Bioinformatics, 36(3), 2020, 653–659
doi: 10.1093/bioinformatics/btz630
Advance Access Publication Date: 22 August 2019
Original Paper

Genome analysis

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 and [2]Columbia Center for Children's Environmental Health, Department of Environmental Health Science, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 3, 2019; revised on July 18, 2019; editorial decision on August 1, 2019; accepted on August 19, 2019

Abstract

Motivation: Deoxyribonucleic acid (DNA) methylation plays a crucial role in human health. Studies have 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 differential methylation 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. Here we introduce the idea of a pseudo-data matrix constructed with cross-product terms between CpGs and environmental factors that are able to capture their interactions. We then develop a powerful and flexible weighted distance-based method with the pseudo-data matrix where association strength was used as weights on CpGs, environmental factors and their interactions to up-weight signals and down-weight noises in distance calculations.

Results: We compared the power of this novel approach and several comparison methods in simulated datasets and the Mothers and Newborns birth cohort of the Columbia Center for Children's Environmental Health to determine whether prenatal polycyclic aromatic hydrocarbons interacts with DNA methylation in association with Attention Deficit Hyperactivity Disorder and Mental Development Index at age 3.

Availability and implementation: An R code for the proposed method $\mathrm{D}^{\mathrm{w-M-E-int}}$ together with a tutorial and a sample dataset is available for downloading from http://www.columbia.edu/~sw2206/softwares.htm.

Contact: sw2206@columbia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Deoxyribonucleic acid (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 (Bakulski *et al.*, 2015; Cardenas *et al.*, 2017; Faulk *et al.*, 2015; Herbstman *et al.*, 2012; Janssen *et al.*, 2013; Nahar *et al.*, 2015; Nye *et al.*, 2016; Perera *et al.*, 2009; Saenen *et al.*, 2017; Sen *et al.*, 2015) 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.* (2012) found that DNA methylation modifies the effect of NO2 on the progression from mild to severe asthma; White *et al.* (2015) found that DNA methylation modifies the risk of PAH–DNA adducts on breast cancer. Despite these findings, due to high dimensionality and low study power, current studies usually focus on finding differential methylation 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.

2 Materials and methods

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 $\mathbf{X}^{\mathrm{M-E}} = [\mathbf{X}^{M}, \mathbf{E}]$, a $2N \times (n+1)$ matrix for main signals of n CpGs and one environmental factor. We normalize each column of $\mathbf{X}^{\mathrm{M-E}}$ to have mean zero and unit standard deviation (SD). The element $x_{ij}^{\mathrm{M-E}}$ harbors the normalized methylation measure of CpG j for subject i, $j = 1, \ldots, n$, and normalized environmental factor E_i of subject i, $j = n+1$, $i = 1, \ldots, 2N$. We then define $\mathbf{X}^{\mathrm{int}}$, a $2N \times n$ pseudo-data matrix with element $x_{ij}^{\mathrm{int}} = x_{ij}^{\mathrm{M}} \times E_i$ harbors the interaction between CpG j and the environmental factor of subject i, $j = 1, \ldots, n$ and $i = 1, \ldots, 2N$. By using $\mathbf{X}^{\mathrm{M-E-int}} = [\mathbf{X}^{\mathrm{M-E}}, \mathbf{X}^{\mathrm{int}}]$, a $2N \times (2n+1)$ pseudo-data matrix, we capture main signals of n CpGs, one environmental factor and n pairwise CpG × E interactions. Here, the proposed method $\mathbf{D}^{\mathrm{w-M-E-int}}$ that tests the null hypothesis that there is no joint effect of methylation, the environmental factor and their interactions on the outcome. The proposed method is very flexible and can be easily adapted to test other hypotheses based on different pseudo-data matrices. Specifically, we are able to test (i) the association between methylation and the outcome by constructing the distance matrix based on the pseudo-data matrix $\mathbf{X}^{\mathrm{M}}$, (ii) the association between interactions and the outcome based on $\mathbf{X}^{\mathrm{int}}$, (iii) the association between the environmental factor and the outcome based on $\mathbf{X}^{\mathrm{E}}$, (iv) the joint effect of methylation and the environmental factor based on $\mathbf{X}^{\mathrm{M-E}}$ and (v) the joint effect of methylation and interaction based on $\mathbf{X}^{\mathrm{M-int}}$.

