The Annals of Applied Statistics
2022, Vol. 16, No. 1, 521–536
https://doi.org/10.1214/21-AOAS1511

IDENTIFYING INTERGENERATIONAL PATTERNS OF CORRELATED METHYLATION SITES

BY XICHEN MOU[1,a], HONGMEI ZHANG[1,b] AND S. HASAN ARSHAD[2,3,c]

[1]*Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis,* [a]*xmou@memphis.edu,* [b]*hzhang6@memphis.edu*

[2]*Allergy and Clinical Immunology, University of Southampton,* [c]*s.h.arshad@soton.ac.uk*

[3]*The David Hide Asthma and Allergy Research Centre, Isle of Wight*

DNA methylation can be transmitted through generations. This paper proposes a clustering method to identify the intergenerational patterns from parents to their offspring. Motivated by the potential of correlation between DNA methylation sites, we use the multivariate generalized beta distribution to model the blockwise correlation structure among the sites. A stochastic EM algorithm is implemented to estimate the parameters, and BIC is applied to determine the optimal number of clusters. Simulations demonstrate the feasibility of the proposed method. We further applied the approach to cluster DNA methylation data generated from a cohort study on asthma and allergic conditions.

1. Introduction. Epigenetics represent DNA modifications that do not change DNA sequences but do influence gene activities. Epigenetic changes such as DNA methylation (DNAm) is potentially inheritable. Different from the intergenerational transmission of inherent properties such as eye color, environmental influences, such as exposure to famine, and anxiety may contribute to heritable epigenetic modifications (Stenz et al. (2018)).

As a common type of epigenetic modification, DNA methylation is a biochemical process that Methyl groups are added to certain segments of DNA molecules. In mammals the study of DNA methylation focuses on 5-CG-3 dinucleotides (CpG sites) which are distributed unevenly on the DNA sequence. Like other epigenetic modifications, there is evidence that DNA methylation can be transmitted through generations (Padmanabhan et al. (2013)). The transmission mechanism has not been fully understood, but the importance of DNA methylation transmission has been recognized in different fields. For instance, in genome-wide association studies of allergic asthma, genetic effects can only explain a rather small proportion of disease risk and fail to explain the observed heritability of allergic phenotypes completely; DNA methylation has the potential to make up for this "missing heritability" (Lee, Park and Park (2011), Lockett et al. (2013)).

Some effort has contributed to understanding DNA methylation transmission patterns. For instance, the methods developed by Han et al. (2015), referred to as HAN for the rest of the article, is based on a beta regression to examine DNA methylation inheritance heterogeneity at different CpG sites between mothers, fathers, and their offspring. HAN grouped CpG sites based on beta regression coefficients. Given one CpG site, HAN built a linear relationship to bridge the population mean of DNA methylation levels between parents and offspring and evaluated the inheritance strength based on the magnitude of regression coefficients. One basic assumption of HAN is the mutual independence of all CpG sites. However, when CpG sites on a chromosome are close in distance, this independence assumption can hardly hold; see, for example, Bell et al. (2011), Eckhardt et al. (2006). The correlation of CpG sites in DNAm has been utilized in DNAm prediction (Zhang et al. (2015)) and imputation (Yu

Received September 2020; revised April 2021.
Key words and phrases. DNA methylation, EM, clustering, generalized beta distribution.

et al. (2020)). To take into account correlations between CpG sites, this study employed the multivariate generalized beta distribution to model the blockwise correlation structure among CpG sites. We propose a stochastic expectation-maximization (EM) algorithm to estimate the parameters and apply the Bayesian information criterion (BIC) to determine the number of clusters and blocks. The proposed method generalizes HAN and thus owns its advantages, such as no reliance on the assumption of normality for DNAm levels, and the ability to evaluate inheritance strength in comparison to alternative methods (Park and Jun (2009), Qin and Self (2006)). By accounting for possible correlations among CpG sites, the proposed model has a better fit to the data. Consequently, the clustering process built upon the proposed model is more sensitive to underlying DNA methylation transmission patterns from one generation to the next.

One important step of utilizing the correlation structure is to partition the correlated CpG sites into blocks. We propose two block-partition approaches: One is focusing on the highly correlated CpG sites that are close in terms of chromosomal coordinates. More generally, CpGs far from each other with respect to coordinates may still be close to each other spatially. Thus, we propose the other approach by removing the constraint of proximity in distance and basing on the observed correlation to find blocks.

The rest of the article is arranged as follows: In Section 2 we present the methodology, including notations, assumptions, and a stochastic EM algorithm for model estimation. In Section 3 we discuss the approach to determine the number of clusters and blocks and then provide six simulation settings to compare the proposed model to HAN. In Section 4 we assess our method on a DNA methylation dataset generated from a cohort study on asthma and allergic conditions conducted on the Isle of Wight, United Kingdom.

2. Methodology.

2.1. *Notations and assumptions.* In this study we define a triad to be a group consisting of two parents and one offspring. Suppose there are I triads. For each member in a triad, we observe J CpG sites which are further divided into B nonoverlapping blocks where CpG sites in a block are correlated in DNA methylation. The length of each block is allowed to be different: There are L_b CpG sites in the bth block, where $b = 1, \dots, B$ and $\sum_{b=1}^{B} L_b = J$. The methylation level of a CpG site was measured by the beta value $M/(M + U + c)$, where M and U are signal intensities of methylated and unmethylated probes, and c (usually equals 100 by default) is an offset constant (Du et al. (2010)). The higher the beta value, the higher the CpG site is methylated. Let $Z1_{ibl}$, $Z2_{ibl}$, and $Z0_{ibl}$ denote the mother's, father's, and offspring's methylation level of the lth CpG site in the bth block of the ith triad, where $0 < Z1_{ibl}, Z2_{ibl}, Z0_{ibl} < 1$, $i = 1 \dots I$, $l = 1 \dots L_b$. In the ith triad, let $\mathbf{Z1}_{ib} = (Z1_{ib1}, \dots, Z1_{ibL_b})$ denote the mother's methylation of CpGs in the bth block. With a little abuse of notation, denote by $\mathbf{Z1}_i = (\mathbf{Z1}_{i1}, \dots, \mathbf{Z1}_{iB})$ the methylation of CpGs in all B blocks, so are $\mathbf{Z2}_{ib}$ and $\mathbf{Z2}_i$ for father, and $\mathbf{Z0}_{ib}$ and $\mathbf{Z0}_i$ for offspring. We further assume CpG sites are independent between blocks while dependent within each block (see Figure 1 for an example of data structure for the ith triad).

To account for within-block dependence, we introduce the multivariate generalized beta distribution specified in Libby and Novick (1982).

DEFINITION 2.1. A random vector $\mathbf{R}$ follows an L-variate generalized beta distribution, denoted by GBeta$(\boldsymbol{\alpha}, \beta)$, when

$$\mathbf{R} = (R_1, \dots, R_L) = \left(\frac{P_1}{P_1 + Q}, \dots, \frac{P_L}{P_L + Q}\right)$$

in which $P_l \sim \text{Gamma}(\alpha_l, 1)$ and $Q \sim \text{Gamma}(\beta, 1)$, where $l = 1 \dots, L$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L)$; P_l's and Q are mutually independent.

