A region-based method for causal mediation analysis of DNA methylation data

Qi Yan[1*], Erick Forno[2], Juan C. Celedón[2] and Wei Chen[2,3]

[1]Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY

[2]Division of Pediatric Pulmonary Medicine, UPMC Children's Hospital of Pittsburgh, University of Pittsburgh, Pittsburgh, PA;

[3]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA;

[*]Correspondence should be addressed to:

Q.Y. (qy2253@cumc.columbia.edu)

Funding

This research was supported by grant HL138098 from the U.S. National Institutes of Health (NIH) to Q.Y. J.C.C. was supported by grants HL079966, HL117191, MD011764 and HL150431 from the U.S. NIH. W.C. was supported by grant HL150431 from the U.S. NIH.

ABSTRACT

Exposure to environmental factors can affect DNA methylation at a CpG site or a genomic region, which can then affect an outcome. In other words, environmental effects on an outcome could be mediated by DNA methylation. To date, single CpG site-based mediation analysis has been employed extensively. More recently, however, there has been considerable interest on studying differentially methylated regions (DMRs), both because DMRs are more likely to have functional effects than single CpG sites and because testing DMRs reduces multiple testing. In this report, we propose a novel causal mediation approach under the counterfactual framework to test the significance of total, direct and indirect effects of predictors on response variable with a methylated region (MR) as the mediator (denoted as MR-Mediation). Functional linear transformation is used to reduce the possible high dimension of the CpG sites in a predefined methylated region and to account for their location information. In our simulation studies, MR-Mediation retained the desired Type I error rates for total, direct and indirect effect tests, for both continuous and binary outcomes. Furthermore, MR-Mediation had better power performance than testing mean methylation level as the mediator in most considered scenarios, especially for indirect effect (i.e., mediated effect) test, which could be more interesting than the other two effect tests. We further illustrate our proposed method by analyzing the methylation mediated effect of exposure to gun violence on total immunoglobulin E (IgE) or atopic asthma among participants in the Epigenetic Variation and Childhood Asthma in Puerto Ricans (EVA-PR) study.

INTRODUCTION

Epigenetic studies are key to understanding regulatory mechanisms of gene expression. Of the available epigenetic markers, DNA methylation is the most stable and widely studied epigenetic modification.[1] DNA methylation refers the addition of a methyl group at the 5′ location of cytosine nucleotides (C), with this modification occurring predominantly at Cs that are immediately followed by a guanine (G) in the 5′ to 3′ direction, denoted CpG. DNA methylation often affects gene transcription and has been associated with complex diseases[2-5] and cancers.[6-10]

DNA methylation analysis has traditionally focused on individual CpG sites. Over the last few years, there has been great interest in developing region-based methods to detect differentially methylated regions (DMRs).[11, 12] Such region-based approach is supported by strong correlation in DNA methylation levels across regions of the genome,[13, 14] as well as by the fact that methylated regions such as CpG islands,[15] CpG island shores,[16] or generic 2-kb regions[17] are more often linked to functionally relevant findings than single CpG sites. Statistically, if there are multiple causative CpG sites that have small individual effects, single-marker testing may have limited power to detect those weak signals. On the other hand, region-based approaches have higher power by combining the effects of multiple CpG sites. Moreover, region-based methods greatly reduce the burden of multiple testing in a genome-wide study.

DMR identification methods could be classified into two categories: 1. Supervised methods that first calculate the p-values for the association between each single CpG site

and the outcome of interest, and then identify the regions with adjacent small p-values, and 2. Unsupervised methods that predefine the regions without using any outcome information, and then test the association between methylation level in the genomic regions and the outcome of interest.

Exposure to environmental factors can affect DNA methylation, which can then affect a disease or condition of interest. In other words, environmental effects on a disease or condition could be mediated by DNA methylation (Figure 1). Mediation analysis investigates the relationship between an exposure and an outcome, while examining how such exposure and outcome relate to the mediator. Although the Baron and Kenny's "four steps" method[18] has been the most commonly used approach to mediation analysis, such approach has been recently extended using a counterfactual framework[19-21] to allow the presence of an interaction between the exposure and mediator, as well as to decompose the estimate of total effect (TE) into estimates of direct (DE) and indirect effects (IE). To date, single CpG site-based mediation analysis has been employed extensively. This approach has indeed shown that DNA methylation can mediate the effects of environmental exposure on complex traits,[22-24] complex diseases[25-27] and cancers.[28, 29]

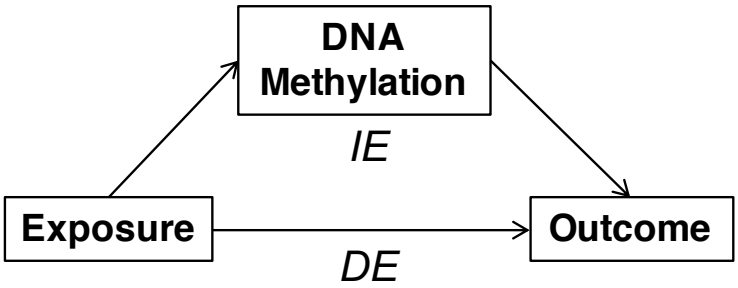

Figure 1. Mediation model with DNA methylation as the mediator.

In this study, we propose a novel causal mediation approach under the counterfactual framework to test the significance of total, direct, and indirect effects of DNA methylation in a genomic region on an outcome of interest. In this approach, a group of CpG sites from a predefined region are utilized as the mediator, and functional linear transformation[30] is used to reduce the possible high dimensions in the region's CpG sites and account for their location information. We denote the method as MR-Mediation. To evaluate the performance of the proposed MR-Mediation statistic, we first conduct extensive simulation studies to assess Type I error rates and power. We then illustrate our proposed method by analyzing whether DNA methylation mediates the association between exposure to gun violence and total immunoglobulin E (IgE) or atopic asthma among children participating in the Epigenetic Variation and Childhood Asthma in Puerto Ricans study (EVA-PR).[31]

METHODS

Functional linear transformation for region-based CpG sites

The idea is to linearly combine a few basis functions to fit a curve that can closely represent the methylation levels of the CpG sites in the region (e.g., an illustrative example is shown in Figure S1). Thus, the dimension can be reduced and the location information can be captured by the fitted curve. Specifically, we consider n subjects who have m CpG sites in a predefined region. The physical locations of the m CpG sites are normalized to a range of $[0, 1]$ (i.e., $0 \leq l_1 \ldots l_m \leq 1$). Thus, we use $M_i(l_j)$ to denote the methylation level of a CpG site at j^{th} location from the ith subject. We assume that the

