15% and 20% outlier samples and two different signal-to-noise ratios 5:45 and 10:40

3.3.1.4 Power for simulation settings with outliers

Power results for simulation settings with outlier samples are summarized in Figure 3.3. We observe that power of all methods increases as the signal-to-noise ratio increases and as the proportion of outlier samples increases as expected, and distance-based methods outperform non-distance-based methods while $EWAS^{DM}$ and $EWAS^{DV}$ have very little power when there are only 10% outlier samples. Among distance-based methods, \mathbf{D}^{w-DM} and \mathbf{D}^{DM} that consider mean signals only have lower power compare to other methods as the mean signals introduced by a few outlier samples are usually too weak to be detected by methods that consider mean signals only. On the other hand, \mathbf{D}^{DM-DV} that considers both mean and variance signals outperforms methods that consider variance signals only, \mathbf{D}^{DV} . The two weighted distance-based methods $\mathbf{D}^{w-DM-DV}$ and \mathbf{D}^{w-DV} are among the best performed methods consistently. This implies the superiority of $\mathbf{D}^{w-DM-DV}$ in the presence of weak signals in both DM and DV.

3.3.1.5 Power for simulation settings with one gene considering correlations among CpGs

The type I error rates under this scenario are summarized in Supplementary Table B.3. The power results are summarized in Supplementary Figure B.1. We note that the power patterns are very similar to those observed in simulations ignoring correlations among CpG sites. This implies that the correlations among neighboring CpGs do not have much impact on the performance of the proposed distance-based methods.

3.3.2 Real data application

We applied the proposed method $\mathbf{D}^{w-DM-DV}$ and all the comparison methods to two GEO 450K DNA methylation data of breast invasive carcinoma (BRCA) (GSE69914 and GSE67919). As we have demonstrated the superior power of $\mathbf{D}^{w-DM-DV}$ over other distance-based methods in the simulation studies, we focused on $\mathbf{D}^{w-DM-DV}$ in the real data application and compared its performance to that of the EWAS method in the main text and included results using all other comparison distance-based methods in the Supplementary File (Section B.2 Real data application). In order for the two EWAS methods, EWAS^{DM} and EWAS^{DV}, to have a fair comparison with $\mathbf{D}^{w-DM-DV}$, we first adjusted multiple comparisons for the number of CpGs in a gene by multiplying the site-level P -values based on DM and DV with the number of CpGs in the gene, and then selected the minimum adjusted DM and DV P -value across all P -values in the gene as the gene-level P -value. We refer to this method as EWAS^{min- P} .

3.3.3 Discovery analysis using the GEO BRCA data

We applied the proposed method $\mathbf{D}^{w-DM-DV}$ and the comparison methods to the GEO 450K DNA methylation data of normal-adjacent tissue of breast invasive carcinoma (BRCA) patients and normal tissue from independent age-matched cancer-free women (GSE69914). In the original GEO BRCA data, there are DNA methylation measures on 485,512 CpGs for 42 tumor and normal-adjacent pairs from breast cancer patients, 50 normal tissue of independent age-matched cancer-free women and 263 additional tumor tissue of independent breast cancer patients. We conducted standard quality control steps where we removed

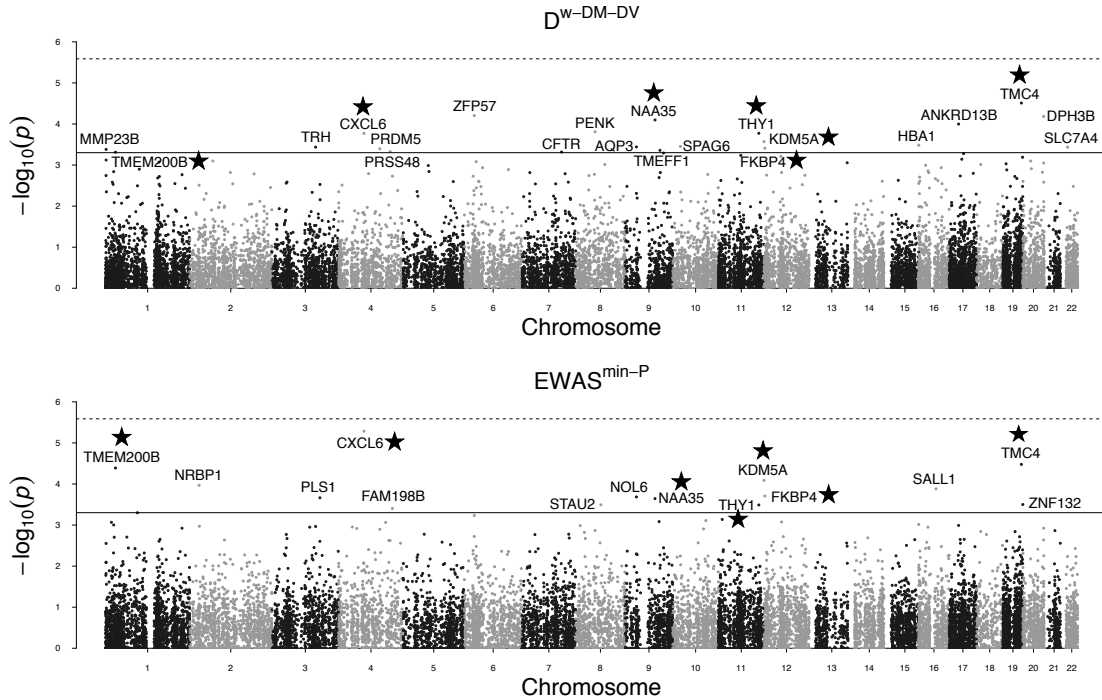


Figure 3.4: Manhattan plots with results from $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\text{min}-P}$. The solid horizontal line is the 0.0005 gene-level P -value threshold. The dashed horizontal line is the Bonferroni adjusted 0.05 significance level ($0.05/19\,271$ genes = 0.0000026 adjusted gene-level P -value threshold). Genes annotated with stars are those identified by both methods at the 0.0005 gene-level P -value threshold.

CpGs on sex chromosomes and those contain either a SNP at the CpG interrogation or at the single nucleotide extension (SBE) based on UCSC dbSNP table version 147 using the R package ‘IlluminaHumanMethylation450kanno.ilmn12.hg19’ (Hansen, 2015). We also required at least 95% CpG coverage per sample and 70% sample coverage per CpG, and only kept CpGs with gene annotations. We ended up with 344,947 CpGs, covering 19,271 genes, from 42 normal-adjacent tissues, 50 normal tissues and 263 independent tumor tissues.

Since Bonferroni adjustment for multiple comparisons of the 19,271 genes is too conservative, especially with the small sample size in the GEO BRCA dataset, we used a less stringent threshold 0.0005 on empirical gene-level P -values obtained from the permutation procedure (Figure 3.4). Our main purpose is to demonstrate the superior performance of the proposed method $\mathbf{D}^{w-DM-DV}$ over several comparison methods, especially the EWAS methods. Results using $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\text{min}-P}$ comparing 42 normal-adjacent tis-

sues to 50 normal tissues are shown in the Manhattan plots in Figure 3.4. At the 0.0005 threshold for gene-level P -values, $\mathbf{D}^{w-DM-DV}$ identified 21 genes (Table 3.2), of which 18 were previously reported to be associated with breast cancer; $\text{EWAS}^{\min-P}$ identified 14 genes (Table 3.3), of which 9 were previously reported to be associated with breast cancer. There are 7 overlapping genes, *TMC4*, *NAA35*, *THY1*, *CXCL6*, *KDM5A*, *FKBP4*, and *TMEM200B* that were identified by both methods. Except for the *PLS1* gene, the 7 genes uniquely identified by $\text{EWAS}^{\min-P}$ all rank very high in $\mathbf{D}^{w-DM-DV}$ results out of the 19,271 genes (Table 3.3). Except for the *CFTR* gene, the 14 genes uniquely identified by $\mathbf{D}^{w-DM-DV}$ also all rank very high in $\text{EWAS}^{\min-P}$ results. This suggests an overall good consistency between results of $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\min-P}$. At the same 0.0005 gene-level P -value threshold, other comparison methods \mathbf{D}^{w-DM} , \mathbf{D}^{w-DV} , \mathbf{D}^{DM-DV} , \mathbf{D}^{DM} and \mathbf{D}^{DV} identified 11, 9, 2, 6 and 4 genes, of which 6, 7, 1, 3 and 1 genes were also identified by the proposed $\mathbf{D}^{w-DM-DV}$ (Supplementary Tables B.4-B.8), respectively.

We further examined the 14 and 7 genes uniquely identified by $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\min-P}$, respectively. We plotted heatmaps of the original DNA methylation measures of CpG sites on these genes for the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues (Supplementary Figures B.2 and B.3). In general, the 14 genes uniquely identified by $\mathbf{D}^{w-DM-DV}$ are those with multiple CpGs of weak signals, i.e. weak dense signals. Moreover, some of these weak dense signals were mainly due to a few outlier normal-adjacent tissue samples, thus were missed by $\text{EWAS}^{\min-P}$. The 7 genes uniquely identified by $\text{EWAS}^{\min-P}$ are those with just one or two CpGs with very strong signals, i.e. strong sparse signals. We also plotted heatmaps of 7 genes identified by both $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\min-P}$ (Supplementary Figure B.4).

We then investigated the two genes, *CFTR* and *PLS1*, that were uniquely identified by $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\min-P}$, respectively, but ranked the last using the other method among all uniquely identified genes. We similarly plotted the heatmap of the original DNA methylation measures of CpG sites in these two genes (Figure 3.5A). For the *CFTR* gene that has 16 CpGs, it is clear that variation in methylation measures increases in the progression from normal tissues to normal-adjacent tissues and to tumor tissues in multiple CpGs when there are several samples among the 42 normal-adjacent tissue samples that

Table 3.2: Twenty one genes identified by $\mathbf{D}^{w-DM-DV}$ at the 0.0005 gene-level P -value threshold using the GEO BRCA Data

Rank	Gene	# CpG	Cancer	Rank in EWAS ^{min-P}
1	<i>TMC4</i> *	13	Breast Cancer (Krijgsman et al., 2012)	2
2	<i>ZFP57</i>	5	Breast Cancer (Tada et al., 2015)	16
3	<i>DPH3B</i>	5	-	61
4	<i>NAA35</i> *	7	Breast Cancer (Abu-Asab et al., 2013)	10
5	<i>ANKRD13B</i>	22	Breast Cancer (Jönsson et al., 2010)	25
6	<i>PENK</i>	23	Breast Cancer (Legendre et al., 2015)	37
7	<i>THY1</i> *	19	Breast Cancer (Lehmann et al., 2011)	13
8	<i>CXCL6</i> *	7	Breast Cancer (Bièche et al., 2007)	1
9	<i>KDM5A</i> *	2	Breast Cancer (Hou et al., 2012)	4
10	<i>HBA1</i>	7	Breast Cancer (Wolf et al., 2007)	23
11	<i>SPAG6</i>	16	Acute Myeloid Leukemia (Steinbach et al., 2006)	170
12	<i>AQP3</i>	7	Breast Cancer (Cao et al., 2013)	140
13	<i>TRH</i>	16	Breast Cancer (Nicolau et al., 2011)	28
14	<i>SLC7A4</i>	12	Breast Cancer (Xia et al., 2012)	175
15	<i>FKBP4</i> *	18	Breast Cancer (Yang et al., 2011)	7
16	<i>PRDM5</i>	18	Breast Cancer (Deng and Huang, 2004)	36
17	<i>MMP23B</i>	2	Breast Cancer (Giussani et al., 2015)	80
18	<i>TMEFF1</i>	5	Breast Cancer (Matise et al., 2012)	156
19	<i>PRSS48</i>	7	-	64
20	<i>CFTR</i>	16	Breast Cancer (Zhang et al., 2013b)	1055
21	<i>TMEM200B</i> *	20	Breast Cancer (Stirzaker et al., 2015)	3

*Genes identified by both $\mathbf{D}^{w-DM-DV}$ and EWAS^{min-P}.

are very different from the normal samples. On the other hand, for the *PLS1* gene that also has 16 CpGs, it was identified uniquely by EWAS^{min-P} because of one signal CpG site cg00137209 (Figure 3.5A), mainly due to the very small variation in the methylation measures of the normal tissues. We then plotted DNA methylation measures of the top

Table 3.3: Fourteen genes identified by EWAS^{min-P} at the 0.0005 gene-level P -value threshold using the GEO BRCA Data

Rank	Gene	# CpG	Top CpG Signal ^a	Cancer	Rank in $\mathbf{D}^{w-DM-DV}$
1	<i>CXCL6</i> *	7	Variance	Breast Cancer (Bièche et al., 2007)	11
2	<i>TMC4</i> *	13	Variance	Breast Cancer (Krijgsman et al., 2012)	1
3	<i>TMEM200B</i> *	20	Variance	Acute Myeloid Leukemia (Rudenko et al., 2016)	41
4	<i>KDM5A</i> *	2	Variance	Breast Cancer (Hou et al., 2012)	4
5	<i>NRBP1</i>	12	Variance	Breast Cancer (Wei et al., 2015)	110
6	<i>SALL1</i>	44	Variance	Breast Cancer (Wolf et al., 2014)	887
7	<i>FKBP4</i> *	18	Variance	Breast Cancer (Yang et al., 2011)	32
8	<i>NOL6</i>	5	Variance	-	160
9	<i>PLS1</i>	16	Variance	Breast Cancer (Bi et al., 2015)	1069
10	<i>NAA35</i> *	7	Variance	Breast Cancer (Abu-Asab et al., 2013)	6
11	<i>ZNF132</i>	12	Mean	Prostate Cancer (Abildgaard et al., 2012)	118
12	<i>STAU2</i>	39	Variance	Hepatocellular Carcinoma (Castaneda et al., 2007)	666
13	<i>THY1</i> *	19	Variance	Breast Cancer (Lehmann et al., 2011)	14
14	<i>FAM198B</i>	14	Variance	Breast Cancer (Fidalgo et al., 2015)	84

^aMean or variance tests with smaller P -value at the most significant CpG in a gene.

*Genes identified by both $\mathbf{D}^{w-DM-DV}$ and EWAS^{min-P}.

4 P -value ranked CpGs, ranked by CpG site-level P -values from both mean and variance tests each after adjusting for multiple comparisons for the number of CpGs in the *CFTR* gene (Figure 3.5B), which clearly shows elevated methylation levels in the progression to tumor. For the *PLS1* gene, we similarly plotted the DNA methylation measures of the top 2 P -value ranked CpGs (Figure 3.5B), where the #1 ranked CpG cg00137209 is the one that shows strong variance signal due to very small variation in the methylation measures of the normal tissues, when neither CpGs showed any enrichment in methylation measures in the progression to tumor. This suggests that genes uniquely identified by EWAS^{min-P} due to extreme P -values at one or two CpGs may not be reliable, while genes identified uniquely by $\mathbf{D}^{w-DM-DV}$ are generally characterized by multiple signal CpGs, thus are more reliable.

We also plotted the DNA methylation measures of all CpGs in these two genes *CFTR*

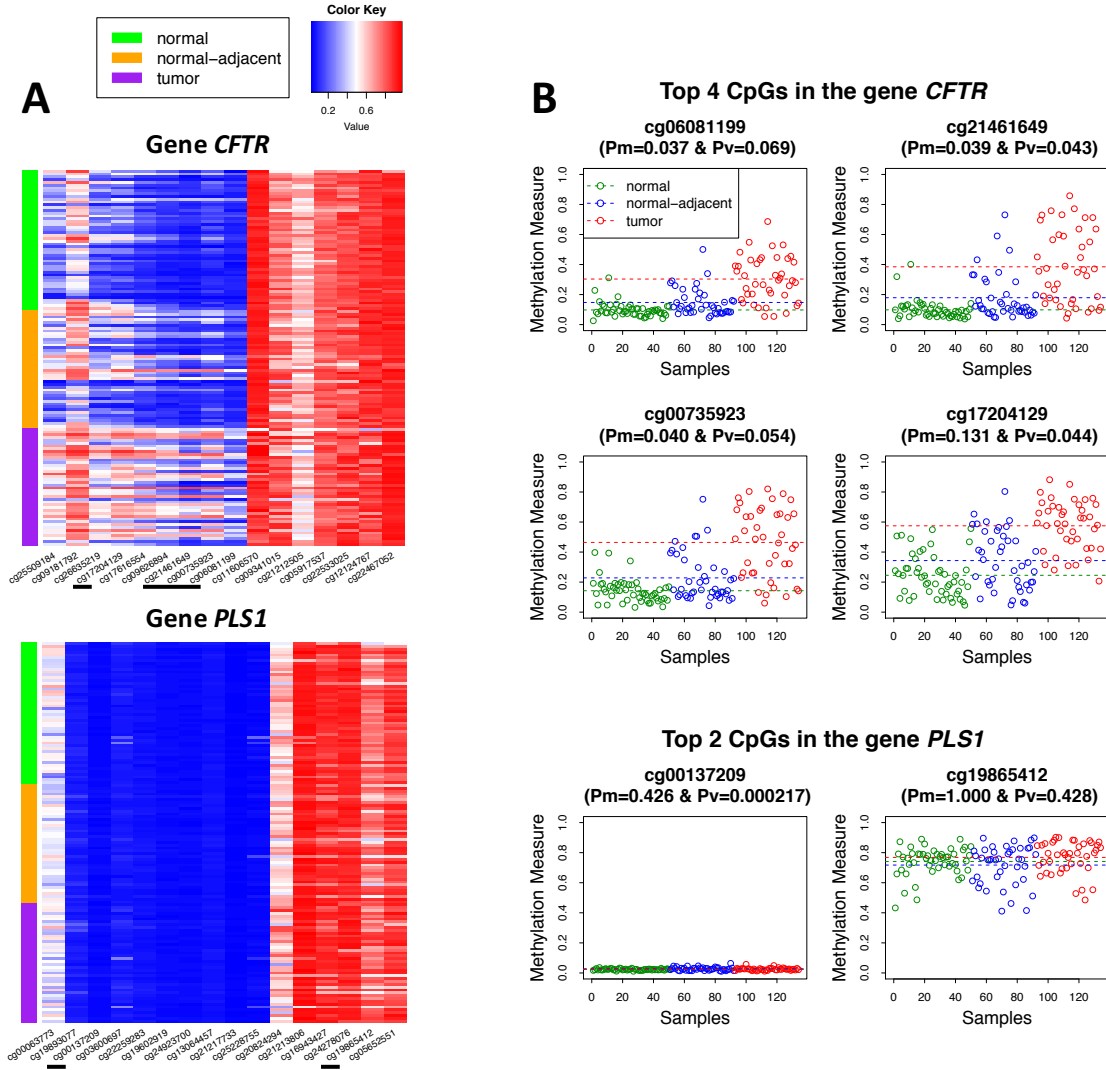


Figure 3.5: (A) Heatmaps of DNA methylation measures of CpGs in the *CFTR* and *PLS1* genes. The CpGs underlined are the top 4 P -value ranked CpGs in the *CFTR* gene and the top 2 P -value ranked CpGs in the *PLS1* gene. (B) DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of the top 4 P -value ranked CpGs in the *CFTR* gene and the top 2 P -value ranked CpGs in the *PLS1* gene. Pm and Pv are P -values from CpG site-level mean and variance tests adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.

and *PLS1* (Supplementary Figures B.5 and B.6, respectively). It is again clear that almost half of the CpGs in the *CFTR* gene have weak mean signals and weak variance signals, thus missed by $\text{EWAS}^{\text{min}-P}$ due to stringent multiple comparisons adjustment. In addition, we plotted the weighted distance matrices of the 50 normal tissues and the 42 normal-adjacent tissues for the *CFTR* gene and the *PLS1* gene (Supplementary Figure B.7). For the *CFTR* gene, we observe little variation in distances among normal samples and increased variation in distances between several pairs of normal and normal-adjacent samples, while for the *PLS1* gene, we observe no clear pattern. We also plotted the DNA methylation measures of CpGs in the *TMC4* gene (Supplementary Figure B.8) that was identified by both $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\text{min}-P}$ and ranked #1 and #2 in the two methods, respectively. There are 13 CpGs in the *TMC4* gene, 3 CpGs have strong variance signals when two of the three CpGs also have mean signals. Thus, the *TMC4* gene was identified by both $\mathbf{D}^{w-DM-DV}$ and $\text{EWAS}^{\text{min}-P}$.

In our previous work on differentially methylated regions (DMRs) using the same GEO BRCA data, we identified 2 DMRs of epigenetic field defects using both mean and variance signals (Wang et al., 2017). The two DMRs cover two genes, *NKX62* and *CCND2*, which rank #113 and #359 in the $\mathbf{D}^{w-DM-DV}$ results. Further investigation revealed that the two DMRs only cover part of the two genes. We therefore broke down the two genes into smaller parts so that there is one part that covers exactly the identified DMR. We then treated these smaller parts as individual regions and repeated $\mathbf{D}^{w-DM-DV}$ across the whole genome. The rank of the *NKX62* part that matches with the DMR moved up to #90 from #113 while the other two parts rank #107 and #4855, respectively. The rank of the *CCND2* part that matches with the other DMR moved up to #154 from #359 and the other part ranks #1116. Overall, the 2 DMR-covered genes previously identified as epigenetic field defects also rank on top in the results of $\mathbf{D}^{w-DM-DV}$. This suggests that we may combine DMR detection techniques with distance-based methods to first better define ‘regions of interest’ using DMR ideas and then assess significance more powerfully with distance-based methods.

We also investigated the relation between the number of CpGs in a gene and the probability that the gene is selected, where we binned genes based on their sizes and calculated

the selection probability of a gene in a bin as the proportion of genes identified out of all genes in the bin. We plotted the selection probabilities against gene sizes (Supplementary Figure B.9) and found that the selection probabilities for different versions of the distance-based methods and EWAS^{min- P} method are not systematically affected by gene sizes.

3.3.4 Validation of the identified epigenetic field defects in the GEO BRCA data

We further validated the 21 genes of epigenetic field defects identified by $\mathbf{D}^{w-DM-DV}$ through comparing methylation measures of the 21-gene-covered CpGs between 263 independent tumor tissues and 42 normal-adjacent tissues to examine if the methylation levels at these CpGs exhibit progression to tumor. Specifically, we performed the two-sample t -test at each of these CpGs and plotted the $-\log_{10}(P\text{-value})$ from the two comparisons, 50 normal tissues versus 42 normal-adjacent tissues and 42 normal-adjacent tissues versus 263 tumor tissues (Supplementary Figure B.10). In general, the majority of these CpGs show more significant signals in the progression from normal tissues to normal-adjacent tissues to tumors.

3.3.5 Replication analysis using an independent data of normal tissues

As epigenetic field defects identified in one set of normal vs. normal-adjacent comparison may be driven by a few ‘outlier’ normal-adjacent samples (Teschendorff et al., 2016a,b; Wang et al., 2017), different epigenetic field defects could be identified in a different set of normal versus normal-adjacent comparison that are driven by different ‘outlier’ normal-adjacent samples. Therefore, we propose to conduct a replication analysis that uses the same normal-adjacent tissue samples but compare to an independent data of normal samples. We used 450K DNA methylation data of 18 normal tissue of 18 breast reduction mammoplasty subjects (GSE67919) (Hair et al., 2015). The original data have methylation measures on 485,577 CpG sites. We followed the same quality control steps as for the discovery GEO BRCA data (GSE69914) and kept the same CpG sites for comparison purposes. We ended up with 344,947 CpGs, covering 19,271 genes, from 18 normal tissues. We then compared these normal samples to the same 42 normal-adjacent tissues from the GEO BRCA data in

a replication analysis.

At the same 0.0005 threshold for gene-level P -values, 7 out of the 21 previously identified genes with epigenetic field defects in the discovery analysis using the GEO BRCA data were replicated by $\mathbf{D}^{w-DM-DV}$. The seven genes are *DPH3B*, *NAA35*, *ANKRD13B*, *CXCL6*, *FKBP4*, *PRSS48* and *CFTR*. We similarly validated these 7 genes by comparing P -values from the two-sample t -tests comparing the 18 replication normal samples to the 42 GEO BRCA normal-adjacent samples and P -values from the two-sample t -tests comparing the 42 GEO BRCA normal-adjacent samples to the 263 independent GEO BRCA tumor samples (Supplementary Figure B.11). All 7 genes, except the *NAA35* and *FKBP4*, exhibit progression to tumor. More details of the replication analysis results using $\mathbf{D}^{w-DM-DV}$, $\text{EWAS}^{\text{min-}P}$ and other comparison distance-based methods were summarized in Supplementary File (Section B.2.3 Replication Analysis) and Supplementary Table B.9 and Supplementary Figures B.12-B.14.

To investigate our hypothesis that different epigenetic field defects may be identified when comparing normal samples to a different set of normal-adjacent samples, we obtained a new set of BRCA normal-adjacent samples ($n = 90$) from the Cancer Genome Atlas (TCGA) project together with their matched tumor samples ($n = 90$). We plotted DNA methylation measures of CpGs in the 7 replicated genes of the 18 replication normal samples, the 50 discovery GEO BRCA normal samples, the 42 discovery GEO BRCA normal-adjacent samples, the 42 discovery GEO BRCA matched tumor samples, and the 90 TCGA normal-adjacent samples and the 90 TCGA matched tumor samples. It is clear that methylation patterns of the TCGA normal-adjacent tissues are very different from that of the discovery GEO BRCA normal-adjacent tissues in most of these CpGs. This supports our hypothesis that methylation patterns can be very different in different pre-cancer tissues (using normal-adjacent tissue as a surrogate) thus different epigenetic field defects may be identified when normal samples are compared to different sets of pre-cancer tissues.

3.4 Discussion

In this study, we developed a weighted epigenetic distance-based method $\mathbf{D}^{w-DM-DV}$ that accumulates both DM (mean) and DV (variance) signals across CpGs in a gene or a genetic region. One known advantage of distance-based methods is, there is no need to preselect outcome-associated features, avoiding the potential to mis-screen features with weak signals. In our proposed weighted epigenetic distance-based method $\mathbf{D}^{w-DM-DV}$, we used CpG site-level association strengths as weights for individual CpGs aiming to up-weight signal CpGs and down-weight noise CpGs. If the feature preselection step could be conducted perfectly, it is equivalent to the case when weight ‘0’ is correctly assigned to noise CpGs and weight ‘1’ is correctly assigned to signal CpGs. Results from simulation studies suggest that when the signal-to-noise ratio in a gene decreases, power of non-weighted epigenetic distance-based methods decreased drastically, while power of the weighted version was well maintained. This suggests that incorporating CpG-site-level association strengths as weights for individual CpGs indeed help to up-weight signal CpGs and down-weight noise CpGs, thus improve the overall study performance. Simulation results also suggest that the weighted epigenetic distance-based methods will be most effective when applied to genes or genetic regions with a small percentage of CpGs having weak signals. This makes the detection of epigenetic field defects, i.e., early epigenetic alterations that are usually infrequent across samples and identifiable as outlier samples, the ideal application of the proposed method $\mathbf{D}^{w-DM-DV}$. Using the GEO BRCA 450K DNA methylation data, $\mathbf{D}^{w-DM-DV}$ identified 21 genes with epigenetic field defects, when 7 out of the 21 genes overlap with the genes identified by $\text{EWAS}^{\text{min-}P}$. Majority of the genes uniquely identified by $\mathbf{D}^{w-DM-DV}$ were previously reported to be associated with breast cancer. Most of the genes uniquely identified by $\text{EWAS}^{\text{min-}P}$ also ranked on top in the $\mathbf{D}^{w-DM-DV}$ results except for the *PLS1* gene. However, further investigations suggested that the *PLS1* gene may not be a real epigenetic field defect. On the other hand, most of the genes uniquely identified by $\mathbf{D}^{w-DM-DV}$ also ranked on top in the $\text{EWAS}^{\text{min-}P}$ results except for the *CFTR* gene, in which the enrichment in the progression to breast cancer was confirmed in further analyses. This suggests that genes identified by $\mathbf{D}^{w-DM-DV}$, which are generally characterized by multiple signal CpGs, are more reliable. It is worth noticing that the 2 DMR-covered genes

identified in our previous work (Wang et al., 2017) also ranked on top in the $\mathbf{D}^{w-DM-DV}$ results. We validated the identified epigenetic field defects by showing a progression to tumor in an independent dataset of tumor tissues. We also conducted a replication analysis by comparing the same set of normal-adjacent tissues to an independent set of normal tissues, and found that 7 out of the 21 genes of epigenetic field defects identified by $\mathbf{D}^{w-DM-DV}$ in the discovery analysis were replicated.

In general, distance-based methods have a better performance than that of site-level EWAS methods when site-level signals are weak. As discussed in our previous work (Wang et al., 2017) and work of others (Teschendorff et al., 2016a,b), epigenetic field defects are often characterized by increased variation in DNA methylation measures due to a few outlier normal-adjacent tissue samples. So the site-level EWAS methods are usually underpowered due to small mean differences as well as stringent multiple comparisons adjustment. Distance-based methods accumulate weak signals to improve power. Distance-based methods are flexible and can be applied to a CpG site, a gene, a pathway, or an entire genome. A closer investigation on what we identified in our previous work (Wang et al., 2017) in DMR detection and the current work suggests that we may take advantages of the techniques in DMR detection and combine that with distance-based methods in future works to more efficiently identify regions of epigenetic field defects.

