OXFORD UNIVERSITY PRESS | Briefings in Bioinformatics

Testing cell-type-specific mediation effects in genome-wide epigenetic studies

Journal:	*Briefings in Bioinformatics*
Manuscript ID	BIB-20-0205.R1
Manuscript Type:	Problem solving protocol
Date Submitted by the Author:	n/a
Complete List of Authors:	Luo, Xiangyu; Renmin University of China, Institute of Statistics and Big Data Schwartz, Joel; Harvard University T H Chan School of Public Health, Department of Environmental Health Baccarelli, Andrea; Columbia University, Environmental Health Sciences Liu, Zhonghua; University of Hong Kong, Statistics and Actuarial Sciences
Keywords:	epigenetic studies, mediation analysis, cell-type specific, multiple testing, DNA methylation, inverse regression

Testing cell-type-specific mediation effects in genome-wide epigenetic studies

Xiangyu Luo[1], Joel Schwartz[2], Andrea Baccarelli[3], Zhonghua Liu[4*]

[1]*Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, China*
[2]*Department of Environmental Health and Epidemiology, Harvard University, Boston, MA, 02115, USA*
[3]*Department of Environmental Health Sciences, Columbia University, New York, NY, 10032, USA*
[4]*Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR, China*

*zhhliu@hku.hk

Xiangyu Luo is an Assistant Professor in the Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Joel Schwartz is a Professor in the Department of Environmental Health, Harvard University, Boston, MA, USA

Andrea Baccarelli is the Leon Hess Professor in the Department of Environmental Health Sciences, Columbia University, New York City, NY, USA

Zhonghua Liu is an Assistant Professor in the Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR, China

Abstract:

Epigenome-wide mediation analysis aims to identify DNA methylation CpG sites that mediate the causal effects of genetic/environmental exposures on health outcomes. However, DNA methylations in the peripheral blood tissues are usually measured at the bulk level based on a heterogeneous population of white blood cells. Using the bulk level DNA methylation data in mediation analysis might cause confounding bias and reduce study power. Therefore, it is crucial to get fine-grained results by detecting mediation CpG sites in a cell-type-specific way. However, there is a lack of methods and software to achieve this goal. We propose a novel method MICS (Mediation In a Cell-type-Specific fashion) to identify cell-type-specific mediation effects in genome-wide epigenetic studies using only the bulk-level DNA methylation data. MICS follows the standard mediation analysis paradigm and consists of three key steps. In step1, we assess the exposure-mediator association for each cell type; in step 2, we assess the mediator-outcome association for each cell type; in step 3, we combine the cell-type-specific exposure-mediator and mediator-outcome associations using a multiple testing procedure named MultiMed [1] to identify significant CpGs with cell-type-specific mediation effects. We conduct simulation studies to demonstrate that our method has correct FDR control. We also apply the MICS procedure to the Normative Aging Study (NAS) and identify nine DNA methylation CpG sites in the lymphocytes that might mediate the effect of cigarette smoking on the lung function.

Availability:

An R package to implement MICS is available at:
https://github.com/XiangyuLuo/MICS

Keywords: DNA methylation, mediation analysis, cell-type specific, multiple testing, inverse regression

1. Introduction

Mediation analysis is a useful statistical method for identifying biological pathways from genetic/environmental exposures to health outcomes [2–4], and has been widely used in both social and biomedical sciences. In epigenetic studies, it is of increasing scientific interest to study the mediator role of DNA methylation in the etiology of human diseases [5–10]. To claim a specific DNA methylation CpG site is a mediator in the causal pathway from an exposure to an outcome, the following two conditions must be satisfied simultaneously: (1) the exposure is associated with the mediator; (2) the mediator is associated with the outcome. For a continuous mediator and a continuous outcome, linear structural equation model (LSEM) is the standard method for detecting the presence of mediation effects [2–4,11]. The LSEM consists of two linear regression models: the mediator and the outcome regressions. The mediator regression links the

exposure to the mediator and can be used to assess whether the exposure is associated with the mediator; and the outcome regression links the exposure, the mediator to the outcome and can be used to assess whether the mediator is associated to the outcome. Both the exposure-mediator and mediator-outcome associations are required to be significant to claim the presence of mediation effects by the mediator, referred to as the joint significant test [12].

In epigenome-wide mediation analysis, as one usually needs to test hundreds of thousands of DNA methylation CpG sites simultaneously in order to discover statistically significant CpG sites that might mediate the effect of an exposure on an outcome, multiple testing correction needs to be performed to control the false positive findings. There are a number of multiple testing procedures developed recently for detecting mediation effects. Boca *et al.* (2014) proposed a permutation method to test multiple mediators simultaneously, however this approach is computationally expensive. Zhang *et al.* (2016) proposed to use the sure independent screening procedure [14] and minimax concave penalty [15] techniques to select promising CpG sites first, and then perform the joint significance test to detect the presence of mediation effects. However, Barfield *et al.* (2017) found that the joint significance test is overly conservative and has very low power in genome-wide epigenetic studies after multiple testing correction. Sampson *et al.* (2018) proposed a new multiple testing procedure with improved performance for identifying significant mediators with guaranteed control of the false discovery rate (FDR), building on the theory developed by Bogomolov and Heller (2018). This method has been implemented in the R package named MultiMed.