$n \times 1$ vector of the continuous outcome $\mathbf{y}$ follows a linear regression model. For the ith subject,

$$y_i = \mathbf{Z}_i \boldsymbol{\beta}_0 + \int_0^1 \mathrm{M}_i(l)\, dl \times \beta_M + \varepsilon_i \qquad (1)$$

where $\mathbf{Z}_i$ is a $1 \times p$ covariate vector, $\boldsymbol{\beta}_0$ is a $p \times 1$ parameter vector (an intercept and $p - 1$ covariates), $\mathrm{M}_i(l)$ is a function of CpG site locations and includes m CpG site values, β_M is the effect of $\mathrm{M}_i(l)$, and ε_i is the random error. By using the ordinary linear smoother,[32] we let

$$\mathrm{M}_i(l) = \big(M_i(l_1), \dots, M_i(l_m)\big)\boldsymbol{\Phi}[\boldsymbol{\Phi}'\boldsymbol{\Phi}]^{-1}\boldsymbol{\phi}(l)$$

where $\big(M_i(l_1), \dots, M_i(l_m)\big)$ is a $1 \times m$ vector for m CpG sites in the region, $\boldsymbol{\phi}(l)$ is a $K_M \times 1$ vector including K_M basis functions, and $\boldsymbol{\Phi}$ is a $m \times K_M$ matrix containing the values of $\boldsymbol{\phi}(l)$ at $l_1 \dots l_m$. Thus, Formula (1) can be rewritten as

$$y_i = \mathbf{Z}_i \boldsymbol{\beta}_0 + \mathrm{R}_i \beta_M + \varepsilon_i \qquad (2)$$

where $\mathrm{R}_i = \big(M_i(l_1), \dots, M_i(l_m)\big)\boldsymbol{\Phi}[\boldsymbol{\Phi}'\boldsymbol{\Phi}]^{-1} \int_0^1 \boldsymbol{\phi}(l)\, dl$ is a scalar. Therefore, R_i can be viewed as a summary of the methylation level in the region. In this study, we focus on cubic B-spline basis functions.

Association between the mediator and the exposure variable

To test the proportions of the TE, DE and IE of the exposure on the outcome variables that are mediated by DNA methylation, we need to first make sure there exists an association between the mediator and exposure variables. To do so, we consider a linear regression model for n independent subjects,

$$\mathrm{R}_i = \mathbf{Z}_i \boldsymbol{\alpha}_0 + \mathrm{X}_i \alpha_X + \varepsilon_{R_i} \qquad (3)$$

where R_i is a scalar representing the regional methylation level for the ith subject, $\mathbf{Z}_i$ is a $1 \times p$ covariate vector, $\boldsymbol{\alpha}_0$ is a $p \times 1$ parameter matrix (an intercept and $p-1$ covariates), X_i is a scalar for the exposure variable, α_X is the parameter for exposure, and ε_{R_i} is a random error. In this study, we first test $H_0: \alpha_X = 0$, and then further test any methylation mediated effects between the exposure and outcome variables only when $\alpha_X \neq 0$.

Counterfactual approach to the causal mediation model

The total effect (TE) from exposure to outcome can be decomposed into the indirect effect (IE) mediated through DNA methylation and the direct effect (DE) not mediated by DNA methylation, which could be mediated by other mechanisms or a direct link between exposure and outcome. For a continuous outcome, the mediation model for the ith subject can be written as

$$\boldsymbol{E}(y_i|\mathbf{Z}_i, \mathrm{X}_i, \mathrm{R}_i) = \mathbf{Z}_i\boldsymbol{\beta}_0 + \mathrm{X}_i\beta_X + \mathrm{R}_i\beta_M + \mathrm{X}_i\mathrm{R}_i\beta_{XM} \qquad (4)$$

where $\mathbf{Z}_i\boldsymbol{\beta}_0$ is for the covariates, $\mathrm{X}_i\beta_X$ is for the exposure variable, $\mathrm{R}_i\beta_M$ is for the regional methylation level, and $\mathrm{X}_i\mathrm{R}_i\beta_{XM}$ is for the interaction between the exposure variable and the regional methylation level. To use the counterfactual approach to the causal mediation model, four identifiability assumptions need to be satisfied: 1. no unmeasured confounding of the exposure and outcome relationship; 2. no unmeasured confounding of the mediator and outcome relationship; 3. no unmeasured confounding of the exposure and mediator relationship; 4. the exposure must not cause any known confounder of the mediator and outcome relationship.[33] If all assumptions hold, the DE, IE and TE for change in exposure from level x_0 to x_1 are given by

$$DE = \boldsymbol{E}\big(y_i|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_0)\big) - \boldsymbol{E}\big(y_i|\mathbf{Z}_i, x_0, \mathrm{R}_i(x_0)\big)$$

$$= (x_1 - x_0)[\beta_X + (\mathbf{Z}_i\boldsymbol{\alpha}_0 + x_0\alpha_X)\beta_{XM}]$$

$$IE = \boldsymbol{E}\big(y_i|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_1)\big) - \boldsymbol{E}\big(y_i|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_0)\big)$$

$$= (x_1 - x_0)\alpha_X(\beta_M + x_1\beta_{XM})$$

$$TE = \boldsymbol{E}\big(y_i|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_1)\big) - \boldsymbol{E}\big(y_i|\mathbf{Z}_i, x_0, \mathrm{R}_i(x_0)\big)$$

$$= DE + IE$$

Since we only focus on the scenarios with $\alpha_X \neq 0$, conceivably, it can be shown that

$$H_0: DE = 0 \text{ equivalent to } H_0: \beta_X = \beta_{XM} = 0$$

$$H_0: IE = 0 \text{ equivalent to } H_0: \beta_M = \beta_{XM} = 0 \qquad (5)$$

$$H_0: TE = 0 \text{ equivalent to } H_0: \beta_X = \beta_M = \beta_{XM} = 0.$$

Thus, we can test $H_0: DE = 0$ by a *F*-test with degrees of freedom $(2,\ n-p-3)$, $H_0: IE = 0$ by a *F*-test with degrees of freedom $(2,\ n-p-3)$, and $H_0: TE = 0$ by a *F*-test with degrees of freedom $(3,\ n-p-3)$.

Similarly, when the outcome is binary, the mediation model for the ith subject can be written as

$$logit\big(P(y_i = 1|\mathbf{Z}_i, \mathrm{X}_i, \mathrm{R}_i)\big) = \mathbf{Z}_i\boldsymbol{\beta}_0 + \mathrm{X}_i\beta_X + \mathrm{R}_i\beta_M + \mathrm{X}_i\mathrm{R}_i\beta_{XM} \qquad (6).$$

If all identifiability assumptions hold, the DE, IE and TE for change in exposure from level x_0 to x_1 are given by

$$log\{OR^{DE}\} = logit\Big(P\big(y_i = 1|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_0)\big)\Big) - logit\Big(P\big(y_i = 1|\mathbf{Z}_i, x_0, \mathrm{R}_i(x_0)\big)\Big)$$

$$log\{OR^{IE}\} = logit\Big(P\big(y_i = 1|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_1)\big)\Big) - logit\Big(P\big(y_i = 1|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_0)\big)\Big)$$

$$log\{OR^{TE}\} = logit\Big(P\big(y_i = 1|\mathbf{Z}_i, x_1, \mathrm{R}_i(x_1)\big)\Big) - logit\Big(P\big(y_i = 1|\mathbf{Z}_i, x_0, \mathrm{R}_i(x_0)\big)\Big).$$

