Supplementary Data for "A powerful and flexible weighted distance-based method incorporating interactions between DNA methylation and environmental factors on health outcomes"

Ya Wang[1], Min Qian[1], Deliang Tang[2], Julie Herbstman[2], Frederica Perera[2] and Shuang Wang[1*]

1. Department of Biostatistics, Mailman School of Public Health, Columbia University.

2. Columbia Center for Children's Environmental Health, Department of Environmental Health Science, Mailman School of Public Health, Columbia University.

*Corresponding author. Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA. E-mail: sw2206@columbia.edu

1 Additional Simulation Studies

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 S1. Type I error rates in simulation settings with multiple genes of different sizes

Method	**Pooled F statistics**	**Not pool F statistics**
$\mathbf{D}^{\text{w-M-E-int}}$	0.044	0.040
$\mathbf{D}^{\text{w-M}}$	0.052	0.049
$\mathbf{D}^{\text{w-int}}$	0.050	0.046
$\mathbf{D}^{\text{M-E-int}}$	0.048	0.040
$\mathbf{D}^{\text{M}}$	0.034	0.053
$\mathbf{D}^{\text{int}}$	0.046	0.039
L^S	-	0.019
L^M	-	0.033

1.2 Simulation settings with different types of signals

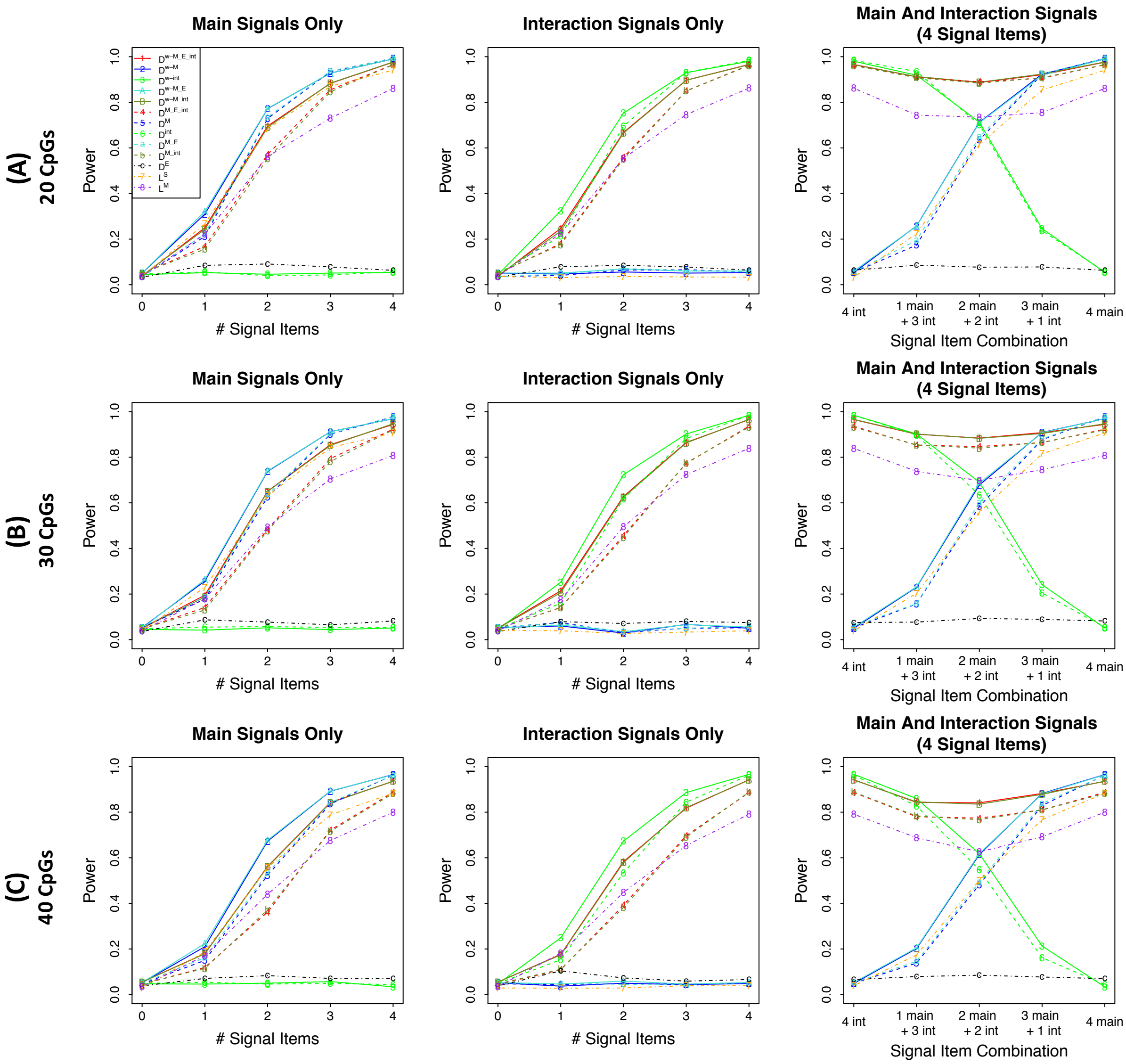

Figure S1. Power results for simulation settings with main signals only, interaction signals only and both main and interaction signals when there are (A) 20 CpGs, (B) 30 CpGs, and (C) 40 CpGs in a gene.