However, the DNA methylation samples are typically measured at the bulk level, and thus the obtained methylome for each sample actually measures the signals aggregated from distinct cell types [5,17,18]. For example, DNA methylation measurements based on peripheral blood samples essentially measure the DNA methylation levels of a mixture of white blood cells [17]. It is plausible that the genetic/environmental exposure variable only affects the methylation level of a CpG site in some but not all of the cell types, which in turn affects the outcome. Therefore, identification of the exact cell types that mediate the effect of the exposure on the outcome can help us gain novel biological insight into the disease etiology. However, detection of mediating CpG site in a cell-type-specific fashion is challenging because the cell-type-specific methylation levels are typically not measured, but only the bulk-level DNA methylation measurements are available. To the best of our knowledge, no existing methods can achieve this goal.

Recently, several methods have been proposed to detect cell-type-specific CpG sites associated with one phenotype of interest, including a linear regression method with the cell type proportion and phenotype interaction terms [19], a hierarchical deconvolution model [20], and a tensor composition approach [21]. However, those methods are developed for association studies and thus cannot be directly applied to mediation analysis. Therefore, there is a pressing need of statistical methods for testing

the cell-type-specific mediation effects in genome-wide epigenetic studies.

In this paper, we propose a novel statistical method named MICS to discover CpG sites with cell-type-specific mediation effects using only the bulk-level DNA methylation data. Following the standard mediation analysis paradigm, we need to assess the exposure-mediator and mediator-outcome associations for each cell type. Inspired by previous work [19,20], the cell-type-specific exposure-mediator associations can be assessed by combining multiple cell-type-specific latent mediator regression models. A new challenge is that the cell-type-specific mediator-outcome associations cannot be assessed using the same trick. Leveraging the classical Frisch-Waugh-Lovell theorem [22,23], we propose an inverse regression framework to assess the mediator-outcome associations for each cell type. The exposure-mediator and mediator-outcome associations are further combined using a multiple comparison procedure that has correct false discovery rate (FDR) control [1,16]. Our simulation studies show that the MICS procedure has correct FDR control and can identify cell-type-specific mediation effects with good power. An application to the Normative Aging Study (NAS) identifies nine CpG sites in lymphocytes that might mediate the effects of cigarette smoking on the lung function. We have implemented the MICS method into an R package which is now actively maintained and freely available.

The rest of our paper is organized as follows. We first define basic notation and introduce our new method MICS in Section 2. Second, we conduct simulation studies to evaluate the performance of MICS in Section 3. In Section 4, we apply MICS to the NAS data to demonstrate the usefulness of our method. We conclude the paper with discussions in Section 5.

2. Methods

In this section, we present the MICS approach to test cell-type-specific mediation effects of a particular CpG site in the causal pathway from an exposure to an outcome in genome-wide epigenetic studies. Assume that there are m CpG sites and K cell types. In an overview, MICS consists of three steps, as illustrated in Figure 1. In step 1, we calculate p-values for the exposure-mediator associations for each cell type at each CpG site, and obtain an m by K p-value matrix $\mathbf{Pval}^{(1)} = \left(pval_{jk}^{(1)}\right)_{1 \le j \le m,\, 1 \le k \le K}$. In step 2, we calculate p-values for the mediator-outcome associations for each cell type at each CpG site and obtain another m by K p-value matrix $\mathbf{Pval}^{(2)}$. In step 3, we use the MultiMed multiple comparison procedure to combine the two p-value matrices, which can theoretically guarantee the FDR control [1,16].

[Insert Figure 1 here]

2.1 Basic Notation

Denote the measured methylation beta value matrix by $\boldsymbol{O}$ with m CpG sites in rows and n samples in columns. Specifically, O_{ji} is the beta value of the CpG site j in sample i. The beta value O_{ji} is usually a mixture of methylation values from multiple cell types, such as in blood. Denote the proportion of cell-type k in sample i by w_{ki}, which is now routinely calculated using the reference-based –Houseman method [17], implemented in the R package *minfi* [24]. Hence, O_{ji} can be represented as a weighted sum of cell-type-specific methylation beta values with weights being the corresponding sample's cell type proportions. Denote the unmeasured methylation value of CpG site j in cell-type k for sample i by u_{ijk}. We have the following expression

$$O_{ji} = \sum_{k=1}^{K} u_{ijk} w_{ki}. \quad (1)$$

We emphasize that the beta values rather than the M values should be used in MICS because the M values would break the linearity relationship between the aggregated methylation value O_{ji} and the cell-type-specific methylation values u_{ijk} [17], which will substantially increase the model complexity. The exposure vector $\boldsymbol{s} = (s_1, s_2, ..., s_n)$ records the exposure variable of each sample, and the outcome vector is represented by $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. In addition, we use $\boldsymbol{x_i} = (x_{1i}, x_{2i}, ..., x_{Di})$ to represent the D dimensional covariates for the ith sample. Following the standard mediation analysis paradigm [2–4], our proposed MICS procedure also consists of three key steps: (1) assess the exposure-mediator association; (2) assess the mediator-outcome association; (3) combine the exposure-mediator and mediator-outcome associations to assess mediation effects. Since the analysis in the first two steps will be performed for each CpG site separately, we will suppress the CpG site index j for notation simplicity in Sections 2.2 and 2.3.