172 According to Valeri and VanderWeele (2013),[33] a rare outcome assumption (i.e., low
173 disease prevalence) is needed so as to have the odds ratio (OR) in the case-control
174 design equivalent to the OR in the population, which is further approximately equivalent
175 to the relative risk (RR) in the population. We also adopt the rare outcome assumption in
176 this study. Then, the DE, IE and TE can be derived as

177 $$log\{OR^{DE}\} \approx (x_1 - x_0)[\beta_X + (\mathbf{Z}_i\boldsymbol{\alpha}_0 + x_0\alpha_X + \beta_M\sigma_R^2)\beta_{XM}] + \frac{1}{2}(x_1^2 - x_0^2)\beta_{XM}^2\sigma_R^2$$

178 $$log\{OR^{IE}\} \approx (x_1 - x_0)\alpha_X(\beta_M + x_1\beta_{XM})$$

179 $$log\{OR^{TE}\} = log\{OR^{DE}\} + log\{OR^{IE}\},$$

180 where σ_R^2 is the variance of random error from the mediator regression, Formula (3). The
181 existence of σ_R^2 is due to $exp(\varepsilon_R) \sim lognormal(0, \sigma_R^2)$. The null hypotheses for testing
182 $DE = 0$, $IE = 0$ and $TE = 0$ for binary outcome are the same as given for continuous
183 outcome. We can test $H_0: DE = 0$ by a χ^2-distributed Rao's score statistic with 2 degrees
184 of freedom, $H_0: IE = 0$ with 2 degrees of freedom, and $H_0: TE = 0$ with 3 degrees of
185 freedom. Please note that the mediator regression, Formula (3), is run only for controls
186 to account for the case-control design.[33] In other words, because of the rare outcome
187 assumption, the exposure distribution in the controls from the case-control design is
188 approximately equivalent to the exposure distribution in the whole population.

189

190 *The Epigenetic Variation and Childhood Asthma in Puerto Ricans study (EVA-PR)*

191 EVA-PR is a case-control study of asthma in subjects aged 9 to 20 years. Subject
192 recruitment and study procedures for EVA-PR have been described elsewhere.[31]
193 Genome-wide DNA methylation was measured in 488 nasal epithelial samples using the
194 Infinium HumanMethylation450 BeadChip arrays (Illumina, San Diego, CA). The

195 preprocessing and quality control (QC) procedures were described in our prior study.[31]
196 After quality control, CpG sites with an overall mean β-value within [0.1, 0.9] were kept,
197 leaving 227,901 CpG sites in the final dataset. Methylation M-value was calculated by
198 $\log_2$(β-value/(1–β-value)). This methylation dataset was used in both the following
199 simulation studies and in real data analysis.
200
201 The child's lifetime exposure to gun violence was treated as the exposure variable in the
202 following real data analysis. Lifetime exposure to gun violence was derived from the
203 exposure to violence scale[34-36] and analyzed as a binary variable (having heard gunshots
204 at least twice vs. no more than once).[37] Atopy was defined as an IgE ≥0.35 IU/mL to ≥1
205 of five common allergens (Der p 1, Bla g 2, Fel d 1, Can f 1, and Mus m 1). Asthma was
206 defined as a physician's diagnosis plus at least one episode of wheeze in the previous
207 year. Atopic asthma was defined as the presence of both atopy and asthma, and thus
208 treated as the binary outcome variable in the real data analysis; controls were non-
209 asthmatic subjects. Total plasma IgE was treated as continuous outcome after $\log_{10}$-
210 transformation.
211
212 *Simulation studies*
213 We used the CpG sites from 488 nasal epithelial samples in EVA-PR for our simulation
214 studies. The CpG sites were grouped into corresponding genes, and the kept CpG sites
215 also needed to be within 5 kb distance from its nearest gene. The histogram plot (Figure
216 S2) showed that most of the genes included less than 20 CpG sites, although some genes
217 could have several hundred CpG sites. Thus, we selected different sizes of genes in our

simulation studies. Specifically, we selected (1) 705 genes with 10 CpG sites; (2) 306 genes with 15 CpG sites; (3) 152 genes with 20 CpG sites; and (4) 21 genes with 40 CpG sites.

Furthermore, we need to choose the basis functions and the number of basis functions to reduce the dimension of each gene. To simulate the exposure $\mathbf{X}$ and outcome $\boldsymbol{y}$ variables, we used the same cubic (i.e., $order = 4$) B-spline basis functions for $\mathrm{M}_i(l)$. We considered $K_M = \{4, \ldots, 8\}$ for all genes.

For each gene, we simulated one 488×1 vector for binary exposure $\mathbf{X}$ via the model:

$$logit\big(P(\mathbf{X} = 1)\big) = \mathbf{M}\boldsymbol{\alpha}_X^*$$

where $\mathbf{M}$ included 15% consecutive CpG sites covering the longest distance in the gene, each $\alpha_X^* = 1.2/m_{15\%}$ and $m_{15\%}$ is the number of 15% CpG sites in the gene. The purpose is to make the exposure and the mediator (i.e., methylation) associated. Then, we can evaluate the power of testing $\alpha_X = 0$ in Formula (3).

For each gene, we then simulated continuous and binary outcomes $\mathbf{y}$ for 488 samples via the models:

$$\mathbf{y} = \beta_0 + \mathbf{Z}_1\beta_1 + \mathbf{Z}_2\beta_2 + \mathbf{X}\beta_X + \mathbf{M}\boldsymbol{\beta}_M + \mathbf{XM}\boldsymbol{\beta}_{XM} + \boldsymbol{\varepsilon} \text{ and}$$

$$logit\big(P(\mathbf{y} = 1)\big) = \beta_{0^*} + \mathbf{Z}_1\beta_1 + \mathbf{Z}_2\beta_2 + \mathbf{X}\beta_X + \mathbf{M}\boldsymbol{\beta}_M + \mathbf{XM}\boldsymbol{\beta}_{XM}$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{X}$ and $\mathbf{M}$ were the same as described above, $\boldsymbol{\varepsilon}$ was generated from standard normal distribution, $\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$ and $\beta_{0^*} = -1$. We considered (1) all $m_{15\%}$ effective CpG sites in the same direction (100%+) and (2) 50% effective CpG

sites in one direction and the other 50% effective CpG sites in the opposite direction (50%+/50%-). We then considered eight settings for β_X, $\boldsymbol{\beta}_M$ and $\boldsymbol{\beta}_{XM}$ to simulate $\mathbf{y}$, and these eight settings (Table 1) were used to evaluate Type I error rates and power of testing DE, IE and TE with the null hypotheses described in Formula (5). As shown in Table 1, Setting (1) was used to assess Type I error rates of testing DE, IE and TE; Setting (2) was used to assess Type I error rates of testing DE and power of testing IE and TE; Setting (3) was used to assess Type I error rates of testing IE and power of testing DE and TE; and Settings (4) ~ (8) were used to assess power of testing DE, IE and TE. Please note that Setting (2) may not be the true null hypothesis for DE. Although $\beta_X = 0$, there could still be implicit $\mathbf{X}$ effect, because $\mathbf{X}$ and $\mathbf{M}$ are associated, and $\boldsymbol{\beta}_M \neq 0$. Analogously, Setting (3) may not be the true null hypothesis for IE.