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

$$d_{st}^{\mathrm{M-E-int}} = \sqrt{\Delta_E^2 + \sum_{j=1}^{n} (\Delta_{\mathrm{M,\ j}}^2 + \Delta_{\mathrm{int,\ j}}^2)} \tag{1}$$

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

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:

$$w_j^{\mathrm{M}} = \frac{-\log_{10}(p_j^{\mathrm{M}})}{-\log_{10}(p_{\mathrm{E}}) + \sum_{j=1}^{n} -\log_{10}(p_j^{\mathrm{M}}) + \sum_{j=1}^{n} -\log_{10}(p_j^{\mathrm{int}})}$$

$$w_j^{\mathrm{int}} = \frac{-\log_{10}(p_j^{\mathrm{int}})}{-\log_{10}(p_{\mathrm{E}}) + \sum_{j=1}^{n} -\log_{10}(p_j^{\mathrm{M}}) + \sum_{j=1}^{n} -\log_{10}(p_j^{\mathrm{int}})} \tag{2}$$

$$w_{\mathrm{E}} = \frac{-\log_{10}(p_{\mathrm{E}})}{-\log_{10}(p_{\mathrm{E}}) + \sum_{j=1}^{n} -\log_{10}(p_j^{\mathrm{M}}) + \sum_{j=1}^{n} -\log_{10}(p_j^{\mathrm{int}})}$$

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

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

$$d_{st}^{\mathrm{w-M-E-int}} = \sqrt{w_{\mathrm{E}} \Delta_E^2 + \sum_{j=1}^{n} (w_j^{\mathrm{M}} \Delta_{\mathrm{M,\ j}}^2 + w_j^{\mathrm{int}} \Delta_{\mathrm{int,\ j}}^2)}. \tag{3}$$

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 $\mathbf{D}^{\mathrm{w-M-E-int}}$ introduced in Equation (3)

$$F = \frac{tr(HGH)}{tr[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})]} \tag{4}$$

where $\mathbf{H} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ is a $2N \times 2N$ projection matrix, Y is a $2N \times 1$ vector with case $(Y=1)$ and control $(Y=0)$ status, $\mathbf{G} = (\mathbf{I} - \frac{1}{2N} 11^T)\mathbf{A}(\mathbf{I} - \frac{1}{2N} 11^T)$ is the Gower's centered matrix, $\mathbf{A} = (a_{st}) = (-\frac{1}{2}(d_{st}^{\mathrm{w-M-E-int}})^2)$, 1 is a $2N$-dimensional column vector with elements 1, and $\mathbf{I}$ is a $2N \times 2N$ identity matrix. In the univariate analysis (e.g. only one CpG without E and their interaction), when the distance matrix is calculated using the standard Euclidean distance, the pseudo-F statistic becomes the standard F-statistic in ANOVA (Zapala and Schork, 2006).

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, $g = 1, \ldots, G$ as:

$$P_g = \frac{\sum_{g'=1}^{G}\{I(F_{g',\text{obs}} \geq F_{g,\text{obs}}) + \sum_{\text{perm}=1}^{999} I(F_{g',\text{perm}} \geq F_{g,\text{obs}})\}}{G \times (1 + 999)}. \quad (5)$$

In the real data applications, we have $G = 18\ 633$ 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 1.1 and Table S1).

2.2 Comparison methods

We compare the performance of the proposed method $\mathbf{D}^{\text{w-M-E-int}}$ that considers both main (M and E) and interaction signals with weights to that of several comparison methods, including the weighted distance-based methods considering (i) methylation signals only $\mathbf{D}^{\text{w-M}}$, (ii) interaction signals only $\mathbf{D}^{\text{w-int}}$, (iii) the distance-based methods without weights considering both main (M and E) and interaction signals $\mathbf{D}^{\text{M-E-int}}$, (iv) methylation signals only $\mathbf{D}^{\text{M}}$, (v) interaction signals only $\mathbf{D}^{\text{int}}$ and (vi) the site-level EWAS methods via logistic regressions on each CpG considering methylation signals only L^S or (vii) both main (M and E) 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 significant methylation 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. 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.