2.2 MICS Step 1: exposure-mediator association

First, we aim to test for the association between the exposure variable and the methylation levels for each cell type at each CpG site. The unmeasured (latent) cell-type-specific methylation values u_{ik} for the ith sample and the kth cell type can be modeled using the following latent mediator regression

$$u_{ik} = \beta_{0,k} + \beta_{s,k} s_i + \sum_{d=1}^{D} \beta_{x,dk} x_{id} + \epsilon_{ik}, \quad (2)$$

where the random error term ϵ_{ik} is assumed to have mean zero and variance σ_k^2. Note that this regression model cannot be fitted directly because the response variable u_{ik} is not observed in the data. Nevertheless, the association between the exposure variable s and the DNA methylation level in the kth cell type is captured by the regression parameter $\beta_{s,k}$ and can be assessed by testing the following hypothesis

$$H_0: \beta_{s,k} = 0 \ \ vs \ H_1: \beta_{s,k} \neq 0.$$

If the cell-type-specific DNA methylation values u_{ik} in Equation (2) are measured, then it is straightforward to fit the regression model (2) and perform the corresponding hypothesis testing to test whether β_{jk} is zero. However, those cell-type-specific methylation values u_{ik} are typically unmeasured, imposing one major challenge for assessing the association between the exposure variable and the methylation levels of those K cell types. We circumvent this difficulty using the trick as in Zheng *et al.* (2018) and Luo, Yang and Wei (2019).

We note that, given pre-computed cell type proportions w_{ki}, we can connect the measured bulk-level methylation levels O_i to the exposure variable s_i and covariates x_{id} without the need to know those cell-type-specific methylation levels u_{ik}. Specifically, we combine Equations (1) and (2) to obtain (with j suppressed) the following equation

$$O_i = \sum_{k=1}^{K} \beta_{0,k} w_{ki} + \sum_{k=1}^{K} \beta_{s,k} s_i w_{ki} + \sum_{d=1}^{D} \sum_{k=1}^{K} \beta_{x,dk} x_{id} w_{ki} + \sum_{k=1}^{K} w_{ki} \epsilon_{ik} \quad (3)$$

where the error term $\sum_{k=1}^{K} w_{ki}\epsilon_{ik}$ can be shown to satisfy $E(\sum_{k=1}^{K} w_{ki}\epsilon_{ik}) = 0$ and $Var(\sum_{k=1}^{K} w_{ki}\epsilon_{ik}) = \sum_{k=1}^{K} w_{ki}^2 \sigma_k^2$. The model (3) can thus be fitted because O_i is measured and can be treated as the response variable, and $\{w_{ki}: 1 \le k \le K\}$, $\{s_i w_{ki} : 1 \le k \le K\}$, $\{x_{id} w_{ki}: 1 \le k \le K\}$, $(i = 1,...,n)$ are all observed and can be treated as the regressors. To handle the heteroscedasticity of the linear model, we can use the weighted least squares (WLS) to fit the above model. Therefore, we can detect the association between the exposure variable and the kth cell type at a given CpG site by testing whether $\beta_{s,k}$ is zero using the standard t test in linear regression models.
We denote the p-value for testing $H_0: \beta_{s,k} = 0$ as $pval_k^{(1)}$, where the superscript indicates that this is the p-value from step 1. We then perform this t test to scan the whole epigenome for all the m CpG sites and obtain the m by K cell-type-specific p-value matrix **Pval**$^{(1)}$ for the exposure-mediator associations.

2.3 MICS Step 2: mediator-outcome association

In step 2, our goal is to detect any association between the mediator and outcome for each cell type. Specifically, for the kth cell type, we propose to use the following outcome regression model

$$y_i = \gamma_{0,k} + \gamma_{s,k} s_i + \gamma_{u,k} u_{ik} + \sum_{d=1}^{D} \gamma_{x,dk} x_{id} + \delta_{ik}, \quad (4)$$

where the regression parameter $\gamma_{u,k}$ captures the association between the mediator and the outcome. Therefore, the mediator-outcome association can be inferred by testing the following hypothesis

$$H_0: \gamma_{u,k} = 0 \ \ vs\ H_1: \gamma_{u,k} \neq 0.$$

Unfortunately, the cell-type-specific methylation level u_{ik} is unknown, and thus this regression model cannot be fitted, and the corresponding hypothesis testing cannot be

performed. If we apply the same trick as in step 1 by using Equation (1), then the term $\gamma_{u,k}u_{ik}$ in Equation (4) will become $\sum_{k=1}^{K}\gamma_{u,k}u_{ik}w_{ki}$ which still contains the unknown quantity u_{ik} and thus cannot be used for model fitting and hypothesis testing. Leveraging the classical Frisch-Waugh-Lovell theorem [22,23], we observe that testing whether $H_0: \gamma_{u,k} = 0$ is equivalent to testing whether $H_0: \tilde{\gamma}_{y,k} = 0$ in the following inverse regression model

$$u_{ik} = \tilde{\gamma}_{0,k} + \tilde{\gamma}_{s,k}s_i + \tilde{\gamma}_{y,k}y_i + \sum_{d=1}^{D}\tilde{\gamma}_{x,dk}x_{id} + \tilde{\delta}_{ik}, \quad (5)$$