Table 1. The eight settings for β_X, $\boldsymbol{\beta}_M$ and $\boldsymbol{\beta}_{XM}$ and their corresponding null hypothesis (H_0) and alternative hypothesis (H_a) for DE, IE and TE

	Settings*				DE	IE	TE
	Continuous y		Binary y				
	100%+	50%+/50%-	100%+	50%+/50%-			
(1)	$\beta_X = \boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0$		$\beta_X = \boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0$		H_0	H_0	H_0
(2)	$\beta_X = \boldsymbol{\beta}_{XM} = 0, \boldsymbol{\beta}_M = 0.2$	$\beta_X = \boldsymbol{\beta}_{XM} = 0, \boldsymbol{\beta}_M = 0.2/-0.2$	$\beta_X = \boldsymbol{\beta}_{XM} = 0, \boldsymbol{\beta}_M = 0.4$	$\beta_X = \boldsymbol{\beta}_{XM} = 0, \boldsymbol{\beta}_M = 0.4/-0.4$	H_0#	H_a	H_a
(3)	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0, \beta_X = 0.2$	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0, \beta_X = 0.2/-0.2$	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0, \beta_X = 0.4$	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0, \beta_X = 0.4/-0.4$	H_a	H_0#	H_a
(4)	$\beta_X = \boldsymbol{\beta}_M = 0, \boldsymbol{\beta}_{XM} = 0.2$	$\beta_X = \boldsymbol{\beta}_M = 0, \boldsymbol{\beta}_{XM} = 0.2/-0.2$	$\beta_X = \boldsymbol{\beta}_M = 0, \boldsymbol{\beta}_{XM} = 0.4$	$\beta_X = \boldsymbol{\beta}_M = 0, \boldsymbol{\beta}_{XM} = 0.4/-0.4$	H_a	H_a	H_a
(5)	$\beta_X = \boldsymbol{\beta}_{XM} = 0.2, \boldsymbol{\beta}_M = 0$	$\beta_X = \boldsymbol{\beta}_{XM} = 0.2/-0.2, \boldsymbol{\beta}_M = 0$	$\beta_X = \boldsymbol{\beta}_{XM} = 0.4, \boldsymbol{\beta}_M = 0$	$\beta_X = \boldsymbol{\beta}_{XM} = 0.4/-0.4, \boldsymbol{\beta}_M = 0$	H_a	H_a	H_a
(6)	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.2, \beta_X = 0$	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.2/-0.2, \beta_X = 0$	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.4, \beta_X = 0$	$\boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.4/-0.4, \beta_X = 0$	H_a	H_a	H_a
(7)	$\beta_X = \boldsymbol{\beta}_M = 0.2, \boldsymbol{\beta}_{XM} = 0$	$\beta_X = \boldsymbol{\beta}_M = 0.2/-0.2, \boldsymbol{\beta}_{XM} = 0$	$\beta_X = \boldsymbol{\beta}_M = 0.4, \boldsymbol{\beta}_{XM} = 0$	$\beta_X = \boldsymbol{\beta}_M = 0.4/-0.4, \boldsymbol{\beta}_{XM} = 0$	H_a	H_a	H_a
(8)	$\beta_X = \boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.2$	$\beta_X = \boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.2/-0.2$	$\beta_X = \boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.4$	$\beta_X = \boldsymbol{\beta}_M = \boldsymbol{\beta}_{XM} = 0.4/-0.4$	H_a	H_a	H_a

* $\boldsymbol{\beta}_M$ and $\boldsymbol{\beta}_{XM}$ are vectors for 15% consecutive CpG sites. For 100%+, each element has the same assigned value. For 50%+/50%-, 50% elements have the same positive value and 50% elements have the same negative value. If the number of CpG sites is odd, one CpG site is assigned to have no effect.

Setting (2) may not be the true null hypothesis for DE. Although $\beta_X = 0$, there could still be implicit $\mathbf{X}$ effect, because $\mathbf{X}$ and $\mathbf{M}$ are associated, and $\boldsymbol{\beta}_M \neq 0$. Analogously, Setting (3) may not be the true null hypothesis for IE.

262 For each gene, we simulated 10 sets of $\mathbf{y}$ for genes with 10 CpG sites (total is
263 705×10=7,050); 30 sets of $\mathbf{y}$ for genes with 15 CpG sites (total is 306×30=9,180); 50 sets
264 of $\mathbf{y}$ for genes with 20 CpG sites (total is 152×50=7,600); and 400 sets of $\mathbf{y}$ for genes with
265 40 CpG sites (total is 21×400=8,400). Thus, we simulated 32,230 datasets for continuous
266 outcome and another 32,230 datasets for binary outcome. After simulating the exposure
267 $\mathbf{X}$ and outcome $\mathbf{y}$ variables, we used two approaches to conduct the analyses: (1) basis
268 function fitted regional methylation level as the mediator (i.e., MR-Mediation) and (2)
269 mean CpG methylation level as the mediator. We assessed the Type I error rates of
270 testing DE, IE and TE only when the p-values of α_X were less than 0.05, and we evaluated
271 the power at the significance level that the p-values of the corresponding effect (i.e., DE,
272 IE or TE) and α_X were both less than 0.05.
273

274 **RESULTS**

275 *Simulation of the Type I error rate*

276 For both continuous and binary outcomes, both approaches retained the desired Type I
277 error rates for DE, IE and TE across all considered gene sizes and basis numbers in
278 Setting (1) (Figures 2, S3-S5 and S8-S10). Since Setting (2) might not be the true null
279 hypothesis for DE, both MR-Mediation and the mean CpG methylation approach had
280 inflated Type I error rates for scenarios with 100%+ effective CpG sites (Figure S6 and
281 S11) and correct Type I error rates for scenarios with 50%+/50%- effective CpG sites
282 (Figure S13 and S15). Although Setting (3) might not be the true null hypothesis for IE,
283 both approaches retained the desired Type I error rates for all simulated conditions for

both continuous and binary outcomes (Figures S7, S12, S14 and S16). Thus, both approaches are valid under their true null hypothesis.