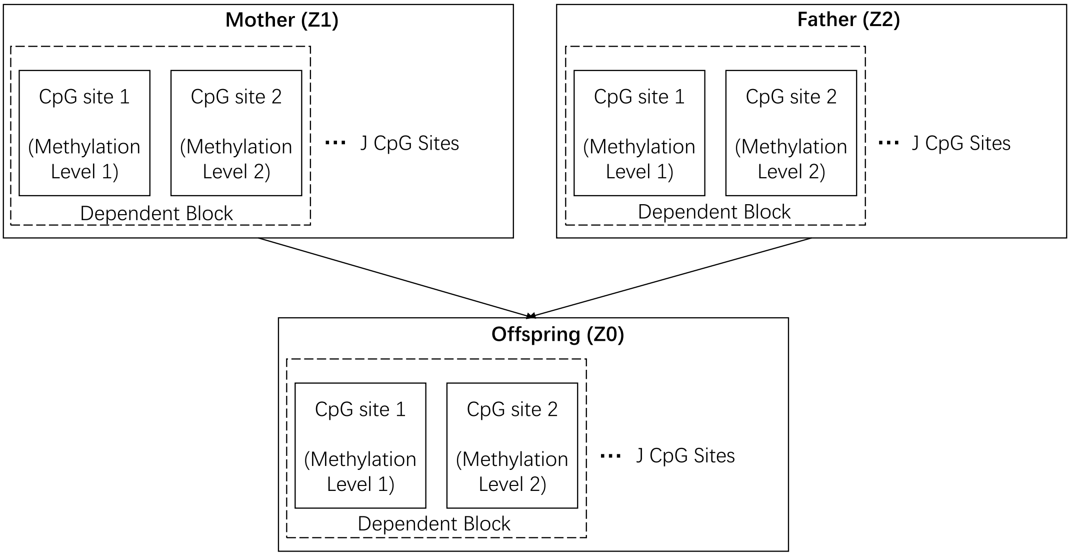

FIG. 1. *An example of data structure of the ith triad. In this example we assume there are two CpG sites in the first block.*

The closed-form density function of $\mathbf{R} \sim \text{GBeta}(\boldsymbol{\alpha}, \beta)$ can be expressed as

$$f(\mathbf{R}|\boldsymbol{\alpha},\beta) = \frac{\Gamma(\sum_{l=1}^{L}\alpha_l+\beta)\prod_{l=1}^{L}\{(\frac{R_l}{1-R_l})^{\alpha_l-1}(\frac{1}{1-R_l})^2\}}{\Gamma(\beta)\prod_{l=1}^{L}\Gamma(\alpha_l)\{1+\sum_{l=1}^{L}(\frac{R_l}{1-R_l})\}^{\sum_{l=1}^{L}\alpha_l+\beta}}, \tag{1}$$

where α_l's and β are positive. We propose to use $f(\mathbf{R}|\boldsymbol{\alpha},\beta)$ to model the distribution of DNA methylation for CpG sites in block b. There are two distinct advantages of this utilization: (1) The marginal distribution of DNA methylation at each CpG site is still beta distribution, which is consistent with prior literature, for example, Houseman et al. (2008); (2) The CpG sites within each block are dependent. Following this definition, the DNAm levels of the L_b CpG sites of the bth block for mother, father, and offspring in the ith triad are denoted by

$$\begin{aligned} \mathbf{Z1}_{ib} &\sim \text{GBeta}(\boldsymbol{\alpha}_b^M, \beta_b^M), \\ \mathbf{Z2}_{ib} &\sim \text{GBeta}(\boldsymbol{\alpha}_b^F, \beta_b^F), \\ \mathbf{Z0}_{ib} &\sim \text{GBeta}(\boldsymbol{\alpha}_b^O, \beta_b^O), \end{aligned} \tag{2}$$

where $\mathbf{Z1}_{ib}$, $\mathbf{Z2}_{ib}$, $\mathbf{Z0}_{ib}$ are random vectors of length L_b denoting mother, father, and offspring's DNAm levels; $\boldsymbol{\alpha}_b^M = (\alpha_{b1}^M, \dots, \alpha_{bL_b}^M)$, $\boldsymbol{\alpha}_b^F = (\alpha_{b1}^F, \dots, \alpha_{bL_b}^F)$, $\boldsymbol{\alpha}_b^O = (\alpha_{b1}^O, \dots, \alpha_{bL_b}^O)$; $\alpha_{bl}^M, \alpha_{bl}^F, \alpha_{bl}^O, \beta_{bl}^M, \beta_{bl}^F, \beta_{bl}^O > 0$; $l = 1, \dots, L_b$; $b = 1, \dots, B$.

2.2. *The clustering method.* Now, we present our method to group CpG sites, based on the inheritance pattern passing from parents to their offspring. Same as in HAN, given one CpG site, for example, the lth CpG site of the bth block, we assume a linear relationship in the logit mean of the methylation level between parents and offspring, which can be expressed as

$$O_{bl} = \gamma_0 + \gamma_1 M_{bl} + \gamma_2 F_{bl},$$

where $O_{bl} = \log(\alpha_{bl}^O) - \log(\beta_{bl}^O)$, $M_{bl} = \log(\alpha_{bl}^M) - \log(\beta_{bl}^M)$, $F_{bl} = \log(\alpha_{bl}^F) - \log(\beta_{bl}^F)$, and γ_0, γ_1, γ_2 are coefficients. The inheritance pattern is fully determined by the vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$, and the inheritance strength can be evaluated by the magnitude of γ_1 and γ_2: If $|\gamma_1| > |\gamma_2|$, the offspring's methylation inherits more from mother and vice versa. If $\gamma_1 =$

$\gamma_2 = 0$, the offspring's methylation has no relationship with that of parents at the population level.

Since it is possible that some CpG sites follow a similar inheritance pattern, we further assume $\boldsymbol{\gamma}$ is equal to one of the K vectors, denoted by $\boldsymbol{\gamma}_k = (\gamma_{0k}, \gamma_{1k}, \gamma_{2k})$ for a certain number of CpG sites, where $k = 1, \dots, K$ and $K \ll J$. That is, we cluster the J CpG sites into K clusters, and, if the lth CpG site of the bth block has a intergenerational pattern k, then we have

$$O_{bl} = \gamma_{0k} + \gamma_{1k} M_{bl} + \gamma_{2k} F_{bl}. \tag{3}$$

Given the observations $\mathbf{Z1}_i$, $\mathbf{Z2}_i$, and $\mathbf{Z0}_i$, our goal is to label each CpG site with one of the K transmission patterns correctly and estimate the coefficients $\boldsymbol{\gamma}_k$ for $k = 1, \dots, K$.