This inverse regression idea has been used previously in genetic association studies [25–29]. Now we can use the same trick as in step 1 by combing Equations (1) and (5), then we obtain the following estimable inverse regression model

$$O_i = \sum_{k=1}^{K}\tilde{\gamma}_{0,k}w_{ki} + \sum_{k=1}^{K}\tilde{\gamma}_{s,k}s_iw_{ki} + \sum_{k=1}^{K}\tilde{\gamma}_{y,k}y_iw_{ki} + \sum_{d=1}^{D}\sum_{k=1}^{K}\tilde{\gamma}_{x,dk}x_{id}w_{ki} + \sum_{k=1}^{K}w_{ki}\tilde{\delta}_{ik}.$$

Similarly, we can use the WLS method to fit this model, and use the t test in the linear model framework to test whether $\tilde{\gamma}_{y,k}$'s are zeros, and –the p-value $pval_k^{(2)}$ for $\tilde{\gamma}_{y,k} = 0$ can be obtained, where the superscript represents step 2. –The collection of $pval_k^{(2)}$ for all the m CpG sites is denoted by the $m \times K$ matrix $\mathbf{Pval}^{(2)}$.

2.4 MICS Step 3: identifying significant mediators

After we obtain the p-values for assessing exposure-mediator and mediator-outcome associations for each cell type across all m CpG sites, we can infer the presence of mediation effects for a particular cell type at a given CpG site using the method developed by Sampson *et al.* (2018) and implemented in the R Bioconductor package MultiMed [30]. Here, we briefly describe this method, and the theoretical properties that guarantee FDR control can be found in the original papers [1,16].

The MultiMed procedure for identifying significant mediators start with two vectors of p-values with the same length m, $P^{(1)} = (p_{11},...,p_{1m})$ for the exposure-mediator associations and $P^{(2)} = (p_{21},...,p_{2m})$ for the mediator-outcome associations. For a given significance level α, define $\omega_{s1} = \{j: p_{1j} < \alpha/2\}$ and $S_1 = C(\omega_{s1})$ where $C(\cdot)$ is the cardinality of a set and $j = 1,...,m$ indexes the CpG site. Similarly, define $\omega_{s2} = \{j: p_{2j} < \alpha/2\}$ and $S_2 = C(\omega_{s2})$. Define the following multiple comparison procedure (MCP)

$$MCP_S = \{j: j \in \omega_{s1} \cap \omega_{s2}, p_{1j} < \frac{0.5\alpha}{S_2}, p_{2j} < 0.5\alpha/S_1\}$$

and then define the subset-adjusted p-value as $p_{sj} = 2\max(S_2p_{1j}, S_1p_{2j})$ if $p_{1j} < \alpha/2$ and $p_{2j} < \alpha/2$; $p_{sj} = 1$ otherwise. We then claim the CpG site $j \in \omega_{s1} \cap \omega_{s2}$

to be a significant mediator if

$$p_{Dj} = \min_{j':p_{sj'} \geq p_{sj}} p_{sj'}/rank(p_{sj'}) \leq \alpha,$$

where p_{Dj} is the FDR-adjusted p-value and $rank(p_{sj})$ is the rank of p_{sj} for the CpG site $j \in \omega_{s1} \cap \omega_{s2}$.

3. Simulation Study

We performed simulation studies to demonstrate the capability of MICS for detecting cell-type-specific mediation effects. We set sample size $n = 400$, the total number of CpG sites $m = 5000$ and the total number of cell types $K = 3$. To illustrate the usefulness of our proposed MICS method over traditional bulk-level mediation analysis, we designed two classes of CpG sites: (1) the mediation effects occur only in cell type 1 for CpG sites 1-5; (2) the mediation effects exist in cell types 1 and 2 for CpG sites 6-10, and the two types of effects have similar magnitudes but in different directions. All the other CpG sites have null mediation effect. For those CpG sites in Class 1, we expect that both MICS and bulk-level approaches can detect them, but MICS can further detect which cell type is actually driving the mediation effect. For those CpG sites in Class 2, since the combined mediation signals from cell types 1 and 2 almost cancel out, we expect that bulk-level approaches fail to identify them, but MICS can successfully detect them in a cell-type-specific fashion.