In summary, we proposed a new weighted distance-based method $\mathbf{D}^{w-DM-DV}$ that considers both DM and DV in DNA methylation and incorporates site-level association strengths as weights on individual CpGs to up-weight signal CpGs and down-weight noise CpGs to further boost the overall study power. The $\mathbf{D}^{w-DM-DV}$ method is especially powerful in detecting epigenetic field defects when methylation alterations between normal tissues and normal-adjacent tissues are usually minimum.

Chapter 4

A Powerful and Flexible Weighted Distance-Based Method Incorporating Interactions Between DNA Methylation and Environmental Factors on Health Outcomes

4.1 Introduction

DNA methylation has been associated with cancers (Das and Singal, 2004; Ehrlich, 2002; Esteller and Herman, 2002; Kulis and Esteller, 2010) and a wide range of human diseases (Feinberg, 2007; Jager et al., 2014; Mill and Petronis, 2007, 2008; Mill et al., 2008; Nestler, 2014; Schanen, 2006). Studies have also demonstrated associations between DNA methylation and environmental factors (Herbstman et al., 2012; Perera et al., 2009; Faulk et al., 2015; Nahar et al., 2015; Janssen et al., 2013; Saenen et al., 2016; Sen et al., 2015; Nye et al., 2016; Bakulski et al., 2015; Cardenas et al., 2017) such as prenatal exposure to polycyclic

aromatic hydrocarbons (PAH) (Herbstman et al., 2012; Perera et al., 2009), Bisphenol A (Faulk et al., 2015; Nahar et al., 2015). In addition, there is evidence supporting the idea that DNA methylation may modify the risk of environmental factors on health outcomes. For example, Fu et al. found that DNA methylation modifies the effect of NO_2 on the progression from mild to severe asthma (Fu et al., 2012); White et al. found that DNA methylation modifies the risk of PAH-DNA adducts on breast cancer (White et al., 2015). Despite these findings, due to high dimensionality and low study power, current studies usually focus on finding differential methylation (DM) on health outcomes at CpG level or gene level combining multiple CpGs and/or finding environmental effects on health outcomes but ignoring their interactions.

Here, we developed a weighted epigenetic distance-based method with a pseudo-data matrix constructed with cross-product terms between DNA methylation and environmental factors that are able to capture their interactions on health outcomes. The distances between pairs of subjects can then be calculated combining the original data matrix with DNA methylation measures and environmental factors together with the pseudo-data matrix with interactions. Using this approach, we can identify both main and interaction effects. We focused on interactions between DNA methylation of CpGs in a gene and an environmental factor on health outcomes, but the proposed method can be readily adapted to interactions among CpGs in a gene on health outcomes. We conducted simulation studies and showed that, when there are both main and interaction effects between DNA methylation and environmental factors, the proposed novel approach that incorporates interactions through a pseudo-data matrix has much better power than comparison methods that consider either main effects or interaction effects. Most importantly, the power of the proposed method is not affected by the source of the signals, i.e., if the signals are main or interaction effects. This makes this approach very attractive due to the known low power of interaction detection.

We applied the proposed method to the data from the Mothers and Newborns (MN) birth cohort of the Columbia Center for Children’s Environmental Health (CCCEH) to identify effects of gene-level DNA methylation, prenatal PAH and their interactions on Attention Deficit Hyperactivity Disorder (ADHD) at age 3. We identified some main effects of

DNA methylation and some interactions with prenatal PAH which were missed by comparison methods. Some of these findings were further replicated in the CCCEH Sibling cohort. We similarly applied the proposed method to the Mental Development Index (MDI) at age 3 and observed a similar pattern in results in both discovery and replication analyses.

4.2 Methods

4.2.1 The proposed method

The proposed weighted distance-based method incorporating DNA methylation by environment interactions has three steps: 1) introducing a pseudo-data matrix constructed with cross-product terms between DNA methylation of CpGs in a gene and environmental factors that captures their interactions, on which a gene-level weighted distance matrix incorporating interactions is defined; 2) calculating the pseudo- F statistic; and 3) assessing the statistical significance empirically using permutations. We focus on binary outcomes and illustrate the method at the gene-level while it can be readily adapted to other types of outcomes and to genetic region or pathway-level.

Step 1: A pseudo-data matrix and a weighted distance matrix incorporating interactions

Here we focus on binary outcomes with equal number of cases and controls and consider one gene with n CpGs. Denote \mathbf{X}^m as a $2N \times n$ matrix with DNA methylation measures for N cases ($Y = 1$) and N controls ($Y = 0$) of n CpGs. Denote \mathbf{E} as a $2N \times 1$ vector with measures of an environment factor. Define $\mathbf{X}^{\text{main}} = [\mathbf{X}^m, \mathbf{E}]$, a $2N \times (n + 1)$ matrix for main signals of n CpGs and one environmental factor. We normalize each column of \mathbf{X}^{main} to have mean zero and unit standard deviation (SD). The element x_{ij}^{main} harbors the normalized methylation measure of CpG j for subject i , $j = 1, \dots, n$, and normalized environmental factor E_i of subject i , $j = n + 1$, $i = 1, \dots, 2N$. We then define \mathbf{X}^{int} , a $2N \times n$ pseudo-data matrix with element $x_{ij}^{\text{int}} = x_{ij}^{\text{main}} \times E_i$ harbors the interaction between CpG j and the environmental factor of subject i , $j = 1, \dots, n$, and $i = 1, \dots, 2N$. By using $\mathbf{X}^{\text{main-int}} = [\mathbf{X}^{\text{main}}, \mathbf{X}^{\text{int}}]$, a $2N \times (2n + 1)$ pseudo-data matrix, we capture main signals of

n CpGs, one environmental factor and n pairwise CpG \times E interactions.

With $\mathbf{X}^{\text{main-int}}$, we first define a non-weighted $2N \times 2N$ distance matrix $\mathbf{D}^{\text{main-int}}$ with element $d_{st}^{\text{main-int}}$ capturing Euclidean distance between individuals s and t , $s, t = 1, \dots, 2N$ on DNA methylation, the environmental factor and their interactions as

$$d_{st}^{\text{main-int}} = \sqrt{\frac{1}{2n+1} \Delta_E^2 + \sum_{j=1}^n \left(\frac{1}{2n+1} \Delta_{\text{main},j}^2 + \frac{1}{2n+1} \Delta_{\text{int},j}^2 \right)} \quad (4.1)$$

where $\Delta_E^2 = (E_s - E_t)^2$, $\Delta_{\text{main},j}^2 = (X_{sj}^{\text{main}} - X_{tj}^{\text{main}})^2$, and $\Delta_{\text{int},j}^2 = (X_{sj}^{\text{int}} - X_{tj}^{\text{int}})^2$.

We then incorporate association strength at CpG site-level as weights to up-weight signals (both main and interaction signals) and down-weight noises in calculating distances. We define weights for main and interaction signals at CpG j and the main signal of the environmental factor as follows:

$$\begin{aligned} w_j^{\text{main}} &= \frac{-\log_{10}(p_j^{\text{main}})}{-\log_{10}(p_E^{\text{main}}) + \sum_{j=1}^n -\log_{10}(p_j^{\text{main}}) + \sum_{j=1}^n -\log_{10}(p_j^{\text{int}})} \\ w_j^{\text{int}} &= \frac{-\log_{10}(p_j^{\text{int}})}{-\log_{10}(p_E^{\text{main}}) + \sum_{j=1}^n -\log_{10}(p_j^{\text{main}}) + \sum_{j=1}^n -\log_{10}(p_j^{\text{int}})} \\ w_E^{\text{main}} &= \frac{-\log_{10}(p_E^{\text{main}})}{-\log_{10}(p_E^{\text{main}}) + \sum_{j=1}^n -\log_{10}(p_j^{\text{main}}) + \sum_{j=1}^n -\log_{10}(p_j^{\text{int}})} \end{aligned} \quad (4.2)$$

where p_j^{main} and p_j^{int} are P -values testing $\beta_{1j} = 0$ and $\beta_{3j} = 0$ in the logistic model $\text{logit}P(Y_i = 1) = \beta_{0j} + \beta_{1j}x_{ij} + \beta_{2j}E_i + \beta_{3j}x_{ij} \times E_i$, and p_E^{main} is the P -value testing $\beta_{1E} = 0$ in the logistic model $\text{logit}P(Y_i = 1) = \beta_{0E} + \beta_{1E}E_i$.

The corresponding weighted distance matrix $\mathbf{D}^{\text{w-main-int}}$ with element $d_{st}^{\text{w-main-int}}$ is defined as

$$d_{st}^{\text{w-main-int}} = \sqrt{w_E^{\text{main}} \Delta_E^2 + \sum_{j=1}^n (w_j^{\text{main}} \Delta_{\text{main},j}^2 + w_j^{\text{int}} \Delta_{\text{int},j}^2)} \quad (4.3)$$

Step 2: The pseudo- F statistic

To test the association between case/control status and DNA methylation distances within a gene and an environmental factor together with their interactions, we calculate a pseudo- F

statistic based on the weighted distance matrix $\mathbf{D}^{\text{w-main-int}}$ introduced in equation 4.3

$$F^{\text{w-main-int}} = \frac{\text{tr}(\mathbf{HGH})}{\text{tr}[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})]} \quad (4.4)$$

where $\mathbf{H} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$, $\mathbf{G} = (\mathbf{I} - \frac{1}{2N}\mathbf{1}\mathbf{1}^T)\mathbf{A}(\mathbf{I} - \frac{1}{2N}\mathbf{1}\mathbf{1}^T)$, $\mathbf{A} = (a_{st}) = \left(-\frac{1}{2}(d_{st}^{\text{w-main-int}})^2\right)$, \mathbf{Y} is a $2N \times 1$ vector with case/control status, $\mathbf{1}$ is a $2N$ -dimensional column vector with elements 1, and \mathbf{I} is a $2N \times 2N$ identity matrix.

Step 3: The statistical significance

Permutation procedures are used to assess statistical significance, where we randomly shuffle the outcome and repeat Steps 1-2 on the permuted data. When we test G genes ($G > 1$) in a study, we pool G pseudo- F statistics from observed and permuted data to compute empirical P -values in order to have more granular P -values (Friedman et al., 2001). We repeat the permutations 999 times, and calculate the empirical P -value for gene g , $g = 1, \dots, G$ as:

$$P_g^{\text{w-main-int}} = \frac{\sum_{g'=1}^G \{1 + \sum_{\text{perm}=1}^{999} I(F_{g',\text{perm}}^{\text{w-main-int}} \geq F_g^{\text{w-main-int}})\}}{G \times (1 + 999)} \quad (4.5)$$

To investigate if genes with different sizes, i.e., number of CpGs, will have different distributions for pseudo- F statistics under the null hypothesis, we conducted simulation studies to compare the type I error rates when the P -value for each gene is calculated based on pooled pseudo- F statistics of all G genes across all permutations (Supplementary Section C.1.1).

4.2.2 Comparison methods

We compare the performance of the proposed method $\mathbf{D}^{\text{w-main-int}}$ that considers both main and interaction signals with weights to that of several comparison methods, including 1) the weighted distance-based methods considering main signals only $\mathbf{D}^{\text{w-main}}$, 2) interaction signals only $\mathbf{D}^{\text{w-int}}$, 3) the distance-based methods without weights considering both main and interaction signals $\mathbf{D}^{\text{main-int}}$, 4) main signals only \mathbf{D}^{main} , 5) interaction signals only \mathbf{D}^{int} , and 6) the site-level EWAS methods via logistic regressions on each CpG considering main signals only L^S or 7) both main and interaction signals L^M . For L^S , a simple logistic model is fitted for each CpG in the gene one by one and a separate simple logistic model

for the environmental factor. A significant main effect of the gene is claimed if any simple logistic model is significant after Bonferroni adjustment for testing the number of CpGs in the gene plus one environmental factor. For L^M , a multiple logistic model with one CpG, the environmental factor and their interaction is fitted for each CpG in a gene, and the gene is considered significant if any multiple logistic model is significant after Bonferroni adjustment for the number of CpGs in the gene.

4.3 Simulation studies

We conducted simulation studies to evaluate type I error rate and power of the proposed method $\mathbf{D}^{\text{w-main-int}}$ and the comparison methods where we only considered one gene with multiple CpGs for illustration purpose. Type I error rate is defined as the proportion of simulations the gene is significant when the data are generated under the null hypothesis of no association. Power is defined as the proportion of simulations the gene is significant when the data are generated with a gene with multiple CpGs of different types of signals. We conducted 1,000 simulations in each simulation setting.

4.3.1 Simulation setup

We simulated methylation M -values \mathbf{X} , which are logit2 transformation of β -values (Du et al., 2010), for samples at multiple CpGs in a gene using multivariate normal distributions. We only considered one gene but with different number of correlated CpGs. The methylation M -values of n CpGs of subject i are generated by

$$\mathbf{X}_i \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Delta}^T \boldsymbol{\Sigma} \boldsymbol{\Delta})$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ determines means and $\boldsymbol{\Delta} = \text{diag}(\sigma_1, \dots, \sigma_n)$ determines SDs, and $\boldsymbol{\Sigma}$ determines correlations among n CpGs in a gene, where we assume an AR(1) correlation with $\rho = 0.5$, i.e., $\Sigma_{uv} = \rho^{|u-v|}$. The environmental factor of subject i is generated from $E_i \sim \text{Bernoulli}(p)$ with p the probability of being exposed. We set $p = 0.5$. After normalizing each column of \mathbf{X} and \mathbf{E} , we calculated pairwise interactions between CpGs and the environmental factor for subject i as $\mathbf{Z}_i = \mathbf{X}_i \times E_i$.

Finally, based on the generated \mathbf{X}_i , E_i , and \mathbf{Z}_i , Y_i is generated from the following Bernoulli distribution

$$Y_i \sim \text{Bernoulli}(p(\mathbf{X}_i, E_i, \mathbf{Z}_i)) \tag{4.6}$$

$$p(\mathbf{X}_i, E_i, \mathbf{Z}_i) = \frac{\exp(\beta_{\mathbf{X}}^T \mathbf{X}_i + \beta_E E_i + \beta_{\mathbf{Z}}^T \mathbf{Z}_i)}{1 + \exp(\beta_{\mathbf{X}}^T \mathbf{X}_i + \beta_E E_i + \beta_{\mathbf{Z}}^T \mathbf{Z}_i)}$$

where $\beta_{\mathbf{X}}^T$, β_E and $\beta_{\mathbf{Z}}^T$ are the effects of n CpGs, one environmental factor and n pairwise CpG \times E interactions on outcome Y .

In each simulation, we set $\mu_j \sim N(-0.47, 3.56)$, $j = 1, \dots, n$, for n CpGs, where -0.47 and 3.56 are the mean and SD of DNA methylation means of all CpGs with gene information from the 432 samples in the CCCEH MN cohort. We set $\sigma_j \sim N(0.62, 0.21)$, $j = 1, \dots, n$, where 0.62 and 0.21 are the mean and SD of methylation SDs. We generated 100 cases and 100 controls. We set all β 's to be 0 to evaluate type I error rates and considered multiple scenarios when signal CpGs have main signals only, interaction signals only, and both main and interaction signals to evaluate power with null CpGs having $\beta = 0$.

4.3.1.1 Simulation settings with different types of signals

We set a gene with 30 CpGs with 1~4 CpGs having 1) main signals only, 2) interaction signals only, and 3) both main and interaction signals. Detailed simulation setups are in Table 4.1.

4.3.1.2 Simulation settings with fixed number of signal items from different number of signal CpGs

A signal item represents a signal in the data matrix $\mathbf{X}^{\text{main-int}}$ regardless it is a main/interaction signal. Because we consider interaction signals as another type of signal compared to main signals, we investigated power when the same signal composition is from different number of signal CpGs. Detailed simulation setups are in Supplementary Table C.2.

Table 4.1: Simulation settings with different types of signals

Scenario	Number of signal items ^a	Simulation setup ^b
Main signals only	1 signal CpG	$\beta_{X_1} = 0.4$
	2 signal CpGs	$\beta_{X_1} = \beta_{X_3} = 0.4$
	3 signal CpGs	$\beta_{X_1} = \beta_{X_3} = \beta_{X_5} = 0.4$
	4 signal CpGs	$\beta_{X_1} = \beta_{X_3} = \beta_{X_5} = \beta_{X_7} = 0.4$
Interaction signals only	1 signal CpG	$\beta_{Z_1} = 0.4$
	2 signal CpGs	$\beta_{Z_1} = \beta_{Z_3} = 0.4$
	3 signal CpGs	$\beta_{Z_1} = \beta_{Z_3} = \beta_{Z_5} = 0.4$
	4 signal CpGs	$\beta_{Z_1} = \beta_{Z_3} = \beta_{Z_5} = \beta_{Z_7} = 0.4$
Both main and interaction signals with fixed number of signal CpGs	1 signal CpG with main signals	
	3 signal CpGs with interaction signals (main-to-interaction signal ratio = 1:3)	$\beta_{X_1} = \beta_{Z_3} = \beta_{Z_5} = \beta_{Z_7} = 0.4$
	2 signal CpGs with main signals	
	2 signal CpGs with interaction signals (main-to-interaction signal ratio = 2:2)	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_5} = \beta_{Z_7} = 0.4$
	3 signal CpGs with main signals	
	1 signal CpG with interaction signals (main-to-interaction signal ratio = 3:1)	$\beta_{X_1} = \beta_{X_3} = \beta_{X_5} = \beta_{Z_7} = 0.4$

^aA signal item represents a signal in the data matrix $\mathbf{X}^{\text{main-int}}$ no matter it is a main signal or an interaction signal.

^b X represents DNA methylation main effects, Z represents DNA methylation by environment interaction effects.

4.3.2 Simulation results

4.3.2.1 Type I error rate

Type I error rates are well controlled at the 0.05 significance level in all simulation settings for all methods (Table 4.2).

Table 4.2: Type I error rates

Methods	20 CpGs*	30 CpGs	40 CpGs
$\mathbf{D}^{\text{w-main-int}}$	0.042	0.049	0.053
$\mathbf{D}^{\text{w-main}}$	0.050	0.052	0.057
$\mathbf{D}^{\text{w-int}}$	0.044	0.045	0.047
$\mathbf{D}^{\text{main-int}}$	0.047	0.051	0.045
\mathbf{D}^{main}	0.039	0.055	0.049
\mathbf{D}^{int}	0.046	0.046	0.049
L^S	0.036	0.038	0.027
L^M	0.037	0.039	0.035

*Number of CpGs in a gene.

4.3.2.2 Simulation settings with different types of signals

As summarized in Figure 4.1, when there are only main signals, $\mathbf{D}^{\text{w-int}}$ and \mathbf{D}^{int} that only consider interaction signals have no power, as expected. $\mathbf{D}^{\text{w-main-int}}$ is slightly less powerful than $\mathbf{D}^{\text{w-main}}$ and similar to L^S . This is because the overall main signals are diluted by the inclusion of pseudo-data for interactions when there are no interaction signals. $\mathbf{D}^{\text{main-int}}$ performs similarly as L^M , while both of them perform inferior to $\mathbf{D}^{\text{w-main-int}}$ with weights. In general, the weighted versions $\mathbf{D}^{\text{w-main-int}}$ and $\mathbf{D}^{\text{w-main}}$ outperform the corresponding non-weighted versions, suggesting that incorporating association strength weights in calculating distances indeed helps up-weight signals and down-weight noises thus improves the overall power.

When there are only interaction signals, $\mathbf{D}^{\text{w-main}}$, \mathbf{D}^{main} and L^S that only consider main signals have no power, as expected. $\mathbf{D}^{\text{w-main-int}}$ is slightly less powerful than $\mathbf{D}^{\text{w-int}}$ when both of them outperform the corresponding non-weighted versions. $\mathbf{D}^{\text{main-int}}$ performs similarly as L^M .

When there are both main and interaction signals, we fixed the number of signal items and the number of signal CpGs to be 4 but varying the main-to-interaction signal ratio, i.e.,

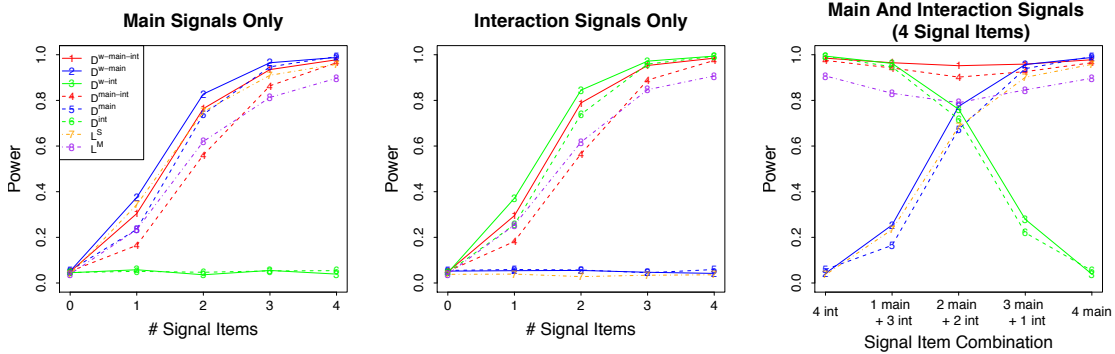


Figure 4.1: Power results for simulation settings with main signals only, interaction signals only and both main and interaction signals when there are 30 CpGs in a gene.

the ratio between the number of main signal CpGs and the number of interaction signal CpGs. As the main-to-interaction signal ratio increases, the power of $\mathbf{D}^{w\text{-main}}$, \mathbf{D}^{main} and L^S that only consider main signals increases, while that of $\mathbf{D}^{w\text{-int}}$ and \mathbf{D}^{int} that only consider interaction signals decreases, and that of $\mathbf{D}^{w\text{-main-int}}$, $\mathbf{D}^{\text{main-int}}$ and L^M that consider both main and interaction signals remains the same. Importantly, $\mathbf{D}^{w\text{-main-int}}$ consistently has the largest power, which implies that the performance of $\mathbf{D}^{w\text{-main-int}}$ is not affected by signal types. Again, the weighted versions outperform the non-weighted versions.

We also considered when there are 20 or 40 CpGs in a gene and summarized results in Supplementary Figure C.1. We found that when we fix the number of signal CpGs but increase the number of noise CpGs in a gene, power of non-weighted methods decreases, while power of weighted versions is well maintained. This suggests that adding weights is effective, especially when a smaller percent of CpGs in a gene are signals. This is consistent with that was observed in our previous work (Wang et al., 2018).

4.3.2.3 Simulation settings with fixed number of signal items from different number of signal CpGs

Power results for simulation settings with fixed number of signal items from different number of signal CpGs are summarized in Supplementary Materials (Section C.1.3) and Supplementary Figure C.2. Overall, the power of distance-based methods increases as the number of signals CpGs increases.

4.4 Real data applications

4.4.1 CCCEH birth cohorts

Between 1998 and 2006, 727 pregnant women residing in Washington Heights, Harlem and the South Bronx were recruited in prenatal clinics to participate in the CCCEH Mothers and Newborns (MN) prospective cohort study. During the 3rd trimester of pregnancy, women were asked to wear a small backpack containing a personal monitor during the daytime for 48 hours. The collected samples were then analyzed for 8 carcinogenic PAHs (Perera et al., 2003). The PAH metric used in the analysis is the sum of 8 carcinogenic PAHs and was dichotomized at the median in the parent population (2.26 ng/m³). In-person postnatal questionnaires were given when the child was 6 months and annually thereafter with developmental questionnaires and assessments were administered every 1-2 years. We have also measured DNA methylation in the white blood cells of umbilical cord blood.

Beginning in March 2008, pregnant women enrolled in the CCCEH MN Study were invited to participate in the CCCEH Sibling Study. Similar to the parent study, women were enrolled if they had a prenatal visit by the 20th week of pregnancy, and were not active smokers or illicit drug users. The same protocol was followed as in the MN cohort. Children were followed until age 7, with assessments of early childhood developmental and behavioral outcomes and cord blood DNA methylation.

4.4.2 Neurodevelopment outcomes

We investigated the associations between prenatal PAH and DNA methylation on neurodevelopmental outcomes when their interactions are considered. We assessed two neurodevelopment outcomes at age of 3: i) Child Behavior Checklist (CBCL) DSM-IV-oriented Attention Deficit Hyperactivity Disorder (ADHD) (Association, 2013) and ii) the Bayley Scales of Infant Development Mental Development Index (MDI) (Bayley, 1993).

Since ADHD diagnosis at age 3 may not be clinically reliable and the main purpose is to demonstrate the superior performance of the proposed method over comparison methods, we dichotomized ADHD at T-score of 50 (high ADHD group T-score > 50 and low with T-score ≤ 50), which is the median of the normed population derived from the raw scores

(Achenbach and Rescorla, 2000). Note that a T-score of 50 was assigned to those with raw scores below the population median, i.e., no differentiation for those below the population median, while a percentile-type T-score was assigned to those above the population median. We performed the discovery analysis using the MN cohort and the replication analysis using the Sibling cohort.

For the MDI outcome, children are dichotomized as normal ($\text{MDI} \geq 85$) or moderately to severely delayed ($\text{MDI} < 85$) (Perera et al., 2006). Since there is only one case of moderately to severely delayed child in the Sibling cohort, to conduct discovery and replication analyses, we randomly split the MN cohort using 2/3 samples for the discovery analysis and 1/3 for the replication analysis.

4.4.3 DNA methylation data processing

We conducted standard data processing steps for DNA methylation with details in Supplementary Materials (Section C.2.1).

4.4.4 Risk of PAH, DNA methylation and their interactions on ADHD

There are 328 samples with complete data of DNA methylation, prenatal PAH and ADHD in the discovery MN cohort, and 43 samples with complete data in the replication Sibling cohort.

4.4.4.1 Discovery analysis in the MN cohort

Since the main purpose is to demonstrate the power of the proposed method $\mathbf{D}^{\text{w-main-int}}$ over comparison methods, instead of using the Bonferroni adjustment for 18,633 genes, we used a subjective threshold of 0.005 on the empirical gene-level P -values obtained from the permutation procedure. At the 0.005 threshold, $\mathbf{D}^{\text{w-main-int}}$ identified 10 genes in the discovery analysis, with 7 due to main signals only and 3 due to interaction signals only (Table 4.3).

Table 4.3: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 10 genes by the proposed method $\mathbf{D}^{\text{w-main-int}}$ at the 0.005 gene-level P -value threshold

Rank in $\mathbf{D}^{\text{w-main-int}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main}}$	Rank in $\mathbf{D}^{\text{w-int}}$
1	<i>LOC84931</i> *	9	1	1513
2	<i>SERPINB3</i>	1	2	18316
3	<i>CYP2E1</i> *	13	6041	1
4	<i>MIR518E</i>	1	15105	2
5	<i>KIR3DP1</i>	1	18630	3
6	<i>KRTAP20-1</i>	1	10	18472
7	<i>IGJ</i>	1	4	18286
8	<i>ADAM32</i>	11	5	15841
9	<i>HIST1H2BJ</i> *	4	3	14178
10	<i>CXCL9</i>	1	11	16665

* Genes replicated in the replication analysis.

4.4.4.2 Replication analysis in the Sibling cohort

Due to the small sample size of the Sibling cohort, we used a gene-level P -value threshold of 0.1 in the replication analysis. Among the 10 genes identified in the discovery MN cohort, 3 (*LOC84931*, *CYP2E1* and *HIST1H2BJ*) were replicated in the replication Sibling cohort. In both discovery and replication analyses, gene *CYP2E1* was identified due to interaction signals, and genes *LOC84931* and *HIST1H2BJ* were identified due to main signals.

Figure 4.2 plots boxplots of methylation measures of the 13 CpGs in gene *CYP2E1*, identified and replicated due to interaction signals, stratifying by PAH and ADHD. Eight out of the 13 CpGs have clear interaction signals in the discovery data, when all 8 showed interaction signals in the same direction in the replication data. It was reported that prenatal exposure to serotonin reuptake inhibitor antidepressants modifies the association between DNA methylation at regulatory region of *CYP2E1* and 3rd trimester maternal depressed mood

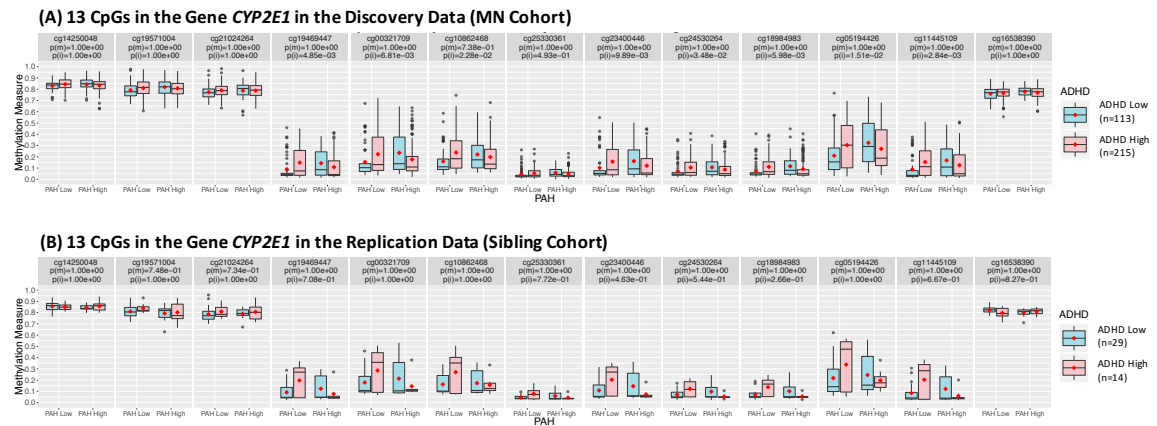


Figure 4.2: Boxplot of DNA methylation measures of the 13 CpGs in gene *CYP2E1* stratified by PAH and ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *CYP2E1*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\logitP(Y = 1) = \beta_0 + \beta_1 \text{CpG} + \beta_2 E + \beta_3 \text{CpG} \times E$.

symptoms (Gurnot et al., 2015). Elevated DNA methylation in the promoter-regulatory region of the gene *CYP2E1* was also reported to be associated with severe psychosocial deprivation in early childhood and socio-cognitive impairment (Kumsta et al., 2016). We similarly plotted for genes *LOC84931* and *HIST1H2BJ* (Supplementary Figures C.3, C.4).