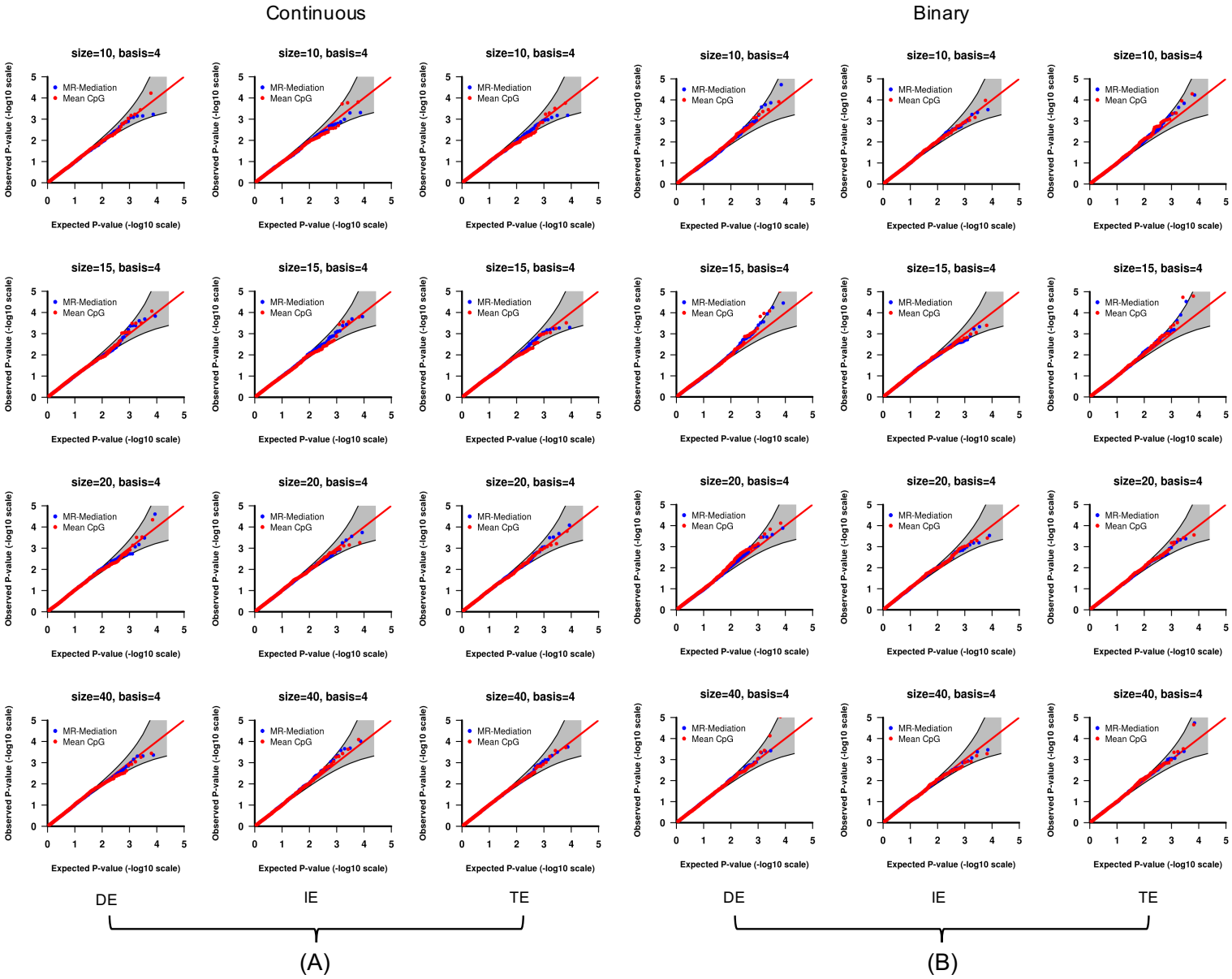

Figure 2. QQ plot of the *p*-values from Setting (1) with (A) continuous outcome and (B) binary outcome. The 1st column shows DE results, the 2nd column shows IE results and the 3rd column shows TE results. Each row shows one gene size (i.e., 10, 15, 20 or 40 CpG sites). Two approaches were compared: (1) MR-Mediation; and (2) Mean CpG methylation approach. A 95% pointwise confidence band (gray area) was computed under the assumption that the *p*-values were drawn independently from a uniform [0, 1] distribution.

294 *Simulation of power*

295 In all considered scenarios, the power of MR-Mediation decreased as the basis number
296 increased when the gene size was relatively small, and remained the similar power as
297 the basis number increased when the gene size was large (Figures S17-S28). When the
298 basis number was 4, MR-Mediation was the most powerful method in all scenarios for
299 testing IE and more powerful than the mean CpG methylation method in almost all
300 scenarios for testing TE and DE (Figures 3 and 4). In real data studies, one could be more
301 interested in testing IE than TE and DE. Thus, MR-Mediation with the basis number equal
302 to 4 is recommended.