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

1.3.1 Simulation setup

Table S2. 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: 2 CpGs with main + interaction signals	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_1} = \beta_{Z_3} = 0.3$
3 signal CpGs: 1 CpG with main + interaction signals; 1 CpG with main signal only; 1 CpG with interaction signal only	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_3} = \beta_{Z_5} = 0.3$
4 signal CpGs: 2 CpGs with main signal only; 2 CpGs with interaction signal only	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_5} = \beta_{Z_7} = 0.3$

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

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 all distance-based methods as well as L^S increases slightly as expected, while that of L^M decreases. And the weighted distance-based methods always perform better than the non-weighted versions.

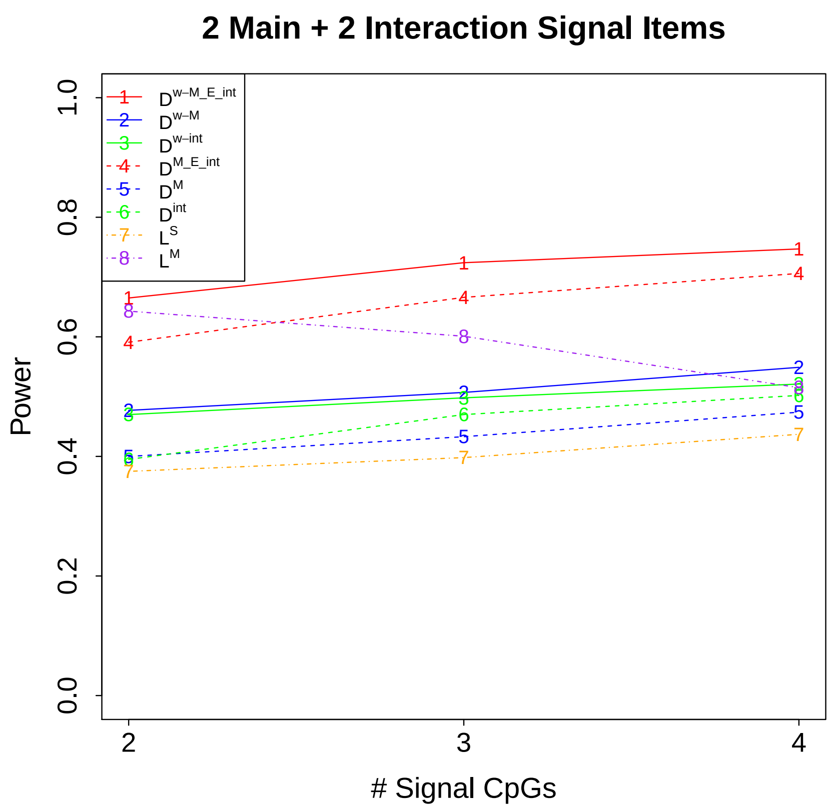

Figure S2. Power results for simulation settings where there are 2 main signal items and 2 interaction signal items coming from 2, 3 and 4 signal CpGs, respectively when there are 30 CpGs in a gene.

2 Real data applications

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'[1]. 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[2]. We then calibrated the 450K data to EPIC distribution[3], 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 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[4] 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.