4.4.4.3 Results of the comparison methods

At the same 0.005 P -value threshold, the comparison methods identified different number of genes (Supplementary Tables C.3-C.9), when all these genes rank within top 3% of the proposed method results. The comparison methods have replication rates 0-40% with an average 14% (Supplementary Table C.10). Detailed results are in Supplementary Materials (Section C.2.2.2).

4.4.5 Risk of PAH, DNA methylation and their interactions on MDI

Two-third MN samples ($n=216$) were used for the discovery analysis and 1/3 ($n=94$) for the replication analysis.

Table 4.4: Application examining prenatal PAH, DNA methylation and their interactions on child MDI at age 3 identified 7 genes by the proposed method at the 0.005 gene-level P -value threshold

Rank in $\mathbf{D}^{\text{w-main-int}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main}}$	Rank in $\mathbf{D}^{\text{w-int}}$
1	<i>UROS</i>	2	2	18516
2	<i>FAM35A</i> *	7	1	15325
3	<i>DIRC1</i> *	3	6	17815
4	<i>MIR521-1</i>	1	16	18302
5	<i>C8orf80</i> *	4	3	2329
6	<i>THSD1P</i> *	5	7	15099
7	<i>C19orf77</i>	9	5	647

*Genes replicated in the replication analysis.

4.4.5.1 Discovery analysis in the discovery data

At the same 0.005 P -value threshold, the proposed method $\mathbf{D}^{\text{w-main-int}}$ identified 7 genes in the discovery analysis, with 5 due to main signals only and 2 due to both main and interaction signals (Table 4.4).

4.4.5.2 Replication analysis in the replication data

At the same 0.1 gene-level P -value threshold for replication, 3 genes, *FAM35A*, *DIRC1* and *THSD1P*, were replicated in the replication analysis due to main signals out of the 5 genes identified in the discovery analysis due to main signals only. Gene *C8orf80* was replicated due to interaction signals, out of the 2 genes identified in the discovery analysis due to both main and interaction signals. That is, the replication rate is 57% with 4 out of 7 genes replicated. Figure 4.3 plots boxplots of DNA methylation measures of the 4 CpGs in gene *C8orf80* stratified by PAH and MDI status that was identified due to both main and interaction signals and replicated due to interaction signals. We similarly plotted genes *FAM35A*, *DIRC1* and *THSD1P* (Supplementary Figures C.5-C.7).

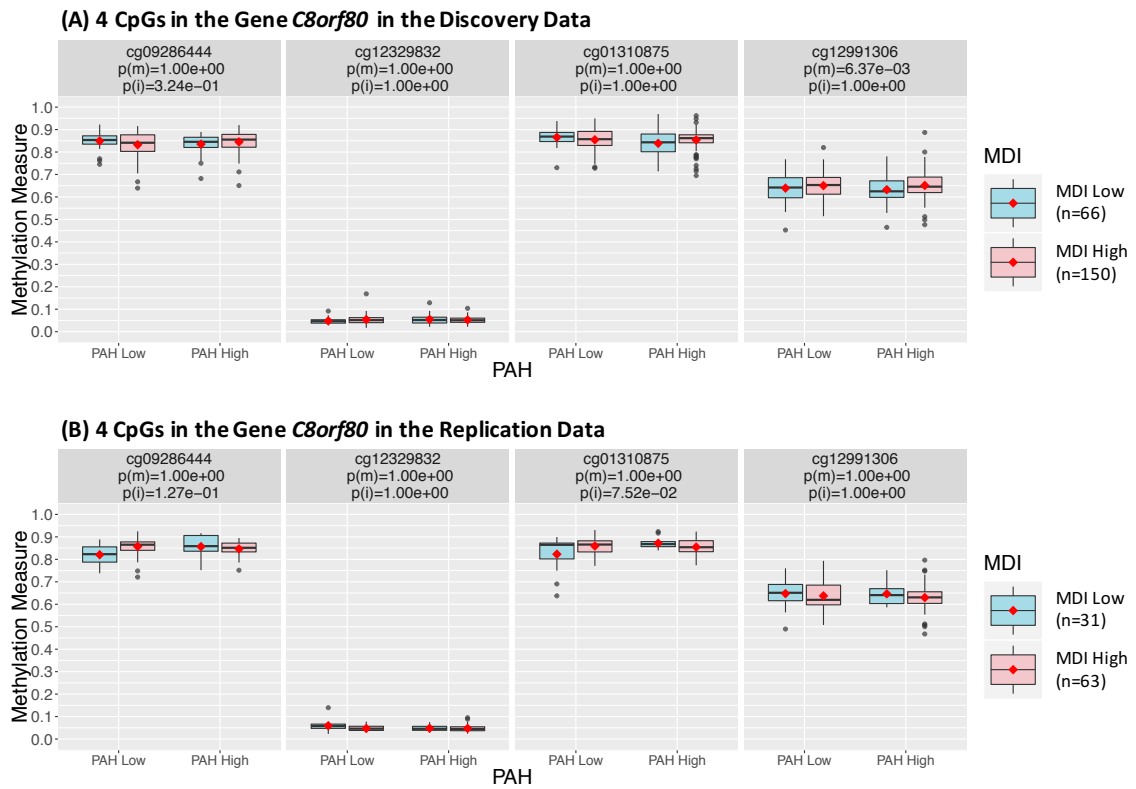


Figure 4.3: Boxplots of DNA methylation measures of the 4 CpGs in gene *C8orf80* stratified by PAH and MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *C8orf80*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2E + \beta_3\text{CpG} \times E$.

4.4.5.3 Results of the comparison methods

All genes identified by the comparison methods rank within top 2% of the proposed method results. The comparison methods have replication rates 0-25% with an average 9% (Supplementary Table C.11 and details in Supplementary Materials Section C.2.3).

4.5 Discussion

We developed a novel weighted distance-based method $\mathbf{D}^{\text{w-main-int}}$ that considered interactions between CpGs in a gene and an environmental factor through constructing a pseudo-data matrix with their cross-product terms. The proposed approach is powerful and flexible with several advantages. First, the weighted distance matrix $\mathbf{D}^{\text{w-main-int}}$ always has a dimension $N \times N$ with N being the sample size regardless the added dimensionality from pairwise interactions. Second, by calculating distances between pairs of individuals across CpGs and their interactions with an environmental factor, weak main/interaction signals are accumulated, boosting the study power. Third, incorporating association strength weights in calculating distances helps up-weight signals and down-weight noises thus further improves the overall power, especially when a small percent of CpGs in a gene are signals. Most importantly, simulation results suggest that when the main-to-interaction signal ratio decreases, i.e., when the number of main signals decreases or the number of interaction signals increases but fixing the total number of signal items, the proposed method $\mathbf{D}^{\text{w-main-int}}$ maintains similar power and almost achieves the highest power among all comparison methods, while the comparison methods have power drop. This makes the proposed approach especially attractive due to the known low power in detecting interactions.

In the application to the CCCEH MN and Sibling cohorts examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3, $\mathbf{D}^{\text{w-main-int}}$ identified 10 genes in the discovery data with 3 replicated in the replication data, while the comparison methods have an average replication rate 14%. In another application on child MDI at age 3, $\mathbf{D}^{\text{w-main-int}}$ identified 7 genes in the discovery data with 4 replicated in the replication data, while the comparison methods have an average replication rate 9%.

In general, the proposed method that considers both main and interaction signals has

a superior performance than methods that consider only one type of signals when there are both. The weighted versions are always more powerful than non-weighted versions, especially when a small percentage of CpGs in a gene have weak signals. The proposed method was developed for DNA methylation by environment interactions but can be readily extended to CpG by CpG interactions similarly using a pseudo-data matrix constructed with cross-product terms between CpGs. However, the dimension of this pseudo-data matrix capturing pairwise CpG by CpG interactions goes up exponentially, which could easily outnumber the dimension of CpGs in the gene. We need to take extra caution to balance between main or interaction signals, especially when assigning weights.

Chapter 5

Conclusions

In this dissertation, new statistical methods for DNA methylation data were developed focusing on detection of differential variation and interactions in order to fill in the gaps in these areas. These work were especially motivated by recent observations that DV contributes to early epigenetic alterations in carcinogenesis, namely epigenetic field defects. Identifying epigenetic field defects is important for early cancer detection, however, existing statistical methods are not powerful enough to detect these early epigenetic alterations comparing normal tissues and normal-adjacent tissues to tumor tissues when the differences are usually minimum.

We developed two methods to detect epigenetic field defects with weak CpG site-level signals. The first method utilizes the dependency among neighboring CpGs to identify differentially methylated regions through combining both DM and DV signals, where in combining the two types of signals, we weighted them differently to balance the contribution of DM and DV signals to the combined score. The superior performance of the proposed method was demonstrated through simulation studies and applications to 450K DNA methylation data of tumor and normal-adjacent tissues of breast invasive carcinoma (BRCA) and kidney renal clear cell carcinoma (KIRC) from The Cancer Genome Atlas (TCGA) project. We identified some cancer-related genes that were missed by the DMR detection methods that use only DM signals or DV signals. The application to an independent 450K DNA methylation data of BRCA tumor and normal-adjacent tissues from Gene Expression Omnibus (GEO) allowed us to replicate some of the detected DMRs and we

concluded that DMRs detected using variance signals are reproducible. Further application to the DNA methylation data of GEO BRCA normal-adjacent tissues from breast cancer patients and normal tissues from age-matched cancer-free women identified epigenetic field defects in two DMRs. Further comparisons between tumor tissues from breast cancer patients to normal tissues from age-matched cancer-free women confirmed that the epigenetic field defects are enriched in the progression to breast cancer. Most importantly, the epigenetic field defects were only identified by the developed new DMR detection method that uses DM and DV combined signals.

The second method accumulates weak DM and DV signals at CpG site-level across CpGs in a gene to detect genes with epigenetic field defects. This is achieved through constructing a pseudo-data matrix with centered quadratic terms of DNA methylation measures that captures DV signals. The epigenetic distances between pairs of subjects can then be calculated combining the original data matrix with measures of DNA methylations together with the pseudo-data matrix with quadratic terms of DNA methylation. Using this approach, we accumulate weak DM and DV signals at CpG site-level across CpGs in a gene. CpG site-level association strengths were added as weights to up-weight signal CpGs and down-weight noise CpGs to further boost the study power. We demonstrated the superior performance of the proposed method through simulation studies and an application to the the same 450K DNA methylation data of normal-adjacent tissues of BRCA patients and normal tissue from independent age-matched cancer-free women from GEO. The proposed method identified genes with epigenetic field defects that were missed by standard EWAS methods and non-weighted distance-based methods, with many of these epigenetic field defects being previously reported to be associated with breast cancer. We further confirmed their enrichment in the progression to breast cancer and replicated some of these identified epigenetic field defects in an independent data.

Other than the known crucial role of DNA methylation in human health, studies have also demonstrated associations between DNA methylation and environmental factors with evidence also supporting the idea that DNA methylation may modify the risk of environmental factors on health outcomes. However, due to high dimensionality and low study power, current studies usually focus on finding DM on health outcomes at CpG level or gene level

combining multiple CpGs and/or finding environmental effects on health outcomes but ignoring their interactions on health outcomes. We developed a powerful and flexible weighted distance-based method that incorporates interactions between DNA methylation and environmental factors on health outcomes. This is achieved through constructing a pseudo-data matrix with cross-product terms between DNA methylation and environmental factors that capture interactions between them. The distances between pairs of subjects can be similarly calculated combining the original data matrix with measures of DNA methylation and environmental factors together with the pseudo-data matrix with interactions. Using this approach, we can identify both main and interaction effects. CpG site-level association strengths were added as weights to up-weight signal CpGs and down-weight noise CpGs to further boost the study power. The proposed approach is powerful and flexible with several advantages. First, the weighted distance matrix always has a dimension of $N \times N$ with N being the sample size regardless the added dimensionality from pairwise interactions. Second, by calculating distances between pairs of individuals across CpGs and their interactions with an environmental factor, weak main/interaction signals are accumulated, boosting the study power. Most importantly, simulation results suggest that when the main-to-interaction signal ratio decreases, i.e., when the number of main signals decreases or the number of interaction signals increases but fixing the total number of signal items, the proposed method maintains similar power and almost achieves the highest power among all comparison methods, while the comparison methods have power drop. That is, the power of the proposed method is not affected by the source of the signals, i.e., if the signals are main or interaction effects. This makes the proposed approach especially attractive due to the known low power in detecting interactions. In the application to the data from the Mothers and Newborns (MN) birth cohort of the Columbia Center for Children's Environmental Health (CCCEH) to identify effects of gene-level DNA methylation, prenatal PAH and their interactions on Attention Deficit Hyperactivity Disorder (ADHD) at age 3, we identified some main effects of DNA methylation and some interactions with prenatal PAH which were missed by comparison methods. Some of these findings were further replicated in the CCCEH Sibling cohort. We similarly applied the proposed method to the Mental Development Index (MDI) at age 3 and observed a similar pattern in results in both discovery

and replication analyses.

Since the weighted distance matrix always has a dimension of $N \times N$ with N being the sample size regardless the added dimensionality from pairwise interaction terms, it can also be extended to CpG by CpG interactions similarly using a pseudo-data matrix constructed with cross-product terms between CpGs.

Part I

Appendices

Appendix A

Appendix to Accounting for Differential Variability in Detecting Differentially Methylated Regions

A.1 Investigation of the distance limits to define clusters

We investigated the relationship between the choice of distance limit, the maximal distance between two neighboring sites to be included in a cluster, and the distribution of the difference in the combined signals scores (before smoothing) between neighboring CpG sites using TCGA BRCA data of tumor and normal-adjacent tissues.

Within each chromosome, we ordered the combined signal scores by their genomic locations and calculated the differences in the combined signals scores between neighboring CpG sites. We then change the distance limits from 300 bp to 2,000 bp (300 bps, 500 bps, 700 bps, 1,000 bps, 1,500 bps, and 2,000 bps) and plotted the distribution of the differences for neighboring CpG sites whose distance is less than the specified distance limit (Figure A.1). We found that the mean and SD in the difference in combined signal scores between neighboring CpG sites increases as the distance limits increases. In the developed algorithm, users could choose other distance limits as an option.

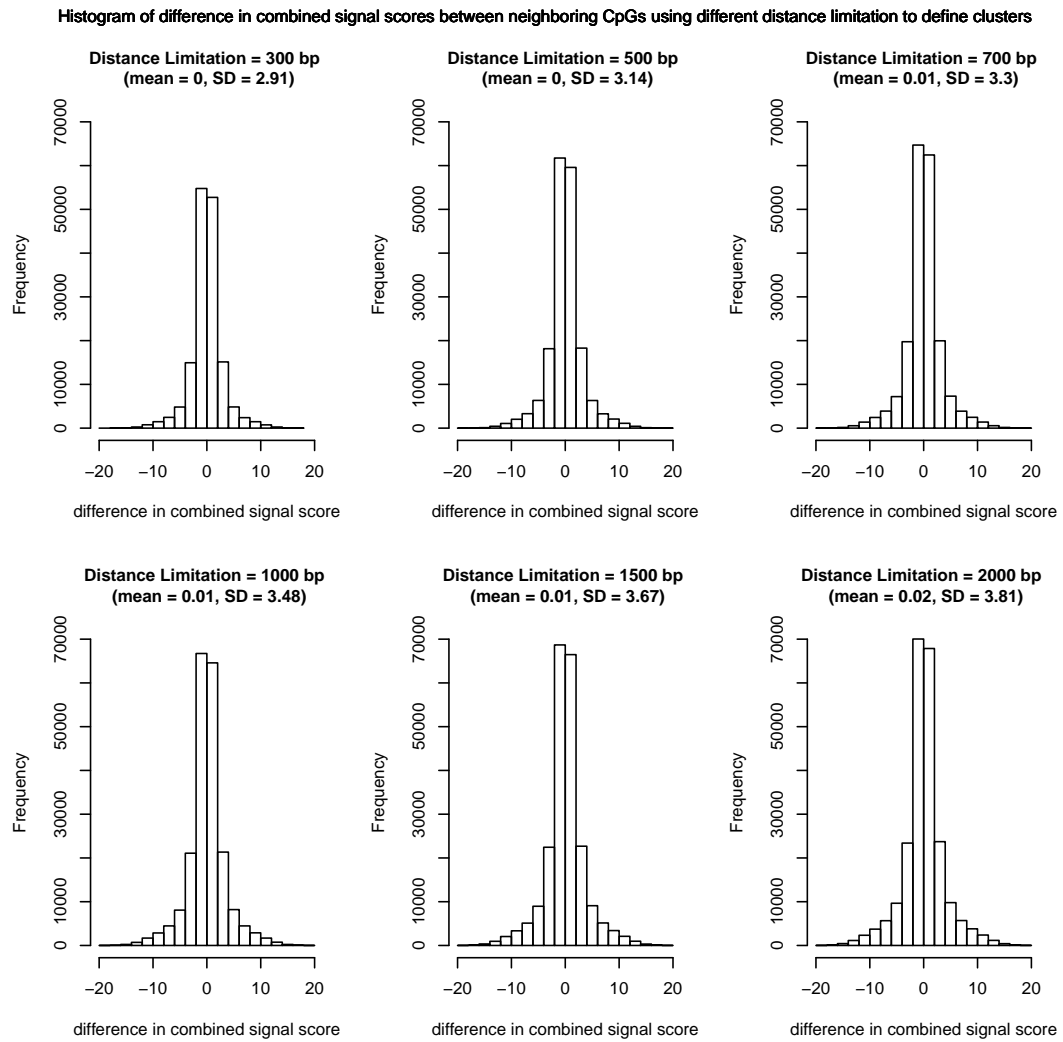


Figure A.1: Histogram of the difference in the combined signal scores for neighboring CpG sites with the choice of difference distance limits.

A.2 Simulation studies for case-control designs

We adapted the proposed new DMR detection method for case-control designs and conducted simulation studies parallel as for matched case-control designs in the main text to evaluate the type I errors and the performance. We compared the performance of the new method with those DMR detection methods that consider 1) mean signals only using two-sided two-sample t -test, the bump hunting method (Jaffe et al., 2012c), the modified bump hunting method which divide the regression coefficient estimates from the original bump hunting method by their standard errors; and 2) variance signals only using one-sided F -test, where we applied the same smoothing step and the same significance assessment step.

A.2.1 Simulation setup

To simulate DNA methylation measures of tumor and normal tissues, we assume logit2 transformed methylation measures, M -values, of sample s follows a multivariate normal distribution

$$M_{s,k} \sim N_{l_k}(\boldsymbol{\mu}, \Delta^T \Sigma \Delta) \quad (\text{A.1})$$

where l_k is the size of the k -th cluster, and the mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{l_k})^T$ and diagonal matrix $\Delta = \text{diag}(\sqrt{\delta_1}, \dots, \sqrt{\delta_{l_k}})$ controls the mean and variance signals. Σ is a variance-covariance matrix ($l_k \times l_k$) considering correlations among l_k CpG sites within the k -th pre-defined cluster. Here we assume an $AR(1)$ correlation with correlation coefficient ρ , i.e., $\Sigma_{mn} = \sigma \times \rho^{|m-n|}$. We set $\rho = 0.5$ in the simulation studies similarly as in matched case-control designs, and set $\sigma = 0.25$. In each simulation, we generated methylation M -values of 10,000 sites from 100 cancer patients and 100 normal controls, where the genomic locations of these 10,000 sites are the first 10,000 sites of Chromosome 1 on the Illumina 450K array.

Since DNA methylation measures are known to be associated with variables such as age (Christensen et al., 2009; Teschendorff et al., 2010) and gender (Liu et al., 2010), we work on methylation residuals after adjusting for such confounders. We investigated type I errors to examine if using methylation residuals controls potential spurious DMRs due to unbalanced distribution of confounders, such as gender. More specifically, we set 50% of

cancer patients to be female while only 20% of normal controls to be female. We simulated 10 spurious DMRs each having 10 CpG sites, within which we set $\mu = 1$ for both tumor and normal tissues in the female group, while $\mu = 0$ for both tumor and normal tissues in the male group. For all other sites, we set $\mu = 0$ for tumor and normal tissues in both gender groups. We considered two scenarios where we applied the new method on: 1) methylation residuals obtained from regressing methylation M -values on gender using linear models, and 2) methylation M -values directly ignoring gender. We conducted 1,000 simulations in each scenario.

In sections to evaluate the performance of the new method, we assume confounders are already accounted for when methylation residuals are used. We simulated 10 true DMRs with different region sizes varying from 3 to 15 CpG sites, and we considered scenarios when each CpG site in the true DMRs has 1) mean signals only, 2) variance signals only, and 3) both mean and variance signals. For all other null sites, we set $\mu = 0$ and $\sigma = 0.25$. For each simulation scenario, we conducted 1,000 simulations.

A.2.2 Simulation results

The type I errors of the new method that considers both mean and variance signals, the two-sample t -test that considers mean signals only, and the F -test that considers variance signals only were all well controlled at 0.039, 0.023 and 0.059 when applied to methylation residuals. When applied to the methylation measures ignoring the gender effect, the type I errors were all inflated at 1.000, 1.000 and 0.997. The type I errors of bump hunting (Jaffe et al., 2012c) and modified bump hunting that directly adjust for gender effect were both well controlled at 0.041. The region size was set at $L \geq 3$ CpG sites.

The ROC curves from the setting with 10 true DMRs having different region sizes are shown in Figure A.2. Similar patterns as in matched case-control designs are observed.

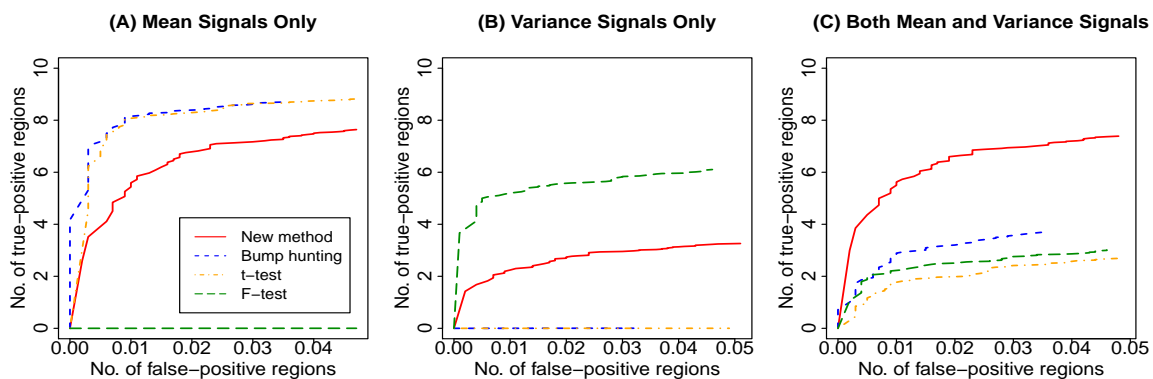


Figure A.2: ROC curves from simulation studies when 10 true DMRs have different region sizes varying from 3 to 15 CpG sites with: (A) mean signals only; (B) variance signals only; and (C) both mean and variance signals. DMRs were defined as regions with minimum region size $L \geq 3$ CpG sites.

A.3 Real data application

Table A.1: Significant DMRs Identified in the TCGA KIRC Data (160 matched pairs)

DMRs ($L^a \geq 3$)	New method	Wilcoxon signed-rank test	Paired t -test	DMRcate	Probe Lasso	Pitman-Morgan test	KS test ^b
Total No. of DMRs (Total No. of DMR-covered CpG sites)	2164 (30558)	146 (4162)	2953 (33786)	23332 (153192)	7996 (41195)	1457 (25070)	1697 (28954)
Mean(SD) number of CpG sites per DMR	14 (7)	29 (7)	11 (6)	7(5)	5 (5)	17 (10)	17 (9)
Mean(SD) number of base pairs per DMR	2575 (1962)	5007 (2489)	2074 (1464)	1207 (1076)	772 (1166)	3226 (2139)	3239 (2127)
No. of overlapping DMRs ^c	-	146	1716	2164	1185	1292	1332

^a L : minimum region size, i.e., minimum number of CpG sites

^bKolmogorov-Smirnov test

^cNo. of overlapping DMRs: a DMR identified by the new method is considered to overlap if this DMR has any overlap with DMRs identified by each comparison method.

Table A.2: 7 Cancer-Related Genes Identified in the Top 10 Ranked DMRs in TCGA KIRC Data^a

Cancer	Gene
Breast Cancer	<i>MCF2L2</i> (Legendre et al., 2015)
Colorectal Cancer	<i>GAD2</i> (Li et al., 2012)
Endometrial Carcinoma	<i>PAX2</i> (Wu et al., 2005)
Hepatocellular Carcinoma	<i>DCAF4L2</i> (Song et al., 2013)
Melanoma	<i>GPR98</i> (Harvey et al., 2013)
Pancreatic Cancer	<i>FOXL1</i> (Zhang et al., 2013a)
Stomach Cancer	<i>RIMS2</i> (Ewing et al., 2015)

^a: There are 10 genes in the top 10 ranked DMRs out of 100 significant DMRs that were uniquely identified by the new method (compared with all five competing methods except for DMRcate), and 7 genes were previously reported to be cancer-related.

Table A.3: 11 Cancer-Related Genes Identified in the Top 10 Ranked DMRs in GEO BRCA Data (Tumor vs. Normal-adjacent)^a

Cancer	Gene
Breast Cancer	<i>CBX8</i> (Lee et al., 2013), <i>NXP1</i> (Faryna et al., 2012)
Colorectal Cancer	<i>VIM</i> (Costa et al., 2010), <i>WNT1</i> (He et al., 2005)
Gastric Cancer	<i>FOXD3</i> (Cheng et al., 2013), <i>RASGRF1</i> (Takamaru et al., 2012)
Lung Cancer	<i>C6orf176</i> (Chen et al., 2016)
Ovarian Cancer	<i>HIST1H3G</i> (Zhang and Luo, 2016), <i>HIST1H2BI</i> (Hong et al., 2010), <i>VCAN</i> (Ghosh et al., 2010)
Prostate Cancer	<i>GFRA1</i> (Huber et al., 2015)

^a: There are 12 genes in the top 10 ranked DMRs out of 37 significant DMRs that were uniquely identified by the new method (compared with all five competing methods except for DMRcate), and 11 genes were previously reported to be cancer-related.

Table A.4: Significant DMRs Identified in the GEO BRCA Data (Tumor vs. Normal)

DMRs ($L^a \geq 3$)	New method	t -test	Bump hunting	Modified bump hunting	F -test
Total No. of DMRs (Total No. ofDMR-covered CpG sites)	830 (15692)	2097 (28384)	683 (11445)	860 (14600)	94 (2537)
Mean (SD) number of CpG sites per DMR	19 (9)	14 (7)	17 (8)	17 (8)	27 (12)
Mean (SD) number of base pairs per DMR	3276 (1955)	2418 (1539)	2898 (1739)	2974 (1754)	4047 (2819)
No. of overlapping DMRs ^b	-	811	498	529	85

^a L : minimum region size, i.e., minimum number of CpG sites

^bNo. of overlapping DMRs: number of DMRs identified by the new method that has any overlap with DMRs identified by each comparison method.