We propose an EM algorithm to achieve this goal. We start by introducing the latent cluster assignment S_{bl}, where $b = 1, \dots, B$ and $l = 1, \dots, L_b$; $S_{bl} = k$ means the lth CpG site in the bth block has a transmission pattern k. Let π_k denote the probability that one CpG site is in a cluster with transmission pattern k, that is, $P(S_{bl} = k) = \pi_k$. Note that π_k is independent of the index b and l. Utilizing the definitions in (2), let $\mathcal{A}^M = (\boldsymbol{\alpha}_1^M, \dots, \boldsymbol{\alpha}_B^M)$, $\mathcal{A}^F = (\boldsymbol{\alpha}_1^F, \dots, \boldsymbol{\alpha}_B^F)$, $\mathcal{A}^O = (\boldsymbol{\alpha}_1^O, \dots, \boldsymbol{\alpha}_B^O)$, $\mathcal{A} = (\mathcal{A}^M, \mathcal{A}^F, \mathcal{A}^O)$. Similarly, let $\mathcal{B}^M = (\beta_1^M, \dots, \beta_B^M)$, $\mathcal{B}^F = (\beta_1^F, \dots, \beta_B^F)$, $\mathcal{B}^O = (\beta_1^O, \dots, \beta_B^O)$, $\mathcal{B} = (\mathcal{B}^M, \mathcal{B}^F, \mathcal{B}^O)$; $\boldsymbol{\gamma}_k = (\gamma_{0k}, \gamma_{1k}, \gamma_{2k})$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K)$; $\boldsymbol{\theta} = (\mathcal{A}, \mathcal{B}, \boldsymbol{\gamma})$, the ensemble of all parameters mentioned above. Let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$. Lastly, denote by $\mathbf{S}_b = (S_{b1}, \dots, S_{bL_b})$, the cluster assignments of L_b CpG sites in the bth block; $\mathbf{S} = (\mathbf{S}_1, \dots, \mathbf{S}_B)$, the cluster assignments of all CpGs across block $b = 1, \dots, B$; $\mathcal{Z} = (\mathbf{Z1}_i, \mathbf{Z2}_i, \mathbf{Z0}_i)$, the ensemble of all observed DNAm levels across I families, where $i = 1, \dots, I$. Recalling that we write $\mathbf{Z0}_i = (\mathbf{Z0}_{i1}, \dots, \mathbf{Z0}_{iB})$, the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\pi}; \mathcal{Z}, \mathbf{S})$ can be expressed as

$$\begin{aligned}
&P(\mathbf{S}|\boldsymbol{\pi}) \times \prod_{i=1}^{I} P(\mathcal{Z}|\mathbf{S}, \boldsymbol{\theta}) \\
&\quad = \prod_{b=1}^{B} \prod_{l=1}^{L_b} \pi_{S_{bl}} \times \prod_{i=1}^{I} \left\{ P(\mathbf{Z1}_i|\mathcal{A}^M, \mathcal{B}^M) P(\mathbf{Z2}_i|\mathcal{A}^F, \mathcal{B}^F) \prod_{b=1}^{B} P(\mathbf{Z0}_{ib}|S_{b1}, \dots, S_{bL_b}, \boldsymbol{\theta}) \right\} \\
&\quad = \prod_{b=1}^{B} \prod_{l=1}^{L_b} \prod_{k=1}^{K} \pi_k^{\mathbb{I}(S_{bl}=k)} \times \prod_{i=1}^{I} \{ P(\mathbf{Z1}_i|\mathcal{A}^M, \mathcal{B}^M) P(\mathbf{Z2}_i|\mathcal{A}^F, \mathcal{B}^F) \} \\
&\quad \quad \times \prod_{i=1}^{I} \prod_{b=1}^{B} \prod_{k_1=1}^{K} \cdots \prod_{k_{L_b}=1}^{K} P(\mathbf{Z0}_{ib}|k_1, \dots, k_{L_b}, \boldsymbol{\theta})^{\mathbb{I}(S_{b1}=k_1, \dots, S_{bL_b}=k_{L_b})},
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The last term

$$P(\mathbf{Z0}_{ib}|k_1, \dots, k_{L_b}, \boldsymbol{\theta}) = f(\mathbf{Z0}_{ib}|\boldsymbol{\alpha}_b^O, \beta_b^O),$$

where f is the density of generalized beta distribution specified in (1). The parameters $\boldsymbol{\alpha}_b^O$ and β_b^O are dependent on $\boldsymbol{\alpha}_b^M$, β_b^M, $\boldsymbol{\alpha}_b^F$, β_b^F, and $k_1, \dots, k_{L_b}$ through the link function (3). In the Supplementary Material (Mou, Zhang and Arshad (2022a)), we show that the likelihood is reduced to HAN if assuming independence over all CpG sites, that is, $L_b = 1$ for $b = 1, \dots, B$.

The common way to label the latent cluster assignments is the EM algorithm, where the cluster assignments S_{bl}'s are treated as "missing data." However, challenges arise when evaluating the Q function

$$\mathbf{Q}(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) = E[\log\{L(\boldsymbol{\theta}, \boldsymbol{\pi}; \mathcal{Z}, \mathbf{S})\}|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}] \tag{4}$$

in the Expectation step, where $\boldsymbol{\theta}^{(t-1)}$ and $\boldsymbol{\pi}^{(t-1)}$ is the estimation in previous iteration. The Q function involves the calculation of

$$\sum_{k_1=1}^{K}\cdots\sum_{k_{L_b}=1}^{K} E\big[\mathbb{I}(S_{b1}=k_1,\dots,S_{bL_b}=k_{L_b})\log\big\{P(\mathbf{Z0}_{ib}|k_1,\dots,k_{L_b},\boldsymbol{\theta})\big\}|\mathcal{Z},\boldsymbol{\theta}^{(t-1)}\big]. \tag{5}$$

This summation contains K^{L_b} terms, which increases exponentially with respect to L_b, and thus is computationally infeasible for large block sizes. A similar challenge is seen in the Q function when evaluating

$$\begin{aligned} &E\big\{\mathbb{I}(S_{bl}=k)\log(\pi_k)|\mathcal{Z},\boldsymbol{\theta}^{(t-1)}\big\} = P\big(S_{bl}=k|\mathcal{Z},\boldsymbol{\theta}^{(t-1)}\big)\log(\pi_k) \\ &\quad = \sum_{k_1=1}^{K}\cdots\sum_{k_{l-1}=1}^{K}\sum_{k_{l+1}=1}^{K}\cdots\sum_{k_{L_b}=1}^{K} P\big(S_{b1}=k_1,\dots,S_{bL_b}=k_{L_b}|\mathcal{Z},\boldsymbol{\theta}^{(t-1)}\big)\log(\pi_k). \end{aligned} \tag{6}$$

To overcome these challenges, we use Markov chain Monte Carlo simulations, in particular, a Gibbs sampler to approximate these values in our algorithm. This technique is often referred to as a stochastic EM algorithm (Nielsen (2000)). Additionally, to improve the computational efficiency we use the maximum likelihood estimation $\hat{\alpha}_{bl}^{M}$, $\hat{\alpha}_{bl}^{F}$, $\hat{\beta}_{b}^{M}$, $\hat{\beta}_{b}^{F}$, $\hat{\beta}_{b}^{O}$ of α_{bl}^{M}, α_{bl}^{F}, β_{b}^{M}, β_{b}^{F}, β_{b}^{O} and $\hat{M}_{bl}=\log(\hat{\alpha}_{bl}^{M})-\log(\hat{\beta}_{b}^{M})$, $\hat{F}_{bl}=\log(\hat{\alpha}_{bl}^{F})-\log(\hat{\beta}_{b}^{F})$ instead of updating them in each iteration. Furthermore, the Q function in (4) is written as $\mathbf{Q}(\boldsymbol{\gamma},\boldsymbol{\pi}|\boldsymbol{\gamma}^{(t-1)},\boldsymbol{\pi}^{(t-1)})$ since $\mathcal{A}$ and $\mathcal{B}$ in $\boldsymbol{\theta}$ are not updated. The detailed algorithm is shown in Algorithm 1.