Note that we can also consider $\mathbf{D}^{\text{w-M-E}}$, $\mathbf{D}^{\text{w-M-int}}$, $\mathbf{D}^{\text{M-E}}$, $\mathbf{D}^{\text{M-int}}$ and $\mathbf{D}^{\text{E}}$ models, which we included in Supplementary Materials to show the flexibility of the proposed method.

3 Simulation studies

We conducted simulation studies to evaluate Type I error and power of the proposed method $\mathbf{D}^{\text{w-M-E-int}}$ and the comparison methods where we only considered one gene with multiple CpGs for illustration purpose. Type I error 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 1000 simulations in each simulation setting.

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

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

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ determines means, $\Delta = \text{diag}(\sqrt{\sigma_1}, \ldots, \sqrt{\sigma_n})$ determines variances 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

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p(\mathbf{X}_i, E_i, \mathbf{Z}_i)) \\ p(\mathbf{X}_i, E_i, \mathbf{Z}_i) &= \frac{\exp(\boldsymbol{\beta}_{\mathbf{X}}^T\mathbf{X}_i + \beta_E E_i + \boldsymbol{\beta}_{\mathbf{Z}}^T\mathbf{Z}_i)}{1 + \exp(\boldsymbol{\beta}_{\mathbf{X}}^T\mathbf{X}_i + \beta_E E_i + \boldsymbol{\beta}_{\mathbf{Z}}^T\mathbf{Z}_i)} \end{aligned} \quad (6)$$

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

In each simulation, we set $\mu_j \sim N(-0.47, 3.56)$, $j = 1, \ldots, n$, 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 $\sigma_j \sim N(0.62, 0.21)$, $j = 1, \ldots, n$, 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 $\beta = 0$.

3.1.1 Simulation settings with different types of signals

We set a gene with 30 CpGs with 1∼4 CpGs having (i) methylation main signals only, (ii) interaction signals only and (iii) both methylation main and interaction signals. Detailed simulation setups are in Table 1.

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

A signal item represents a methylation signal in the data matrix $\mathbf{X}^{\text{M-E-int}}$ 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 S2.

3.2 Simulation results

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 2).

3.2.2 Simulation settings with different types of signals

As summarized in Figure 1, when there are only main signals, $\mathbf{D}^{\text{w-int}}$ and $\mathbf{D}^{\text{int}}$ that only consider interaction signals have no power, as expected. $\mathbf{D}^{\text{w-M-E-int}}$ is slightly less powerful than $\mathbf{D}^{\text{w-M}}$ 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. $\mathbf{D}^{\text{M-E-int}}$ performs similarly as L^M, while both of them perform inferior to $\mathbf{D}^{\text{w-M-E-int}}$ with weights. In general, the weighted versions $\mathbf{D}^{\text{w-M-E-int}}$ and $\mathbf{D}^{\text{w-M}}$ 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, $\mathbf{D}^{\text{w-M}}$, $\mathbf{D}^{\text{M}}$ and L^S that only consider main signals have no power, as expected. $\mathbf{D}^{\text{w-M-E-int}}$ is slightly less powerful than $\mathbf{D}^{\text{w-int}}$ when both of them outperform the corresponding non-weighted versions. $\mathbf{D}^{\text{M-E-int}}$ 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. 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 $\mathbf{D}^{\text{w-M}}$, $\mathbf{D}^{\text{M}}$ and L^S that only consider main signals increases, while that of $\mathbf{D}^{\text{w-int}}$ and $\mathbf{D}^{\text{int}}$ that only consider interaction signals decreases, and that of $\mathbf{D}^{\text{w-M-E-int}}$, $\mathbf{D}^{\text{M-E-int}}$ and L^M