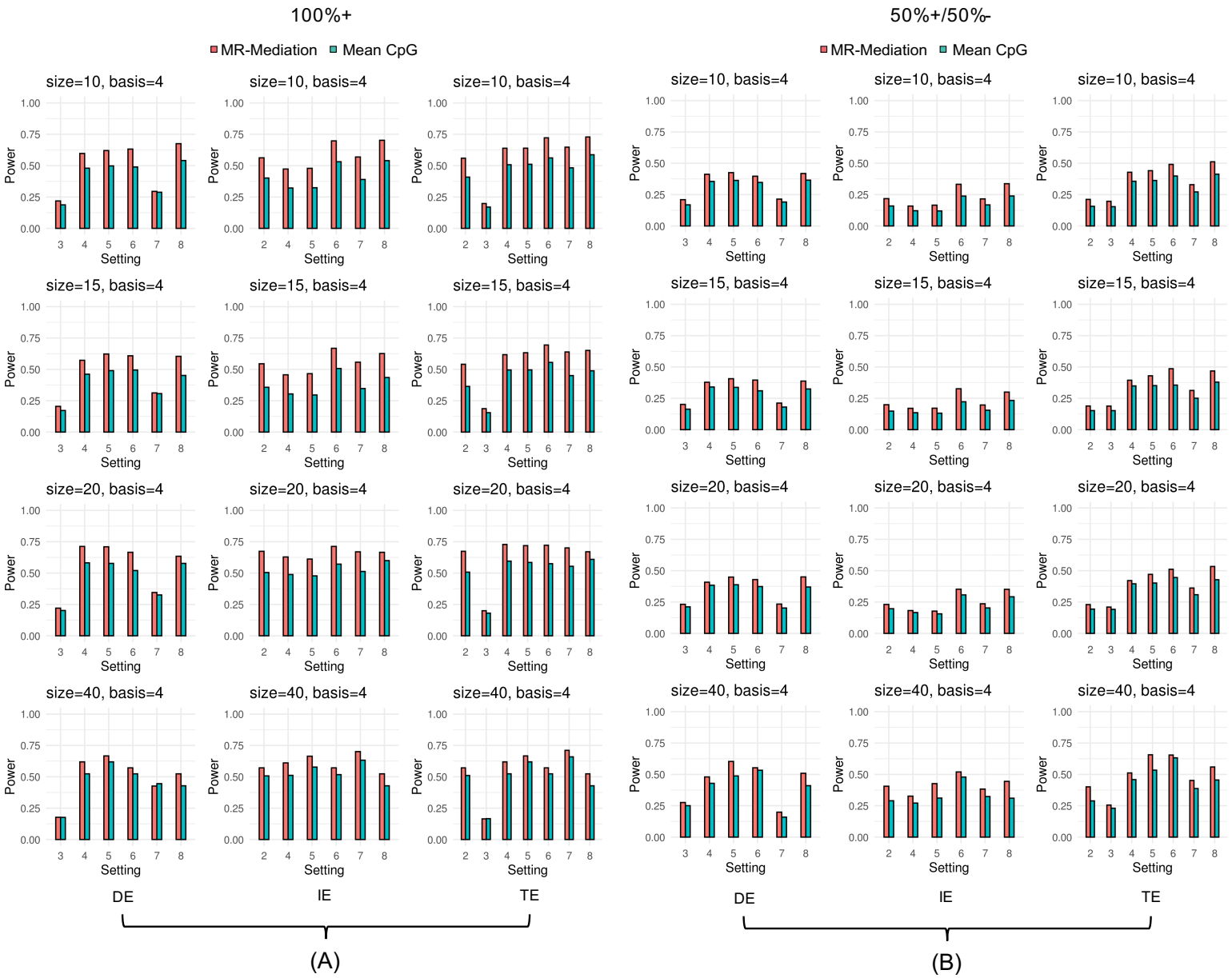

Figure 3. Power comparison for continuous outcome. (A) all positively effective CpG sites (100%+) and (B) 50% positively and 50% negatively effective CpG sites (50%+/50%-). The 1st column shows DE results, the 2nd column shows IE results and the 3rd column shows TE results. Each row shows one gene size (i.e., 10, 15, 20 or 40 CpG sites). Two approaches were compared: (1) MR-Mediation; and (2) Mean CpG methylation approach.

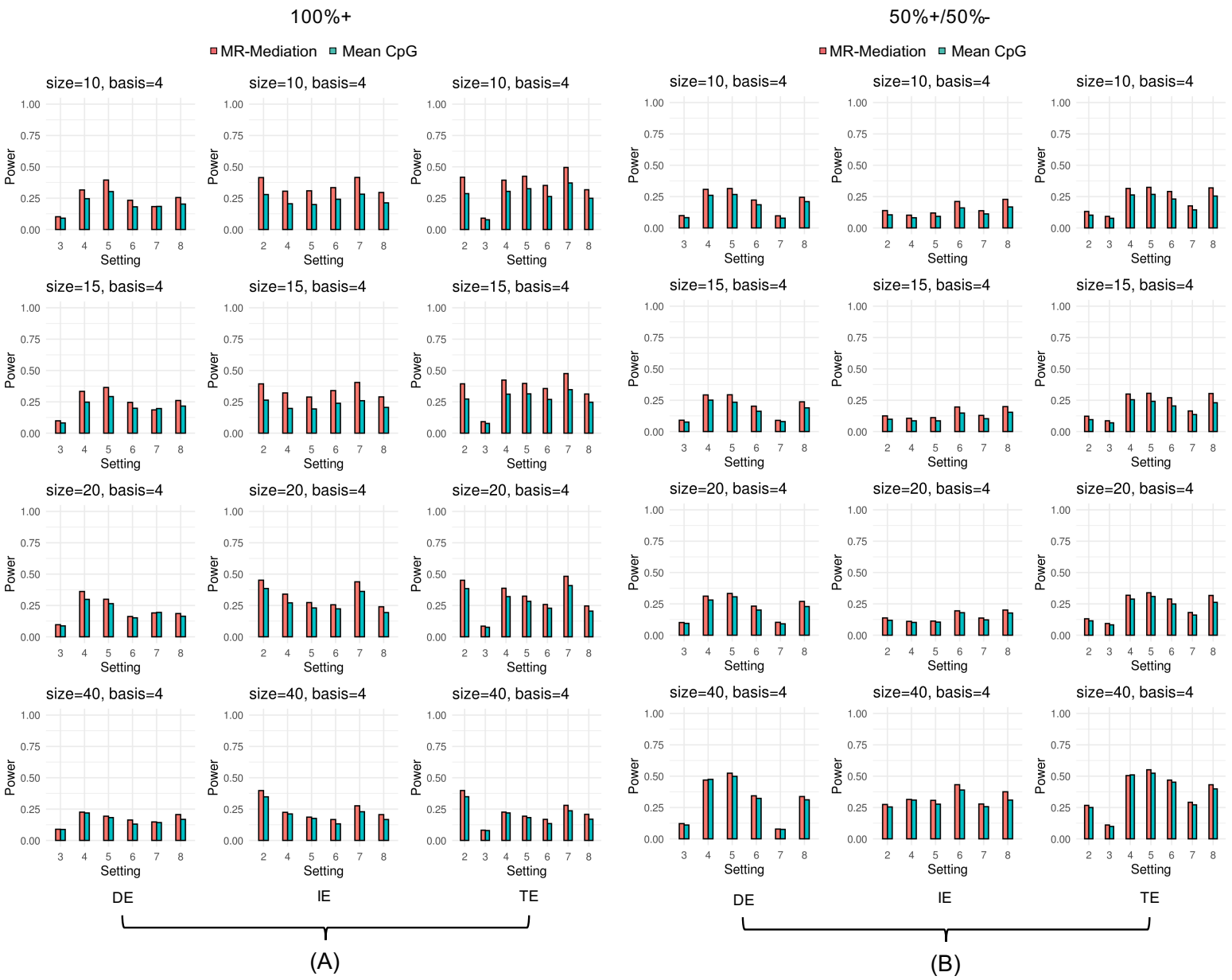

Figure 4. Power comparison for binary outcome. (A) all positively effective CpG sites (100%+) and (B) 50% positively and 50% negatively effective CpG sites (50%+/50%-). The 1st column shows DE results, the 2nd column shows IE results and the 3rd column shows TE results. Each row shows one gene size (i.e., 10, 15, 20 or 40 CpG sites). Two approaches were compared: (1) MR-Mediation; and (2) Mean CpG methylation approach.

Results in EVA-PR

We used the proposed MR-Mediation method to study whether the association between exposure to gun violence and atopic asthma was mediated by DNA methylation, where atopic asthma is a binary outcome. In this study, we considered 5,720 genes including at

319 least 10 CpG sties in 407 subjects without any missing covariates. Note that subjects who
320 had asthma but no atopy were excluded from this analysis, as atopic asthma is the most
321 common type of asthma in children. This analysis was adjusted for age, sex, annual
322 household income (a measure of socioeconomic status), the top five principal
323 components from genotypic data, methylation batch, and latent factors between exposure
324 to gun violence and methylation, and between atopic asthma and methylation, estimated
325 from *sva*.[38] The mediator regression with four cubic B-spline basis functions, Formula (3),
326 was run only for controls to account for the case-control design and we selected the top
327 20 genes for the further mediation analysis. We focused on the IE test to see if the
328 association between exposure to gun violence and atopic asthma was mediated by
329 methylation. We found that *CFD* on chromosome 19 was the top gene associated with
330 exposure to gun violence ($P = 1.93\times10^{-5}$ and $FDR = 0.1102$, Table S1), although it did
331 not reach the genome-wide significance level, and then it could mediate the effect of
332 exposure to gun violence on atopic asthma risk ($P = 3.09\times10^{-2}$). The *CFD* gene has been
333 reported to be associated with asthma.[39] The expression of *CFD* can be reduced by IL-
334 17A, which is required for development of airway hyperresponsiveness (a key
335 intermediate phenotype of asthma). In addition, *TBC1D14* ($P = 7.28\times10^{-3}$) on
336 chromosome 4, *FAM120B* ($P = 3.33\times10^{-2}$) on chromosome 6, *CRB1* ($P = 1.75\times10^{-2}$) on
337 chromosome 10, *LOC339166* ($P = 1.42\times10^{-2}$) on chromosome 17, *ZGPAT* ($P = 2.55\times10^{-}$
338 ^{3}) on chromosome 20, *MED16* ($P = 2.05\times10^{-2}$) on chromosome 19, *PLEKHG6* ($P =$
339 1.24×10^{-3}) on chromosome 12, and *TNNI1* ($P = 8.85\times10^{-3}$) on chromosome 1 could also
340 mediate the effect of exposure to gun violence on atopic asthma risk at the nominal level
341 in the top 20 gene list.