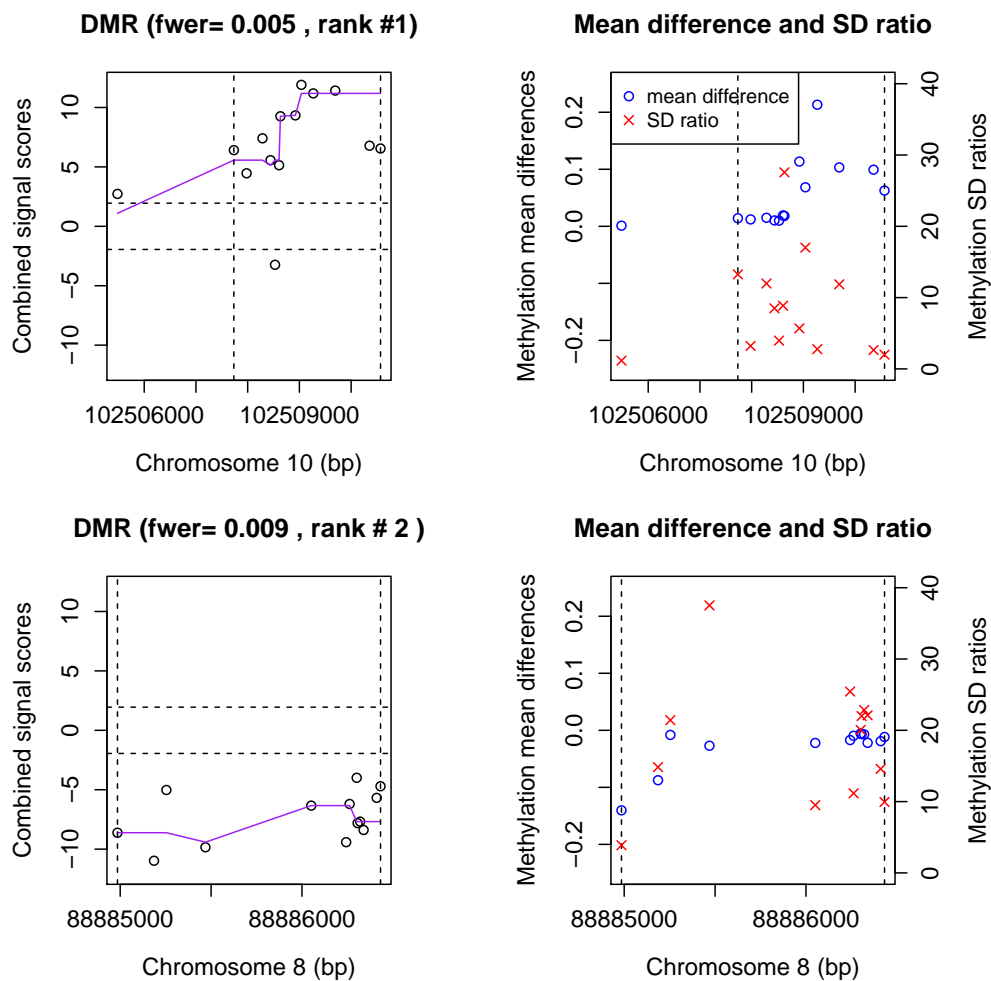


Figure A.3: DMR #1 (top row) and #2 (bottom row) located on chromosomes 10 and 8 (out of 170 DMRs) that were identified uniquely by the new method in the TCGA KIRC data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal-adjacent tissues.

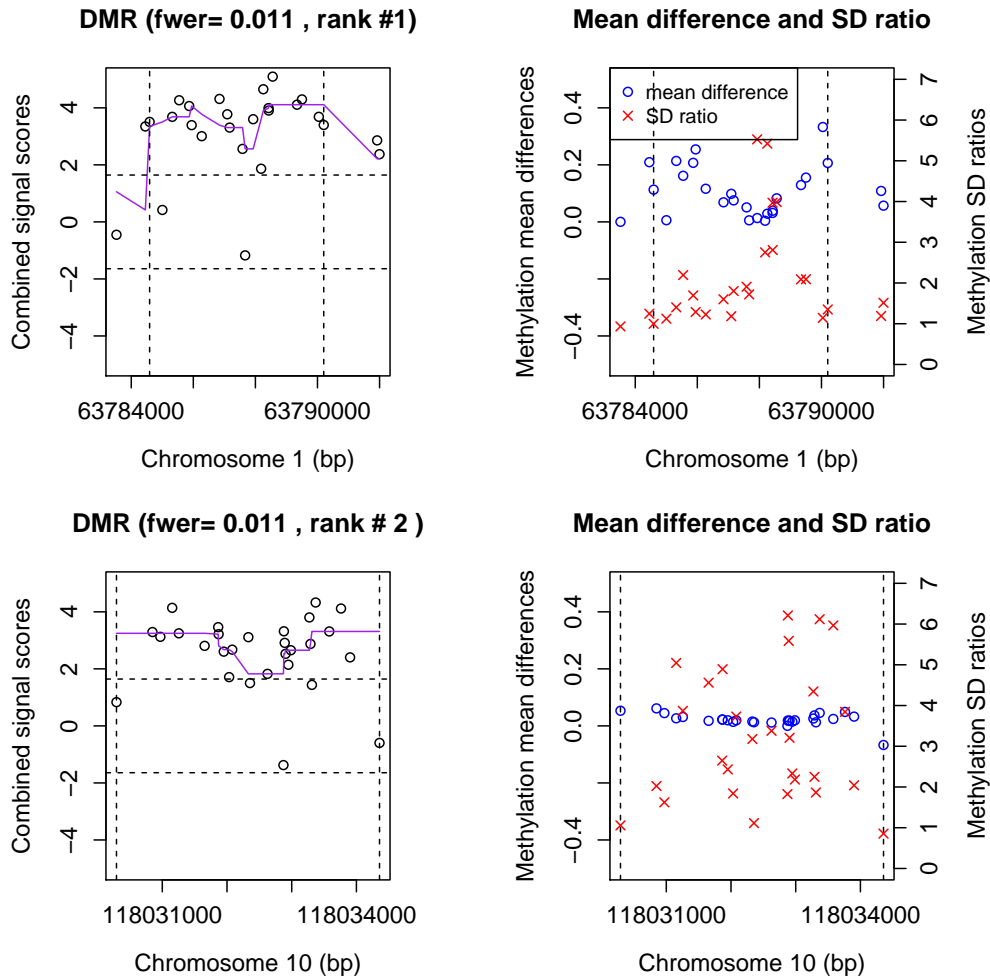


Figure A.4: DMR #1 (top row) and #2 (bottom row) located on chromosomes 1 and 10 (out of 89 DMRs) that were identified uniquely by the new method in the GEO BRCA tumor vs. normal-adjacent data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal-adjacent tissues. There are 3 gene, *SGCE*, *PEG10* and *PHOX2B* in these 2 DMRs. *SGCE* was reported to be associated with colorectal cancer (Ortega et al., 2010), *PEG10* was reported to be associated with hepatocellular carcinoma (Ip et al., 2007), and *PHOX2B* was reported to be associated with neuroblastoma (De Pontual et al., 2007)

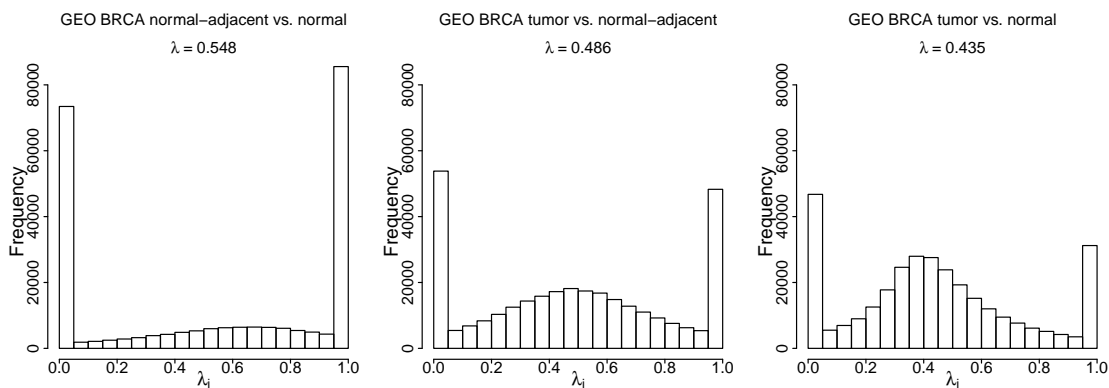


Figure A.5: Distributions of genome-wide site-level scale parameter λ_i in the GEO BRCA data. From left to right shows distribution of λ_i in (1) normal-adjacent vs. normal, (2) tumor vs. normal-adjacent, and (3) tumor vs. normal comparisons.

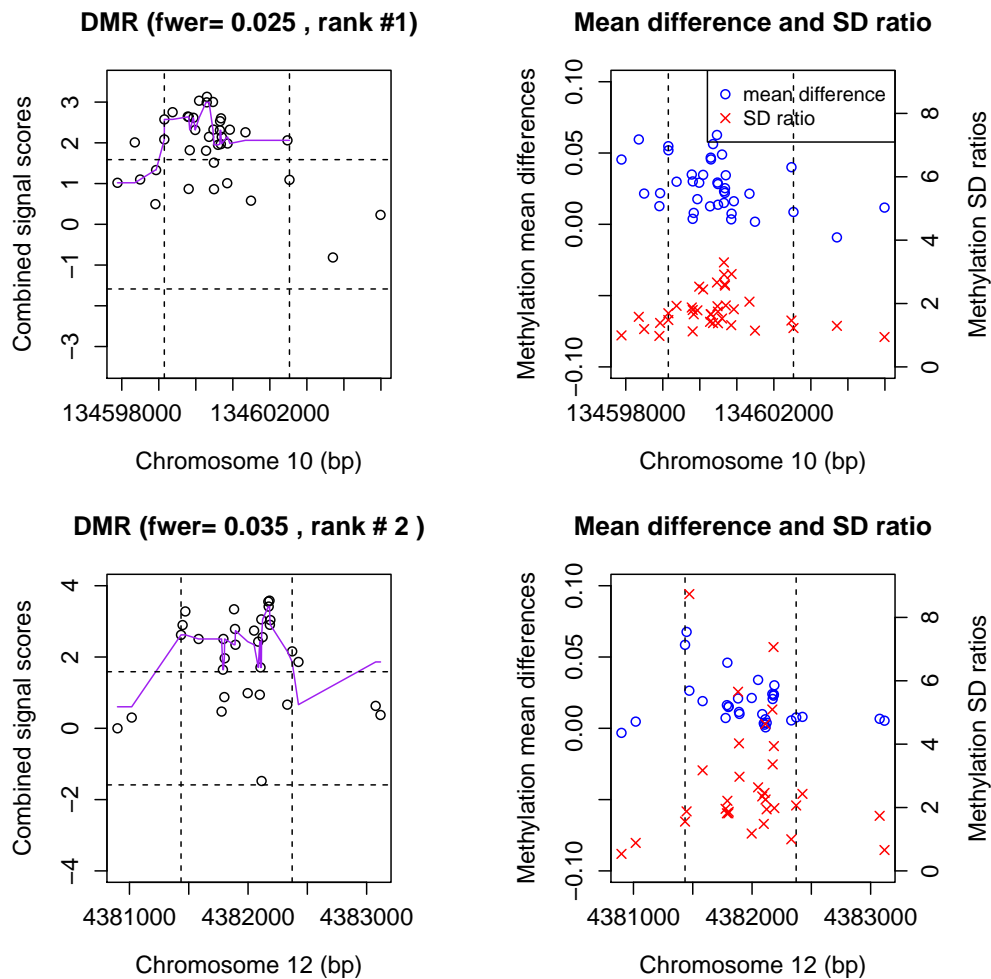


Figure A.6: DMR #1 (top row) and #2 (bottom row) located on chromosomes 10 and 12 that were identified uniquely by the new method in the GEO BRCA normal-adjacent vs. normal data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between normal-adjacent and normal tissues.

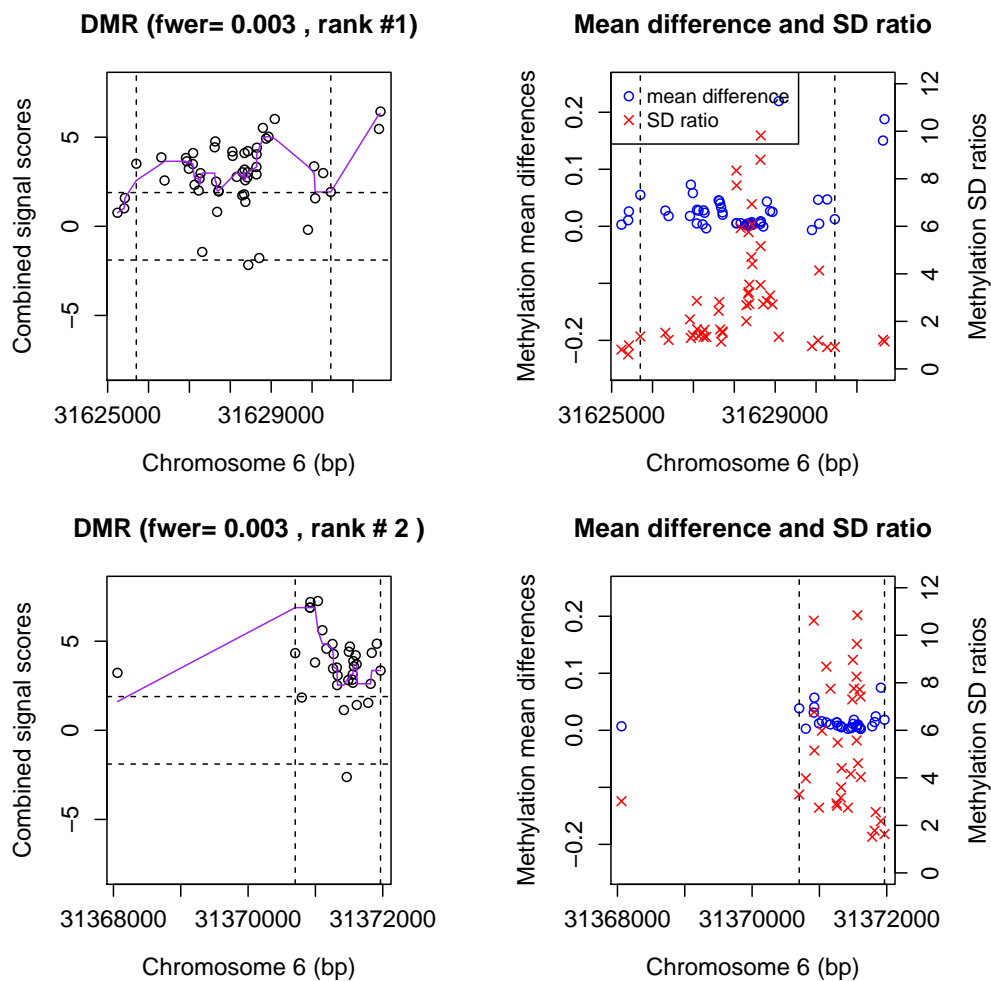


Figure A.7: DMR #1 (top row) and #2 (bottom row) located on chromosomes 6 (out of 15 DMRs) that were identified uniquely by the new method in the GEO BRCA tumor vs. normal data. The vertical dash lines define the boundaries of the DMRs. Left column shows the combined signal scores of sites in the identified DMRs before (circles) and after (curve) smoothing, where the horizontal dotted line defines the threshold k to define a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the identified DMRs between tumor and normal tissues.

Appendix B

Appendix to Detection of Epigenetic Field Defects Using a Weighted Epigenetic Distance-Based Method

B.1 Additional Simulation Studies

B.1.1 Effects of gene sizes in Type I errors

To investigate if genes with different sizes, i.e., number of CpGs, will have different distributions for pseudo- F statistics under the null hypothesis, we conducted simulation studies to evaluate type I error rates of the proposed method and those of the comparison methods. Specifically, we simulated methylation measures for 16 genes that consist of 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, and 100 CpGs, respectively. When calculating the P -value for each gene, we (1) pool all pseudo- F statistics of the 16 genes across all permutations, and (2) only use pseudo- F statistics of that particular gene across all permutations. Type I error rate is defined as the proportion of simulations with any significant genes when the data is generated under the null hypothesis of no genes are associated with case-control status.

Table B.1: Type I error rates in simulation settings with multiple genes of different sizes

Method	Pooled F statistics	Not pool F statistics
$D^{w-DM-DV}$	0.054	0.050
D^{w-DM}	0.046	0.051
D^{w-DV}	0.055	0.054
D^{DM-DV}	0.054	0.057
D^{DM}	0.041	0.055
D^{DV}	0.055	0.054
$EWAS^{DM}$	-	0.050
$EWAS^{DV}$	-	0.029

B.1.2 Values of shape parameters in simulations

We set $a_0 = 46.36$ and $b_0 = 52.28$ for all CpGs in controls and noise CpGs in cases such that a beta distribution $\text{Beta}(a_0, b_0)$ has a mean 0.47 and a SD 0.05, mimicing the real GEO methylation data. The values of a_1 and b_1 for signal CpGs in cases and the mean and SD of the corresponding beta distributions are summarized in Supplementary Table B.2.

Table B.2: Values of a_1 and b_1 for signal CpGs

Scenario	a_1	b_1	Beta distribution	Beta distribution	Mean	SD
			mean	SD	difference ¹	ratio ²
Variance signal only	29.5	33.27	0.47	0.0625	0	1.25
	20.34	22.94	0.47	0.075	0	1.50
	14.82	16.71	0.47	0.0875	0	1.75
	11.24	12.67	0.47	0.10	0	2
	8.78	9.90	0.47	0.1125	0	2.25
	7.02	7.92	0.47	0.125	0	2.50
Mean signal only	48.49	50.47	0.49	0.05	0.02	1
	50.47	48.49	0.51	0.05	0.04	1
	52.28	46.36	0.53	0.05	0.06	1
	53.90	44.10	0.55	0.05	0.08	1
	55.31	41.73	0.57	0.05	0.10	1

¹Mean difference in the signal CpG between cases and controls.

²SD ratio in the signal CpG between cases and controls.

B.1.3 Simulation settings with one gene considering correlations among CpGs

We conducted simulation studies to investigate the impact of correlations among neighboring CpGs on the performance of the proposed distance-based method. We simulated DNA methylation M -values which are the logit2 transformation of methylation β -values, and considered AR(1) correlation among CpGs in a gene with a correlation coefficient $\rho=0.5$. Here we only considered one gene for illustration purposes, and conducted simulation studies parallel as that for methylation β -values in the main text. More specifically, We considered one gene with different signal-to-noise ratios ranging from 1:0, 1:24, 1:49, 3:47, to 5:45. We considered scenarios when signal CpGs have different mean or variance signals by varying means and SDs of a normal distribution used to generate methylation M -values. We considered scenarios when mean differences in methylation M -values between cases and controls are $0.25 \times SD_0$, $0.5 \times SD_0$, $0.75 \times SD_0$, $1 \times SD_0$ and $1.25 \times SD_0$ where SD_0 is the SD in controls. We also considered scenarios when ratios of SDs for cases and controls are 1.25, 1.50, 1.75, 2 and 2.25.

Type I error rates are well controlled at the 0.05 significance level in all scenarios (Supplementart Table B.3). Power results are summarized in Supplementary Figure B.1 where we note that the power patterns are very similar to those observed in simulations based on methylation β -values without considering correlations among CpG sites. This implies that the correlations among neighboring CpGs in a gene do not have much impact on the performance of the proposed distance-based method, neither does the use of methylation M -values or β -values.

Table B.3: Type I error rates in simulation settings with AR(1) correlation among neighboring CpGs with $\rho = 0.5$

Method	1 CpG	25 CpGs	50 CpGs
$D^{w-DM-DV}$	0.044	0.044	0.042
D^{w-DM}	0.057	0.050	0.040
D^{w-DV}	0.042	0.053	0.042
D^{DM-DV}	0.044	0.043	0.042
D^{DM}	0.057	0.048	0.043
D^{DV}	0.043	0.057	0.048
$EWAS^{DM}$	0.054	0.044	0.046
$EWAS^{DV}$	0.052	0.039	0.032

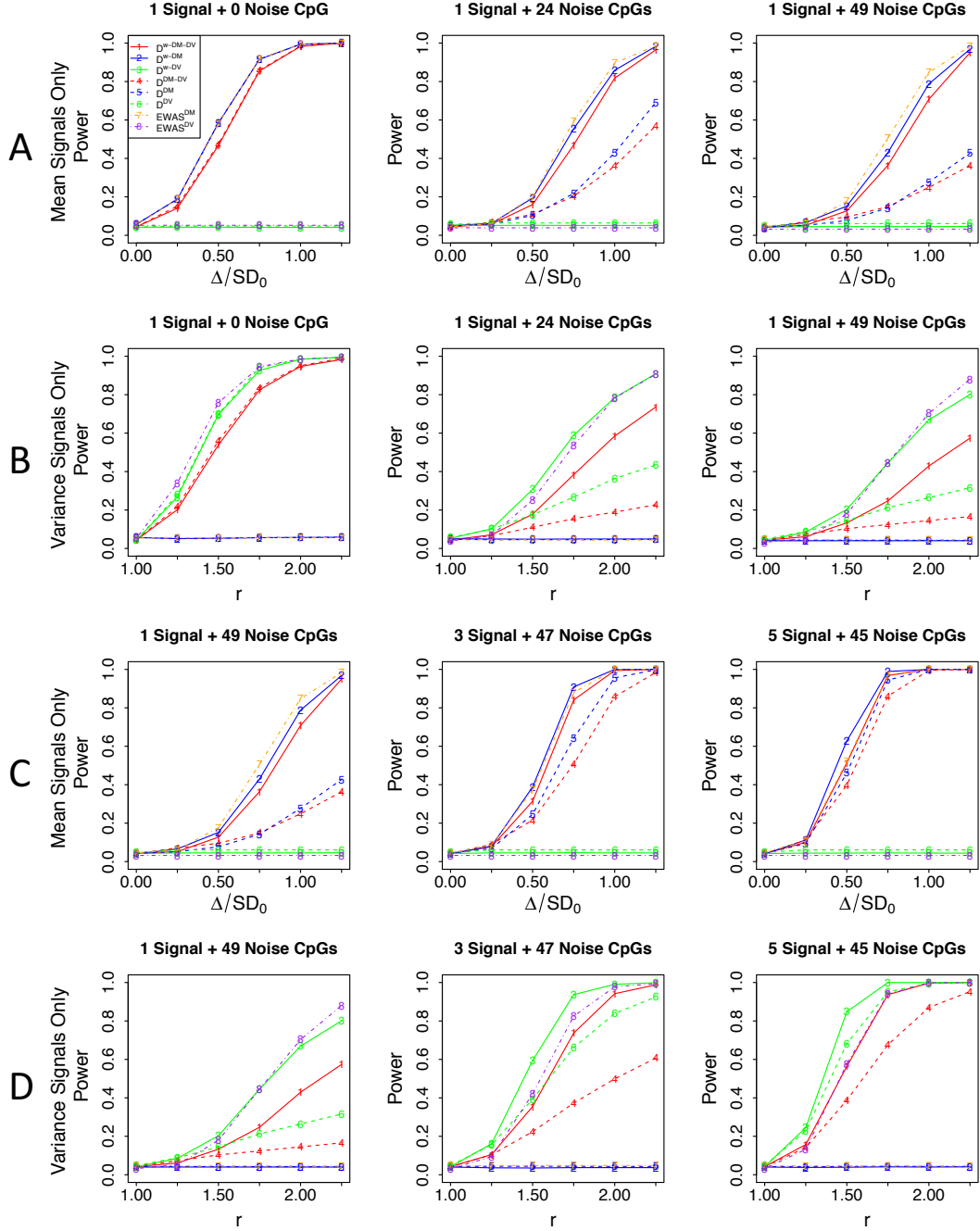


Figure B.1: Power results for simulation settings with one gene considering AR(1) correlation among neighboring CpGs with correlation coefficient $\rho=0.5$. The signal gene has one signal CpG and increasing number of total CpGs, i.e., decreasing signal-to-noise ratios from 1:0, 1:24 to 1:49 (panel A for mean signals only, panel B for variance signals only), or with a fixed total number of CpGs 50 and increasing signal-to-noise ratios from 1:49, 3:47, to 5:45 (panel C for mean signals only, panel D for variance signals only).

B.2 Real data application

B.2.1 Discovery analysis

Table B.4: 11 genes identified by D^{w-DM} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data

Rank	Gene	# CpG	Rank in $D^{w-DM-DV}$	Rank in $EWAS^{min-P}$
1	<i>ZFP57</i> *	5	2	16
2	<i>RGL3</i>	21	45	66
3	<i>ANKRD13B</i> *	22	5	25
4	<i>PENK</i> *	23	6	37
5	<i>MMP23B</i> *	2	17	80
6	<i>MIR564</i>	9	86	55
7	<i>HBA1</i> *	7	10	23
8	<i>SSTR4</i>	9	29	197
9	<i>PPP3R1</i>	11	30	299
10	<i>TRH</i> *	16	13	28
11	<i>SOX1</i>	28	33	86

*Genes also identified by $D^{w-DM-DV}$.

Table B.5: 9 genes identified by D^{w-DV} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data

Rank	Gene	# CpG	Rank in $D^{w-DM-DV}$	Rank in $EWAS^{min-P}$
1	<i>KDM5A</i> *	2	9	4
2	<i>CXCL6</i> *	7	8	1
3	<i>DPH3B</i> *	5	3	61
4	<i>TMC4</i> *	13	1	2
5	<i>ANGPTL3</i>	3	158	50
6	<i>IL4R</i>	12	46	31
7	<i>NAA35</i> *	7	4	10
8	<i>TMEFF1</i> *	5	18	156
9	<i>THY1</i> *	19	7	13

*Genes also identified by $D^{w-DM-DV}$.

Table B.6: 2 significant genes identified by D^{DM-DV} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data

Rank	Gene	# CpG	Rank in $D^{w-DM-DV}$	Rank in $EWAS^{min-P}$
1	<i>MMP23B</i> *	2	17	80
2	<i>ZNF154</i>	12	26	202

*Genes also identified by $D^{w-DM-DV}$.

Table B.7: 6 genes identified by D^{DM} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data

Rank	Gene	# CpG	Rank in $D^{w-DM-DV}$	Rank in $EWAS^{min-P}$
1	<i>MMP23B</i> *	2	17	80
2	<i>ZNF154</i>	12	26	202
3	<i>TRH</i> *	16	13	28
4	<i>SOX17</i>	18	59	609
5	<i>WFDC3</i>	2	392	106
6	<i>SPAG6</i> *	16	11	170

*Genes also identified by $D^{w-DM-DV}$.

Table B.8: 4 significant genes identified by D^{DV} at the 0.0005 gene-level P -value threshold in the GEO BRCA Data

Rank	Gene	# CpG	Rank in $D^{w-DM-DV}$	Rank in $EWAS^{min-P}$
1	<i>C7orf11</i>	1	386	126
2	<i>MIR1305</i>	1	443	78
3	<i>KDM5A</i> *	2	9	4
4	<i>ANGPTL3</i>	3	158	50

*Genes also identified by $D^{w-DM-DV}$.

14 genes uniquely identified by $D^{w-DM-DV}$

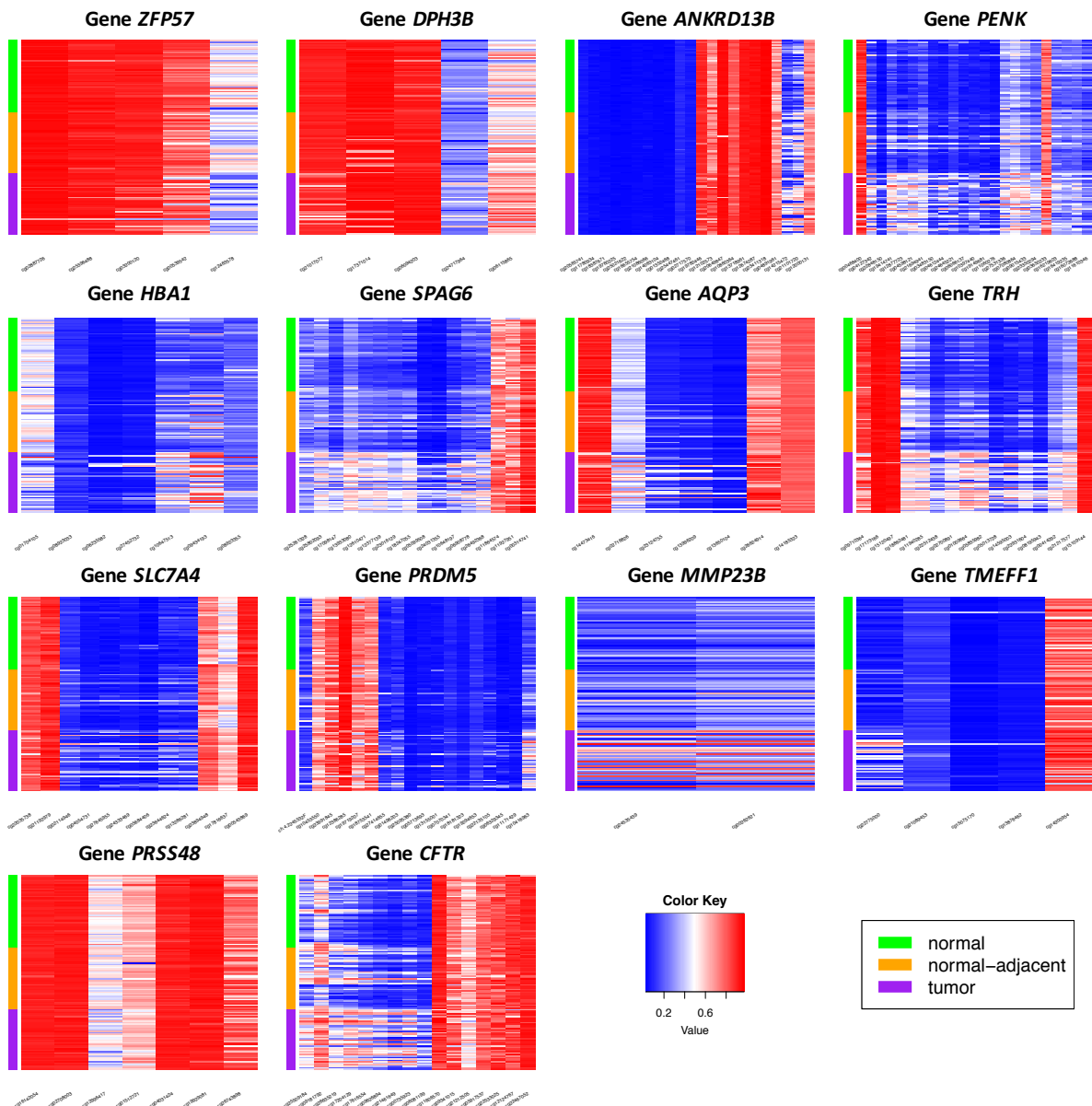


Figure B.2: Heatmaps of original DNA methylation measures of the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues for 14 genes uniquely identified by $D^{w-DM-DV}$.