Table 1. Simulation settings with different types of signals

Scenario	Number of signal items[a]	Simulation setup[b,c]
Main signals only	1 signal CpG	$\beta_{X_1} = 0.35$
	2 signal CpGs	$\beta_{X_1} = \beta_{X_3} = 0.35$
	3 signal CpGs	$\beta_{X_1} = \beta_{X_3} = \beta_{X_5} = 0.35$
	4 signal CpGs	$\beta_{X_1} = \beta_{X_3} = \beta_{X_5} = \beta_{X_7} = 0.35$
Interaction signals only	1 signal CpG	$\beta_{Z_1} = 0.35$
	2 signal CpGs	$\beta_{Z_1} = \beta_{Z_3} = 0.35$
	3 signal CpGs	$\beta_{Z_1} = \beta_{Z_3} = \beta_{Z_5} = 0.35$
	4 signal CpGs	$\beta_{Z_1} = \beta_{Z_3} = \beta_{Z_5} = \beta_{Z_7} = 0.35$
Both main and interaction signals with fixed number of signal CpGs	3 signal CpGs with interaction signals 1 signal CpG with main signals (main-to-interaction signal ratio = 1:3)	$\beta_{X_1} = \beta_{Z_3} = \beta_{Z_5} = \beta_{Z_7} = 0.35$
	2 signal CpGs with interaction signals 2 signal CpGs with main signals (main-to-interaction signal ratio = 2:2)	$\beta_{X_1} = \beta_{X_3} = \beta_{Z_5} = \beta_{Z_7} = 0.35$
	1 signal CpG with interaction signals 3 signal CpGs with main signals (main-to-interaction signal ratio = 3:1)	$\beta_{X_1} = \beta_{X_3} = \beta_{X_5} = \beta_{Z_7} = 0.35$

[a] A signal item represents a methylation signal in the data matrix $\mathbf{X}^{\mathrm{M-E-int}}$ 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.

[c] In each model, we also set $\beta_E = 0.1$.

Table 2. Type I error rates

Methods	20 CpGs[a]	30 CpGs	40 CpGs
$\mathbf{D}^{\mathrm{w-M-E-int}}$	0.037	0.051	0.055
$\mathbf{D}^{\mathrm{w-M}}$	0.051	0.053	0.052
$\mathbf{D}^{\mathrm{w-int}}$	0.044	0.045	0.047
$\mathbf{D}^{\mathrm{M-E-int}}$	0.048	0.048	0.045
$\mathbf{D}^{\mathrm{M}}$	0.036	0.053	0.047
$\mathbf{D}^{\mathrm{int}}$	0.047	0.048	0.049
L^S	0.036	0.041	0.029
L^M	0.037	0.039	0.035

[a] Number of CpGs in a gene.

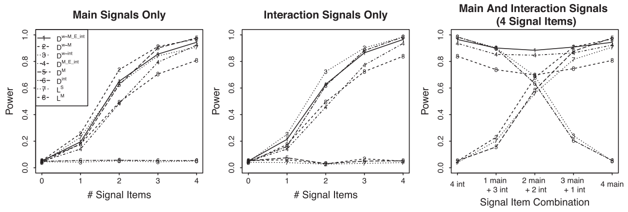

Fig. 1. Power results for simulation settings with methylation signals only, interaction signals only and both methylation and interaction signals when there are 30 CpGs in a gene

that consider both main and interaction signals remains the same. Importantly, $\mathbf{D}^{\mathrm{w-M-E-int}}$ consistently has the largest power, which implies that the performance of $\mathbf{D}^{\mathrm{w-M-E-int}}$ is not affected by signal types. Again, the weighted versions outperform the non-weighted versions.

The power of other possible comparison methods as well as those of the simulation with 20 or 40 CpGs in a gene was summarized in Supplementary Figure S1. 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.*, 2019).

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 Material Section 1.3 and Figure S2. Overall, the power of distance-based methods increases as the number of signal CpGs increases.

4 Real data applications

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 MN prospective cohort study. During the third trimester of pregnancy, women were asked to wear a small backpack containing a personal monitor during the daytime for 48 h. The collected samples were then analyzed for eight carcinogenic PAHs (Perera *et al.*, 2003). The PAH metric used in the analysis is the sum of eight carcinogenic PAHs and was dichotomized at the median in the parent population (2.26 ng/m^3). 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.

4.2 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 DSM-IV-oriented ADHD (American Psychiatric Association, 2013) and (ii) the Bayley Scales of Infant Development 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.

4.3 DNA methylation data processing

We conducted standard data processing steps for DNA methylation with details in Supplementary Material Section 2.1.