342

343 We also applied MR-Mediation to study whether the association between exposure to
344 gun violence and total IgE was mediated by DNA methylation, where total IgE is a
345 continuous outcome. We also used 5,720 genes including at least 10 CpG sties in 473
346 subjects without any missing covariates. This analysis was adjusted for age, sex, annual
347 household income, the top five principal components from genotypic data, methylation
348 batch, and latent factors between exposure to gun violence and methylation, and between
349 total IgE and methylation, estimated from *sva*.[38] We selected the top 20 genes by *p*-values
350 from the association between exposure to gun violence and gene-based methylation
351 (Formula (3)) and conducted mediation analysis (Formula (4)). Again, we focused on the
352 IE test to see if the association between exposure to gun violence and total IgE was
353 mediated by methylation. The results showed that *PIP5K1C* (P = 4.33×10^{-3}) on
354 chromosome 19 could mediate the effect of exposure to gun violence on total IgE change
355 at the nominal level in the top 20 gene list (Table S2).

356

357 **DISCUSSION**

358 In this study, we developed a novel causal mediation approach, MR-Mediation, under the
359 counterfactual framework to test the significance of DE, IE and TE. We implemented MR-
360 Mediation in R (http://www.r-project.org) and the R package (https://cran.r-
361 project.org/web/packages/MRmediation/index.html) is available. The major advantage of
362 the counterfactual framework to mediation analysis is that it allows the decomposition of
363 TE into DE and IE, even in models with nonlinearities and interactions. The counterfactual

364 framework extends the widely used Baron and Kenny mediation approach by allowing an
365 interaction term between exposure and mediator in the outcome regression.
366
367 We focused on testing the significance of DE, IE and TE instead of estimating their effects.
368 For example, in IE, we test $H_0: \beta_M = \beta_{XM} = 0$. When we reject the null hypothesis, it is
369 possible that $\beta_M = 0$ and $\beta_{XM} \neq 0$, which could not be known from the proposed test.
370 Then the estimates of β_M may not be reasonable to be used in calculating IE. The
371 parameters of covariates even need to be significant to calculate the estimates of DE and
372 TE. In this approach, a group of CpG sites from a predefined region are utilized as the
373 mediator, and the functional transformation is used to reduce the possible high dimension
374 in the region-based CpG sites and account for their location information. The region-
375 based analysis methods can improve power by combining weak signals and by reducing
376 the multiple testing penalty. Users can use different ways to define regions of interest in
377 addition to genes used in this study, such as, CpG islands and CpG shores.
378
379 For binary outcomes, we used the rare disease assumption, so that the OR in the case-
380 control design is equal to the OR in the population, which is approximately equal to the
381 RR in the population. When the disease is common in the population, the OR does not
382 approximate the RR. In that case, a log link should be used in the generalized linear
383 model for binary outcomes instead of a logit link, but the log link often fails to work with a
384 binomial distribution. The Zou's modified Poisson regression[40] could be applied in this
385 situation, although this is an approximate approach.
386

In the simulation studies, we show that MR-Mediation retains the correct Type I error rate for testing DE, IE and TE for both continuous and binary outcomes. In power comparison, we are more interested in IE, and MR-Mediation with basis=4 achieves better power than the mean CpG methylation approach. Our real data results show that MR-Mediation could help to identify potential methylation mediated effects. The genome-wide data analysis could be completed within an hour using one CPU.

Since MR-Mediation is a joint significance test, the exposure and regional methylation relationship and the regional methylation and outcome relationship are tested in two steps. It's possible that CpG sites associated with the exposure and CpG sites associated with the outcome are different, which cannot be detected by the region-based mediator method. A further mediation analysis using single CpG sites may be needed. The single CpG site mediation function is also available in our R package. In addition, the summarized methylation level in the region can be viewed as the area under the curve. However, the areas could be still the same, even if curves are different. In future studies, we will consider to automatically partition the area into several sub-areas and each sub-area has a summarized methylation value. In such methods, $\boldsymbol{\alpha}_X$, $\boldsymbol{\beta}_M$ and $\boldsymbol{\beta}_{XM}$ are vectors instead of scalars as in this study. It's possible that $\boldsymbol{\alpha}_X(\boldsymbol{\beta}_M + x_1\boldsymbol{\beta}_{XM}) = 0$, even if $\boldsymbol{\alpha}_X \neq 0$ and $\boldsymbol{\beta}_M \neq 0$ or $\boldsymbol{\beta}_{XM} \neq 0$. We will need to address this potential problem.

MR-Mediation can be extended to high dimensional predictors, such as multiple SNPs from a region, so as to investigate DNA methylation-mediated genetic effects on an outcome. These SNPs can be from the same region as the CpG sites. Such region-based

410 approach on both SNPs and CpG sites can boost statistical power, because the region-
411 based approach combines small individual effects and reduce multiple testing. Moreover,
412 the region-based CpG sites can also be treated as predictors and gene expression can
413 be treated as a mediator in order to understand the regulatory mechanisms of DNA
414 methylation on gene expression that further affect diseases or traits.
415