3. Simulation evidence.

In this section we first discuss the selection of the number of clusters K and the block size L. Then, we use simulations to assess the performance of the proposed method (GBClust, short for generalized beta clustering) and compare it to HAN which assumes CpG sites are mutually independent. It is worth noting that the purpose of this section is to evaluate the finite-sample performance of both methods under various block sizes and correlation strengths. Thus, for conciseness we assume all CpG blocks are of the same size L throughout this section. The block sizes may be different with each other in the real data analysis; we will illustrate the method of partitioning blocks in Section 4.

3.1. *Selection of K and L.*

In this study, BIC is applied to select the optimal K and L. The BIC of our model is given by

$$-2l+(2J+3B+4K-1)\log(3\times I\times J),$$

where $l=\log\{L(\boldsymbol{\theta},\boldsymbol{\pi};\mathcal{Z},\mathbf{S})\}$, $2J+3B+4K-1$ is the number of free parameters and $3\times I\times J$ is the number of observations. We use a grid search to find the optimal K and L corespondent to the minimum BIC value.

To assess whether the proposed BIC criterion could identify the correct K and L, we simulated $J=5000$ CpG sites, each of which was randomly assigned to one of $K=4$ clusters. The coefficients of each cluster $\boldsymbol{\gamma}_1=(0.4,0,0.2)$, $\boldsymbol{\gamma}_2=(-0.01,-0.3,0)$, $\boldsymbol{\gamma}_3=(0.15,0,0)$, and $\boldsymbol{\gamma}_4=(-0.26,-0.2,0.14)$ correspond to patterns where DNA methylation is mainly inherited from father, mother, neither, and both, respectively. The probability that a CpG site falls in each of the four clusters is given by $\boldsymbol{\pi}=(0.25,0.25,0.25,0.25)$, that is, approximately 25% of CpG sites fall in each cluster. The $L=5$ neighboring CpG sites in a block are correlated, following the generalized beta distribution described in Section 2. Parameters of the generalized beta distribution $\mathcal{A}$ and $\mathcal{B}$ are randomly selected such that CpG sites follow different marginal beta distributions. The number of triads $I=50$; both I and J are comparable to the real data analysis in Section 4. In this simulation study we generated 100 Monte

Algorithm 1: The Stochastic EM algorithm

```
Result: $\boldsymbol{\gamma}^{(t)}, \mathbf{S}^{(t)}$
1. Initialize $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\pi}^{(0)}$ and $\mathbf{S}^{(0)}$. Set $t = 1$.
2. For $b = 1, \dots, B$:
   (a) Initialize $\boldsymbol{\alpha}_b^{O(t-1)} = (\alpha_{bl}^{O(t-1)})$, in which $\alpha_{bl}^{O(t-1)}$ satisfies
           $$\log(\alpha_{bl}^{O(t-1)}) - \log(\hat{\beta}_b^O) = \gamma_{0k^*}^{(t-1)} + \gamma_{1k^*}^{(t-1)} \hat{M}_{bl} + \gamma_{2k^*}^{(t-1)} \hat{F}_{bl},$$
       where $l = 1, \dots, L_b$ and $k^* = S_{bl}^{(t-1)}$.
   (b) For $l = 1, \dots, L_b$:
       i. For $k = 1, \dots, K$:
          A. Set $\mathbf{S}_{b,k}^{(t-1)} = (S_{b1}^{(t)}, \dots, S_{b(l-1)}^{(t)}, k, S_{b(l+1)}^{(t-1)}, \dots, S_{bL_b}^{(t-1)})$.
          B. Set $\boldsymbol{\alpha}_{b,k}^{O(t-1)} = (\alpha_{b1}^{O(t)}, \dots, \alpha_{b(l-1)}^{O(t)}, \alpha_{bl,k}^{O(t)}, \alpha_{b(l+1)}^{O(t-1)}, \dots, \alpha_{bL_b}^{O(t-1)})$,
             where $\alpha_{bl,k}^{O(t)}$ satisfies
                 $$\log(\alpha_{bl,k}^{O(t)}) - \log(\hat{\beta}_b^O) = \gamma_{0k}^{(t-1)} + \gamma_{1k}^{(t-1)} \hat{M}_{bl} + \gamma_{2k}^{(t-1)} \hat{F}_{bl}.$$
          C. Calculate
                 $$p_k = \pi_k^{(t-1)} \prod_{i=1}^{I} P(\mathbf{Z0}_{ib} | \mathbf{S}_{b,k}^{(t-1)}, \boldsymbol{\alpha}_{b,k}^{O(t-1)}, \hat{\beta}_b^O)$$
             and $q_k = p_k / \sum_{k=1}^{K} p_k$.
       ii. Sample $S_{bl}^{(t)} \sim \text{Multinomial}(q_1, \dots, q_K)$. Set $\alpha_{bl}^{O(t)} = \alpha_{bl,k^\dagger}^{O(t)}$, where $k^\dagger = S_{bl}^{(t)}$.
3. Approximate (5) and (6) with $\log\{P(\mathbf{Z0}_{ib} | S_{b1}^{(t)}, \dots, S_{bL_b}^{(t)}, \boldsymbol{\theta})\}$ and $\log(\pi_{S_{bl}^{(t)}})$,
   respectively, in $\mathbf{Q}(\boldsymbol{\gamma}, \boldsymbol{\pi} | \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\pi}^{(t-1)})$. Set
       $$(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\pi}^{(t)}) = \operatorname{argmax} \mathbf{Q}(\boldsymbol{\gamma}, \boldsymbol{\pi} | \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}).$$
4. Increase $t$ by 1. Repeat Step 2 to Step 4 until convergence.
```

Carlo (MC) replicates of this setting. In each MC replicate we calculate the BIC values for $K = 2, 3, 4, 5, 6$ and $L = 1, 2, 5, 10, 20, 25$ for computational efficiency (in total, 30 combinations). As an illustration, Figure 2 presents the average BIC values over 100 MC replicates

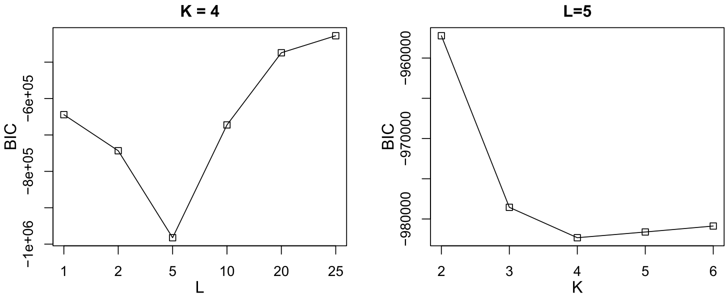

FIG. 2. *Scree plot of BIC at different K's and L's. Left: Given true $K = 4$, the average BIC over* 100 *MC replicates with $L = 1, 2, 5, 10, 20, 25$. Right: Given true $L = 5$, the average BIC over* 100 *MC replicates with $K = 2, 3, 4, 5, 6$.*

TABLE 1
Occurrence frequencies of optimal parameters selected from the BIC method. No values of L other than 5 are selected and thus omitted

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
$L = 5$	0	0	96	4	0

given $K = 4$ and $L = 5$ separately. In the left plot, fixing $K = 4$, BIC drops significantly at the beginning and reaches the lowest at $L = 5$. Similarly, fixing $L = 5$, BIC reaches the minimum at $K = 4$ in the right plot. After checking other K's and L's combinations, the minimum BIC value is achieved at $K = 4$ and $L = 5$ which is the underlying setting in this simulation study. Table 1 summarizes the frequencies of optimal K's and L's in the 100 MC replicates. The result supports using BIC to select K and L: 96 out of 100 MC replicates choose the correct $K = 4$ and $L = 5$, and the remaining four replicates select the correct $L = 5$ but a bit larger $K = 5$. We repeated the same procedure when the underlying $L = 1, 2, 10$ and obtained similar results as when $L = 5$: At least 99 out of 100 MC replicates choose the correct K's and L's in these settings.