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

2.2.1 Replication analysis in the Sibling cohort

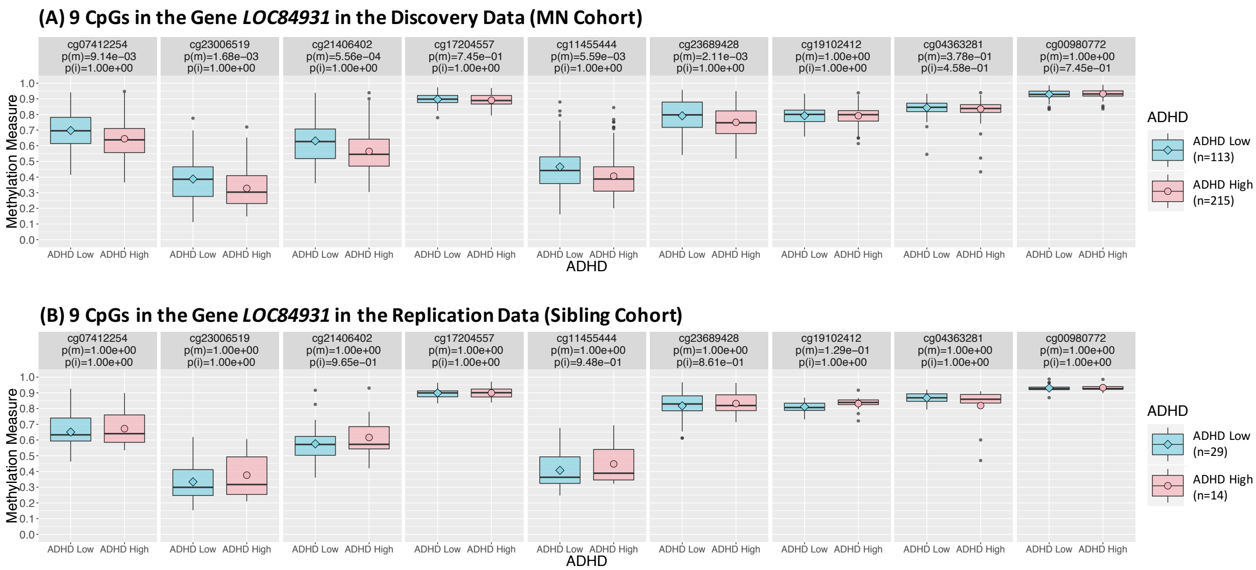

Figure S3. Boxplot of DNA methylation measures of the 9 CpGs in gene *LOC84931* stratified by ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *LOC84931*) P-values testing $\beta_{M1} = 0$ in the logistic model: $\text{logit}P(Y = 1) = \beta_{M0} + \beta_{M1}\text{CpG}$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2 E + \beta_3\text{CpG} \times E$, respectively.

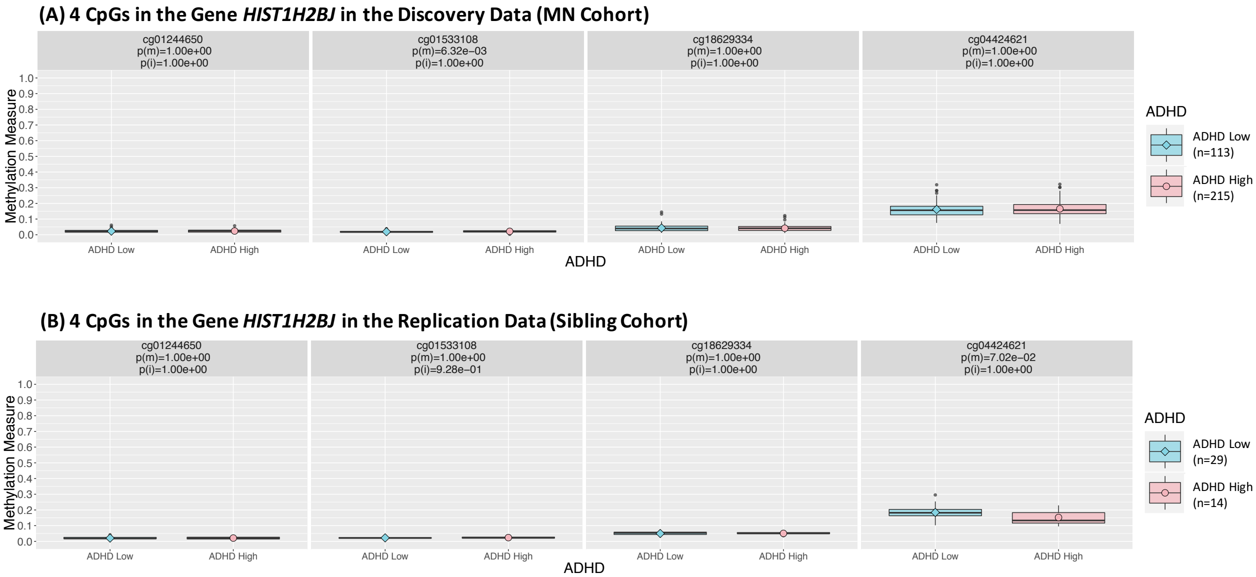

Figure S4. Boxplot of DNA methylation measures of the 4 CpGs in gene *HIST1H2BJ* stratified by ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *HIST1H2BJ*) P-values testing $\beta_{M1} = 0$ in the logistic model: $\text{logit}P(Y = 1) = \beta_{M0} + \beta_{M1}\text{CpG}$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2 E + \beta_3\text{CpG} \times E$, respectively.