416 **Disclosure of Potential Conflict of Interest**

417 No potential conflicts of interest were disclosed.
418

419 **Supplemental Material**

420 Supplemental data for this article can be accessed on the publisher's website.

421 **REFERENCES**

422 1. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW,
423 Rigoutsos I, Loring J, et al. Dynamic changes in the human methylome during
424 differentiation. Genome Res 2010; 20:320-31.
425 2. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol
426 2010; 28:1057-68.
427 3. Melotte V, Lentjes MH, van den Bosch SM, Hellebrekers DM, de Hoon JP, Wouters
428 KA, Daenen KL, Partouns-Hendriks IE, Stessels F, Louwagie J, et al. N-Myc downstream-
429 regulated gene 4 (NDRG4): a candidate tumor suppressor gene and potential biomarker
430 for colorectal cancer. J Natl Cancer Inst 2009; 101:916-27.
431 4. Schmidt B, Liebenberg V, Dietrich D, Schlegel T, Kneip C, Seegebarth A,
432 Flemming N, Seemann S, Distler J, Lewin J, et al. SHOX2 DNA methylation is a biomarker
433 for the diagnosis of lung cancer based on bronchial aspirates. BMC Cancer 2010; 10:600.
434 5. Jain S, Chen S, Chang KC, Lin YJ, Hu CT, Boldbaatar B, Hamilton JP, Lin SY,
435 Chang TT, Chen SH, et al. Impact of the location of CpG methylation within the GSTP1
436 gene on its specificity as a DNA marker for hepatocellular carcinoma. PLoS One 2012;
437 7:e35789.
438 6. Lord J, Cruchaga C. The epigenetic landscape of Alzheimer's disease. Nat
439 Neurosci 2014; 17:1138-40.
440 7. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, Eaton ML,
441 Keenan BT, Ernst J, McCabe C, et al. Alzheimer's disease: early alterations in brain DNA
442 methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci 2014; 17:1156-63.
443 8. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, Troakes C,
444 Al-Sarraj S, Burrage J, Macdonald R, et al. Methylomic profiling implicates cortical
445 deregulation of ANK1 in Alzheimer's disease. Nat Neurosci 2014; 17:1164-70.
446 9. Pidsley R, Viana J, Hannon E, Spiers H, Troakes C, Al-Saraj S, Mechawar N,
447 Turecki G, Schalkwyk LC, Bray NJ, et al. Methylomic profiling of human brain tissue
448 supports a neurodevelopmental origin for schizophrenia. Genome Biol 2014; 15:483.
449 10. Jaffe AE, Gao Y, Deep-Soboslay A, Tao R, Hyde TM, Weinberger DR, Kleinman
450 JE. Mapping DNA methylation across development, genotype and schizophrenia in the
451 human frontal cortex. Nat Neurosci 2016; 19:40-7.
452 11. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining,
453 analyzing, grouping and correcting spatially correlated P-values. Bioinformatics 2012;
454 28:2986-8.
455 12. Butcher LM, Beck S. Probe Lasso: a novel method to rope in differentially
456 methylated regions with 450K DNA methylation data. Methods 2015; 72:21-8.
457 13. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox
458 TV, Davies R, Down TA, et al. DNA methylation profiling of human chromosomes 6, 20
459 and 22. Nat Genet 2006; 38:1378-85.
460 14. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA,
461 Wen B, Feinberg AP. Comprehensive high-throughput arrays for relative methylation
462 (CHARM). Genome Res 2008; 18:780-90.
463 15. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome
464 integrates intrinsic and environmental signals. Nat Genet 2003; 33 Suppl:245-54.

16. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009; 41:178-86.
17. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009; 462:315-22.
18. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of personality and social psychology 1986; 51:1173-82.
19. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods 2010; 15:309-34.
20. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992; 3:143-55.
21. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol 2010; 172:1339-48.
22. Bozack AK, Cardenas A, Quamruzzaman Q, Rahman M, Mostofa G, Christiani DC, Kile ML. DNA methylation in cord blood as mediator of the association between prenatal arsenic exposure and gestational age. Epigenetics 2018; 13:923-40.
23. Cardenas A, Lutz SM, Everson TM, Perron P, Bouchard L, Hivert MF. Mediation by Placental DNA Methylation of the Association of Prenatal Maternal Smoking and Birth Weight. Am J Epidemiol 2019; 188:1878-86.
24. Barfield R, Shen J, Just AC, Vokonas PS, Schwartz J, Baccarelli AA, VanderWeele TJ, Lin X. Testing for the indirect effect under the null for genome-wide mediation analyses. Genet Epidemiol 2017; 41:824-33.
25. Tobi EW, Slieker RC, Luijk R, Dekkers KF, Stein AD, Xu KM, Biobank-based Integrative Omics Studies C, Slagboom PE, van Zwet EW, Lumey LH, et al. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. Sci Adv 2018; 4:eaao4364.
26. Yang CF, Karmaus WJJ, Yang CC, Chen ML, Wang IJ. Bisphenol a Exposure, DNA Methylation, and Asthma in Children. Int J Environ Res Public Health 2020; 17.
27. Neophytou AM, Oh SS, Hu D, Huntsman S, Eng C, Rodriguez-Santana JR, Kumar R, Balmes JR, Eisen EA, Burchard EG. In utero tobacco smoke exposure, DNA methylation, and asthma in Latino children. Environ Epidemiol 2019; 3:e048.
28. Wu D, Yang H, Winham SJ, Natanzon Y, Koestler DC, Luo T, Fridley BL, Goode EL, Zhang Y, Cui Y. Mediation analysis of alcohol consumption, DNA methylation, and epithelial ovarian cancer. J Hum Genet 2018; 63:339-48.
29. Battram T, Richmond RC, Baglietto L, Haycock PC, Perduca V, Bojesen SE, Gaunt TR, Hemani G, Guida F, Carreras-Torres R, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. Int J Epidemiol 2019; 48:1493-504.
30. Fan R, Wang Y, Mills JL, Wilson AF, Bailey-Wilson JE, Xiong M. Functional linear models for association analysis of quantitative traits. Genet Epidemiol 2013; 37:726-42.
31. Forno E, Wang T, Qi C, Yan Q, Xu CJ, Boutaoui N, Han YY, Weeks DE, Jiang Y, Rosser F, et al. DNA methylation in nasal epithelium, atopy, and atopic asthma in children: a genome-wide study. Lancet Respir Med 2019; 7:336-46.
32. Ramsay JO, Silverman BW. Functional Data Analysis. Springer, New York 1996.

511 33. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator
512 interactions and causal interpretation: theoretical assumptions and implementation with
513 SAS and SPSS macros. Psychol Methods 2013; 18:137-50.
514 34. Sternthal MJ, Jun HJ, Earls F, Wright RJ. Community violence and urban
515 childhood asthma: a multilevel analysis. Eur Respir J 2010; 36:1400-9.
516 35. Suglia SF, Ryan L, Wright RJ. Creation of a community violence exposure scale:
517 accounting for what, who, where, and how often. Journal of traumatic stress 2008;
518 21:479-86.
519 36. Thomson CC, Roberts K, Curran A, Ryan L, Wright RJ. Caretaker-child
520 concordance for child's exposure to violence in a preadolescent inner-city population.
521 Arch Pediatr Adolesc Med 2002; 156:818-23.
522 37. Ramratnam S, Han Y, Rosas-Salazar C, Forno E, Brehm J, Rosser F, Marsland
523 A, Colón-Semidey A, Alvarez M, Miller G, et al. Exposure to gun violence and asthma
524 among children in Puerto Rico. Respir Med 2015; In Press.
525 38. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for
526 removing batch effects and other unwanted variation in high-throughput experiments.
527 Bioinformatics 2012; 28:882-3.
528 39. Mathews JA, Wurmbrand AP, Ribeiro L, Neto FL, Shore SA. Induction of IL-17A
529 Precedes Development of Airway Hyperresponsiveness during Diet-Induced Obesity and
530 Correlates with Complement Factor D. Front Immunol 2014; 5:440.
531 40. Zou G. A modified poisson regression approach to prospective studies with binary
532 data. Am J Epidemiol 2004; 159:702-6.