7 genes uniquely identified by $EWAS^{min-P}$

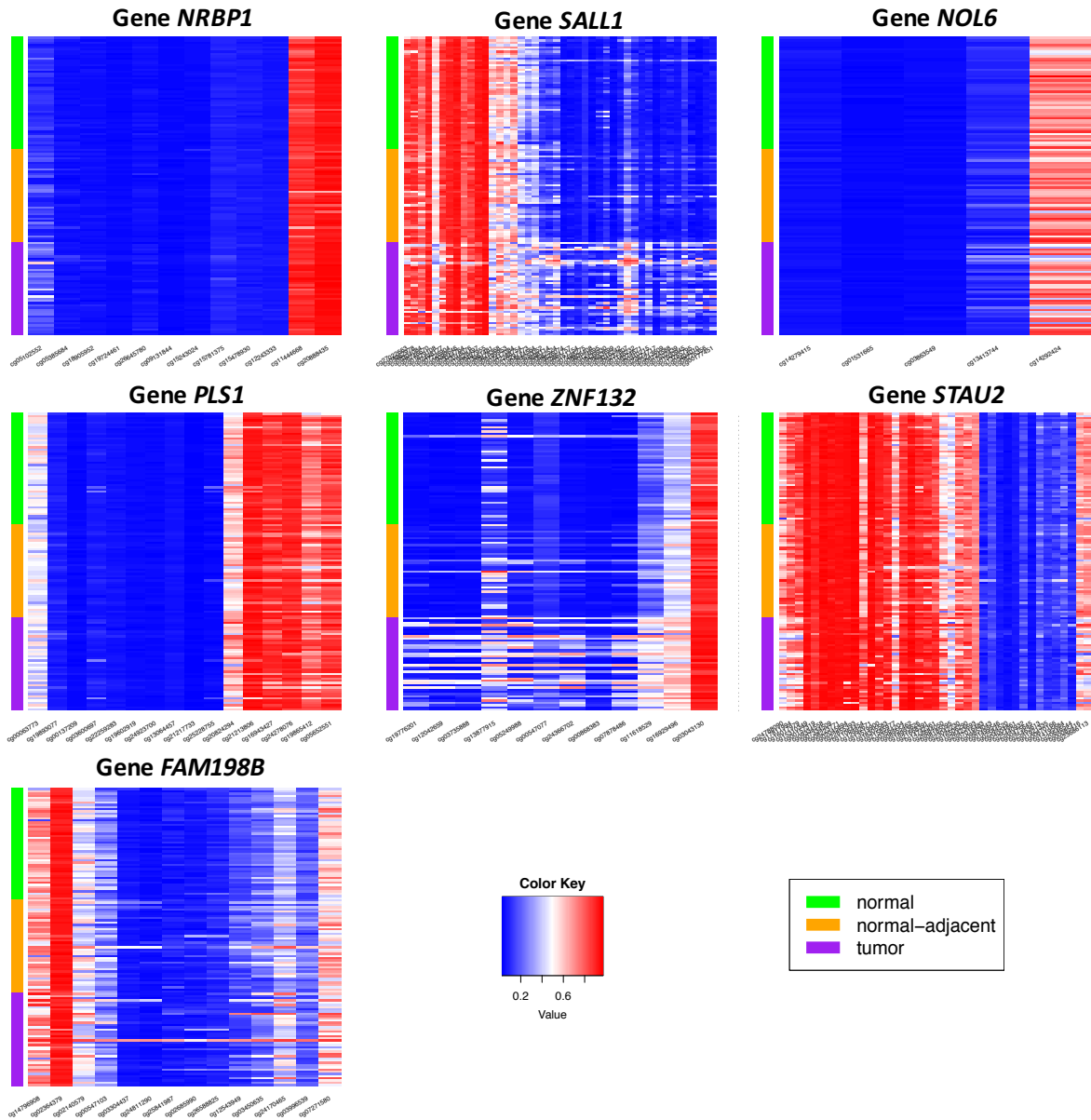


Figure B.3: Heatmaps of original DNA methylation measures of the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues for 7 genes uniquely identified by $EWAS^{min-P}$.

7 genes identified by both $D^{w-DM-DV}$ and $EWAS^{min-P}$

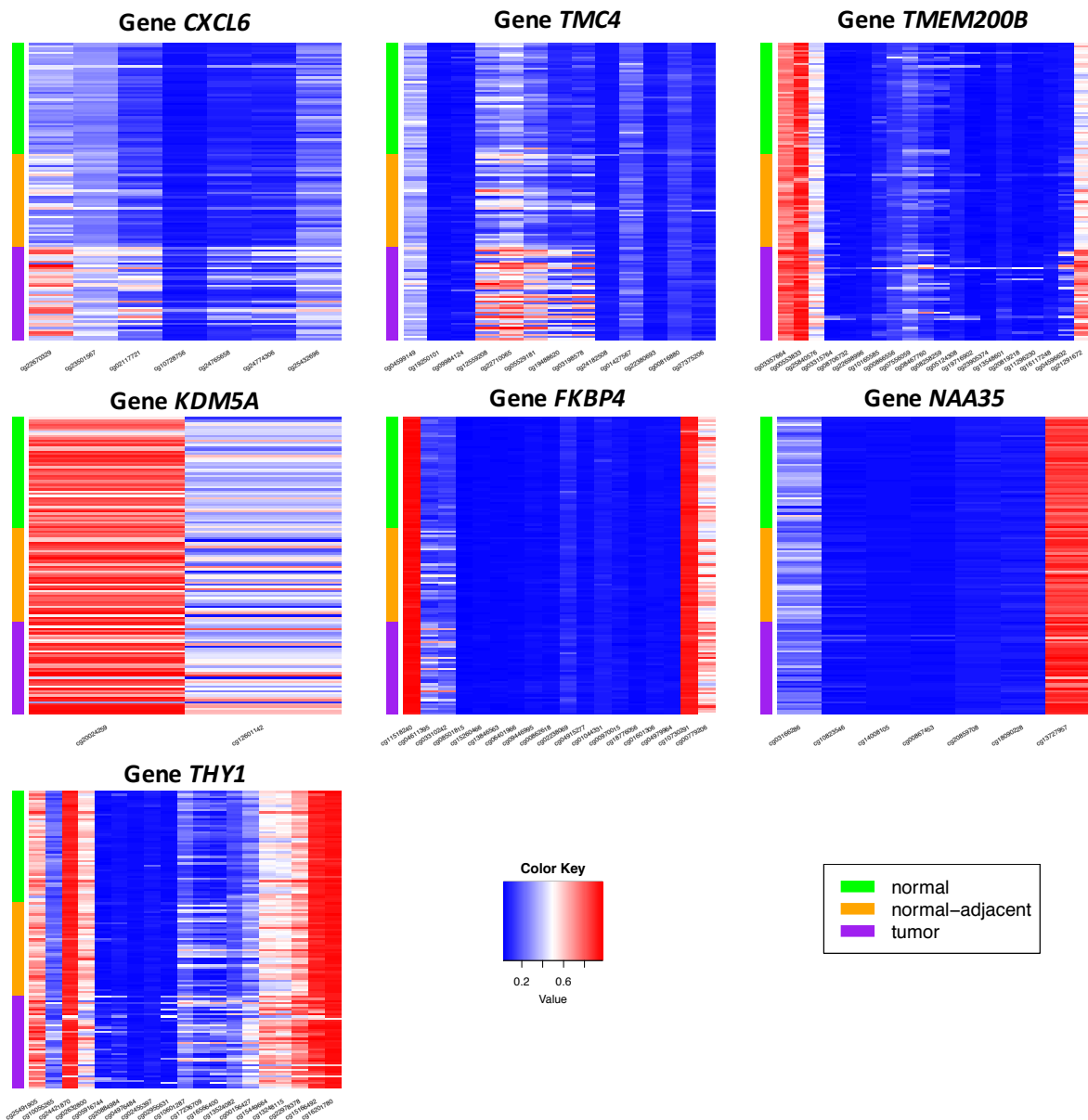


Figure B.4: Heatmaps of original DNA methylation measures of the 50 normal tissues, 42 normal-adjacent tissues together with the 42 matched tumor tissues for 7 genes identified by both $D^{w-DM-DV}$ and $EWAS^{min-P}$.

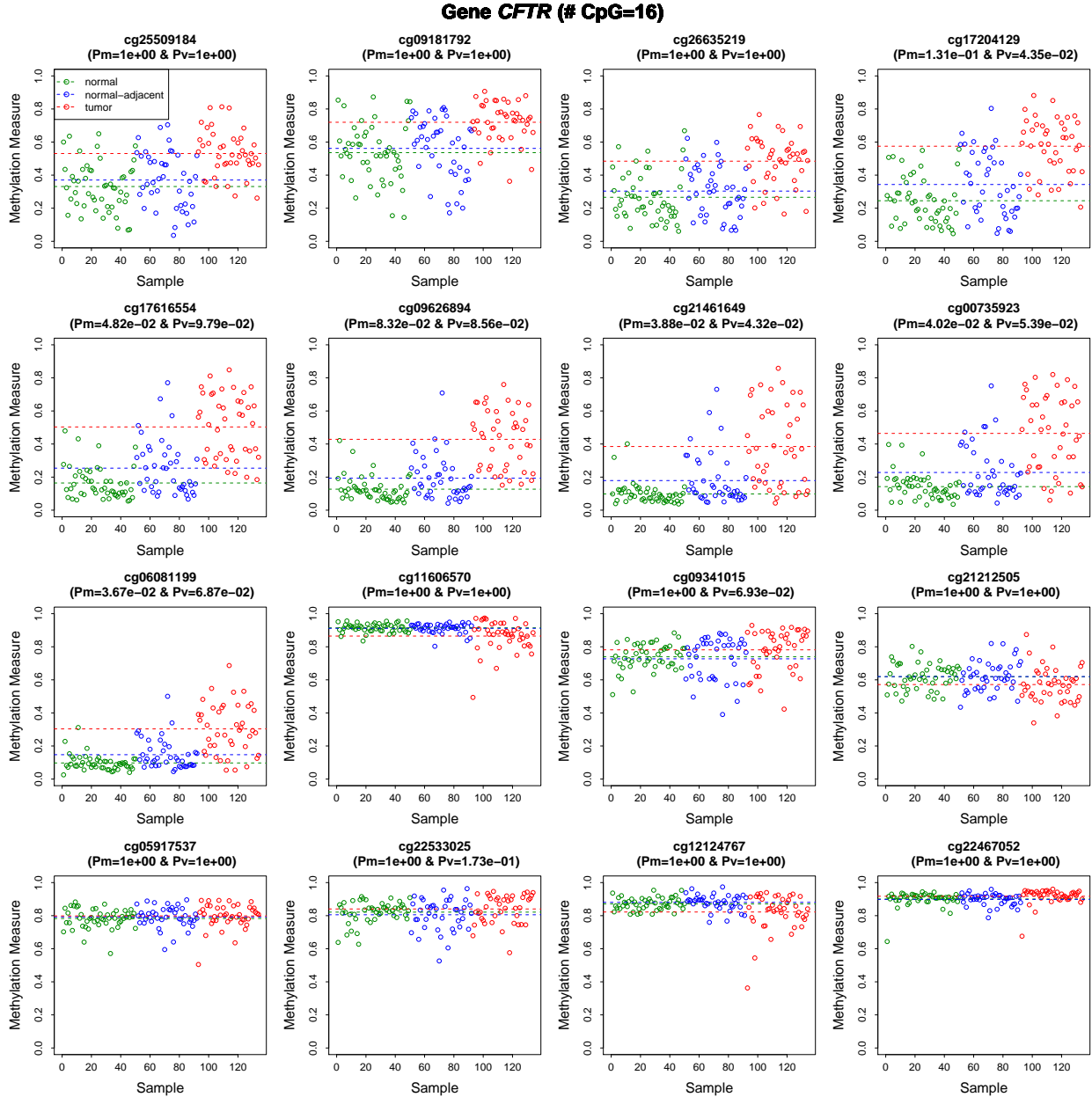


Figure B.5: DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of 16 CpGs in the *CFTR* gene that was uniquely identified by $D^{w-DM-DV}$, but ranked the last using $EWAS^{min-P}$ among all uniquely identified genes. Pm and Pv are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.

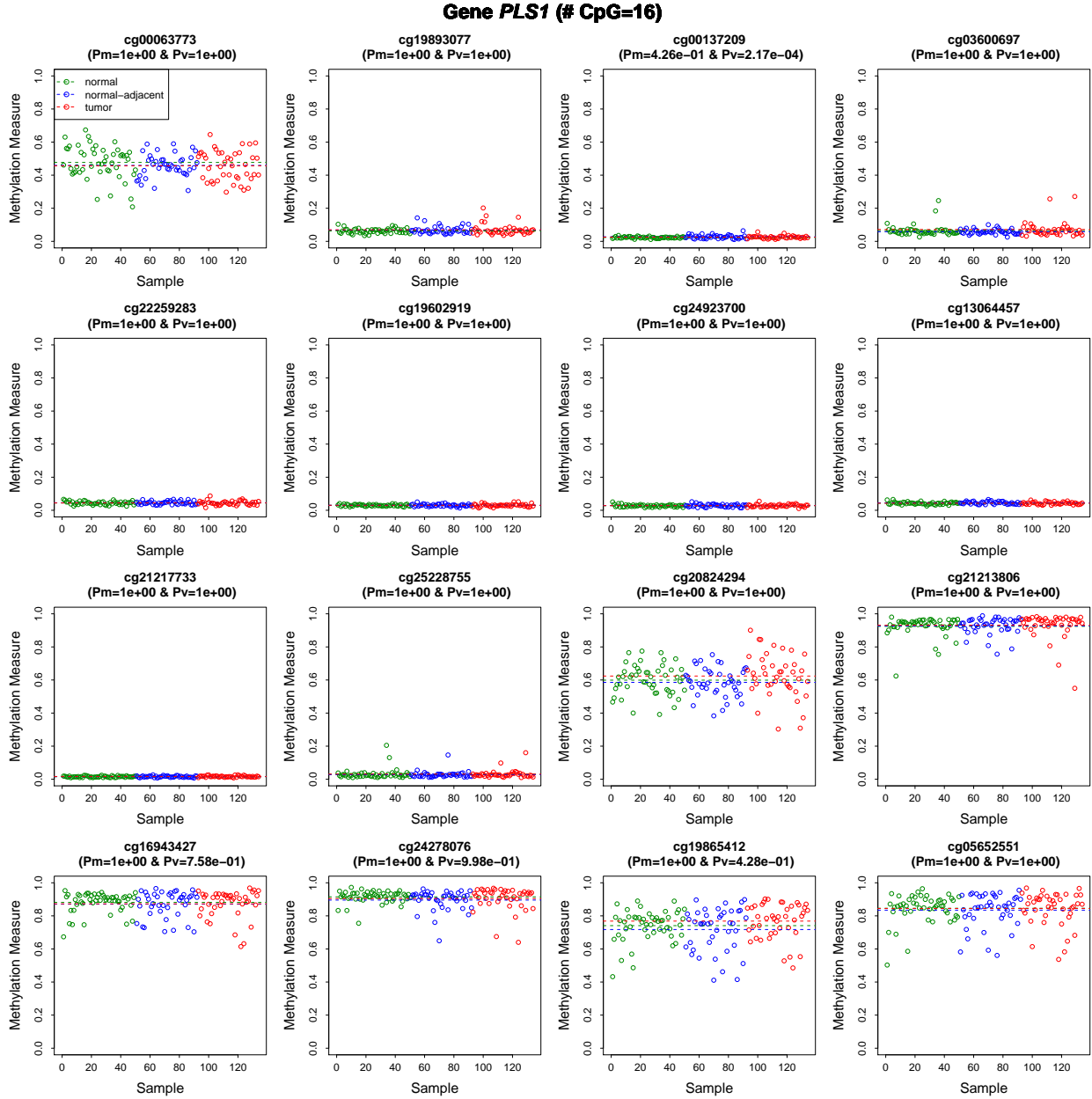


Figure B.6: DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of 16 CpGs in the *PLS1* gene that was uniquely identified by $EWAS^{min-P}$, but ranked the last using $D^{w-DM-DV}$ among all uniquely identified genes. Pm and Pv are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.

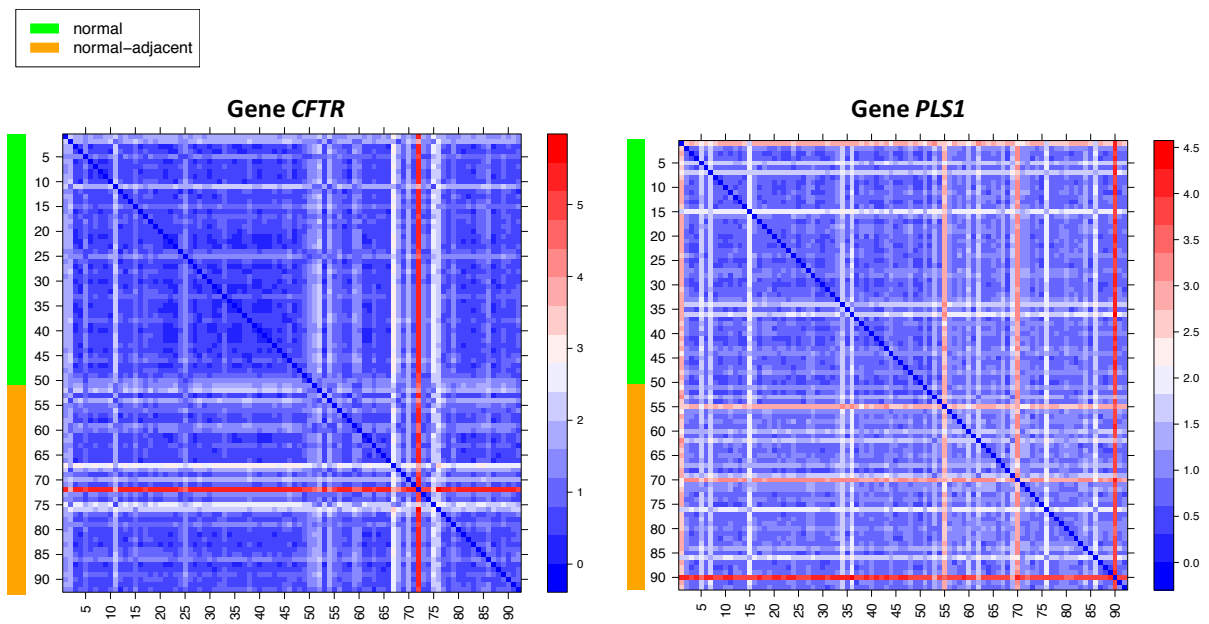


Figure B.7: Weighted distance matrices for genes *CFTR* and *PLS1*.

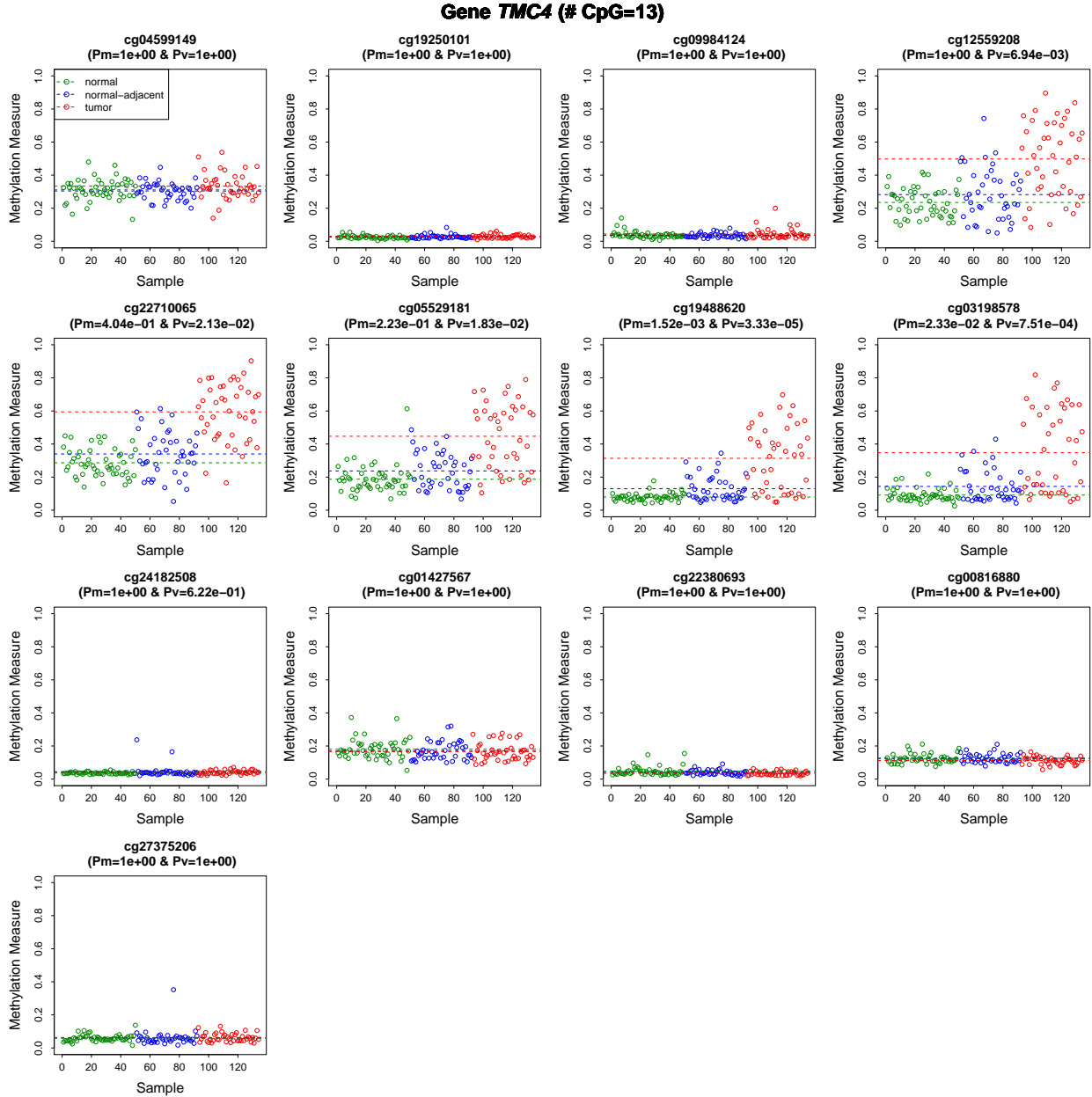


Figure B.8: DNA methylation measures of 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors of 13 CpGs in the *TMC4* gene that was identified by both $D^{w-DM-DV}$ and $EWAS^{min-P}$ and ranked on #1 and #2, respectively. Pm and Pv are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene. The three horizontal lines represent mean methylation levels of the three groups of normal tissues, normal-adjacent tissues and matched tumors.

Selection Probability by Gene Size

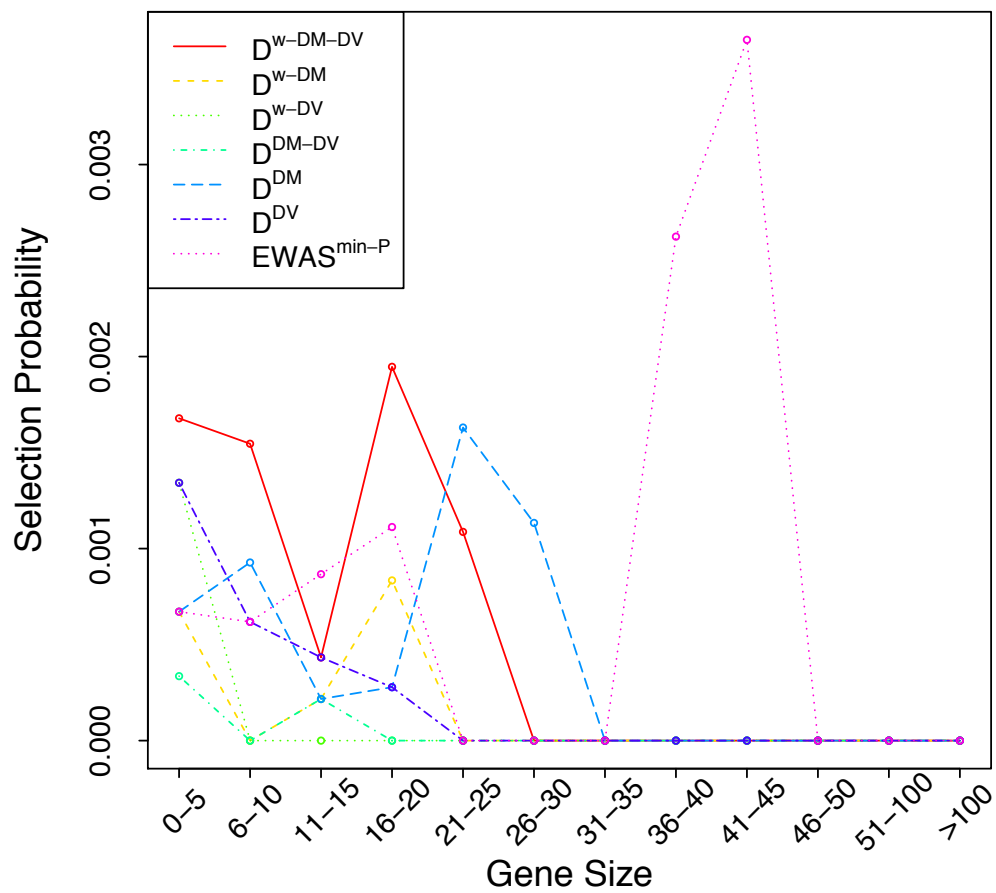


Figure B.9: The selection probability of identifying a gene out of all genes of the same size.

B.2.2 Validation analysis

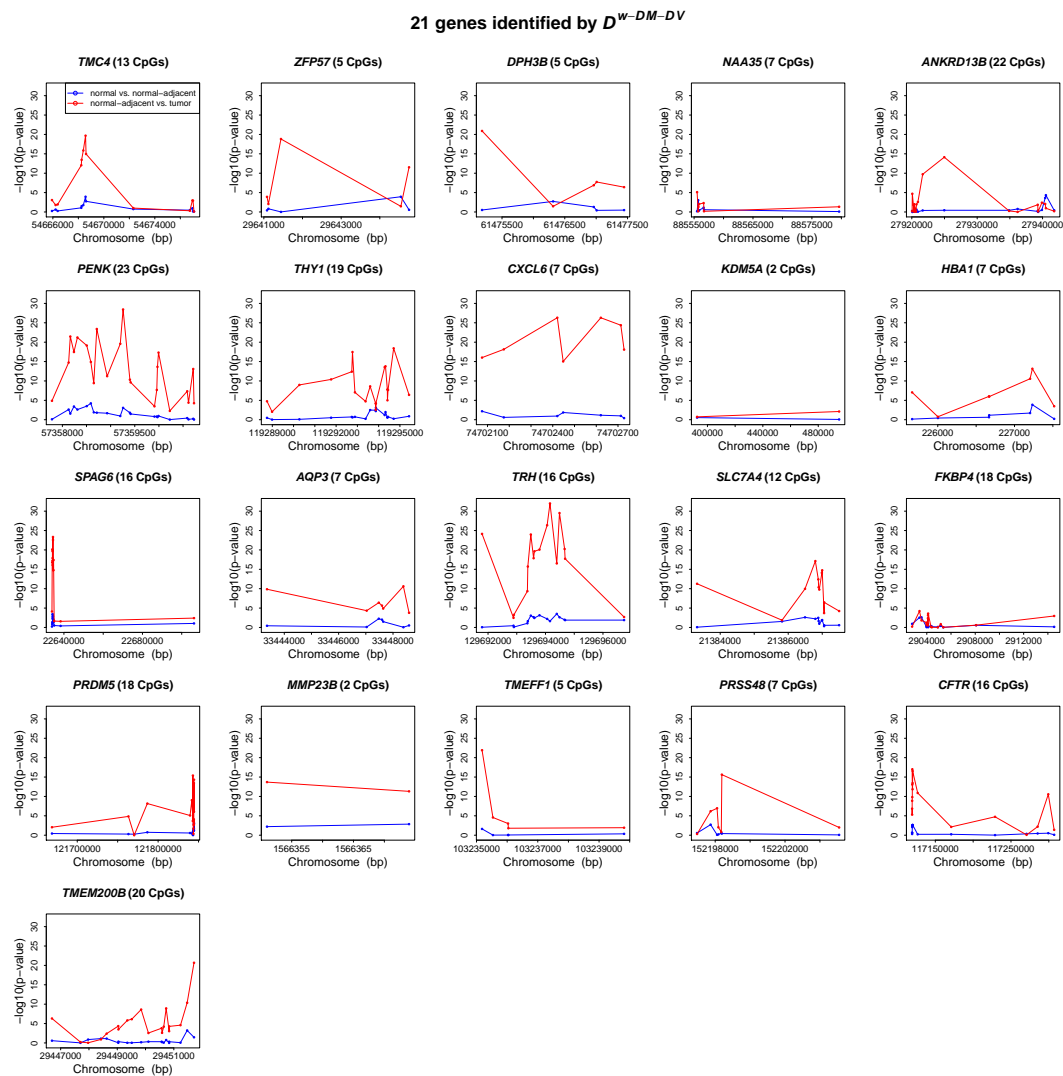


Figure B.10: $-\log_{10}(p\text{-value})$ from CpG site-level t -tests in (1) normal-adjacent versus normal comparison and (2) unmatched tumor versus normal-adjacent comparison in the GEO BRCA data for 21 genes identified by $D^{w-DM-DV}$.