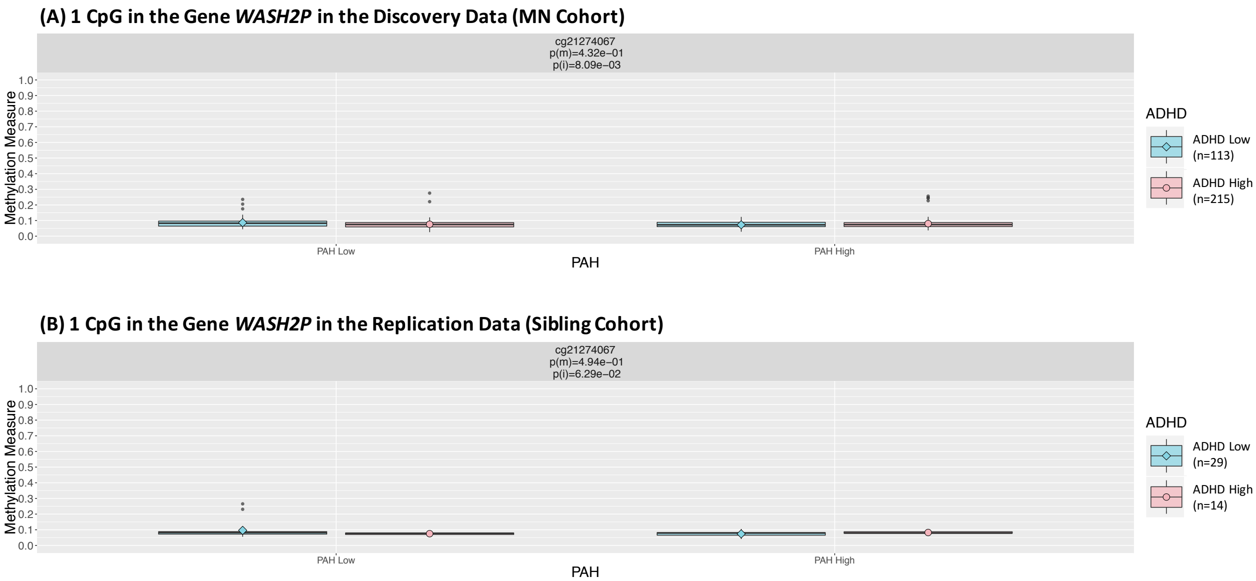

Figure S5. Boxplot of DNA methylation measures of the 1 CpG in gene *WASH2P* stratified by PAH and ADHD status in the (A) discovery analysis using the MN cohort, and the (B) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *WASH2P*) P-values testing $\beta_{M1} = 0$ in the logistic model: $\text{logit}P(Y = 1) = \beta_{M0} + \beta_{M1}\text{CpG}$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2 E + \beta_3\text{CpG} \times E$, respectively.

2.2.2 Results from the comparison methods

Table S3. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 29 genes by $\mathbf{D}^{\text{w-M}}$ at the 0.005 gene-level P-value threshold

Rank in $\mathbf{D}^{\text{w-M}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w-M-E-int}}$
1	*LOC84931**	9	1
2	*SERPINB3**	1	2
3	*HIST1H2BJ* *	4	11
4	*IGJ**	1	7
5	*ADAM32**	11	8
6	*TRIM38*	7	28
7	*NDUFA5**	9	17
8	*KRTAP20-1**	1	6
9	*CXCL9**	1	10
10	*BICD1**	14	16
11	*SPDYC*	9	20
12	*RNF187*	12	33
13	*LOC284578*	3	24
14	*PLA2G4D*	14	18
15	*IL7R*	2	35
16	*PLOD2*	11	42
17	*SPACA1**	6	12
18	*ASZ1*	9	37
19	*TLK2*	4	49
20	*FSCB*	2	31
21	*XPO5*	1	22
22	*LYRM1**	3	13
23	*KIAA0776*	8	34
24	*TBCK*	8	55
25	*CDH20*	7	19
26	*LOC100271836*	1	47
27	*IFNG*	7	36
28	*REP15*	1	30
29	*MIR548I2*	2	78

*Genes also identifed by $\mathbf{D}^{\text{w-M-E-int}}$.

Table S4. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 16 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 $\mathbf{D}^{\text{w-M-E-int}}$**
1	*CYP2E1**	13	3
2	*MIR518E**	1	5
3	*KIR3DP1* *	1	4
4	*GBAP1*	6	32
5	*MAS1**	2	15
6	*ARHGEF15*	9	90
7	*LRIT2*	7	29
8	*OR8G1**	1	9
9	*WASH2P**	1	14
10	*OR2AE1*	3	43
11	*OR2T27*	1	26
12	*HNMT*	5	39
13	*TNFRSF10B*	11	48
14	*MIR604*	2	21
15	*C1orf190*	6	38
16	*PTER*	10	83

*Genes also identifed by $\mathbf{D}^{\text{w-M-E-int}}$.

Table S5. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 6 genes by $\mathbf{D}^{\text{M-E-int}}$ at the 0.005 gene-level P-value threshold

Rank in $\mathbf{D}^{\text{M-E-int}}$	**Gene**	**# CpG**	**Rank in $\mathbf{D}^{\text{w-M-E-int}}$**
1	*LOC84931**	9	1
2	*CYP2E1**	13	3
3	*MIR518E**	1	5
4	*SERPINB3**	1	2
5	*SPACA1**	6	12
6	*IGJ**	1	7

*Genes also identifed by $\mathbf{D}^{\text{w-M-E-int}}$.