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.

4.4.1 Discovery analysis in the MN cohort

Since the main purpose is to demonstrate the power of the proposed method $\mathbf{D}^{\text{w}-\text{M}-\text{E}-\text{int}}$ 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, $\mathbf{D}^{\text{w}-\text{M}-\text{E}-\text{int}}$ identified 17 genes in the discovery analysis, with 11 due to main signals only and 6 due to interaction signals only (Table 3).

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 17 genes identified in the discovery MN cohort, 4 (*LOC84931*, *CYP2E1*, *HIST1H2BJ* and *WASH2P*) were replicated in the replication Sibling cohort. In both discovery and replication analyses, genes *CYP2E1* and *WASH2P* were identified due to interaction signals, and genes *LOC84931* and *HIST1H2BJ* were identified due to main signals.

Table 3. Application examining prenatal PAH, DNA methylation and their interactions on child ADHD at age 3 identified 17 genes by the proposed method $\mathbf{D}^{\text{w}-\text{M}-\text{E}-\text{int}}$ at the 0.005 gene-level *P*-value threshold

Rank in $\mathbf{D}^{\text{w}-\text{M}-\text{E}-\text{int}}$	Gene	# CpG	Rank in $\mathbf{D}^{\text{w}-\text{M}}$	Rank in $\mathbf{D}^{\text{w}-\text{int}}$
1	*LOC84931*[a]	9	1	1513
2	*SERPINB3*	1	2	18 316
3	*CYP2E1*[a]	13	6013	1
4	*KIR3DP1*	1	18 535	3
5	*MIR518E*	1	13 908	2
6	*KRTAP20-1*	1	8	18 472
7	*IGJ*	1	4	18 286
8	*ADAM32*	11	5	15 841
9	*OR8G1*	1	17 560	8
10	*CXCL9*	1	9	16 665
11	*HIST1H2BJ*[a]	4	3	14 178
12	*SPACA1*	6	17	17 866
13	*LYRM1*	3	22	12 866
14	*WASH2P*[a]	1	15 711	9
15	*MAS1*	2	4067	5
16	*BICD1*	14	10	3485
17	*NDUFA5*	9	7	9318

[a] Genes replicated in the replication analysis.

Figure 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 third trimester maternal depressed mood 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*, *HIST1H2BJ* and *WASH2P* (Supplementary Figs S3–S5).

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 S3–S9), when all these genes rank within top 4% of the proposed method results. The comparison methods have replication rates 0–33% with an average 12% (Supplementary Table S10). Detailed results are in Supplementary Material Section 2.2.

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.

4.5.1 Discovery analysis in the discovery data

At the same 0.005 *P*-value threshold, the proposed method $\mathbf{D}^{\text{w}-\text{M}-\text{E}-\text{int}}$ identified seven genes in the discovery analysis, with five due to main signals only and two due to both main and interaction signals (Table 4).

4.5.2 Replication analysis in the replication data

At the same 0.1 gene-level *P*-value threshold for replication, three genes, *FAM35A*, *DIRC1* and *THSD1P*, were replicated in the replication analysis due to main signals out of the five genes identified in the discovery analysis due to main signals only. Gene *C8orf80* was replicated due to interaction signals, out of the two genes identified in the discovery analysis due to both main and interaction signals. That is, the replication rate is 57% with four out of seven genes replicated. Figure 3 plots boxplots of DNA methylation measures of the four CpGs in gene *C8orf80* stratified by PAH and MDI status that was identified due to both main and interaction signals and

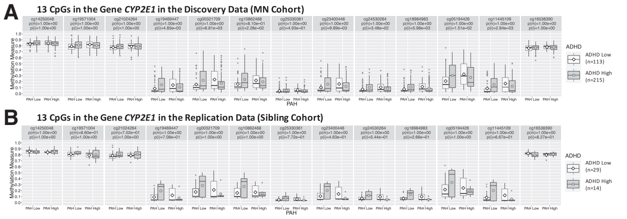

Fig. 2. Boxplot of DNA methylation measures of the 13 CpGs in gene *CYP2E1* stratified by PAH and ADHD status in the (**A**) discovery analysis using the MN cohort, and the (**B**) replication analysis using the Sibling cohort. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *CYP2E1*) *P*-values testing $\beta_{M1} = 0$ 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