B.2.3 Replication analysis

Table B.9: Summary number of genes identified by comparison methods in both the discovery analysis and replication analysis at the 0.0005 gene-level P -value threshold

Method	# of gene replicated / identified in discovery data	Replicated genes
$D^{w-DM-DV}$	7/21	<i>DPH3B, NAA35, ANKRD13B, CXCL6, FKBP4, PRSS48, CFTR</i>
D^{w-DM}	4/11	<i>ANKRD13B, MMP23B, PPP3R1, MIR564</i>
D^{w-DV}	1/9	<i>ANGPTL3</i>
D^{DM-DV}	2/2	<i>MMP23B, ZNF154</i>
D^{DM}	2/6	<i>MMP23B, ZNF154</i>
D^{DV}	1/4	<i>C7orf11</i>
$EWAS^{min-P}$	11/14	<i>NAA35, THY1, CXCL6, FKBP4, TMEM200B, FAM198B, NRBP1, NOL6, STAU2, SALL1, PLS1</i>

7 replicated genes identified by $D^{w-DM-DV}$

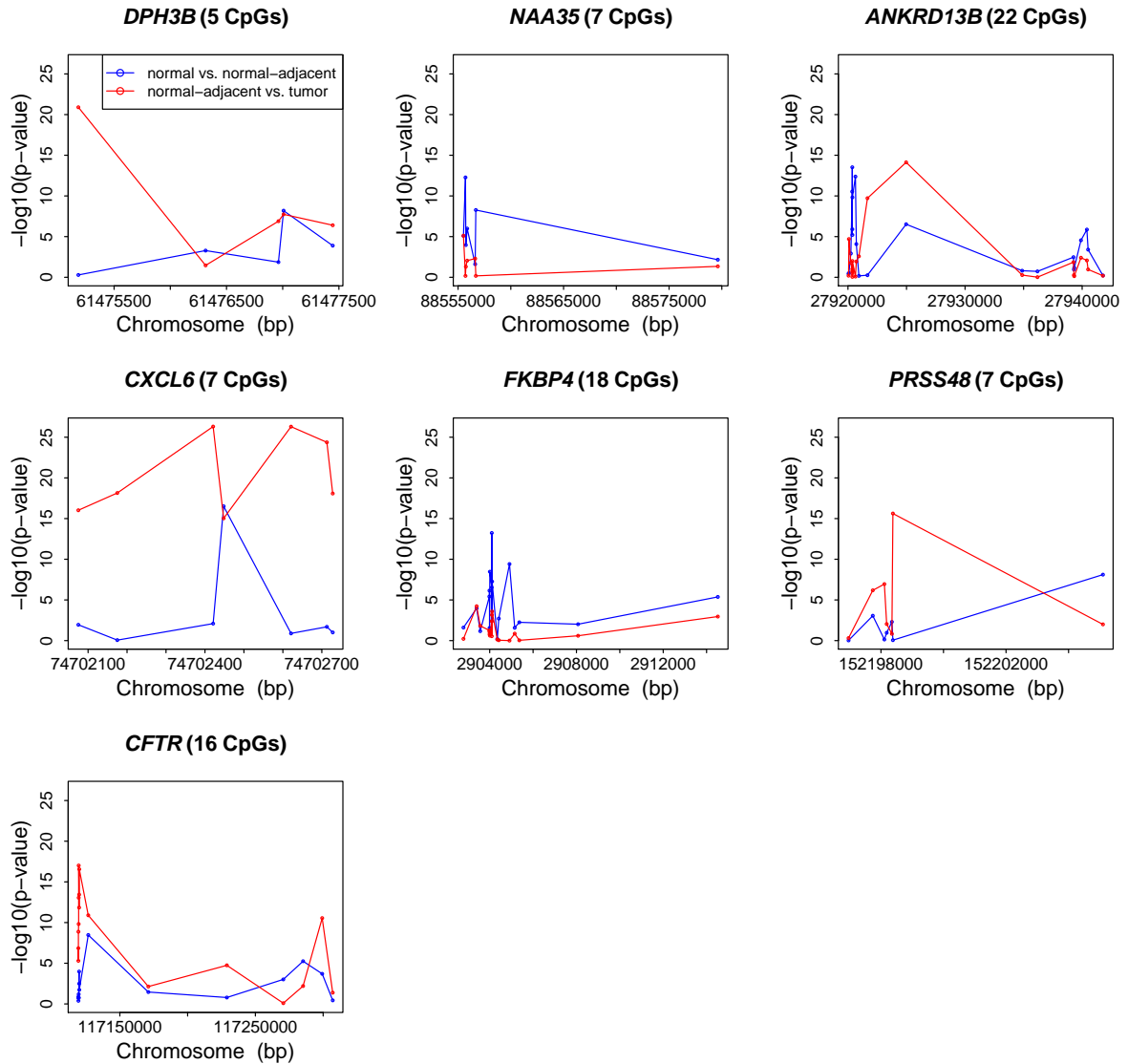


Figure B.11: $-\log_{10}(P\text{-value})$ from CpG site-level t -tests in (1) normal-adjacent versus normal comparison and (2) unmatched tumor versus normal-adjacent comparison in the replication analysis for 7 replicated genes identified by $D^{w-DM-DV}$.

We plotted the DNA methylation measures for 18 normal samples from the replication data (GSE67919), 50 normal samples, 42 normal-adjacent samples and 42 matched tumor samples from the discovery data (GSE69914) (Supplementary Figure B.12) for all CpGs in the *CFTR* gene. In the discovery analysis, the *CFTR* gene was uniquely identified by $\mathbf{D}^{w-DM-DV}$ but ranked the last using $EWAS^{min-P}$ among all $\mathbf{D}^{w-DM-DV}$ uniquely identified genes. The *CFTR* gene was replicated in the replication data by $\mathbf{D}^{w-DM-DV}$ due to weak dense signals similarly as in the discovery analysis. As the second example, we plotted the DNA methylation measures of all CpGs in the *CXCL6* gene, which was identified in the discovery analysis by both $\mathbf{D}^{w-DM-DV}$ and $EWAS^{min-P}$ and was replicated by both methods in the replication data (Supplementary Figure B.13). It is clear that all CpGs in the *CXCL6* gene have weak signals, and some of these weak dense signals were mainly due to a few outlier normal-adjacent tissue samples which was also observed in the discovery analysis. All CpGs in the *CXCL6* gene showed enrichment in methylation measures in the progression to tumor.

Among the 14 genes identified by $EWAS^{min-P}$ in the discovery analysis, 11 were replicated in the replication data, which are *CXCL6*, *TMEM200B*, *NRBP1*, *SALL1*, *FKBP4*, *NOL6*, *PLS1*, *NAA35*, *STAU2*, *THY1*, and *FAM198B*. We similarly plotted the DNA methylation measures of all CpGs in the *PLS1* gene (Supplementary Figure B.14) of the samples in the replication data. In the discovery analysis, the *PLS1* gene was uniquely identified by $EWAS^{min-P}$ but ranked the last using $\mathbf{D}^{w-DM-DV}$ among all $EWAS^{min-P}$ uniquely identified genes, and we found that it was identified mainly due to the strong variance signal at the CpG site cg00137209 as a result of very small variation in the methylation measures of the 50 normal tissues in the discovery data. In the replication analysis, however, the *PLS1* gene was replicated due to the mean signal at a different CpG site cg21213806. This implies that the *PLS1* gene might not be reliable, even it was replicated.

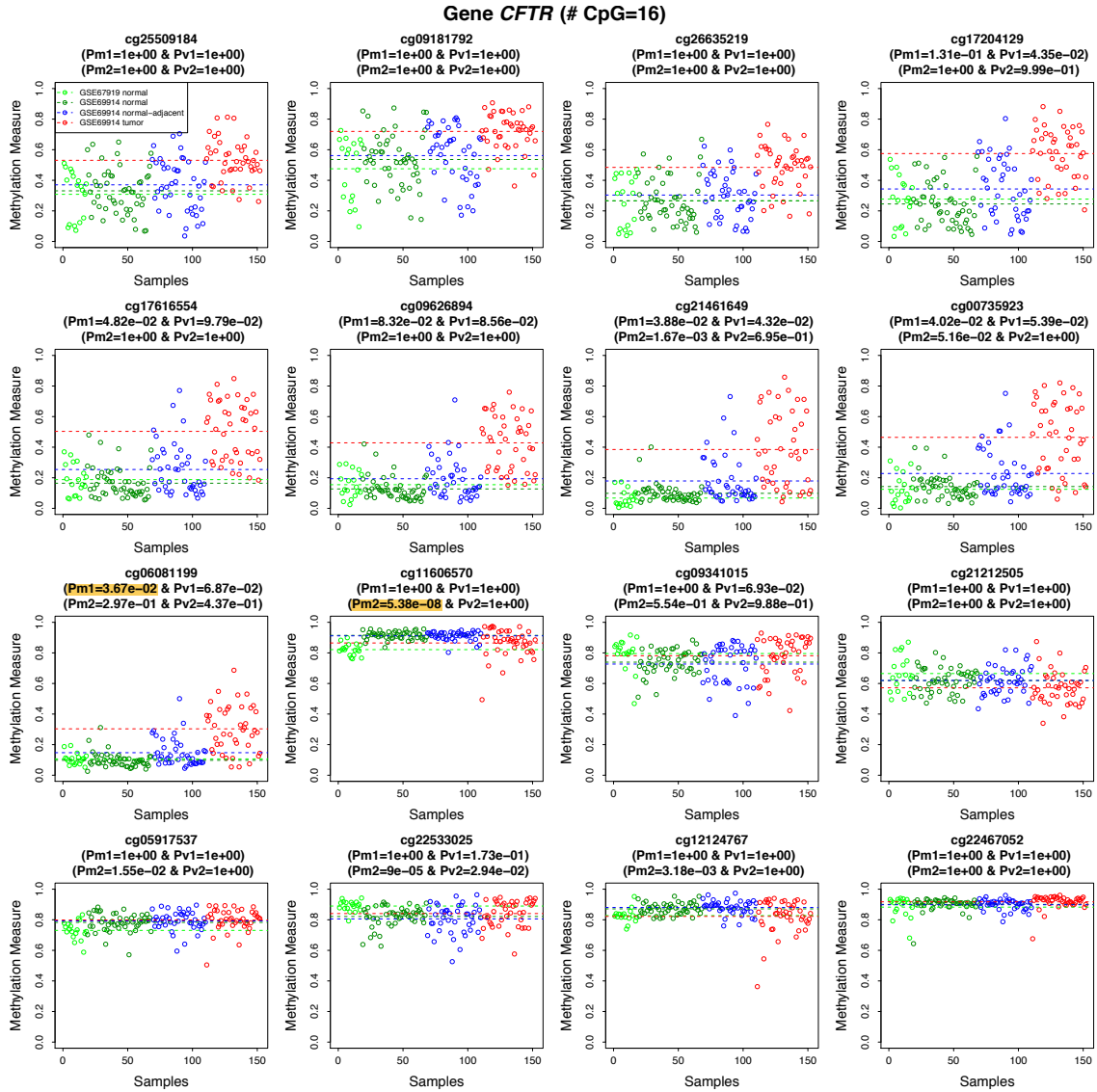


Figure B.12: DNA methylation measures of 18 normal tissues from the replication data (GSE67919), 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors from the discovery data (GSE69914) of 16 CpGs in the *CFTR* gene that was uniquely identified in the discovery analysis and replicated by $D^{w-DM-DV}$. Pm1 and Pv1 are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene from the discovery analysis, and Pm2 and Pv2 are those from the replication analysis. Highlighted are the minimum adjusted DM and DV P -value across all P -values in the gene in each comparison. The four horizontal lines represent mean methylation levels of the four groups of tissues.

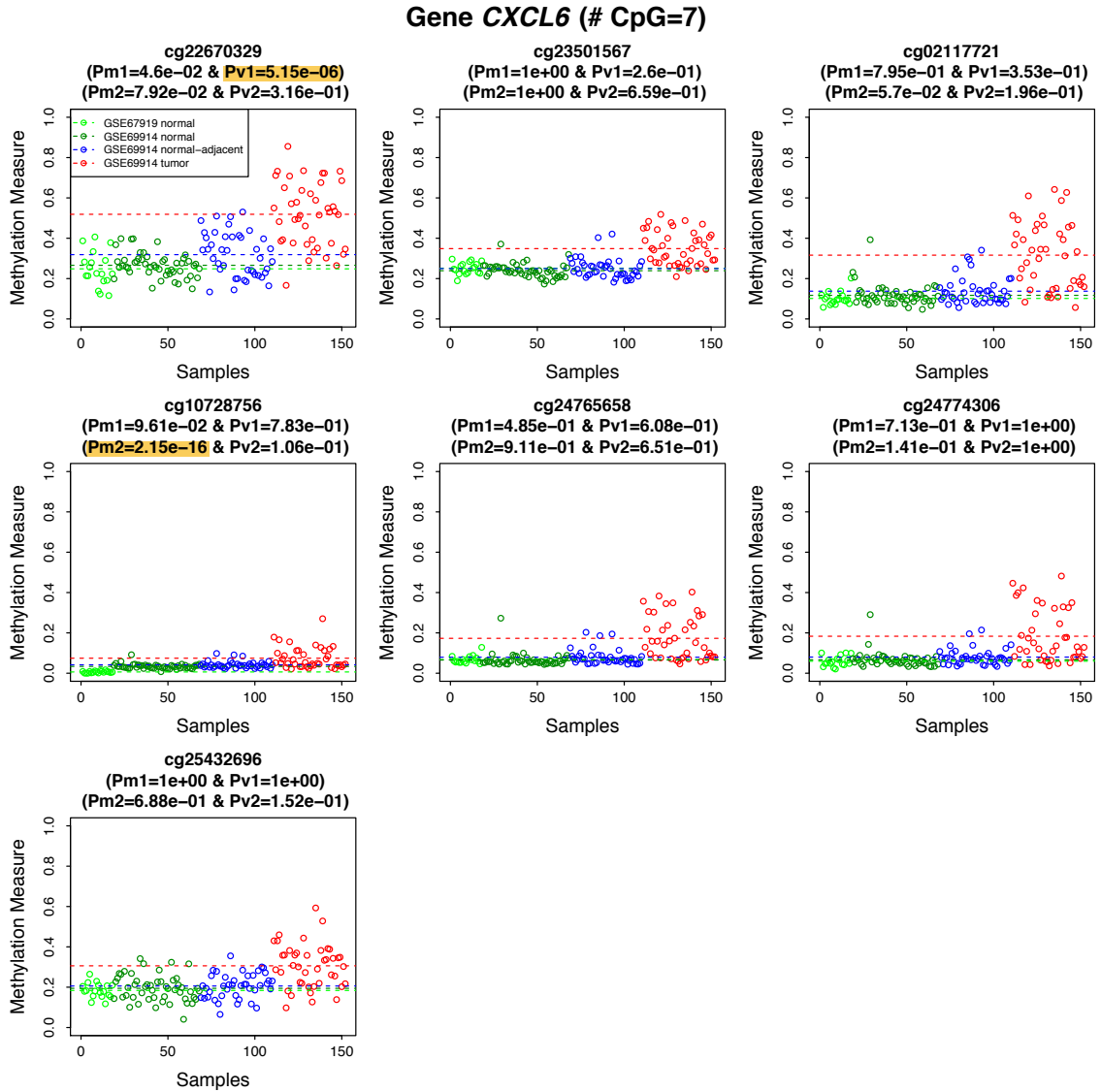


Figure B.13: DNA methylation measures of 18 normal tissues from the replication data (GSE67919), 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors from the discovery data (GSE69914) of 7 CpGs in the *CXCL6* gene which was identified in the discovery analysis and replicated by both $D^{w-DM-DV}$ and $EWAS^{min-P}$. Pm1 and Pv1 are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene from the discovery analysis, and Pm2 and Pv2 are those from the replication analysis. Highlighted are the minimum adjusted DM and DV P -value across all P -values in the gene in each comparison. The four horizontal lines represent mean methylation levels of the four groups of tissues.

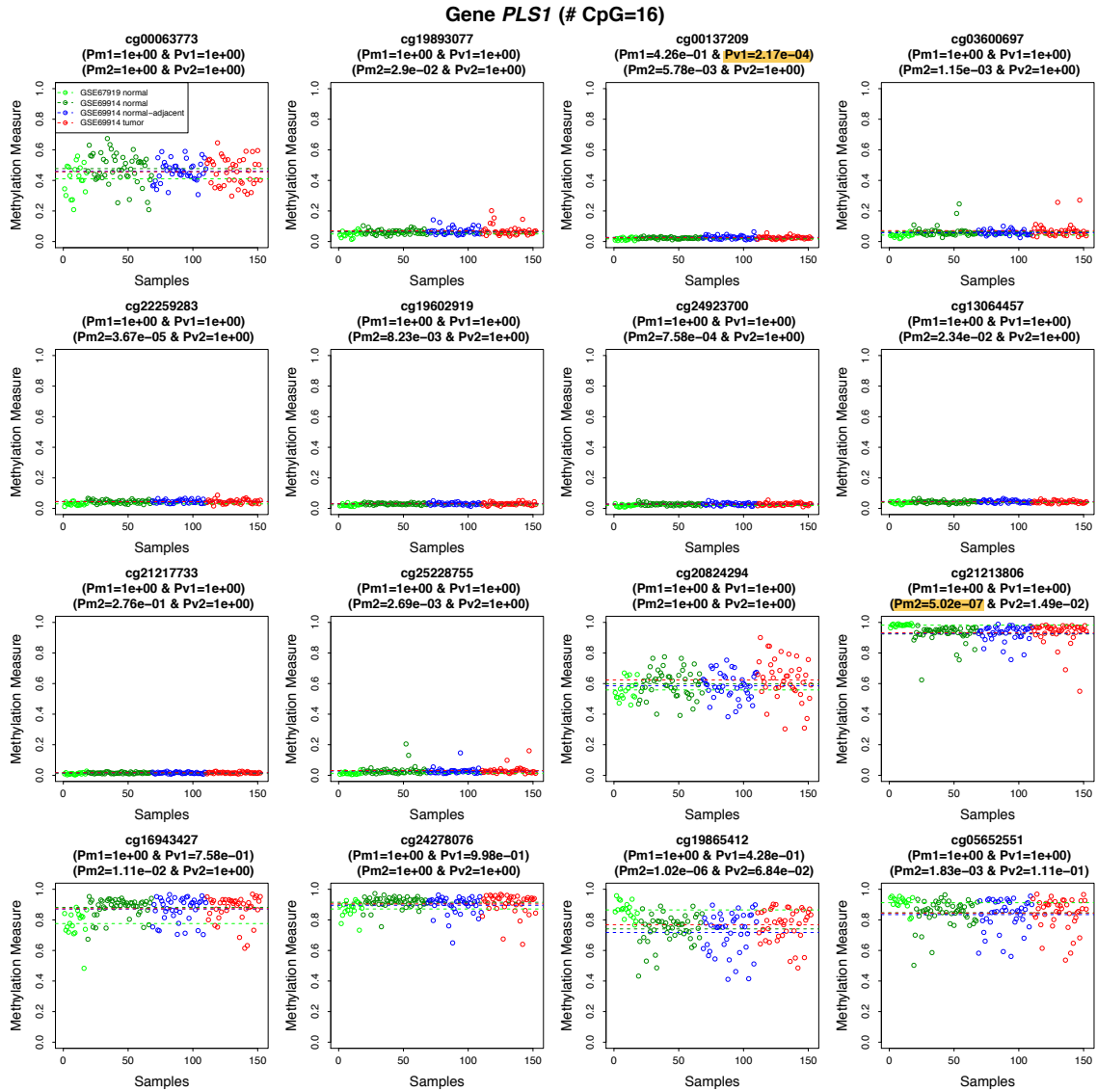


Figure B.14: DNA methylation measures of 18 normal tissues from the replication data (GSE67919), 50 normal tissues, 42 normal-adjacent tissues and 42 matched tumors from the discovery data (GSE69914) of 16 CpGs in the *PLS1* gene that was uniquely identified in the discovery analysis and replicated by $EWAS^{min-P}$. Pm1 and Pv1 are P -values from CpG site-level mean and variance tests that are adjusted for multiple comparisons for the number of CpGs in the gene from the discovery analysis, and Pm2 and Pv2 are those from the replication analysis. Highlighted are the minimum adjusted DM and DV P -value across all P -values in the gene in each comparison. The four horizontal lines represent mean methylation levels of the four groups of tissues.

Appendix C

Appendix to A Powerful and Flexible Weighted Distance-Based Method Incorporating Interactions Between DNA Methylation and Environmental Factors on Health Outcomes

C.1 Additional Simulation Studies

C.1.1 Effects of gene sizes in Type I errors

To investigate if genes with different sizes, i.e., number of CpGs, will have different distributions for pseudo- F statistics under the null hypothesis, we conducted simulation studies to evaluate type I error rates of the proposed method and those of the comparison methods. Specifically, we simulated methylation measures for 16 genes that consist of 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, and 100 CpGs, respectively. When calculating the P -value for each gene, we (1) pool all pseudo- F statistics of the 16 genes across all permutations, and (2) only use pseudo- F statistics of that particular gene across all permutations. Type I error rate is defined as the proportion of simulations with any significant genes when the data is generated under the null hypothesis of no genes are associated with case-control status.

Table C.1: Type I error rates in simulation settings with multiple genes of different sizes

Method	Pooled F statistics	Not pool F statistics
$\mathbf{D}^{\text{w-main-int}}$	0.045	0.043
$\mathbf{D}^{\text{w-main}}$	0.034	0.044
$\mathbf{D}^{\text{w-int}}$	0.048	0.042
$\mathbf{D}^{\text{main-int}}$	0.046	0.044
\mathbf{D}^{main}	0.046	0.051
\mathbf{D}^{int}	0.041	0.035
L^S	-	0.017
L^M	-	0.033

C.1.2 Simulation settings with different types of signals

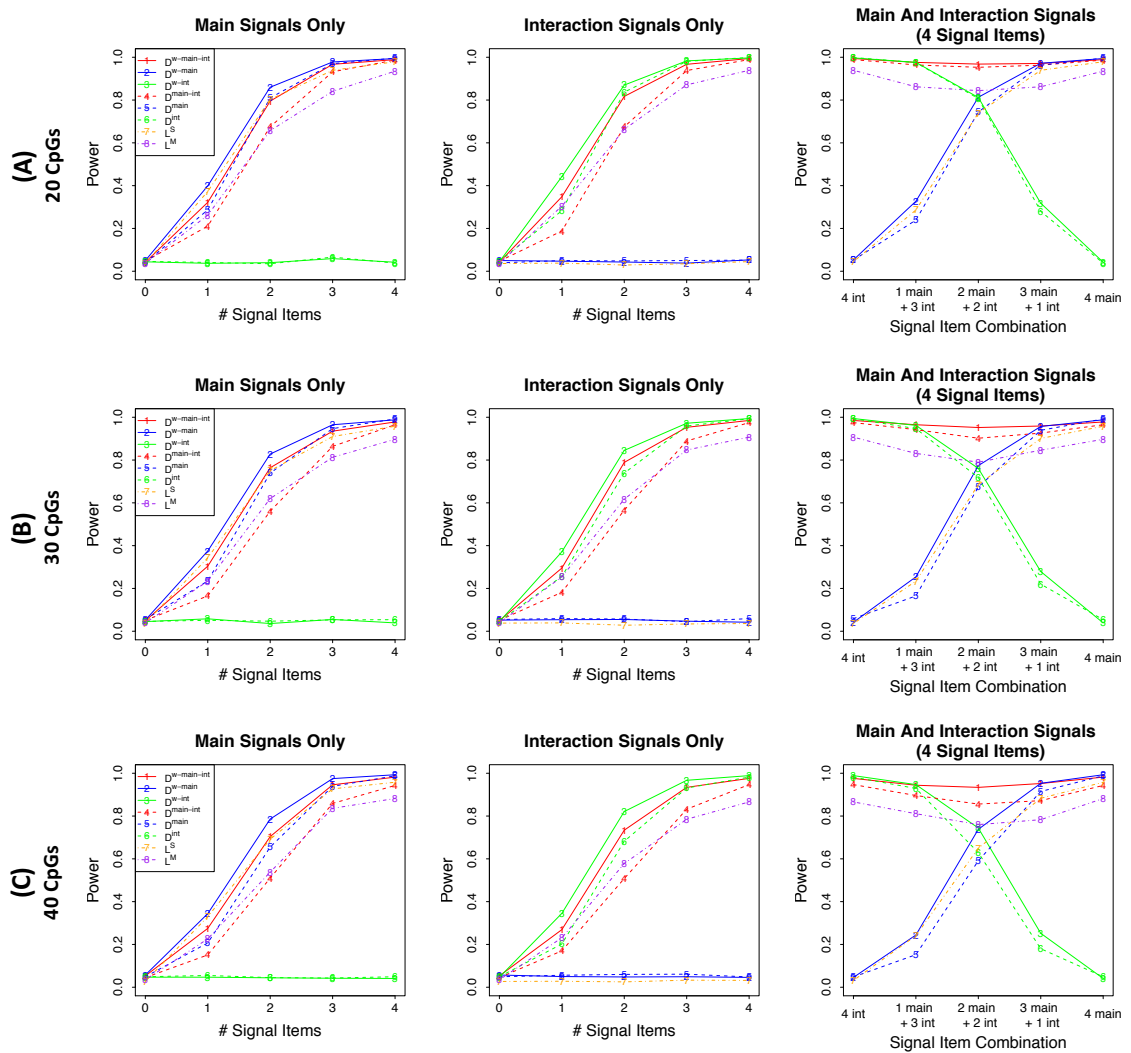


Figure C.1: Power results for simulation settings with main signals only, interaction signals only and both main and interaction signals when there are (A) 20 CpGs, (B) 30 CpGs, and (C) 40 CpGs in a gene.

C.1.3 Simulation settings with fixed number of signal items coming from different number of signal CpGs

C.1.3.1 Simulation setup

Table C.2: Simulation settings with 4 signal items and the same signal composition (2 main
 and 2 interaction signals) but from 2~4 signal CpGs

Number of Signal CpGs and settings	Simulation setup ^a
2 signal CpGs:	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_1} = \beta_{Z_3} = 0.3$
2 CpGs with main + interaction signals	
3 signal CpGs:	
1 CpG with main + interaction signals;	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_3} = \beta_{Z_5} = 0.3$
1 CpG with main signal only;	
1 CpG with interaction signal only	
4 signal CpGs:	
2 CpGs with main signal only;	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_5} = \beta_{Z_7} = 0.3$
2 CpGs with interaction signal only	

^a X represents DNA methylation main effects, Z represents DNA methylation by environment interaction effects.

C.1.3.2 Simulation results

When the 4 signal items are set with 2 main signals and 2 interaction signals and increasing the number of signal CpGs from 2 to 4, the power of $\mathbf{D}^{\text{w-main}}$, \mathbf{D}^{main} and L^S that considers main signals only increases slightly as expected. The power of $\mathbf{D}^{\text{w-main-int}}$ and $\mathbf{D}^{\text{main-int}}$ that considers both main and interaction signals also increases as the number of signal CpGs increases, while that of L^M decreases. For methods $\mathbf{D}^{\text{w-int}}$ and \mathbf{D}^{int} that considers interaction signals only, we observe the largest power when there are 3 signal CpGs, slightly lower/similar power when there are 4 signal CpGs, and lowest power when there are 2 signal CpGs. The non-monotone trend might due to the fact that the 2 CpGs with interaction signals were randomly set in the data matrix $\mathbf{X}^{\text{main-int}}$. That is, the 2 CpGs with main signals were fixed to be the 1st and 3rd CpGs in the gene, while the 2 CpGs with interaction signals were chosen at different locations in the gene in each setting. Overall, the power of $\mathbf{D}^{\text{w-int}}$ and \mathbf{D}^{int} increases as the number of signal CpGs increases, and the weighted versions always perform better than the non-weighted versions.

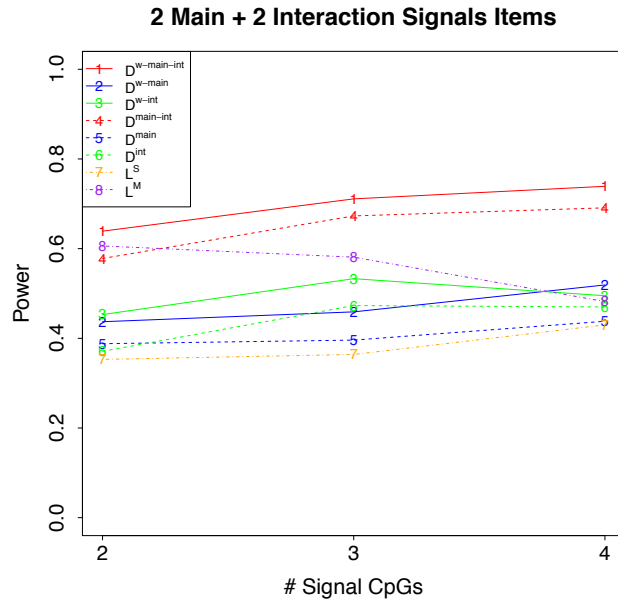


Figure C.2: Power results for simulation settings where there are 2 main signal items and 2 interaction signal items coming from 2, 3 and 4 signal CpGs, respectively when there are 30 CpGs in a gene.