Table S6. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 29 genes by $\mathbf{D}^{\mathrm{M}}$ at the 0.005 gene-level P-value threshold

Rank in $\mathbf{D}^{\mathbf{M}}$	**Gene**	**# CpG**	**Rank in $\mathbf{D}^{\textbf{w-M-E-int}}$**
1	*SERPINB3**	1	2
2	*IGJ**	1	7
3	*LOC84931**	9	1
4	*KRTAP20-1**	1	6
5	*CXCL9**	1	10
6	*XPO5*	1	22
7	*LOC100271836*	1	47
8	*REP15*	1	30
9	*SPACA1**	6	12
10	*SNORD113-5*	1	25
11	*DEFB110*	1	53
12	*KATNA1*	1	23
13	*USP18*	1	54
14	*SLCO1B1*	1	44
15	*KRTAP21-3*	1	45
16	*RUNDC1*	1	60
17	*tAKR*	1	50
18	*FSCB*	2	31
19	*CRISP2*	10	40
20	*PDHX*	1	124
21	*CXADR*	2	71
22	*SNAP29*	1	133
23	*GSTA2*	1	73
24	*TNFSF18*	1	226
25	*KRTAP7-1*	1	122
26	*RBM46*	12	67
27	*KRIT1*	1	178
28	*POLR3G*	1	264
29	*CDH20*	7	19

*Genes also identifed by $\mathbf{D}^{\text{w-M-E-int}}$.

Table S7. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 12 genes by $\mathbf{D}^{\text{int}}$ at the 0.005 gene-level P-value threshold

Rank in $\mathbf{D}^{\text{int}}$	**Gene**	**# CpG**	**Rank in $\mathbf{D}^{\text{w-M-E-int}}$**
1	*MIR518E**	1	5
2	*KIR3DP1**	1	4
3	*CYP2E1**	13	3
4	*OR8G1**	1	9
5	*WASH2P**	1	14
6	*OR2T27*	1	26
7	*SPRYD5*	1	27
8	*UCHL5*	1	52
9	*GK3P*	1	77
10	*MAS1**	2	15
11	*TAS2R3*	1	101
12	*OR14C36*	1	297

*Genes also identifed by $\mathbf{D}^{\text{w-M-E-int}}$.

Table S8. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 6 genes by L^S at the 0.005 gene-level P-value threshold

Rank in L^S	**Gene**	**# CpG**	**Rank in $\mathbf{D}^{\text{w-M-E-int}}$**
1	*LOC84931**	9	1
2	*ADAM32**	11	8
3	*SERPINB3**	1	2
4	*TRIM38*	7	28
5	*IGJ**	1	7
6	*NDUFA5**	9	17

*Genes also identifed by $\mathbf{D}^{\text{w-M-E-int}}$.

Table S9. 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 $\mathbf{D}^{\text{w-M-E-int}}$**
1	*UBASH3B*	23	92
2	*MYH2*	8	41
3	*JARID2*	84	644
4	*TNFRSF10B*	11	48

The seven comparison methods $\mathbf{D}^{\text{w-M}}$, $\mathbf{D}^{\text{w-int}}$, $\mathbf{D}^{\text{M-E-int}}$, $\mathbf{D}^{\text{M}}$, $\mathbf{D}^{\text{int}}$, L^S and L^M identified 29, 16, 6, 29, 12, 6, 4 genes and replicated 3, 2, 2, 1, 3, 0, 0 genes (replication rate ranges 0-33% with a mean of 12%). These results are summarized in Supplementary Table S10. The 2 genes, *LOC84931* and *HIST1H2BJ*, replicated by $\mathbf{D}^{\text{w-M}}$ were also identified and replicated by the proposed method due to main signals. The other 2 genes, *CYP2E1* and *WASH2P*, replicated by both $\mathbf{D}^{\text{w-int}}$ and $\mathbf{D}^{\text{int}}$ were also identified and replicated by the proposed method due to interaction signals. In addition, the 2 genes *LOC84931* and *CYP2E1* were both replicated by $\mathbf{D}^{\text{M-E-int}}$ and the proposed method due to main/interaction signals. The other 3 genes, *SPDYC*, *REP15* and *UCHL5*, replicated by $\mathbf{D}^{\text{w-M}}$, $\mathbf{D}^{\text{M}}$ and $\mathbf{D}^{\text{int}}$, respectively, were ranked #20, #30, and #52 (P-value=0.0059, 0.0068 and 0.0122) in the proposed method $\mathbf{D}^{\text{w-M-E-int}}$ results in the discovery analysis. In general, the genes replicated by the comparison methods were either all replicated or ranked on top in the $\mathbf{D}^{\text{w-M-E-int}}$ results. This suggests that the proposed method that incorporates both main and interaction signals indeed has better performance.