3.2. *Model comparison.* Both GBClust and HAN are evaluated in six settings (S1–S6, listed below). The results are summarized in Tables 2 and 3. In each setting, 100 MC replicates are generated. The performance is measured by the mean and standard deviation (SD) of classification accuracy, defined as the number of CpGs assigned to the correct clusters divided by the total number of CpGs J:

TABLE 2
Accuracies of GBClust and HAN in S1–S5

			$L = 1$	$L = 2$	$L = 5$	$L = 10$	$L = 20$
S1	GBClust	Mean	0.8882	0.8911	0.9152	0.9296	0.9411
		SD	0.0047	0.0087	0.0073	0.0214	0.0220
	HAN	Mean	0.8882	0.8175	0.7142	0.7116	0.7097
		SD	0.0047	0.0501	0.0532	0.0343	0.0367
S2	GBClust	Mean	0.9587	0.9639	0.9531	0.9604	0.9752
		SD	0.0228	0.0033	0.0114	0.0097	0.0076
	HAN	Mean	0.9587	0.9554	0.9442	0.9473	0.9381
		SD	0.0228	0.0035	0.0352	0.0221	0.0467
S3	GBClust	Mean	0.8860	0.8931	0.9178	0.9265	0.9441
		SD	0.0272	0.0057	0.0049	0.0514	0.0048
	HAN	Mean	0.8860	0.8226	0.7139	0.7096	0.7121
		SD	0.0272	0.0475	0.0460	0.0421	0.0396
S4	GBClust	Mean	0.8706	0.9024	0.9163	0.9176	0.9200
		SD	0.0751	0.0309	0.0312	0.0465	0.0523
	HAN	Mean	0.8706	0.8638	0.8010	0.7833	0.7741
		SD	0.0751	0.0259	0.0382	0.0348	0.0288
S5	GBClust	Mean	0.8650	0.8391	0.8761	0.8879	0.8908
		SD	0.0187	0.0520	0.0521	0.0567	0.0814
	HAN	Mean	0.8650	0.7371	0.7006	0.6971	0.6986
		SD	0.0187	0.0600	0.0425	0.0393	0.0380

TABLE 3

Accuracies of GBClust and HAN in S6. The $L = 1$ column corresponds to the independent case. The $L = 5$, 10, and 20 columns correspond to cases when the average pairwise correlation $\rho = 0.1$, 0.5, and 0.9, respectively

			$L = 1$		$L = 5$	$L = 10$	$L = 20$
Setting 1	GBClust	Mean	0.7730	$\rho = 0.1$	0.8214	0.8413	0.8361
		SD	0.0088		0.0203	0.0277	0.0463
	HAN	Mean	0.7730		0.7720	0.7761	0.7768
		SD	0.0088		0.0293	0.0394	0.0368
Setting 2	GBClust	Mean	0.7527	$\rho = 0.5$	0.7626	0.7947	0.8193
		SD	0.0318		0.0418	0.0574	0.0503
	HAN	Mean	0.7527		0.5494	0.5475	0.5289
		SD	0.0318		0.0556	0.0595	0.0582
Setting 3	GBClust	Mean	0.7483	$\rho = 0.9$	0.7592	0.8669	0.9227
		SD	0.0225		0.0389	0.0587	0.0430
	HAN	Mean	0.7483		0.6321	0.6325	0.6267
		SD	0.0225		0.0299	0.0209	0.0238

S1. This scenario aims to study the effect of block size L. Given each of $L = 1, 2, 5, 10, 20$, we simulated $J = 5000$ CpG sites with every L neighbouring CpG sites following the generalized beta distribution. Other settings, such as K, I, $\boldsymbol{\pi}$ and coefficients $\boldsymbol{\gamma}_1$ to $\boldsymbol{\gamma}_4$, are the same as in Section 3.1.

S2. More triads I. The setting is the same as S1, except that the number of triads $I = 100$ instead of 50. This scenario aims to evaluate the consistency of both methods as sample size increases.

S3. Different number of CpG sites J. The setting is the same as S1, except that $J = 10,000$ instead of 5000. This scenario aims to evaluate the performance of both methods when a larger number of CpG sites are available.

S4. Unequal number of CpG sites in each cluster. The setting is the same as S1, except that $\boldsymbol{\pi} = (0.4, 0.3, 0.2, 0.1)$ instead of $(0.25, 0.25, 0.25, 0.25)$. This setting intends to check the robustness of both methods when the number of CpG sites in each cluster is unevenly distributed.

S5. More varieties in transmission pattern. We set $K = 5$, adding an extra cluster with $\boldsymbol{\gamma}_5 = (-0.46, -0.1, 0.1)$ to S1. We reassign the probability of the five clusters $\boldsymbol{\pi}$ to be $(0.2, 0.2, 0.2, 0.2, 0.2)$. This scenario aims to assess the robustness of both methods when more transmission patterns exist.

S6. Different correlation strength. Recall that the methylation levels of the L CpG sites follow the generalized beta distribution which describes dependency among CpGs within each block. We measure the blockwise correlation strength by averaging the correlation of each pair of CpG sites within a block. We constructed three settings with different sets of $\boldsymbol{\theta}$'s such that, when $L > 1$, the average pairwise correlation approximately equals $\rho = 0.1, 0.5, 0.9$ from settings 1 to 3. The goal of this scenario is to examine the performance of GBClust and HAN under different strengths of correlations.

In Table 2, when $L = 1$, both methods give the same result since GBClust is trivialized to HAN when all CpG sites are independent. In S1, as L increases, the performance of GBClust improves while HAN's accuracy deteriorates. This is within expectation, since large blocks of correlated CpGs violate the independence assumption of HAN. In practice, we expect to analyze more than 10,000 CpG sites. When $L = 5$, for instance, the average accuracy of GBClust rises by approximately 20 percentage than HAN, according to Table 2. The percentage of increase at the level of 20% can lead to an average of 2000 CpGs correctly classified, assuming in total 10,000 CpGs under investigation. In S2, as I increased from 50 to 100, the performance of GBClust and HAN is improved with increased accuracy in comparison to S1, indicating consistency of both methods. For S3–S5, GBClust demonstrates its robustness, against different scales of J and unbalanced $\boldsymbol{\pi}$, and its ability to detect various transition patterns. As in S1 and S2, GBClust always outperforms HAN.