Table 4. Application examining prenatal PAH, DNA methylation and their interactions on child MDI at age 3 identified seven genes by the proposed method at the 0.005 gene-level *P*-value threshold

Rank in $\mathbf{D}^{\mathrm{w-M-E-int}}$	Gene	# CpG	Rank in $\mathbf{D}^{\mathrm{w-M}}$	Rank in $\mathbf{D}^{\mathrm{w-int}}$
1	*UROS*	2	2	18 516
2	*FAM35A*[a]	7	1	15 325
3	*DIRC1*[a]	3	5	17 815
4	*THSD1P*[a]	5	7	15 099
5	*C19orf77*	9	6	647
6	*MIR521-1*	1	9	18 302
7	*C8orf80*[a]	4	3	2329

[a] Genes replicated in the replication analysis.

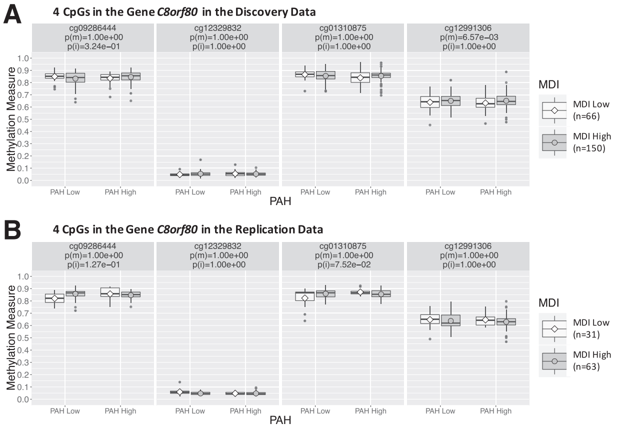

Fig. 3. Boxplots of DNA methylation measures of the four CpGs in gene *C8orf80* stratified by PAH and MDI status in the (**A**) discovery analysis using the 2/3 MN discovery data, and the (**B**) replication analysis using the 1/3 MN replication data. Here $p(m)$ and $p(i)$ are Bonferroni-adjusted (for number of CpGs in gene *C8orf80*) *P*-values testing $\beta_{M1} = 0$ in the logistic model $\mathrm{logit}P(Y = 1) = \beta_{M0} + \beta_{M1}\mathrm{CpG}$ and $\beta_3 = 0$ in the multiple logistic model $\mathrm{logit}P(Y = 1) = \beta_0 + \beta_1\mathrm{CpG} + \beta_2 E + \beta_3\mathrm{CpG} \times E$, respectively

replicated due to interaction signals. We similarly plotted genes *FAM35A*, *DIRC1* and *THSD1P* (Supplementary Figs S6–S8).

4.5.3 Results of the comparison methods

All genes identified by the comparison methods rank within top 3% of the proposed method results. The comparison methods have replication rates 0–25% with an average 16% (Supplementary Table S11 and details in Supplementary Material Section 2.3).

5 Discussion

We developed a novel weighted distance-based method $\mathbf{D}^{\mathrm{w-M-E-int}}$ 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 $\mathbf{D}^{\mathrm{w-M-E-int}}$ 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 $\mathbf{D}^{\mathrm{w-M-E-int}}$ 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, $\mathbf{D}^{\mathrm{w-M-E-int}}$ identified 17 genes in the discovery data with 4 replicated in the replication data, while the comparison methods have an average replication rate 12%. In another application on child MDI at age 3, $\mathbf{D}^{\mathrm{w-M-E-int}}$ identified seven genes in the discovery data with four replicated in the replication data, while the comparison methods have an average replication rate 16%.

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 out-number the dimension of CpGs in the gene. We need to take extra caution to balance between main or interaction signals, especially when assigning weights.

Funding

This work was supported by the National Institutes of Health [5UG3OD023290-02 to Y.W., J.B.H., F.P.P., S.W.].

Conflict of Interest: none declared.

References

Achenbach,T.M. and Rescorla,L.A. (2000) *Manual for the ASEBA Preschool Forms and Profiles*. Vol. **30**. University of Vermont, Research Center for Children, Youth, & Families, Burlington, VT.