Table S10. 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**
$\mathbf{D}^{\text{w-M-E-int}}$	4/17	*LOC84931, CYP2E1, HIST1H2BJ, WASH2P*
$\mathbf{D}^{\text{w-M}}$	3/29	*LOC84931, HIST1H2BJ, SPDYC*
$\mathbf{D}^{\text{w-int}}$	2/16	*CYP2E1, WASH2P*
$\mathbf{D}^{\text{M-E-int}}$	2/6	*LOC84931, CYP2E1*
$\mathbf{D}^{\text{M}}$	1/29	*REP15*
$\mathbf{D}^{\text{int}}$	3/12	*CYP2E1, WASH2P, UCHL5*
L^S	0/6	-
L^M	0/4	-

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

Table S11. 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**
$\mathbf{D}^{\text{w-M-E-int}}$	4/7	*FAM35A, DIRC1, THSD1P, C8orf80*
$\mathbf{D}^{\text{w-M}}$	5/22	*FAM35A, DIRC1, THSD1P, CPA2, LOC407835*
$\mathbf{D}^{\text{w-int}}$	1/6	*KCTD19*
$\mathbf{D}^{\text{M-E-int}}$	0/2	-
$\mathbf{D}^{\text{M}}$	4/18	*MIR519B, C9orf122, KRTAP20-2, VN1R4*
$\mathbf{D}^{\text{int}}$	3/12	*GSTA1, OR4P4, FAM166B*
L^S	1/4	*THSD1P*
L^M	0/3	-

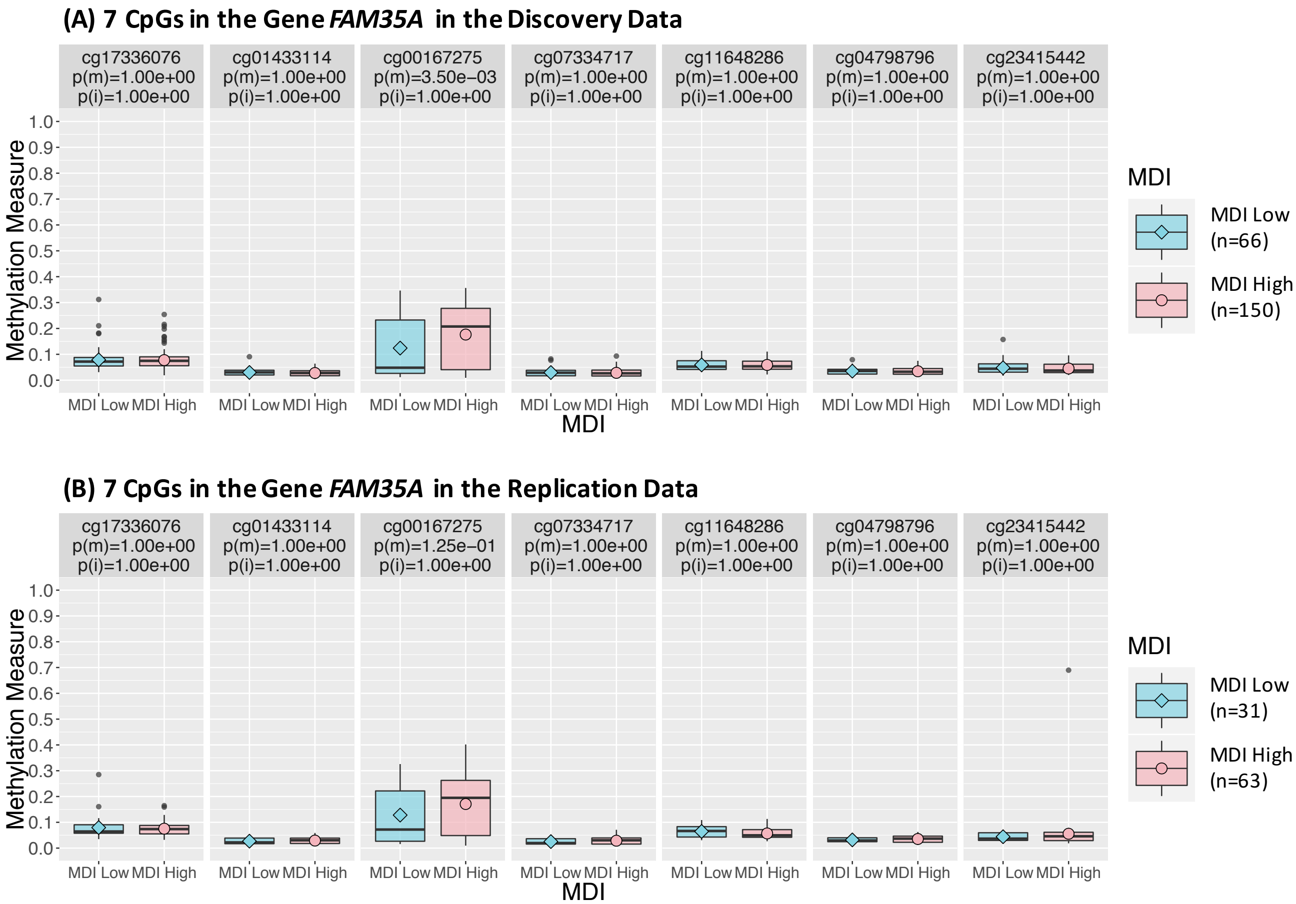

Figure S6. Boxplot of DNA methylation measures of the 7 CpGs in gene *FAM35A* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *FAM35A*) P-values testing $\beta_{M1} = 0$ in the logistic model: $\text{logit}P(Y = 1) = \beta_{M0} + \beta_{M1}\text{CpG}$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2 E + \beta_3\text{CpG} \times E$, respectively.