To mimic the NAS real data, we generate a binary exposure variable s_i from a Bernoulli distribution with success probability 0.5. –We generate a binary covariate x_{i1} from a Bernoulli distribution with success probability 0.5 and a continuous covariate x_{i2} from a uniform distribution on (-0.5, 0.5). Next, we generated cell-type-specific methylation values u_{ijk} $(0 < u_{ijk} < 1)$. For those CpG sites $1 \leq j \leq 5$ of Class 1 in cell type 1, $u_{ij1} = 0.35 + 0.3s_i + 0.03x_{i1} + 0.01x_{i2} + \epsilon$. For those CpG sites $6 \leq j \leq 10$ of Class 2, in cell type 1, we set $u_{ij1} = 0.35 + 0.3s_i + 0.03x_{i1} + 0.01x_{i2} + \epsilon$; and in cell type 2, we set $u_{ij2} = 0.5 - 0.3s_i + 0.03x_{i1} + 0.01x_{i2} + \epsilon$, where the noise terms ϵ is from a uniform distribution on (-0.01, 0.01). In all other cases, s_i has no effect on methylation, $u_{ijk} = \mu_k + 0.03x_{i1} + 0.01x_{i2} + \epsilon$, where $\mu_1 = 0.35, \mu_2 = 0.5$ and $\mu_3 = 0.35$.

Given s_i, x_{i1}, x_{i2} and u_{ijk}, we generate the outcome y_i as follows.

$$\begin{aligned} y_i &= \gamma_0 + \sum_{j=1}^{m}\sum_{k=1}^{K}\gamma_{u,jk}u_{ijk} + \gamma_s s_i + \gamma_{x,1}x_{i1} + \gamma_{x,2}x_{i2} + \delta_i \\ &= 4.3 + 0.8u_{i1,1} + 1.0u_{i2,1} + 0.8u_{i3,1} + 0.8u_{i4,1} + 1.0u_{i5,1} \\ &\quad + 0.6u_{i6,1} + 0.6u_{i7,1} + 0.7u_{i8,1} + 0.8u_{i9,1} + 0.6u_{i10,1} \\ &\quad + 0.7u_{i6,2} + 0.7u_{i7,2} + 0.4u_{i8,2} + 0.9u_{i9,2} + 0.3u_{i10,2} \\ &\quad - 1.5s_i - 0.3x_{i1} - 0.1x_{i2} + \delta_i, \end{aligned}$$

where δ_i is a noise term with standard deviation 0.02. Finally, the cell-type proportion

(w_{1i},w_{2i},w_{3i}) was drawn from Dirichlet distribution with parameters (80,80,40), and the observed bulk-level methylation values O_{ji} was obtained by $\sum_{k=1}^{K} u_{ijk}w_{ki}$.

We conducted both MICS and bulk-level mediation analysis based on the observed data $\{O_{ji}, y_i, s_i, x_{i1}, x_{i2}: 1 \le j \le m, 1 \le i \le n\}$ with FDR less than 0.2. We repeat this experiment 1000 times and summarize the results in Table 1. The reported power of Class 1 CpG sites is calculated as the average powers of those CpG sites 1-5 based on 1000 replications. The power of Class 2 CpG sites is defined similarly. The FDR in cell type k was computed as the average of the proportions of falsely detected CpG sites among the total detected in cell type k.

[Insert Table 1 here]

The MICS procedure has good FDR control as expected. For Class 1 CpG sites, the mediation effects exist only for cell type 1. The bulk-level analysis can also detect Class 1 CpG sites with comparable power, but it cannot tell which cell type is contributing to the mediation effect signal. For Class 2 CpG sites, the overall mediation signals in cell types 1 and 2 are almost canceled out, so the bulk level analysis has very low power to detect them. In contrast, MICS can detect that cell type 1 and 2 have mediation effects with good power.

We further did more simulation studies to assess the sensitivity of MICS to the estimation accuracy of cellular compositions by varying the absolute error from 1% to 5% based on our original simulation set up. In this range, we found that the FDR of MICS is still well controlled at nominal level. The largest change of power is about 2% (on absolute scale) for cell type 1 and is about 3.5% change for cell type 2. Therefore, the MICS is robust to the cell proportion estimations within 5% absolute error range which can be achieved using the Houseman's method [17]. Therefore, we recommend that rare (less than 5%) cell types should be aggregated when using MICS to produce more stable and accurate results.

4. An Application to the Normative Aging Study

In this section, we applied MICS to the Normative Aging Study (NAS) to test the cell-type-specific mediation effects of DNA methylation in the causal pathway from smoking behavior to lung function. The NAS is an ongoing prospective cohort study established in Eastern Massachusetts in 1963 by the U.S. Department of Veteran Affairs (VA) [31]. The men were free of known chronic medical conditions at enrollment and returned for on-sites, follow-up visits every 3-5 years. During these visits, detailed physical examinations were performed, bio-specimens including blood were obtained, and questionnaire data pertaining to diet, smoking status, and additional lifestyle factors that may impact health were collected. DNA methylation was measured using the Illumina Infinium HumanMethylation450 BeadChips on blood samples collected after an overnight fast [32]. Briefly, we removed any probes on the sex chromosomes, non-

CG probes, and probes within 10 bp of a known SNP within the 1000 genomes population. The batch effects were adjusted by the ComBat algorithm [33]. More information about the preprocessing and collection of the data has been published elsewhere [34]. The mediation analysis was done on a total of 449,547 probes from 607 men. The methylation beta-values ranging from 0 (no methylation) to 1 (full methylation) was calculated for each CpG site. The exposure was a binary variable smoking status (current or former smokers versus never smokers), and the outcome was the forced expiratory flow at 25%-75% of the forced expiratory vital capacity ($FEF_{25-75\%}$). We transformed $FEF_{25-75\%}$ using squared root to achieve better normality. We adjusted for age, height, weight, education history, medication history to control for potential confounding bias.