In Table 3, when the correlation is weak ($\rho = 0.1$), GBClust is improved significantly while HAN stays the same, generally, as L increases. However, in the setting of stronger correlation ($\rho = 0.5, 0.9$), GBClust's accuracies still rise while HAN drops significantly. This pattern also follows our expectation: Stronger correlation means a larger deviation from HAN's independence assumption. It is worth noting that the comparison of Table 3 is "horizontal:" Controlling one of the three settings, we evaluate both methods for different L's. It is meaningless to compare the results of Table 3 "vertically," for example, fixing $L = 10$ to compare the performance across different ρ's. Unlike the horizontal comparison, where the performances are solely dependent on L, in the vertical comparison the accuracies are not only determined by ρ but also by other factors, such as the Euclidean distances between $\boldsymbol{\gamma}_k$'s in each setting. In summary, as shown in Tables 2 and 3, GBClust reduces to HAN when $L = 1$. When dependence exists ($L \geq 2$), although results from HAN are acceptable in various situations, GBClust uniformly outperforms HAN across all the settings.

4. Real data analysis. In this section we apply the proposed method to DNA methylation data, measured in whole blood of $I = 41$ triads, such that at least one of the two parents is a participant of a birth cohort. This birth cohort study, carried out on the Isle of Wight in the United Kingdom, was established in 1989–1990 to investigate risk factors of allergic diseases over two generations (Arshad et al. (2018)). DNA methylation was measured using the Infinium Human Methylation 450 BeadChip platform (Bibikova et al. (2011)). The raw data contains methylation levels of 484,000 CpG sites. DNA methylation of these 41 triads (123 subjects) was assessed across seven batches. After excluding unreliable CpG sites in different batches, clearing background noise and correcting for batch effect (Aryee et al. (2014), Johnson, Li and Rabinovic (2007), Wang et al. (2012)), there are 308,000 CpG sites remaining; details on DNA methylation generation and preprocessing are included in the Supplementary Material (Mou, Zhang and Arshad (2022a)). Since the aim of the study is to identify parent-to-offspring transmission patterns, only CpG sites with, at least, a moderate correlation between the two generations are of most interest. Thus, we further removed CpG sites whose mother-offspring and father-offspring Pearson correlation coefficient < 0.5. After the screening, 4063 CpG sites of interest are left which are included in our analysis. We further analyzed the data based on a relaxed correlation cutoff 0.4; see Section 4.3 for details.

4.1. *Two block-partition approaches.* One main challenge of applying the proposed method is to identify the dependent blocks. We provide two approaches with the first based on physical distance and the second utilizing the observed correlation of CpG sites. Approach one is established on the phenomenon that the methylation levels of CpG sites are more likely to be correlated, as their physical locations are closer on a DNA sequence (Bell et al. (2011), Eckhardt et al. (2006)). This phenomenon can be observed in Figure 3, where the DNAm levels of CpG cg02954987 to cg11566975 are highly correlated with each other for offspring, mother, and father. Approach one consists of two steps: First, we order CpG sites by the chromosomal coordinates and calculate the correlation of methylation levels between every CpG site and the one next to it ("neighboring" pairs). Next, a chain of consecutive CpG sites forms a block if the correlation of neighboring CpGs is larger than the threshold $= 0.7$. For instance, the correlation between CpG sites 1 and 2, 2 and 3, 3 and 4, 4 and 5 are 0.2, 0.87,

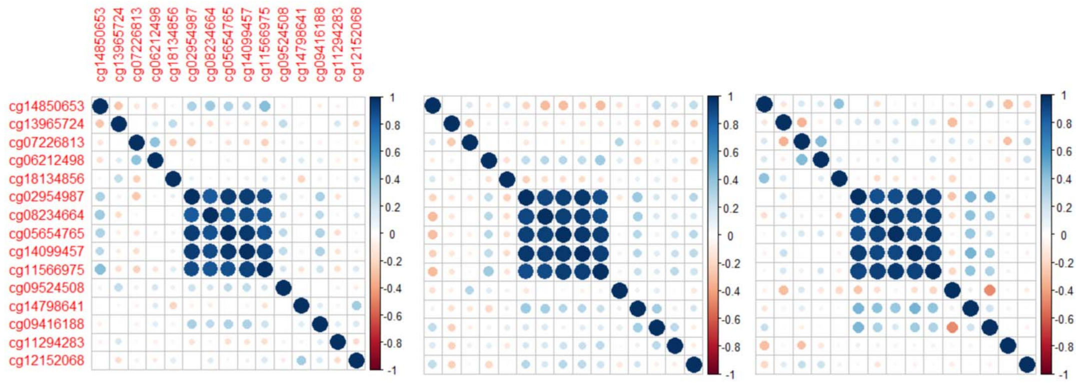

FIG. 3. *From left to right*: *Offspring*, *mother*, *and father's DNAm correlation heatmap of a genomic region near gene LAMB*2 *on Chromosome* 3. *The labels on x and y axes are CpG names designated by Illumnia. The correlation structure is consistent across three family members.*

0.72, 0.43, respectively, then we treat CpG sites 2, 3, and 4 as a dependent block. In Section 4.3 we further investigated smaller correlation thresholds to assess the sensitivity of this approach.

In approach two, we determine the dependent blocks solely on the observed correlation. First, the 4063 CpG sites of interest are partitioned into 2000 subgroups by the k-means clustering method. The number of subgroups is determined based on the phenomenon that the majority of dependent blocks contain 2–5 CpGs in our dataset; by setting 2000 subgroups, which is approximately half of the 4063 CpGs in the analysis, the average block size equals 2 which fits our observations. Other subgroup numbers are assessed in Section 4.3. Following the property of the k-means method, CpG sites within one subgroup (when group size $>$ 1) tend to have similar DNA methylation series. Thus, if a subgroup consists of only one CpG site (group size $=$ 1), then it is treated as an independent site; a subgroup containing multiple CpGs is considered as a potential block candidate. In the next step we screen the block candidates by evaluating the correlation strength. The average pairwise correlation in each subgroup is calculated: If the average correlation $>$ 0.7, we keep it as a dependent block, or, else, we break it into independent sites.

4.2. *Findings from two approaches.* After partitioning the blocks using either approach, the proposed stochastic EM algorithm is applied, and the BIC criterion is used to select the optimal K. We also apply HAN to the data set, assuming all CpG sites are independent. Figure 4 presents the scree plot of the BIC values across varying K's for the proposed approaches and HAN's model. The BIC values of both approaches in Figure 4 are uniformly

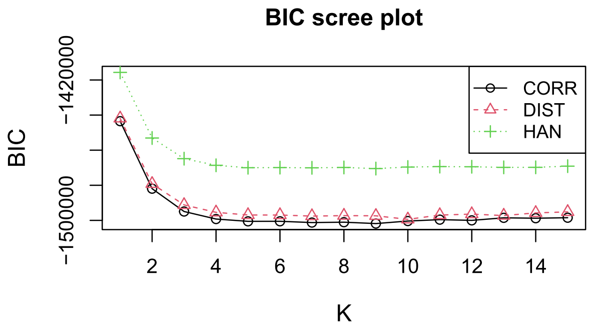

FIG. 4. *BIC scree plot of HAN and GBClust's two block-partition approaches. DIST corresponds to approach one based on distance*; *CORR corresponds to approach two based on correlation.*

TABLE 4

Mean and standard deviations (in parentheses) of coefficient estimates for approach one when intergenerational correlation threshold = 0.5 *and correlation cutoff point of neighboring CpGs* = 0.7, *based on* 100 *starting seeds to accommodate sampling errors*