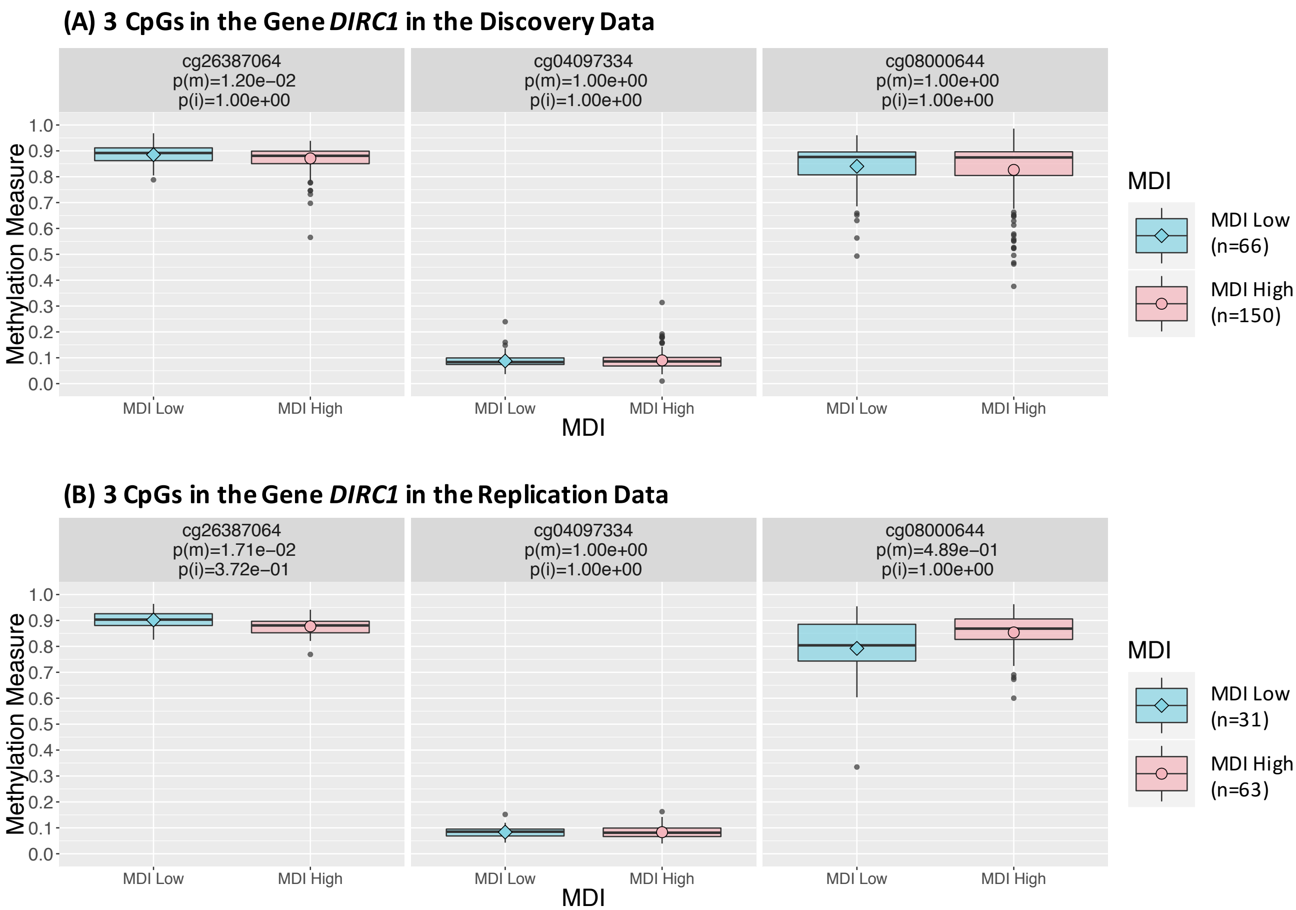

Figure S7. Boxplot of DNA methylation measures of the 3 CpGs in gene *DIRC1* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *DIRC1*) P-values testing $\beta_{M1} = 0$ in the logistic model: $\text{logit}P(Y = 1) = \beta_{M0} + \beta_{M1}\text{CpG}$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y = 1) = \beta_0 + \beta_1\text{CpG} + \beta_2 E + \beta_3\text{CpG} \times E$, respectively.