In this data example, we categorized white blood cells into three main cell types, lymphocyte, monocyte and granulocyte based on the white blood cell lineage relationships. Other cell types with very small proportions were removed to improve numerical stability and accuracy. Cell type proportions were estimated using the Houseman's method based on a set of informative CpG sites [17].

We then applied the proposed MICS procedure to our data set. With FDR less than 0.2, we detected the following nine significant CpG mediators in the lymphocytes: cg03031959 (on gene RASEF of chromosome 9, q=0.11), cg03867465 (on gene MGC45800 of chromosome 4, q=0.11), cg06691963 (on gene FOXP1, q=0.11), cg06926934 (on genes MIR1469 and NR2F2 of chromosome 15, q=0.11), cg09832443 (on gene CREB5, q=0.11), cg15160198 (on gene CYP7B1 of chromosome 8, q=0.11), cg18794577 (on gene GRIN3A of chromosome 9, q=0.11), cg04376185 (on gene SCML1 of chromosome X, q=0.13) and cg09727046 (on gene MIR548H4 of chromosome 15, q=0.13). We did not find any significant CpG sites for monocytes and granulocytes. Detailed annotation information of the detected CpG sites in lymphocytes is provided in Table 2.

[Insert Table 2 here]

For cg03031959, the gene RASEF has been found to be a diagnostic biomarker for lung cancer [35] and is also associated with other lung-related diseases such as chronic obstruct pulmonary disease [36], cigarette smoke-induced pulmonary arterial smooth muscle remodeling [37] and asthma [38]. For cg06691963, the corresponding gene FOXP1 controls and regulates lung development [41,51] and plays a role in non-small cell lung cancer [42]. The biological findings for other significant CpG sites are summarized in the last column of Table 2.

The computation time for the MICS depends on the sample size and the number of CpG sites. In our data set, the sample size is about 600, and the number of CpG sites is 449,547. The total running time is about 2 hours on a 2.5 GHz processor laptop with 16 G memory.

5. Conclusion

In this paper, we first reviewed existing mediation analysis methods in genome-wide epigenetic studies, and then we proposed a new method named MICS to detect cell-type specific mediation effects using the bulk DNA methylation measurements. The standard LSEM cannot be directly applied here because the cell-type specific DNA methylation levels are not available, and the traditional mediator and outcome regression models cannot be fitted. We borrowed the idea from Zheng *et al.* (2018) and Luo, Yang and Wei (2019) which combines multiple cell-type specific latent mediator regressions together so that the combined regression model can be fitted. The same trick cannot be directly used for the mediator-outcome regression because the unknown cell-type specific methylation levels are on the right-hand size of the regression model. We propose to use the inverse regression approach by switching the positions of the outcome and mediator in the regression model. By the classical Frisch-Waugh-Lovell theorem [22,23], this inverse regression approach yields the same inference result for the mediator-outcome associations. The mediation effects are finally assessed by combining the exposure-mediator and mediator-outcome associations for each cell type using the MultiMed approach which guarantees the FDR control [1]. We conduct simulation studies to demonstrate that the proposed MICS procedure has correct FDR control and has good power to detect cell-type specific mediation effects. We further applied our method to the ongoing NAS data and identified nine CpG sites in the lymphocytes that might mediate the effect of smoking on the lung function.

The framework adopted by MICS can be easily extended to other types of genomic data that is similar to the methylation data in terms of being mixed from heterogeneous cell populations, for example, bulk RNA-seq data. In addition, it is future work to adapt the MICS framework to more general settings, for example, when some variables are missing or multivariate exposures. Regarding the practical usage, MICS is accompanied by a user-friendly and computationally efficient R package and can be routinely used for detecting cell-type-specific mediation CpG sites in genome-wide epigenetic studies.

Key Points

- We proposed a method named MICS that can detect cell-type specific mediation effects of a CpG site using only the bulk-level DNA methylation measurements without knowing the cell-type specific methylation levels.
- Simulation study shows that MICS has correct FDR control and good power to detect cell-type specific mediation effects. An application to the NAS data set identified nine CpG sites in the lymphocytes that might mediate the effect of smoking on the lung function.

- Our framework can be extended to other genomic data types, such as the RNA-seq data, to study the mediator roles of gene expressions from genetic/environmental exposures to health outcomes.

Acknowledgements

We thank the High-performance Computing Platform of Renmin University of China for providing computing resources.

Conflict of interest

The authors declare no conflict of interests.