Cluster index	γ_0	γ_1 (mother)	γ_2 (father)	Num of CpGs
1	0.7079 (0.0201)	0.4956 (0.0569)	0.5085 (0.0582)	70
2	0.3951 (0.0132)	0.4788 (0.0471)	0.5461 (0.0480)	207
3	0.1797 (0.0096)	0.5257 (0.0302)	0.5037 (0.0296)	510
4	0.0082 (0.0069)	0.6428 (0.0218)	0.3741 (0.0214)	1403
5	−0.2394 (0.0127)	0.5491 (0.0112)	0.5015 (0.0113)	1873

smaller than HAN across all K's, indicating a better fit. The scree plots of both approaches decrease sharply at the beginning, until $K = 5$, where they reach a plateau. By the elbow method, $K = 5$ is selected as the elbow point and used for the subsequent analysis. In Tables 4 and 5 we repeat the stochastic EM algorithm with 100 starting seeds and present the mean/SD of $\boldsymbol{\gamma}_k$'s and the number of CpG sites in each cluster. In comparison with the two tables, two approaches present comparable estimation of intercepts across the five clusters. Recalling that the magnitudes of γ_1 and γ_2 represent the inheritance strength from mother and father, respectively, cluster 4 in both approaches is maternally dominated since γ_1 is significantly larger than γ_2. Similarly, cluster 5 can be identified as equally inherited from both parents with comparable γ_1 and γ_2. The coefficients of clusters 1 to 3 show some differences in patterns between the two approaches: In approach one, cluster 2 can be identified as paternally dominated with a moderate difference between γ_2 and γ_1, while clusters 1 and 3 are equally inherited. In approach two, clusters 1–3 are significantly paternally dominated. Approach two can be treated as a generalization of approach one by removing the constraint that only neighboring CpGs can form blocks. Supported by patterns shown in Figure 4, where the BIC values of approach two is uniformly better than approach one, approach two shows a better fit to the data. Results from approach two also indicate a stronger potential of paternal dominance in DNA methylation inheritance at the population level. In summary, both approaches lead to consistent results in clusters 4 and 5 which contains a majority of CpGs (81% and 80% CpGs in Tables 4 and 5, respectively); approach two is able to identify more paternally inherited CpG sites in the rest clusters.

When comparing the findings from the proposed method and the results of HAN (Han et al. (2015)), patterns of clusters 4 and 5 are consistent with the patterns of clusters 6 and 4 in the Table 7 from Han's work, which indicates that transmission patterns of these two clusters are insensitive overall with respect to the use of blocks. More importantly, it is interesting to see that a much larger number of CpGs showing paternal-predominance in DNA methylation

TABLE 5

Mean and standard deviations (in parentheses) of coefficient estimates for approach two when correlation threshold = 0.5 *and k-means clusters* = 2000 *based on* 100 *starting seeds to accommodate sampling errors*

Cluster index	γ_0	γ_1 (mother)	γ_2 (father)	Num of CpGs
1	0.6939 (0.0141)	0.3442 (0.0649)	0.6346 (0.0614)	70
2	0.3788 (0.0085)	0.3925 (0.0496)	0.6021 (0.0563)	210
3	0.1619 (0.0065)	0.4323 (0.0315)	0.5740 (0.0334)	514
4	−0.0041 (0.0058)	0.5921 (0.0181)	0.4125 (0.0188)	1751
5	−0.2643 (0.0132)	0.5271 (0.0128)	0.5210 (0.0148)	1518

transmission are identified, using the proposed method, especially when we use approach two to determine dependent blocks. We use results from block-determination approach two (the less restrictive approach) to illustrate the differences. In Table 8 from Han et al. (2015), 53 CpGs were identified as paternal-predominated transmission sites. Our approach identified three clusters (clusters 1 to 3; Table 5) with in total 794 out of 4063 CpG sites (around 15 times as in Han's work) such that DNA methylation transmission at those CpGs was mainly from paternal transmission and each cluster showed different levels of dominance with cluster 1 demonstrating the strongest paternal predominance and cluster 3 the weakest. A closer examination of these three clusters revealed that all of the 53 CpGs in Han's work are also identified by the proposed method as paternal-predominated transmission sites. In particular, 41 (77%) of the 53 CpGs in Han's work are in our cluster 1, and the remaining are in cluster 2. Based on results from the proposed approach, at a much larger portion of CpGs in the genome than what was found in Han's result, DNA methylation transmission was predominantly paternal. Some studies have indicated the stability of DNA methylation inheritance (Hofmeister et al. (2017)). In conjunction with the regulatory functionality of DNA methylation on gene activities, it is rather important to improve the accuracy in the detection of CpG sites such that DNA methylation transmission is paternally or maternally dominated.

4.3. *Further investigations.* We further applied the proposed methods to the DNAm data, based on intergenerational correlation $= 0.4$ instead of 0.5, which results in 14,791 CpG sites. To assess the sensitivity of the two block-partition approaches, we also tested different correlation thresholds of neighbouring CpGs (denoted by $\mathcal{C}$) in approach one and different numbers of k-means clusters (denoted by $\mathcal{N}$) in approach two. The combinations of these settings yield 10 tables (Tables C.1–C.10) which are available in the Supplementary Material (Mou, Zhang and Arshad (2022a)). There are mainly two findings: (1)When intergenerational correlation $= 0.4$, the BIC scree plot finds $K = 6$ clusters, based on the elbow method. The inheritance patterns are in general consistent with the results when intergenerational correlation $= 0.5$. But with the inclusion of more CpG sites, the transition patterns seem to be clearer (Tables C.5–C.10): The majority of CpG sites are equally or maternally inherited, while smaller proportion (approximately 3000 CpG sites in clusters 1–3) is paternally inherited. (2) Different $\mathcal{C}$'s and $\mathcal{N}$'s do not have a significant impact on the patterns of the clusters in most settings. However, as we lower $\mathcal{C}$ to 0.3 (Tables C.2 and C.7) or reduce $\mathcal{N}$ to less than $J/2$ (Tables C.4 and C.10), growing inconsistencies in transmission patterns and large standard deviations of coefficients increase are observed. One possible reason is that both operations encourage the formation of large blocks, while the common block sizes in our dataset ≤ 5, as shown in Figure 3 as an example. The mismatch between theoretical and real block sizes might have weakened the performance. We recommend setting $\mathcal{C} \geq 0.7$ and $\mathcal{N} \geq J/2$ to avoid unnecessary large blocks.