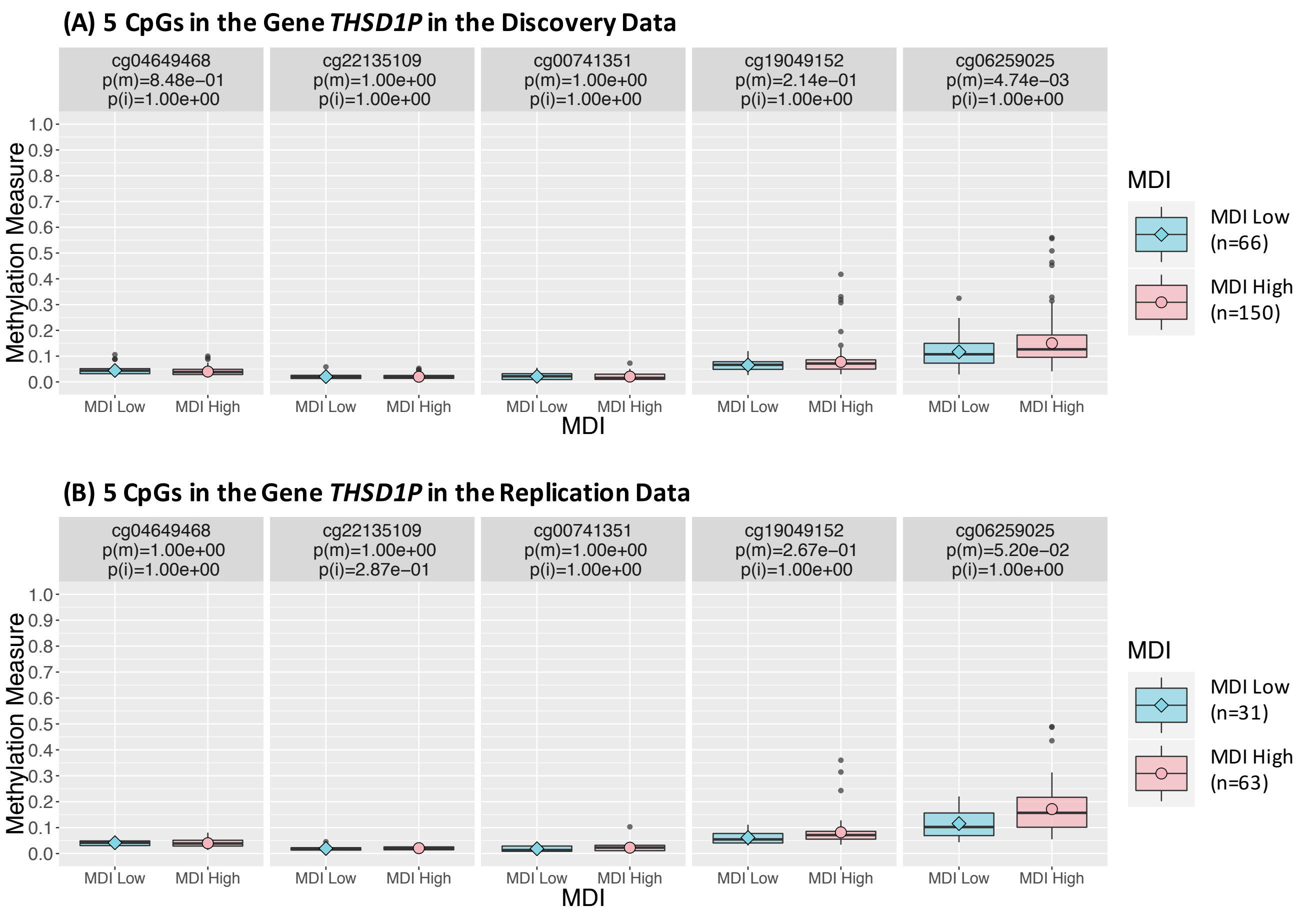

Figure S8. Boxplot of DNA methylation measures of the 5 CpGs in gene *THSD1P* stratified by MDI status in the (A) discovery analysis using the 2/3 MN discovery data, and the (B) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *THSD1P*) P-values testing $\beta_{M1} = 0$ in the logistic model: $\text{logit}P(Y=1) = \beta_{M0} + \beta_{M1}\text{CpG}$ and $\beta_3 = 0$ in the multiple logistic model: $\text{logit}P(Y=1) = \beta_0 + \beta_1\text{CpG} + \beta_2 E + \beta_3\text{CpG} \times E$, respectively.

References

[1] Hansen K. IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for illumina's 450k methylation arrays. *R package, version 0.2*, 1, 2015.

[2] Pidsley R, Wong CC, Volta M, et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, 14(1):293, 2013.

[3] Horvath S. DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):3156, 2013.

[4] Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.