References:

1. Sampson JN, Boca SM, Moore SC, et al. FWER and FDR control when testing multiple mediators. Bioinformatics 2018; 34:2418–2424
2. Baron RM, Kenny DA. The moderator--mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. 1986; 51:1173
3. MacKinnon D. Introduction to statistical mediation analysis. 2012;
4. VanderWeele T. Explanation in causal inference: methods for mediation and interaction. 2015;
5. Liu Y, Aryee MJ, Padyukov L, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat. Biotechnol. 2013; 31:142–147
6. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat. Commun. 2015; 6:
7. Ventham NT, Kennedy NA, Adams AT, et al. Integrative epigenome-wide analysis demonstrates that {DNA} methylation may mediate genetic risk in inflammatory bowel disease. Nat. Commun. 2016; 7:13507
8. Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics 2016; 32:3150–3154
9. Barfield R, Shen J, Just AC, et al. Testing for the indirect effect under the null for genome-wide mediation analyses. Genet. Epidemiol. 2017; 41:824–833
10. Jamieson E, Korologou-Linden R, Wootton RE, et al. Smoking, DNA Methylation, and Lung Function: a Mendelian Randomization Analysis to Investigate Causal Pathways. Am. J. Hum. Genet. 2020; 106:315–326
11. Imai K, Keele L, Yamamoto T. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. Stat. Sci. 2010; 25:51–71
12. MacKinnon DP, Lockwood CM, Hoffman JM, et al. A comparison of methods to

test mediation and other intervening variable effects. Psychol. Methods 2002; 7:83
13. Boca SM, Sinha R, Cross AJ, et al. Testing multiple biological mediators simultaneously. Bioinformatics 2014; 30:214–220
14. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B (Statistical Methodol.) 2008; 70:849–911
15. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 2010; 38:894–942
16. Bogomolov M, Heller R. Assessing replicability of findings across two studies of multiple features. Biometrika 2018; 105:505–516
17. Houseman EA, Accomando WP, Koestler DC, et al. {DNA} methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012; 13:86
18. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014; 15:R31
19. Zheng SC, Breeze CE, Beck S, et al. Identification of differentially methylated cell types in epigenome-wide association studies. Nat. Methods
20. Luo X, Yang C, Wei Y. Detection of cell-type-specific risk-{C}p{G} sites in epigenome-wide association studies. Nat. Commun. 2019; 10:3113
21. Rahmani E, Schweiger R, Rhead B, et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nat. Commun. 2019; 10:3417
22. Frisch R, Waugh F V. Partial time regressions as compared with individual trends. Econom. J. Econom. Soc. 1933; 387–401
23. Lovell MC. Seasonal adjustment of economic time series and multiple regression analysis. J. Am. Stat. Assoc. 1963; 58:993–1010
24. 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 2014; 30:1363–9
25. O'Reilly PF, Hoggart CJ, Pomyen Y, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One 2012; 7:
26. Yan T, Li Q, Li Y, et al. Genetic association with multiple traits in the presence of population stratification. Genet. Epidemiol. 2013; 37:571–580
27. Wang K. Testing genetic association by regressing genotype over multiple phenotypes. PLoS One 2014; 9:
28. Wu B, Pankow JS. Statistical methods for association tests of multiple continuous traits in genome-wide association studies. Ann. Hum. Genet. 2015; 79:282–293
29. Majumdar A, Witte JS, Ghosh S. Semiparametric allelic tests for mapping multiple phenotypes: Binomial regression and mahalanobis distance. Genet. Epidemiol. 2015; 39:635–650
30. Boca SM, Heller R, Sampson JN. MultiMed: Testing multiple biological mediators simultaneously. 2019;
31. Bell B, Rose CL, Damon A. The Normative Aging Study: An Interdisciplinary and Longitudinal Study of Health and Aging. Aging Hum. Dev. 1972; 3:5–17
32. Bibikova M, Barnes B, Tsan C, et al. High density {DNA} methylation array with single {CpG} site resolution. Genomics 2011; 98:288–295

33. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical {B}ayes methods. Biostatistics 2007; 8:118–127
34. Panni T, Mehta AJ, Schwartz JD, et al. Genome-wide analysis of DNA methylation and fine particulate matter air pollution in three study populations: KORA F3, KORA F4, and the normative aging study. Environ. Health Perspect. 2016; 124:983–990
35. Oshita H, Nishino R, Takano A, et al. RASEF is a novel diagnostic biomarker and a therapeutic target for lung cancer. Mol. Cancer Res. 2013; 11:937–951
36. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat. Genet. 2019; 51:494–505
37. Li Q, Wu J, Xu Y, et al. Role of RASEF hypermethylation in cigarette smoke-induced pulmonary arterial smooth muscle remodeling. Respir. Res. 2019; 20:52
38. Hernandez-Pacheco N, Pino-Yanes M, Flores C. Genomic predictors of asthma phenotypes and treatment response. Front. Pediatr. 2019; 7:6
39. Park S-M, Choi E-Y, Bae D-H, et al. The LncRNA EPEL promotes lung cancer cell proliferation through E2F target activation. Cell. Physiol. Biochem. 2018; 45:1270–1283
40. Shinjo K, Okamoto Y, An B, et al. Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma. Carcinogenesis 2012; 33:1277–1285
41. Shu W, Lu MM, Zhang Y, et al. Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. Development 2007; 134:1991–2000
42. Feng J, Zhang X, Zhu H, et al. High expression of FoxP1 is associated with improved survival in patients with non--small cell lung cancer. Am. J. Clin. Pathol. 2012; 138:230–235
43. Baribault C, Ehrlich KC, Ponnaluri VKC, et al. Developmentally linked human DNA hypermethylation is associated with down-modulation, repression, and upregulation of transcription. Epigenetics 2018; 13:275–289
44. Kim W, Giannikou K, Dreier JR, et al. A genome-wide association study implicates NR2F2 in lymphangioleiomyomatosis pathogenesis. Eur. Respir. J. 2019; 53:
45. Leonard MO, Howell K, Madden SF, et al. Hypoxia selectively activates the CREB family of transcription factors in the in vivo lung. Am. J. Respir. Crit. Care Med. 2008; 178:977–983
46. Zhao YD, Yun HZH, Peng J, et al. De novo synthesize of bile acids in pulmonary arterial hypertension lung. Metabolomics 2014; 10:1169–1175
47. Hiramitsu S, Ishikawa T, Lee W-R, et al. Estrogen receptor beta-mediated modulation of lung cancer cell proliferation by 27-hydroxycholesterol. Front. Endocrinol. (Lausanne). 2018; 9:470
48. Jia J, Conlon TM, Sarker RSJ, et al. Cholesterol metabolism promotes B-cell positioning during immune pathogenesis of chronic obstructive pulmonary disease. EMBO Mol. Med. 2018; 10:
49. Ma JZ, Payne TJ, Li MD. Significant association of glutamate receptor, ionotropic