C.2 Real data applications

C.2.1 DNA methylation data processing

DNA methylation in the MN cohort was measured in 432 cord blood samples, for which 168 had data from the 450K array with 485,577 CpG sites and 264 from the EPIC array with 866,895 CpG sites. DNA methylation data in the Sibling cohort was measured from 67 cord blood samples, for which 40 had data from the 450K array and 27 from the EPIC array. For methylation data, we conducted standard quality control steps where we removed CpGs on sex chromosomes and those contain either a single nucleotide polymorphism (SNP) at the CpG interrogation or at the single nucleotide extension (SBE) based on UCSC dbSNP table version 147 using the R package ‘IlluminaHumanMethylation450kanno.ilmn12.hg19’ (Hansen, 2015). We further required at least 95% CpG coverage per sample and 70% sample coverage per CpG, and corrected for the type II probe bias using the ‘wateRmelon’ package (Pidsley et al., 2013). We then calibrated the 450K data to EPIC distribution (Horvath, 2013), and only kept overlapping CpG sites that were covered by both arrays which also had gene annotations, leaving 263,574 common CpG sites covering 18,633 genes in both MN and Sibling methylation datasets. We then transformed the methylation measures to M -values by taking logit2 transformation, and applied linear regression models on M -values at each CpG to adjust for cell proportions estimated from the ‘minfi’ package (Aryee et al., 2014) and obtained the M -value residuals. We then applied the proposed method and all comparison methods to the M -value residuals in the following analyses.

C.2.2 Risk of PAH, DNA methylation and their interactions on ADHD

C.2.2.1 Replication analysis in the Sibling cohort

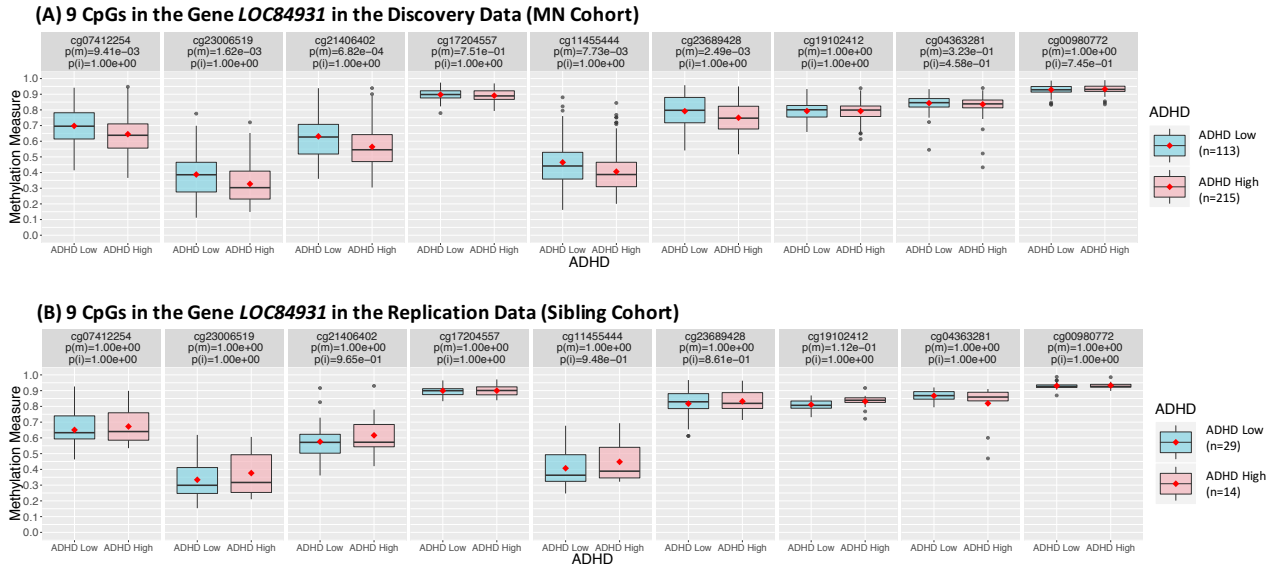


Figure C.3: Boxplot of DNA methylation measures of the 9 CpGs in gene *LOC84931* stratified by ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *LOC84931*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1 \text{CpG} + \beta_2 E + \beta_3 \text{CpG} \times E$.

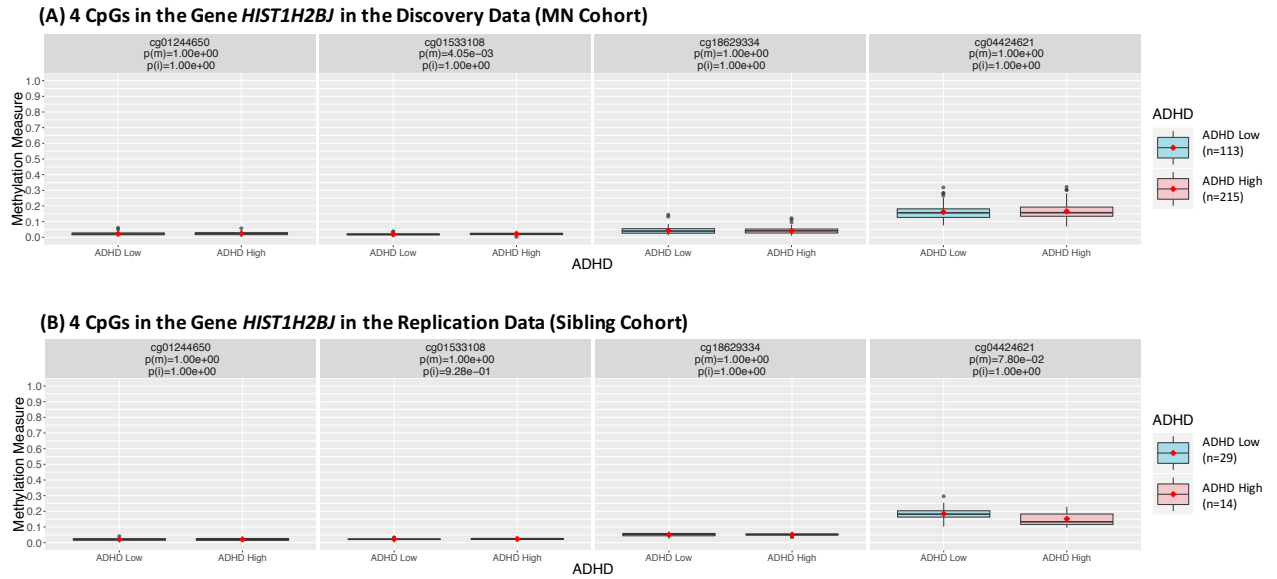


Figure C.4: Boxplot of DNA methylation measures of the 4 CpGs in gene *HIST1H2BJ* stratified by ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *HIST1H2BJ*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1 \text{CpG} + \beta_2 E + \beta_3 \text{CpG} \times E$.

C.2.2.2 Results from the comparison methods

Table C.3: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 11 genes by $\mathbf{D}^{\text{w-main}}$ at the 0.005 gene-level P -value threshold

Rank in $\mathbf{D}^{\text{w-main}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main-int}}$
1	<i>LOC84931</i> *	9	1
2	<i>SERPINB3</i> *	1	2
3	<i>HIST1H2BJ</i> *	4	9
4	<i>IGJ</i> *	1	7
5	<i>ADAM32</i> *	11	8
6	<i>TRIM38</i>	7	33
7	<i>SPDYC</i>	9	17
8	<i>NDUFA5</i>	9	16
9	<i>BICD1</i>	14	15
10	<i>KRTAP20-1</i> *	1	6
11	<i>CXCL9</i> *	1	10

*Genes also identified by $\mathbf{D}^{\text{w-main-int}}$.

Table C.4: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 14 genes by $\mathbf{D}^{\text{w-int}}$ at the 0.005 gene-level P -value threshold

Rank in $\mathbf{D}^{\text{w-int}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main-int}}$
1	<i>CYP2E1</i> *	13	3
2	<i>MIR518E</i> *	1	4
3	<i>KIR3DP1</i> *	1	5
4	<i>GBAP1</i>	6	27
5	<i>MAS1</i>	2	14
6	<i>ARHGEF15</i>	9	100
7	<i>LRIT2</i>	7	24
8	<i>OR8G1</i>	1	22
9	<i>WASH2P</i>	1	12
10	<i>OR2AE1</i>	3	54
11	<i>OR2T27</i>	1	35
12	<i>HNMT</i>	5	37
13	<i>TNFRSF10B</i>	11	49
14	<i>MIR604</i>	2	19

*Genes also identified by $\mathbf{D}^{\text{w-main-int}}$.

Table C.5: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 5 genes by $\mathbf{D}^{\text{main-int}}$ at the 0.005 gene-level P -value threshold

Rank in $\mathbf{D}^{\text{main-int}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main-int}}$
1	<i>LOC84931</i> *	9	1
2	<i>CYP2E1</i> *	13	3
3	<i>MIR518E</i> *	1	4
4	<i>SERPINB3</i> *	1	2
5	<i>SPACA1</i>	6	11

*Genes also identified by $\mathbf{D}^{\text{w-main-int}}$.

Table C.6: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 8 genes by \mathbf{D}^{main} at the 0.005 gene-level P -value threshold

Rank in \mathbf{D}^{main}	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main-int}}$
1	<i>LOC84931</i> *	9	1
2	<i>SPACA1</i>	6	11
3	<i>CRISP2</i>	10	40
4	<i>SERPINB3</i> *	1	2
5	<i>IGJ</i> *	1	7
6	<i>RBM46</i>	12	66
7	<i>KRTAP20-1</i> *	1	6
8	<i>CXCL9</i> *	1	10

*Genes also identified by $\mathbf{D}^{\text{w-main-int}}$.

Table C.7: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 11 genes by \mathbf{D}^{int} at the 0.005 gene-level P -value threshold

Rank in \mathbf{D}^{int}	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-main-int}}$
1	<i>MIR518E</i> *	1	4
2	<i>KIR3DP1</i> *	1	5
3	<i>CYP2E1</i> *	13	3
4	<i>OR8G1</i>	1	22
5	<i>WASH2P</i>	1	12
6	<i>OR2T27</i>	1	35
7	<i>SPRYD5</i>	1	20
8	<i>UCHL5</i>	1	60
9	<i>GK3P</i>	1	78
10	<i>MAS1</i>	2	14
11	<i>TAS2R3</i>	1	82

*Genes also identified by $\mathbf{D}^{\text{w-main-int}}$.

Table C.8: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 4 genes by L^S at the 0.005 gene-level P -value threshold

Rank in L^S	Gene	# CpG	Rank in $D^{w\text{-main-int}}$
1	<i>LOC84931</i> *	9	1
2	<i>ADAM32</i> *	11	8
3	<i>TRIM38</i>	7	33
4	<i>SERPIN3</i> *	1	2

*Genes also identified by $D^{w\text{-main-int}}$.

Table C.9: Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 4 genes by L^M at the 0.005 gene-level P -value threshold

Rank in L^M	Gene	# CpG	Rank in $D^{w\text{-main-int}}$
1	<i>UBASH3B</i>	23	106
2	<i>MYH2</i>	8	41
3	<i>JARID2</i>	84	590
4	<i>TNFRSF10B</i>	11	49

The seven comparison methods $\mathbf{D}^{\text{w-main}}$, $\mathbf{D}^{\text{w-int}}$, $\mathbf{D}^{\text{main-int}}$, \mathbf{D}^{main} , \mathbf{D}^{int} , L^S and L^M identified 11, 14, 5, 8, 11, 4, 4 genes and replicated 2, 2, 2, 0, 3, 0, 0 genes (replication rate ranges 0-40% with a mean of 14%). These results are summarized in Supplementary Table C.10. The 2 genes, *LOC84931* and *HIST1H2BJ*, replicated by $\mathbf{D}^{\text{w-main}}$, as well as the 2 genes, *LOC84931* and *CYP2E1*, replicated by $\mathbf{D}^{\text{main-int}}$, were all identified and replicated by the proposed method due to main/interaction signals. The gene *CYP2E1* was replicated by $\mathbf{D}^{\text{w-int}}$, \mathbf{D}^{int} and $\mathbf{D}^{\text{main-int}}$, and was also identified and replicated by the proposed method due to interaction signals. The other gene, *WASH2P*, that was also replicated by both $\mathbf{D}^{\text{w-int}}$ and \mathbf{D}^{int} , was ranked #12 (P -value=0.0056) in the proposed method $\mathbf{D}^{\text{w-main-int}}$ results in the discovery analysis. In general, the genes replicated by the comparison methods were either all replicated or ranked on top in the $\mathbf{D}^{\text{w-main-int}}$ results. This suggests that the proposed method that incorporates both main and interaction signals indeed has better performance.

Table C.10: Summary number of genes identified at the 0.005 gene-level P -value threshold and replicated at the 0.1 gene-level P -value threshold in the application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3

Method	# of gene replicated / identified in discovery data	Replicated genes
$\mathbf{D}^{\text{w-main-int}}$	3/10	<i>LOC84931, CYP2E1, HIST1H2BJ</i>
$\mathbf{D}^{\text{w-main}}$	2/11	<i>LOC84931, HIST1H2BJ</i>
$\mathbf{D}^{\text{w-int}}$	2/14	<i>CYP2E1, WASH2P</i>
$\mathbf{D}^{\text{main-int}}$	2/5	<i>LOC84931, CYP2E1</i>
\mathbf{D}^{main}	0/8	-
\mathbf{D}^{int}	3/11	<i>CYP2E1, WASH2P, UCHL5</i>
L^S	0/4	-
L^M	0/4	-

C.2.3 Risk of PAH, DNA methylation and their interactions on MDI

Table C.11: Summary number of genes identified at the 0.005 gene-level P -value threshold and replicated at the 0.1 gene-level P -value threshold in the application examining prenatal PAH, DNA methylation and their interactions on child MDI at age 3

Method	# of gene replicated / identified in discovery data	Replicated genes
$D^{w\text{-main-int}}$	4/7	<i>FAM35A, DIRC1, C8orf80, THSD1P</i>
$D^{w\text{-main}}$	3/15	<i>FAM35A, C19orf77, DIRC1</i>
$D^{w\text{-int}}$	1/6	<i>KCTD19</i>
$D^{\text{main-int}}$	0/2	-
D^{main}	0/3	-
D^{int}	3/12	<i>GSTA1, OR4P4, FAM166B</i>
L^S	0/3	-
L^M	0/3	-

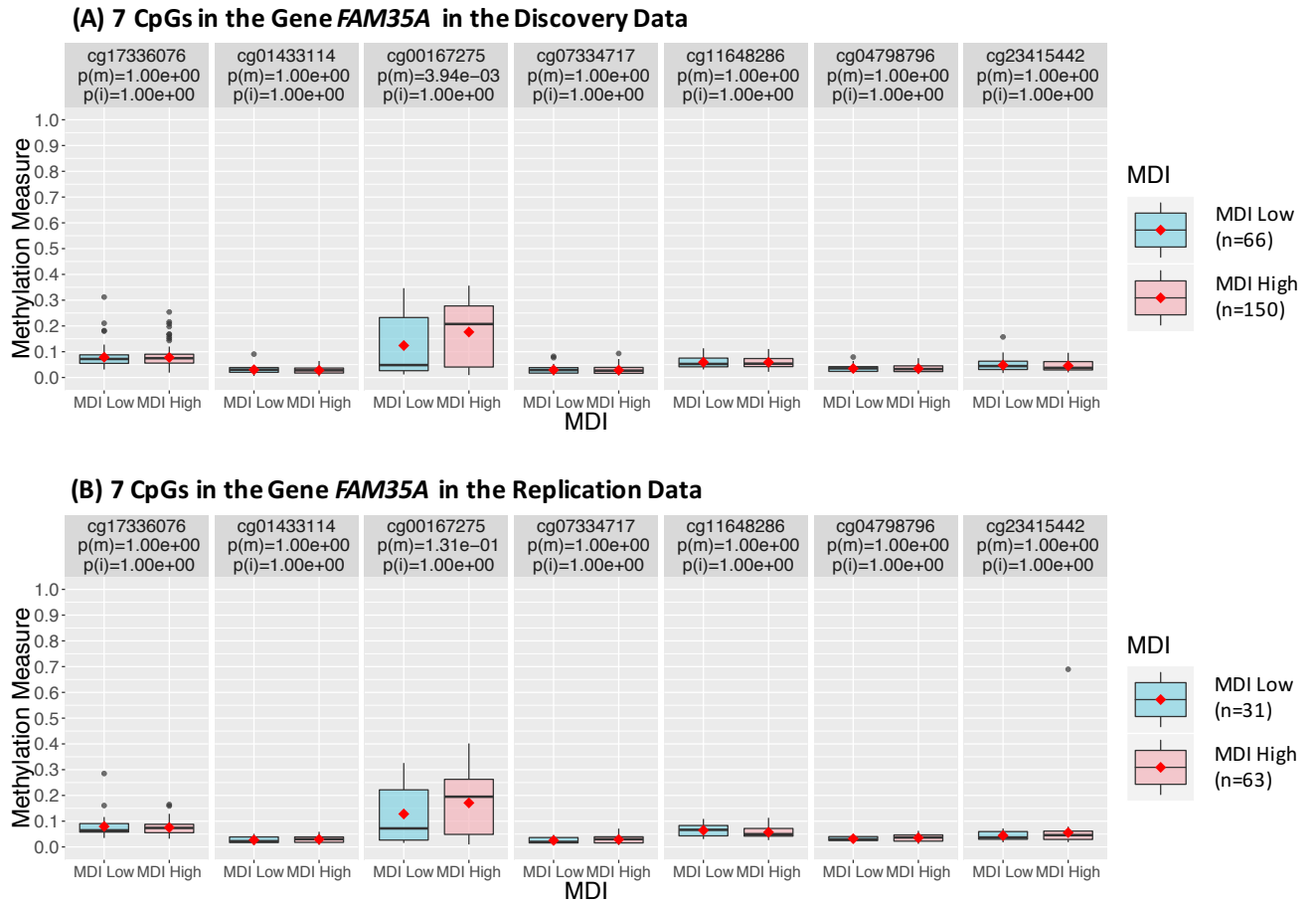


Figure C.5: Boxplot of DNA methylation measures of the 7 CpGs in gene *FAM35A* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *FAM35A*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1 \text{CpG} + \beta_2 E + \beta_3 \text{CpG} \times E$.



Figure C.6: Boxplot of DNA methylation measures of the 3 CpGs in gene *DIRC1* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *DIRC1*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1 \text{CpG} + \beta_2 E + \beta_3 \text{CpG} \times E$.

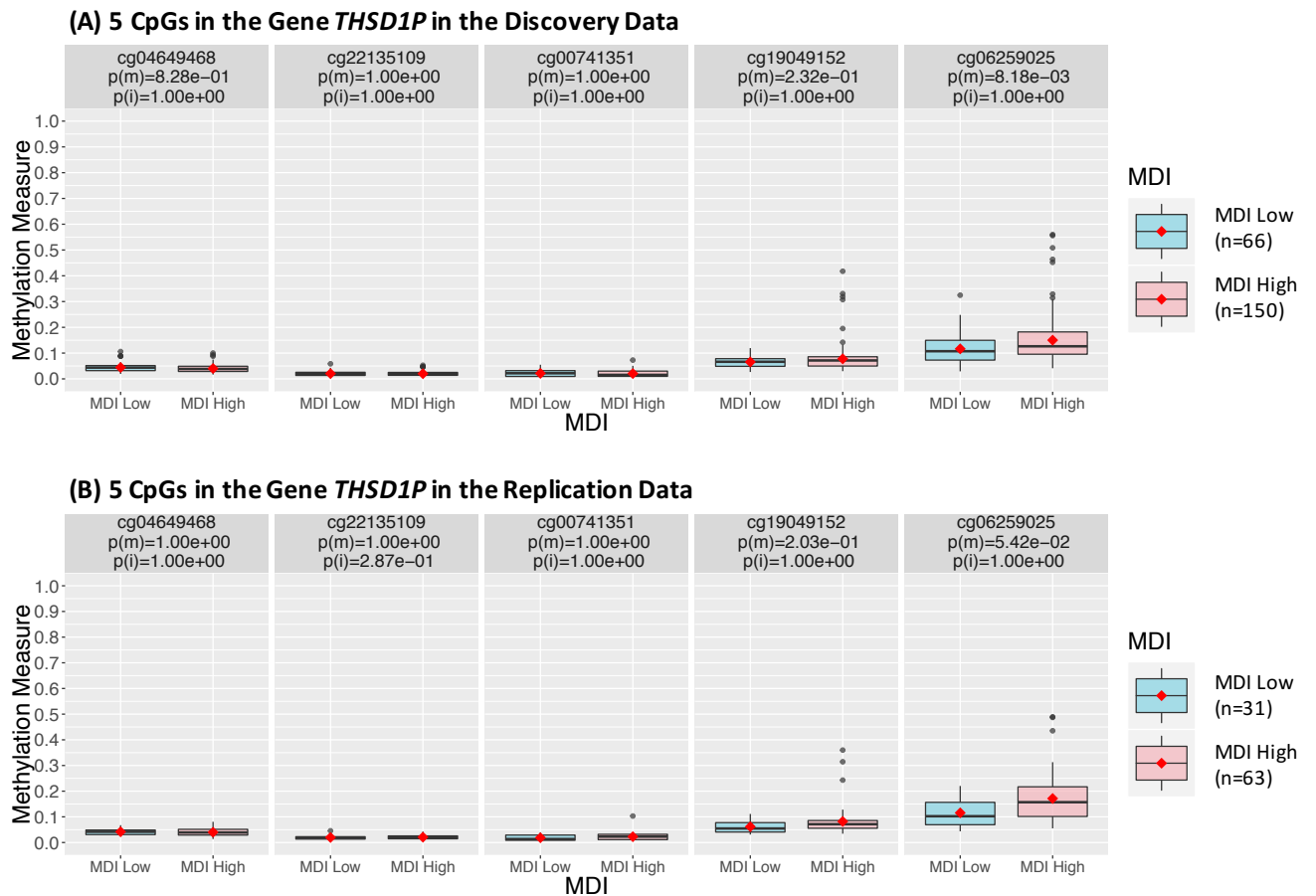


Figure C.7: Boxplot of DNA methylation measures of the 5 CpGs in gene *THSD1P* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *THSD1P*) P -values testing $\beta_1 = 0$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1 \text{CpG} + \beta_2 E + \beta_3 \text{CpG} \times E$.

Part II

Bibliography

Bibliography

- Abildgaard, M. O., Borre, M., Mortensen, M. M., Ulhøi, B. P., Tørring, N., Wild, P., Kristensen, H., Mansilla, F., Ottosen, P. D., Dyrskjøt, L., et al. (2012). Downregulation of zinc finger protein 132 in prostate cancer is associated with aberrant promoter hypermethylation and poor prognosis. *International journal of cancer*, 130(4):885–895.
- Abu-Asab, M., Abu-Asab, N., Loffredo, C., Clarke, R., and Amri, H. (2013). Identifying early events of gene expression in breast cancer with systems biology phylogenetics. *Cytogenetic and genome research*, 139(3):206–214.
- Achenbach, T. M. and Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles*, volume 30. Burlington, VT: University of Vermont, Research center for children, youth, & families.
- Ahn, S. and Wang, T. (2013). A powerful statistical method for identifying differentially methylated markers in complex diseases. *Pacific Symposium on Biocomputing*, pages 69–79.
- Akalın, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, 13:R87.
- Alholle, A., Brini, A., Bauer, J., Gharanei, S., Niada, S., Slater, A., Gentle, D., Maher, E., Jeys, L., Grimer, R., et al. (2015). Genome-wide DNA methylation profiling of recurrent and non-recurrent chordomas. *Epigenetics*, 10(3):213–220.

- Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46.
- Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369.
- Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
- Ayyala, D. N., Frankhouser, D. E., Ganbat, J.-O., Marcucci, G., Bundschuh, R., Yan, P., and Lin, S. (2015). Statistical methods for detecting differentially methylated regions based on MethylCap-seq data. *Briefings in bioinformatics*, page bbv089.
- Bakulski, K. M., Lee, H., Feinberg, J. I., Wells, E. M., Brown, S., Herbstman, J. B., Witter, F. R., Halden, R. U., Caldwell, K., Mortensen, M. E., et al. (2015). Prenatal mercury concentration is associated with changes in DNA methylation at TCEANC2 in newborns. *International journal of epidemiology*, 44(4):1249–1262.
- Bayley, N. (1993). *Bayley scales of infant development: Manual*. Psychological Corporation.
- Baylin, S. B. (2005). DNA methylation and gene silencing in cancer. *Nature clinical practice Oncology*, 2:S4–S11.
- Baylin, S. B., Esteller, M., Rountree, M. R., Bachman, K. E., Schuebel, K., and Herman, J. G. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human molecular genetics*, 10(7):687–692.
- Bernstein, C., Nfonsam, V., Prasad, A. R., and Bernstein, H. (2013). Epigenetic field defects in progression to cancer. *World journal of gastrointestinal oncology*, 5(3):43.
- Bertonha, F. B., de Camargo Barros Filho, M., Kuasne, H., Dos Reis, P. P., da Costa Prando, E., Muñoz, J. J. A. M., Roffé, M., Hajj, G. N. M., Kowalski, L. P., Rainho, C. A., et al. (2015). PHF21B as a candidate tumor suppressor gene in head and neck squamous cell carcinomas. *Molecular oncology*, 9(2):450–462.

- Bi, D., Ning, H., Liu, S., Que, X., and Ding, K. (2015). Gene expression patterns combined with network analysis identify hub genes associated with bladder cancer. *Computational biology and chemistry*, 56:71–83.
- Bièche, I., Chavey, C., Andrieu, C., Busson, M., Vacher, S., Le Corre, L., Guinebretière, J.-M., Burlinckon, S., Lidereau, R., and Lazennec, G. (2007). CXC chemokines located in the 4q21 region are up-regulated in breast cancer. *Endocrine-related cancer*, 14(4):1039–1052.
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.
- Butcher, L. M. and Beck, S. (2015). Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods*, 72:21–28.
- Cao, X.-C., Zhang, W.-R., Cao, W.-F., Liu, B.-W., Zhang, F., Zhao, H.-M., Meng, R., Zhang, L., Niu, R.-F., Hao, X.-S., et al. (2013). Aquaporin3 is required for FGF-2-induced migration of human breast cancers. *PloS one*, 8(2):e56735.
- Cardenas, A., Rifas-Shiman, S. L., Agha, G., Hivert, M.-F., Litonjua, A. A., DeMeo, D. L., Lin, X., Amarasiriwardena, C. J., Oken, E., Gillman, M. W., et al. (2017). Persistent DNA methylation changes associated with prenatal mercury exposure and cognitive performance during childhood. *Scientific Reports*, 7(1):288.
- Castaneda, F., Rosin-Steiner, S., and Jung, K. (2007). Functional genomics analysis of low concentration of ethanol in human hepatocellular carcinoma (HepG2) cells. role of genes involved in transcriptional and translational processes. *International journal of medical sciences*, 4(1):28.
- Chen, Y., Ning, Y., Hong, C., and Wang, S. (2014). Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina Arrays. *Genetic Epidemiology*, 38(1):42–50.
- Chen, Z., Huang, H., Liu, J., Ng, H. K. T., Nadarajah, S., Huang, X., and Deng, Y. (2013).

- Detecting differentially methylated loci for illumina array methylation data based on human ovarian cancer data. *BMC Medical Genomics*, 6:S9.
- Chen, Z., Li, J.-L., Lin, S., Cao, C., Gimbrone, N. T., Yang, R., Fu, D. A., Carper, M. B., Haura, E. B., Schabath, M. B., et al. (2016). cAMP/CREB-regulated LINC00473 marks LKB1-inactivated lung cancer and mediates tumor growth. *The Journal of clinical investigation*, 126(6):2267.
- Cheng, A. S., Li, M. S., Kang, W., Cheng, V. Y., Chou, J.-L., Lau, S. S., Go, M. Y., Lee, C. C., Ling, T. K., Ng, E. K., et al. (2013). Helicobacter pylori causes epigenetic dysregulation of FOXD3 to promote gastric carcinogenesis. *Gastroenterology*, 144(1):122–133.
- Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R., Sugarbaker, D. J., Yeh, R.-F., Wiencke, J. K., and Kelsey, K. T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genetics*, 5(8):e1000602.
- Costa, V. L., Henrique, R., Danielsen, S. A., Duarte-Pereira, S., Eknaes, M., Skotheim, R. I., Rodrigues, Â., Magalhães, J. S., Oliveira, J., Lothe, R. A., et al. (2010). Three epigenetic biomarkers, GDF15, TMEFF2 and VIM, accurately predict bladder cancer from DNA-based analyses of urine samples. *Clinical Cancer Research*, pages clincanres–1312.
- Curradi, M., Izzo, A., Badaracco, G., and Landsberger, N. (2002). Molecular mechanisms of gene silencing mediated by DNA methylation. *Molecular and cellular biology*, 22(9):3157–3173.
- Das, P. M. and Singal, R. (2004). DNA methylation and cancer. *Journal of clinical oncology*, 22(22):4632–4642.
- De Pontual, L., Trochet, D., Bourdeaut, F., Thomas, S., Etchevers, H., Chompret, A., Minard, V., Valteau, D., Brugieres, L., Munnich, A., et al. (2007). Methylation-associated

- PHOX2B gene silencing is a rare event in human neuroblastoma. *European Journal of Cancer*, 43(16):2366–2372.
- Deng, Q. and Huang, S. (2004). PRDM5 is silenced in human cancers and has growth suppressive activities. *Oncogene*, 23(28):4903.
- Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587.
- Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38:1378 – 1385.
- Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal instability and tumors promoted by DNA hypomethylation. *Science*, 300(5618):455–455.
- Ehrlich, M. (2002). DNA methylation in cancer: too much, but also too little. *Oncogene*, 21(35):5400–5413.
- Esteller, M. and Herman, J. G. (2002). Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *The Journal of pathology*, 196(1):1–7.
- Ewing, A. D., Gacita, A., Wood, L. D., Ma, F., Xing, D., Kim, M.-S., Manda, S. S., Abril, G., Pereira, G., Makohon-Moore, A., et al. (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome research*, pages gr-196238.
- Fackler, M. J., Umbricht, C. B., Williams, D., Argani, P., Cruz, L.-A., Merino, V. F., Teo, W. W., Zhang, Z., Huang, P., Visvanathan, K., et al. (2011). Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer research*, 71(19):6195–6207.