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub, Arlington, VA.

Bakulski,K.M. *et al.* (2015) Prenatal mercury concentration is associated with changes in DNA methylation at TCEANC2 in newborns. *Int. J. Epidemiol.*, **44**, 1249–1262.

Bayley,N. (1993) *Bayley Scales of Infant Development: Manual*. Psychological Corporation, San Antonio, TX.

Cardenas,A. *et al.* (2017) Persistent DNA methylation changes associated with prenatal mercury exposure and cognitive performance during childhood. *Sci. Rep.*, **7**, 288.

Das,P.M. and Singal,R. (2004) DNA methylation and cancer. *J. Clin. Oncol.*, **22**, 4632–4642.

Du,P. *et al.* (2010) Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Ehrlich,M. (2002) DNA methylation in cancer: too much, but also too little. *Oncogene*, **21**, 5400–5413.

Esteller,M. and Herman,J.G. (2002) Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J. Pathol.*, **196**, 1–7.

Faulk,C. *et al.* (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. *Environ. Epigenet.*, **1**, dvv006.

Feinberg,A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **47**, 433–440.

Friedman,J. *et al.* (2001) *The Elements of Statistical Learning*. Vol. **1**. Springer Series in Statistics, New York, NY, USA.

Fu,A. *et al.* (2012) An environmental epigenetic study of ADRB 2 5'-UTR methylation and childhood asthma severity. *Clin. Exp. Allergy*, **42**, 1575–1581.

Gurnot,C. *et al.* (2015) Prenatal antidepressant exposure associated with CYP2E1 DNA methylation change in neonates. *Epigenetics*, **10**, 361–372.

Herbstman,J.B. *et al.* (2012) Prenatal exposure to polycyclic aromatic hydrocarbons, benzo [a] pyrene–DNA adducts, and genomic DNA methylation in cord blood. *Environ. Health Perspect.*, **120**, 733.

Jager,P.L.D. *et al.* (2014) Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.*, **17**, 1156–1163.

Janssen,B.G. *et al.* (2013) Placental DNA hypomethylation in association with particulate air pollution in early life. *Part. Fibre Toxicol.*, **10**, 22.

Kulis,M. and Esteller,M. (2010) DNA methylation and cancer. *Adv. Genet.*, **70**, 27–56.

Kumsta,R. *et al.* (2016) Severe psychosocial deprivation in early childhood is associated with increased DNA methylation across a region spanning the transcription start site of CYP2E1. *Transl. Psychiatry*, **6**, e830.

Mill,J. and Petronis,A. (2007) Molecular studies of major depressive disorder: the epigenetic perspective. *Mol. 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. *J. Child Psychol. Psychiatry*, **49**, 1020–1030.

Mill,J. *et al.* (2008) Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *Am. J. Hum. Genet.*, **82**, 696–711.

Nahar,M.S. *et al.* (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.

Nye,M.D. *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. *Environ. Epigenet.*, **2**, dvv009.

Perera,F. *et al.* (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**, e4488.

Perera,F.P. *et al.* (2003) Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. *Environ. Health Perspect.*, **111**, 201.

Perera,F.P. *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. *Environ. Health Perspect.*, **114**, 1287–1292.

Saenen,N.D. *et al.* (2017) Lower placental leptin promoter methylation in association with fine particulate matter air pollution during pregnancy and placental nitrosative stress at birth in the ENVIR ON AGE cohort. *Environ. Health Perspect.*, **125**, 262–268.

Schanen,N.C. (2006) Epigenetics of autism spectrum disorders. *Hum. Mol. Genet.*, **15**, R138–R150.

Sen,A. *et al.* (2015) Lead exposure induces changes in 5-hydroxymethylcytosine clusters in CpG islands in human embryonic stem cells and umbilical cord blood. *Epigenetics*, **10**, 607–621.

Wang,Y. *et al.* (2019) Detection of epigenetic field defects using a weighted epigenetic distance-based method. *Nucleic Acids Res.*, **47**, e66.

White,A.J. *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**, 1791–1802.

Zapala,M.A. and Schork,N.J. (2006) Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci.*, **103**, 19430–19435.