N-methyl-D-aspartate 3A (GRIN3A), with nicotine dependence in European-and African-American smokers. Hum. Genet. 2010; 127:503–512

50. Song M-A, Freudenheim JL, Brasky TM, et al. Biomarkers of Exposure and Effect in the Lungs of Smokers, Nonsmokers, and Electronic Cigarette Users. Cancer Epidemiol. Prev. Biomarkers 2020; 29:443–451
51. Li S, Wang Y, Zhang Y, et al. Foxp1/4 control epithelial cell fate during lung development and regeneration through regulation of anterior gradient 2. Development 2012; 139:2500–2509

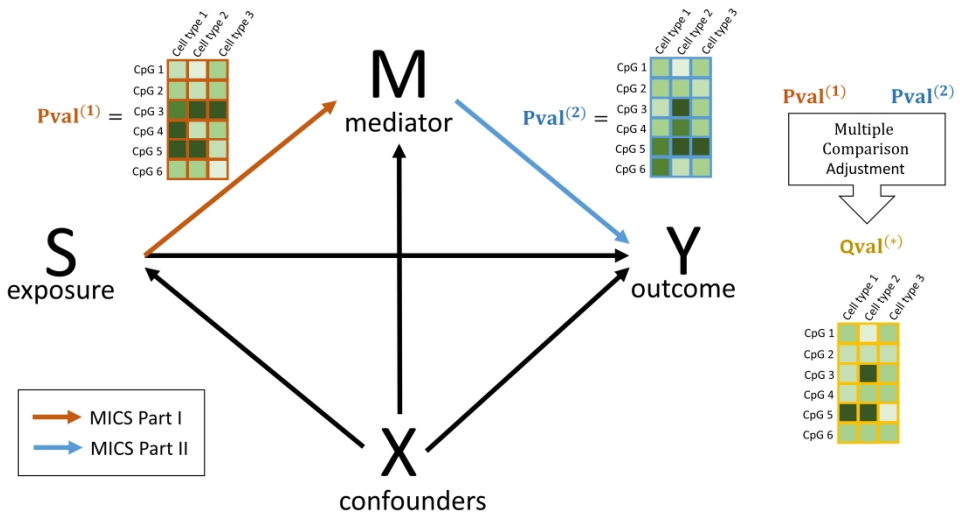

Figure 1. An illustration of the three steps of the proposed MICS method. Step 1: MICS calculates the p-value matrix for the associations between the exposure and the mediator for each cell type. Step 2: MICS calculates the p-value matrix for the associations between the mediators on the outcome for each cell type. Step 3: Combine the cell-type specific exposure-mediator and mediator-outcome p-value matrix using the multiple testing procedure MultiMed.

338x190mm (300 x 300 DPI)

Table 1. Cell-type-specific power and FDR of MICS based on 1000 replications.

	Cell type 1	Cell type 2	Cell type 3
Power on Class 1 CpG sites	0.901	NA	NA
Power on Class 2 CpG sites	0.88	0.94	NA
FDR	0.01	0.02	0.01

Table 2. Nine CpG mediators in the lymphocytes detected by MICS with annotation information.

CpG site ID	q-value	CHR	location	gene	island	previous findings
cg03031959	0.11	9	85677921	RASEF	chr9:85677015-85678321	[35–38]
cg03867465	0.11	4	183062301	MGC45800	chr4:183062278-183062481	[39,40]
cg06691963	0.11	3	71149599	FOXP1	NA	[41,42]
cg06926934	0.11	15	96875675	MIR1469; NR2F2	chr15:96873408-96877721	[43,44]
cg09832443	0.11	7	28612680	CREB5	NA	[45]
cg15160198	0.11	8	65711369	CYP7B1	chr8:65710990-65711722	[46–48]
cg18794577	0.11	9	104501030	GRIN3A	chr9:104499849-104501076	[49]
cg04376185	0.13	X	17755081	SCML1	chrX:17755053-17756648	None
cg09727046	0.13	15	69366554	MIR548H4	chr15:69366321-69366796	[50]