- Fahrner, J. A., Eguchi, S., Herman, J. G., and Baylin, S. B. (2002). Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer research*, 62(24):7213–7218.
- Faryna, M., Konermann, C., Aulmann, S., Bermejo, J. L., Brugger, M., Diederichs, S., Rom, J., Weichenhan, D., Claus, R., Rehli, M., et al. (2012). Genome-wide methylation screen in low-grade breast cancer identifies novel epigenetically altered genes as potential biomarkers for tumor diagnosis. *The FASEB Journal*, 26(12):4937–4950.
- Faulk, C., Kim, J. H., Jones, T. R., McEachin, R. C., Nahar, M. S., Dolinoy, D. C., and Sartor, M. A. (2015). Bisphenol A-associated alterations in genome-wide DNA methylation and gene expression patterns reveal sequence-dependent and non-monotonic effects in human fetal liver. *Environmental epigenetics*, 1(1).
- Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, 47:433–440.
- Feinberg, A. P. and Irizarry, R. A. (2010). Evolution in health and medicine sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *PNAS*, 107(Suppl 1):1757–1764.
- Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–153.
- Fernandez, A. F., Assenov, Y., Martin-Subero, J. I., Balint, B., Siebert, R., Taniguchi, H., Yamamoto, H., Hidalgo, M., Tan, A.-C., Galm, O., et al. (2012). A DNA methylation fingerprint of 1628 human samples. *Genome research*, 22(2):407–419.
- Fidalgo, F., Rodrigues, T. C., Pinilla, M., Silva, A. G., do Socorro Maciel, M., Rosenberg, C., de Andrade, V. P., Carraro, D. M., and Krepischi, A. C. V. (2015). Lymphovascular invasion and histologic grade are associated with specific genomic profiles in invasive carcinomas of the breast. *Tumor Biology*, 36(3):1835–1848.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

- Fu, A., Leaderer, B. P., Gent, J. F., Leaderer, D., and Zhu, Y. (2012). An environmental epigenetic study of ADRB 2 5'-UTR methylation and childhood asthma severity. *Clinical & Experimental Allergy*, 42(11):1575–1581.
- Gervin, K., Hammerø, M., Akselsen, H. E., Moe, R., Nygård, H., Brandt, I., Gjessing, H. K., Harris, J. R., Undlien, D. E., and Lyle, R. (2011). Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Research*, 21(11):1813–1821.
- Ghosh, S., Albitar, L., LeBaron, R., Welch, W. R., Samimi, G., Birrer, M. J., Berkowitz, R. S., and Mok, S. C. (2010). Up-regulation of stromal versican expression in advanced stage serous ovarian cancer. *Gynecologic oncology*, 119(1):114–120.
- Giussani, M., Merlino, G., Cappelletti, V., Tagliabue, E., and Daidone, M. G. (2015). Tumor-extracellular matrix interactions: Identification of tools associated with breast cancer progression. In *Seminars in cancer biology*, volume 35, pages 3–10. Elsevier.
- Gomes, I. M., Maia, C. J., and Santos, C. R. (2012). Steap proteins: from structure to applications in cancer therapy. *Molecular Cancer Research*, 10(5):573–587.
- Gurnot, C., Martin-Subero, I., Mah, S. M., Weikum, W., Goodman, S. J., Brain, U., Werker, J. F., Kobor, M. S., Esteller, M., Oberlander, T. F., et al. (2015). Prenatal antidepressant exposure associated with CYP2E1 DNA methylation change in neonates. *Epigenetics*, 10(5):361–372.
- Hair, B. Y., Xu, Z., Kirk, E. L., Harlid, S., Sandhu, R., Robinson, W. R., Wu, M. C., Olshan, A. F., Conway, K., Taylor, J. A., et al. (2015). Body mass index associated with genome-wide methylation in breast tissue. *Breast cancer research and treatment*, 151(2):453–463.
- Han, F. and Pan, W. (2010). Powerful multi-marker association tests: unifying genomic distance-based regression and logistic regression. *Genetic epidemiology*, 34(7):680–688.
- Hansen, K. (2015). IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for illumina's 450k methylation arrays. *R package, version 0.2*, 1.

- Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83.
- Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., and Feinberg, A. P. (2011a). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775.
- Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011b). Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768.
- Harvey, K. F., Zhang, X., and Thomas, D. M. (2013). The Hippo pathway and human cancer. *Nature Reviews Cancer*, 13(4):246.
- He, B., Reguart, N., You, L., Mazieres, J., Xu, Z., Lee, A. Y., Mikami, I., McCormick, F., and Jablons, D. M. (2005). Blockade of Wnt-1 signaling induces apoptosis in human colorectal cancer cells containing downstream mutations. *Oncogene*, 24(18):3054–3058.
- Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653.
- Herbstman, J. B., Tang, D., Zhu, D., Qu, L., Sjödin, A., Li, Z., Camann, D., and Perera, F. P. (2012). Prenatal exposure to polycyclic aromatic hydrocarbons, benzo [a] pyrene–DNA adducts, and genomic DNA methylation in cord blood. *Environmental health perspectives*, 120(5):733.
- Herman, J. G. and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349(21):2042–2054.
- Hesse, N., Schröder, C., and Rahmann, S. (2015). An optimization approach to detect differentially methylated regions from whole genome bisulfite sequencing data. Technical report, PeerJ PrePrints.

- Hinoue, T., Weisenberger, D. J., Lange, C. P., Shen, H., Byun, H.-M., Van Den Berg, D., Malik, S., Pan, F., Noushmehr, H., and van Dijk, C. M. (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome research*, 22(2):271–282.
- Ho, J. W., Stefani, M., dos Remedios, C. G., and Charleston, M. A. (2008). Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, 24(13):i390–i398.
- Hong, S., Dong, H., Jin, L., and Xiong, M. (2010). Gene co-expression network analysis of two ovarian cancer datasets. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pages 269–274. IEEE.
- Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):3156.
- Hou, J., Wu, J., Dombkowski, A., Zhang, K., Holowatyj, A., Boerner, J. L., and Yang, Z.-Q. (2012). Genomic amplification and a role in drug-resistance for the KDM5A histone demethylase in breast cancer. *American journal of translational research*, 4(3):247.
- Huang, H., Chen, Z., and Huang, X. (2013). Age-adjusted nonparametric detection of differential DNA methylation with case-control designs. *BMC Bioinformatics*, 14.
- Huber, R. M., Lucas, J. M., Gomez-Sarosi, L. A., Coleman, I., Zhao, S., Coleman, R., and Nelson, P. S. (2015). DNA damage induces GDNF secretion in the tumor microenvironment with paracrine effects promoting prostate cancer treatment resistance. *Oncotarget*, 6(4):2134.
- Ip, W.-K., Lai, P. B.-S., Wong, N. L.-Y., Sy, S. M.-H., Beheshti, B., Squire, J. A., and Wong, N. (2007). Identification of PEG10 as a progression related biomarker for hepatocellular carcinoma. *Cancer letters*, 250(2):284–291.
- Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddelloh, J. A., Wen, B., and Feinberg, A. P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, 18(5):780–790.

- Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J., Sabunciyan, S., and Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178–186.
- Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254.
- Jaffe, A. E., Feinberg, A. P., Irizarry, R. A., and Leek, J. T. (2012a). Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, 13(1):166–178.
- Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012b). Bump hunting to identifying differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1):200–209.
- Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012c). Bump hunting to identifying differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1):200–209.
- Jager, P. L. D., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., Eaton, M. L., Keenan, B. T., Ernst, J., McCabe, C., Tang, A., Raj, T., Replogle, J., Brodeur, W., Gabriel, S., Chai, H. S., Younkin, C., Younkin, S. G., Zou, F., Szyf, M., Epstein, C. B., Schneider, J. A., Bernstein, B. E., Meissner, A., Ertekin-Taner, N., Chibnik, L. B., Kellis, M., Mill, J., and Bennett, D. A. (2014). Alzheimer’s disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature Neuroscience*, 17:1156–1163.
- Janssen, B. G., Godderis, L., Pieters, N., Poels, K., Kiciński, M., Cuypers, A., Fierens, F., Penders, J., Plusquin, M., Gyselaers, W., et al. (2013). Placental DNA hypomethylation in association with particulate air pollution in early life. *Particle and fibre toxicology*, 10(1):22.
- Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

- Jönsson, G., Staaf, J., Vallon-Christersson, J., Ringnér, M., Holm, K., Hegardt, C., Gunnarsson, H., Fagerholm, R., Strand, C., Agnarsson, B. A., et al. (2010). Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Research*, 12(3):R42.
- Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., and Hoffmann, S. (2016). metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, 26(2):256–262.
- Katsurano, M., Niwa, T., Yasui, Y., Shigematsu, Y., Yamashita, S., Takeshima, H., Lee, M., Kim, Y., Tanaka, T., and Ushijima, T. (2012). Early-stage formation of an epigenetic field defect in a mouse colitis model, and non-essential roles of T- and B-cells in DNA methylation induction. *Oncogene*, 31(3):342.
- Koukoura, O., Spandidos, D. A., Daponte, A., and Sifakis, S. (2014). DNA methylation profiles in ovarian cancer: implication in diagnosis and therapy. *Molecular Medicine Reports*, 10(1):3–9.
- Krijgsman, O., Roepman, P., Zwart, W., Carroll, J. S., Tian, S., de Snoo, F. A., Bender, R. A., Bernardis, R., and Glas, A. M. (2012). A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response. *Breast cancer research and treatment*, 133(1):37–47.
- Kulis, M. and Esteller, M. (2010). DNA methylation and cancer. *Adv Genet*, 70(10):27–56.
- Kumsta, R., Marzi, S. J., Viana, J., Dempster, E., Crawford, B., Rutter, M., Mill, J., and Sonuga-Barke, E. J. (2016). Severe psychosocial deprivation in early childhood is associated with increased DNA methylation across a region spanning the transcription start site of CYP2E1. *Translational psychiatry*, 6(6):e830.
- Lasseigne, B. N., Burwell, T. C., Patil, M. A., Absher, D. M., Brooks, J. D., and Myers, R. M. (2014). DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma. *BMC medicine*, 12(1):1.

- Lee, S. H., Um, S.-J., and Kim, E.-J. (2013). CBX8 suppresses Sirtinol-induced premature senescence in human breast cancer cells via cooperation with SIRT1. *Cancer letters*, 335(2):397–403.
- Legendre, C., Gooden, G. C., Johnson, K., Martinez, R. A., Liang, W. S., and Salhia, B. (2015). Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clinical epigenetics*, 7(1):100.
- Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., and Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750.
- Li, H., Du, Y., Zhang, D., Wang, L.-N., Yang, C., Liu, B., Wang, W.-J., Shi, L., Hong, W.-G., Zhang, L., et al. (2012). Identification of novel DNA methylation markers in colorectal cancer using MIRA-based microarrays. *Oncology reports*, 28(1):99–104.
- Lister, R., Pelizzola, M., Downen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.
- Liu, H., Liu, X., Zhang, S., Lv, J., Li, S., Shang, S., Jia, S., Wei, Y., Wang, F., Su, J., et al. (2015). Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic acids research*, 44(1):75–94.
- Liu, J., Morgan, M., Hutchison, K., and Calhoun, V. D. (2010). A study of the influence of sex on genome wide methylation. *PLoS One*, 5(4):e10028.
- Lokk, K., Vooder, T., Kolde, R., Vääk, K., Võsa, U., Roosipuu, R., Milani, L., Fischer, K., Koltsina, M., Urgard, E., et al. (2012). Methylation markers of early-stage non-small cell lung cancer. *PloS one*, 7(6):e39813.

- Lund, G., Andersson, L., Lauria, M., Lindholm, M., Fraga, M. F., Villar-Garea, A., Ballesta, E., Esteller, M., and Zaina, S. (2004). DNA methylation polymorphisms precede any histological sign of atherosclerosis in mice lacking apolipoprotein E. *The Journal of Biological Chemistry*, 279:29147–29154.
- Many, A. M. and Brown, A. M. (2010). Mammary stem cells and cancer: roles of Wnt signaling in plain view. *Breast Cancer Research*, 12(5):313.
- Matise, L. A., Palmer, T. D., Ashby, W. J., Nashabi, A., Chytil, A., Aakre, M., Pickup, M. W., Gorska, A. E., Zijlstra, A., and Moses, H. L. (2012). Lack of transforming growth factor- β signaling promotes collective cancer cell invasion through tumor-stromal crosstalk. *Breast Cancer Research*, 14(4):R98.
- Mayo, T. R., Schweikert, G., and Sanguinetti, G. (2014). M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, 31(6):809–816.
- McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297.
- Mill, J. and Petronis, A. (2007). Molecular studies of major depressive disorder: the epigenetic perspective. *Molecular Psychiatry*, 12:799–814.
- Mill, J. and Petronis, A. (2008). Pre- and peri-natal environmental risks for attention-deficit hyperactivity disorder (ADHD): the potential role of epigenetic processes in mediating susceptibility. *Journal of Child Psychology and Psychiatry*, 49(10):1020–1030.
- Mill, J., Tang, T., Kaminsky, Z., Khare, T., Yazdanpanah, S., Bouchard, L., Jia, P., Asadzadeh, A., Flanagan, J., Schumacher, A., Wang, S.-C., and Petronis, A. (2008). Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *The American Journal of Human Genetics*, 82(3):696–711.
- Morgan, W. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, 31(1/2):13–19.
- Nahar, M. S., Liao, C., Kannan, K., Harris, C., and Dolinoy, D. C. (2015). In utero

- bisphenol a concentration, metabolism, and global DNA methylation across matched placenta, kidney, and liver in the human fetus. *Chemosphere*, 124:54–60.
- Nestler, E. J. (2014). Epigenetic mechanisms of drug addiction. *Neuropharmacology*, 76:259–268.
- Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.
- Nye, M. D., King, K. E., Darrah, T. H., Maguire, R., Jima, D. D., Huang, Z., Mendez, M. A., Fry, R. C., Jirtle, R. L., Murphy, S. K., et al. (2016). Maternal blood lead concentrations, DNA methylation of MEG3 DMR regulating the DLK1/MEG3 imprinted domain and early growth in a multiethnic cohort. *Environmental epigenetics*, 2(1):dvv009.
- Ortega, P., Moran, A., Fernandez-Marcelo, T., De Juan, C., Frias, C., Lopez-Asenjo, J.-A., Sanchez-Pernaute, A., Torres, A., Diaz-Rubio, E., Iniesta, P., et al. (2010). MMP-7 and SGCE as distinctive molecular factors in sporadic colorectal cancers from the mutator phenotype pathway. *International journal of oncology*, 36(5):1209.
- Perera, F., Tang, W.-y., Herbstman, J., Tang, D., Levin, L., Miller, R., and Ho, S.-m. (2009). Relation of DNA methylation of 5'-CpG island of ACSL3 to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. *PloS one*, 4(2):e4488.
- Perera, F. P., Rauh, V., Tsai, W.-Y., Kinney, P., Camann, D., Barr, D., Bernert, T., Garfinkel, R., Tu, Y.-H., Diaz, D., et al. (2003). Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. *Environmental health perspectives*, 111(2):201.
- Perera, F. P., Rauh, V., Whyatt, R. M., Tsai, W.-Y., Tang, D., Diaz, D., Hoepner, L., Barr, D., Tu, Y.-H., Camann, D., et al. (2006). Effect of prenatal exposure to airborne polycyclic aromatic hydrocarbons on neurodevelopment in the first 3 years of life among inner-city children. *Environmental health perspectives*, 114(8):1287–1292.

- Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., Lord, R. V., Clark, S. J., and Molloy, P. L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin*, 8(1):6.
- Phillips, T. (2008). The role of methylation in gene expression. *Nature Education*, 1(1):116.
- Phipson, B. and Oshlack, A. (2014). Diffvar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome biology*, 15(9):465.
- Pidsley, R., Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, 14(1):293.
- Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, 31(1-2):9–12.
- Raza, K. and Jaiswal, R. (2013). Reconstruction and analysis of cancer-specific gene regulatory networks from gene expression profiles. *arXiv preprint arXiv:1305.5750*.
- Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610.
- Ruan, P., Shen, J., Santella, R. M., Zhou, S., and Wang, S. (2016). NEpiC: a network-assisted algorithm for epigenetic studies using mean and variance combined signals. *Nucleic Acids Research*, page gkw546.
- Rudenko, V., Kazakova, S., Tanas, A., Popa, A., Nemirovchenko, V., Kuznetsova, E., Zaletaev, D., and Strelnikov, V. (2016). Identification of aberrant DNA methylation in pediatric acute myeloid leukaemia by multiplex methylation sensitive PCR. *Annals of Oncology*, 27(suppl.6).
- Ruike, Y., Imanaka, Y., Sato, F., Shimizu, K., and Tsujimoto, G. (2010). Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC genomics*, 11(1):137.
- Saenen, N. D., Vrijens, K., Janssen, B. G., Roels, H. A., Neven, K. Y., Vanden Berghe, W., Gyselaers, W., Vanpoucke, C., Lefebvre, W., De Boever, P., et al. (2016). Lower

- placental leptin promoter methylation in association with fine particulate matter air pollution during pregnancy and placental nitrosative stress at birth in the ENVIRON AGE cohort. *Environmental health perspectives*, 125(2):262–268.
- Saito, Y. and Mituyama, T. (2015). Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions. *BMC genomics*, 16(Suppl 12):S3.
- Saito, Y., Tsuji, J., and Mituyama, T. (2014). Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic acids research*, page gkt1373.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.
- Schanen, N. C. (2006). Epigenetics of autism spectrum disorders. *Human Molecular Genetics*, 15(2):138–150.
- Sen, A., Cingolani, P., Senut, M.-C., Land, S., Mercado-Garcia, A., Tellez-Rojo, M. M., Baccarelli, A. A., Wright, R. O., and Ruden, D. M. (2015). Lead exposure induces changes in 5-hydroxymethylcytosine clusters in CpG islands in human embryonic stem cells and umbilical cord blood. *Epigenetics*, 10(7):607–621.
- Sharma, G., Mirza, S., Prasad, C. P., Srivastava, A., Gupta, S. D., and Ralhan, R. (2007). Promoter hypermethylation of p16 INK4A, p14 ARF, CyclinD2 and Slit2 in serum and tumor DNA from breast cancer patients. *Life sciences*, 80(20):1873–1881.
- Shen, J., Wang, S., Zhang, Y.-J., Kappil, M., Wu, H.-C., Kibriya, M. G., Wang, Q., Jasmine, F., Ahsan, H., Lee, P.-H., Yu, M.-W., Chen, C.-J., and Santella, R. M. (2012). Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology*, 55(6):1799–1808.
- Slaughter, D. P., Southwick, H. W., and Smejkal, W. (1953). “Field cancerization” in oral stratified squamous epithelium. clinical implications of multicentric origin. *Cancer*, 6(5):963–968.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential

- expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25.
- Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L., and Baccarelli, A. A. (2013). A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, page btt498.
- Song, M.-A., Tiirikainen, M., Kwee, S., Okimoto, G., Yu, H., and Wong, L. L. (2013). Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. *PloS one*, 8(2):e55761.
- Steinbach, D., Schramm, A., Eggert, A., Onda, M., Dawczynski, K., Rump, A., Pastan, I., Wittig, S., Pfaffendorf, N., Voigt, A., et al. (2006). Identification of a set of seven genes for the monitoring of minimal residual disease in pediatric acute myeloid leukemia. *Clinical cancer research*, 12(8):2434–2441.
- Stirzaker, C., Zotenko, E., Song, J. Z., Qu, W., Nair, S. S., Locke, W. J., Stone, A., Armstrong, N. J., Robinson, M. D., Dobrovic, A., et al. (2015). Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. *Nature communications*, 6:5899.
- Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.
- Sun, H. and Wang, S. (2012). Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375.
- Sun, H. and Wang, S. (2013). Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Statistics in Medicine*, 32(12):2127–2139.
- Sun, H., Wang, Y., Chen, Y., Li, Y., and Wang, S. (2017). pETM: a penalized Exponential Tilt Model for analysis of correlated high-dimensional DNA methylation data. *Bioinformatics*, 33(12):1765–1772.

- Tada, Y., Yamaguchi, Y., Kinjo, T., Song, X., Akagi, T., Takamura, H., Ohta, T., Yokota, T., and Koide, H. (2015). The stem cell transcription factor ZFP57 induces IGF2 expression to promote anchorage-independent growth in cancer cells. *Oncogene*, 34(6):752–760.
- Takamaru, H., Yamamoto, E., Suzuki, H., Nojima, M., Maruyama, R., Yamano, H.-o., Yoshikawa, K., Kimura, T., Harada, T., Ashida, M., et al. (2012). Aberrant methylation of RASGRF1 is associated with an epigenetic field defect and increased risk of gastric cancer. *Cancer Prevention Research*, 5(10):1203–1212.
- Teschendorff, A. E., Gao, Y., Jones, A., Ruebner, M., Beckmann, M. W., Wachter, D. L., Fasching, P. A., and Widschwendter, M. (2016a). DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nature communications*, 7.
- Teschendorff, A. E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J. J., Kitchener, H. C., and Widschwendter, M. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome medicine*, 4(3):24.
- Teschendorff, A. E., Jones, A., and Widschwendter, M. (2016b). Stochastic epigenetic outliers can define field defects in cancer. *BMC bioinformatics*, 17(1):1.
- Teschendorff, A. E., Liu, X., Caren, H., Pollard, S. M., Beck, S., Widschwendter, M., and Chen, L. (2014). The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Comput Biol*, 10(7):e1003709.
- Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., and Jacobs, I. J. (2009). An epigenetic signature in peripheral blood predicts active ovarian cancer. *PloS one*, 4(12):e8274.
- Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., Savage, D. A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gaythe, S. A., Jones, A., Beck, S., Wagne, W., Laird, P. W., Jacobs, I. J., and Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440–446.

- Teschendorff, A. E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, 28(11):1487–1494.
- Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.
- Virmani, A., Rathi, A., Heda, S., Sugio, K., Lewis, C., Tonk, V., Takahashi, T., Roth, J. A., Minna, J. D., Euhus, D. M., et al. (2003). Aberrant methylation of the cyclin D2 promoter in primary small cell, nonsmall cell lung and breast cancers. *International journal of cancer*, 107(3):341–345.
- Wang, J., Li, J., Gu, J., Yu, J., Guo, S., Zhu, Y., and Ye, D. (2015). Abnormal methylation status of FBXW10 and SMPD3, and associations with clinical characteristics in clear cell renal cell carcinoma. *Oncology letters*, 10(5):3073–3080.
- Wang, Y., Qian, M., Ruan, P., Teschendorff, A. E., and Wang, S. (2018). Detection of epigenetic field defects using a weighted epigenetic distance-based method. *Nucleic acids research*, 47(1):e6–e6.
- Wang, Y., Teschendorff, A. E., Widschwendter, M., and Wang, S. (2017). Accounting for differential variability in detecting differentially methylated regions. *Briefings in Bioinformatics*.
- Wei, H., Wang, H., Ji, Q., Sun, J., Tao, L., and Zhou, X. (2015). NRBP1 is downregulated in breast cancer and NRBP1 overexpression inhibits cancer cell proliferation through wnt/ β -catenin signaling pathway. *OncoTargets and therapy*, 8:3721.
- Wen, Y., Chen, F., Zhang, Q., Zhuang, Y., and Li, Z. (2016). Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics*, 32(22):3396–3404.
- Wessel, J. and Schork, N. J. (2006). Generalized genomic distance-based regression method-

- ology for multilocus association analysis. *The American Journal of Human Genetics*, 79(5):792–806.
- Wettenhall, J. M. and Smyth, G. K. (2004). limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, 20(18):3705–3706.
- White, A. J., Chen, J., McCullough, L. E., Xu, X., Cho, Y. H., Teitelbaum, S. L., Neugut, A. I., Terry, M. B., Hibshoosh, H., Santella, R. M., et al. (2015). Polycyclic aromatic hydrocarbon (PAH)–DNA adducts and breast cancer: modification by gene promoter methylation in a population-based study. *Cancer Causes & Control*, 26(12):1791–1802.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.
- Wolf, I., Bose, S., Desmond, J. C., Lin, B. T., Williamson, E. A., Karlan, B. Y., and Koeffler, H. P. (2007). Unmasking of epigenetically silenced genes reveals dna promoter methylation and reduced expression of *ptch* in breast cancer. *Breast cancer research and treatment*, 105(2):139–155.
- Wolf, J., Müller-Decker, K., Flechtenmacher, C., Zhang, F., Shahmoradgoli, M., Mills, G., Hoheisel, J., and Boettcher, M. (2014). An in vivo RNAi screen identifies *SALL1* as a tumor suppressor in human breast cancer with a role in *CDH1* regulation. *Oncogene*, 33(33):4273.
- Wu, H., Chen, Y., Liang, J., Shi, B., Wu, G., Zhang, Y., Wang, D., Li, R., Yi, X., Zhang, H., et al. (2005). Hypomethylation-linked activation of *PAX2* mediates tamoxifen-stimulated endometrial carcinogenesis. *Nature*, 438(7070):981–987.
- Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P., and Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic acids research*, page gkv715.
- Xia, T.-S., Wang, G.-Z., Ding, Q., Liu, X.-A., Zhou, W.-B., Zhang, Y.-F., Zha, X.-M., Du, Q., Ni, X.-J., Wang, J., et al. (2012). Bone metastasis in a novel breast cancer mouse

- model containing human breast and human bone. *Breast cancer research and treatment*, 132(2):471–486.
- Yang, W. S., Moon, H.-G., Kim, H. S., Choi, E.-J., Yu, M.-H., Noh, D.-Y., and Lee, C. (2011). Proteomic approach reveals FKBP4 and S100A9 as potential prediction markers of therapeutic response to neoadjuvant chemotherapy in patients with breast cancer. *Journal of proteome research*, 11(2):1078–1088.
- Yip, W.-K., Fier, H., DeMeo, D. L., Aryee, M., Laird, N., and Lange, C. (2014). A novel method for detecting association between DNA methylation and diseases using spatial information. *Genetic epidemiology*, 38(8):714–721.
- Yong-Deok, K., Eun-Hyoung, J., Yeon-Sun, K., Kang-Mi, P., Jin-Yong, L., Sung-Hwan, C., Tae-Yun, K., Tae-Sung, P., Soung-Min, K., Myung-Jin, K., et al. (2015). Molecular genetic study of novel biomarkers for early diagnosis of oral squamous cell carcinoma. *Medicina oral, patologia oral y cirugia bucal*, 20(2):e167.
- Zapala, M. A. and Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the national academy of sciences*, 103(51):19430–19435.
- Zhang, G., He, P., Gaedcke, J., Ghadimi, B. M., Ried, T., Yfantis, H. G., Lee, D. H., Hanna, N., Alexander, H. R., and Hussain, S. P. (2013a). FOXL1, a novel candidate tumor suppressor, inhibits tumor aggressiveness and predicts outcome in human pancreatic cancer. *Cancer research*.
- Zhang, J. T., Jiang, X. H., Xie, C., Cheng, H., Da Dong, J., Wang, Y., Fok, K. L., Zhang, X. H., Sun, T. T., Tsang, L. L., et al. (2013b). Downregulation of CFTR promotes epithelial-to-mesenchymal transition and is associated with poor prognosis of breast cancer. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1833(12):2961–2969.
- Zhang, M. and Luo, S. (2016). Gene expression profiling of epithelial ovarian cancer reveals key genes and pathways associated with chemotherapy resistance. *Genet Mol Res*, 15(1):11.

- Zhang, P., Wen, X., Gu, F., Deng, X., Li, J., Dong, J., Jiao, J., and Tian, Y. (2013c). Methylation profiling of serum DNA from hepatocellular carcinoma patients using an Infinium Human Methylation 450 BeadChip. *Hepatology international*, 7(3):893–900.
- Zhao, J., Liang, Q., Cheung, K.-F., Kang, W., Lung, R. W., Tong, J. H., To, K. F., Sung, J. J., and Yu, J. (2013). Genome-wide identification of Epstein-Barr virus-driven promoter methylation profiles of human genes in gastric cancer cells. *Cancer*, 119(2):304–312.