4.4. *Computational efficiency and source code.* The computing time of the proposed methods can be divided to two parts: preliminary calculation and stochastic EM algorithms. Preliminary calculation includes data loading, block partition, and pre-estimation of parameters in which, block partition occupies the majority of time. After the preliminary calculation stage the stochastic EM algorithm (Algorithm 1 in Section 2.2) will be applied to assign CpG sites to clusters and estimate the coefficients. The time cost is related to multiple factors: The number of CpG sites J, the number of triads I, the number of clusters K that CpG sites are assigned to, the maximum allowed iterations in the stochastic EM algorithm, and the block-partition approaches we are using. The real data analysis in this study was run on Xeon gold 6148 CPU (frequency $= 2400$ MHz) on a computing cluster. In this real data application, for $J = 4063$ CpG sites, $I = 41$ triads, and $K = 5$ clusters, maximum iterations = 1000, the computing time is less than five minutes in both block-partition approaches; in the Supplementary Material (Mou, Zhang and Arshad (2022a)), we also investigated the scenarios when intergenerational correlation threshold = 0.4 ($J = 14{,}791$, $I = 41$, $K = 6$, maximum iterations = 1000); it consumes less than 25 minutes to finish with memory cost less than 4000 MB for both block-partition approaches. Note that in Tables 4 and 5, although it is not mandatory, we ran the estimation procedure 100 times each with a different seed to account for sampling errors. This procedure was carried out via parallel computing on a high-performance computing (HPC) cluster. To systematically evaluate the computational complexity, we employed both block-partition approaches on a DNAm dataset irrelevant to the current study. It has a total of 54 triads. We used the same HPC cluster and recorded the time spent in different J's and K's. The details are presented in the Supplementary Material (Mou, Zhang and Arshad (2022a)). There are three findings: (1) The number of clusters K does not have a significant influence on computing time. (2) The computing time for approach one increases approximately linearly with the increase of the number of CpG sites J: When $K = 5$, as J increases from 4435 to 105,883 (24 times), time cost increases from four minutes to 104 minutes (26 times). (3) The computing time for approach two is slower as sample size increases. Time cost increases from five minutes to 758 minutes when J increases from 4435 to 105,883. While the time spent in block partition for approach one is negligible, the k-means method applied in approach two is time-consuming: The computational complexity of k-means is $O(J^2)$, that is, time will quadruple when J doubles. In Table D.1 and Figure D.1, one can observe the time spent in k-means method increases dramatically as J increases. Thus, from the perspective of computational efficiency, approach one may suit large dataset better. Source codes in the form of R code that implement our method is available in both Supplementary Material (Mou, Zhang and Arshad (2022b)) and Github (https://github.com/abc1m2x3c/GBClustering).

5. Summary and discussion. This article proposes a clustering method to group CpG sites by the parent-to-offspring transmission pattern of DNA methylation. For CpG sites in the same cluster, their inheritance strength can be estimated by the coefficients of maternal/paternal transmission. Understanding the strength of intergenerational transmission can potentially help the prediction of health conditions and prevention of new onset of diseases. Accounting for the potential correlation between CpG sites, the proposed method improves the classification accuracy of HAN and leads to a better model fit, as was shown in the simulation evidence and real data analysis results.

Parental effects may differ in the intergenerational inheritance transmission to offsprings. For instance, childhood allergies have an asymmetric association with maternal and paternal allergic conditions (Arshad et al. (2012)). However, the role that epigenetics plays in this phenomenon is unclear. Paternal effects can be achieved through environmentally-induced epigenetic variations (Curley, Mashoodh and Champagne (2011), Soubry et al. (2014)), and mother-offspring interactions can cause DNA methylation changes in offspring (Kappeler and Meaney (2010)). It will be informative to identify CpG sites in the whole genome showing parent-specific dominance in DNA methylation inheritance. We hope our work paves a way to this direction.

The main challenge of applying the proposed method is to group correlated CpGs into blocks. In our analysis of data collected from a cohort study on asthma and allergic conditions, we propose two approaches to achieve this goal: one is based on physical distance, and the other approach relies solely on observed correlations. Both approaches showed improved BIC compared to the BIC in HAN. Approach two can be considered as a generalized version of approach one by removing the constraints on CpG sites' locations. However, one concern is that some high correlations identified by approach two may stem from random noise (spurious correlation), due to the high-dimensional data structure (Clarke et al. (2008)); thus, there might be a potential risk of overfitting. On the other hand, approach one is based on the physical structure of the DNA molecule and is more stringent for practitioners. In terms of choosing between the two approaches for block design, we believe it is study-specific in the applied fields. Since the correlation-based approach takes into account spatial distances, if a study is not limited to neighboring CpGs, findings from approach two (based on the observed correlation) are deemed to be comprehensive. However, as discussed in Section 4.4, the quadratic time complexity of k-means clustering prevents approach two from being effectively used when J is large. Thus, the choice between the two approaches may also be influenced by computing capacity.

The idea of using a generalized beta distribution to model the correlation between CpG sites may be applied in other problems; see, for example, Houseman et al. (2008). When DNA methylation is measured by the beta value $M/(M+U+c)$, it is common to model the methylation data using beta distribution. By generalizing the univariate beta distribution in the process of model building, one is expected to observe higher accuracy and improved goodness of fit, as shown in this study. In the proposed method, CpG sites in a block share the same β, while different α's are allowed. This may limit the flexibility of model fitting. However, this is the price we pay to introduce correlation and improve the goodness of fit. A promising future study is to apply a more generalized version of multivariate beta distribution to allow different β's which may need more theoretical work and requires intensive computing.

Besides analyzing the triads' data, the proposed method has the potential to be extended to other scenarios. For instance, one can analyze the transgenerational inheritance pattern of a large pedigree dataset by breaking down it into triads. One concern in this direction is the potential violation of the independence assumption between families. One possible way to deal with this type of situation when inferring clusters is to utilize the concept of composite likelihoods, but sensitivity of findings to the dependence between families certainly deserves an in-depth investigation.

Acknowledgments. We are thankful to the High Performance Computing facility at the University of Memphis. We also thank the Editor, Associate Editor, and four anonymous reviewers for their constructive suggestions that greatly improved this paper.

To whom correspondence should be addressed: hzhang6@memphis.edu.

Funding. The work is supported by National Institutes of Health (R01 AI121226 and R01 AI091905).

SUPPLEMENTARY MATERIAL

Supplement to "Identifying intergenerational patterns of correlated methylation sites" (DOI: 10.1214/21-AOAS1511SUPPA; .pdf). This supplementary material consists of four appendices. Appendix A shows the equivalence of the proposed method with Han et al. (2015) when CpG sites are independent with each other. Appendix B demonstrates the DNA methylation generation and preprocessing for the real-world dataset in the main paper. Appendix C contains supplementary data analysis results. Appendix D presents additional evaluation of computational efficiency for the proposed methods.

Computer codes for "Identifying intergenerational patterns of correlated methylation sites" (DOI: 10.1214/21-AOAS1511SUPPB; .zip). Source codes in the form of R code and examples for the proposed methods.

REFERENCES

Arshad, S. H., Karmaus, W., Raza, A., Kurukulaaratchy, R. J., Matthews, S. M., Holloway, J. W., Sadeghnejad, A., Zhang, H., Roberts, G. et al. (2012). The effect of parental allergy on childhood allergic diseases depends on the sex of the child. *J. Allergy Clin. Immunol.* **130** 427–434.

Arshad, S. H., Holloway, J. W., Karmaus, W., Zhang, H., Ewart, S., Mansfield, L., Matthews, S., Hodgekiss, C., Roberts, G. et al. (2018). Cohort profile: The Isle of Wight whole population birth cohort (IOWBC). *Int. J. Epidemiol.* **47** 1043–1044i.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. and Irizarry, R. A. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* **30** 1363–1369.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y. and Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12** R10. https://doi.org/10.1186/gb-2011-12-1-r10

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P. et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* **98** 288–295.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A. and Wang, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **8** 37–49.

Curley, J. P., Mashoodh, R. and Champagne, F. A. (2011). Epigenetics and the origins of paternal effects. *Horm. Behav.* **59** 306–314.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L. and Lin, S. M. (2010). Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinform.* **11** 587.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R. et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38** 1378–1385.

Han, S., Zhang, H., Lockett, G. A., Mukherjee, N., Holloway, J. W. and Karmaus, W. (2015). Identifying heterogeneous transgenerational DNA methylation sites via clustering in beta regression. *Ann. Appl. Stat.* **9** 2052–2072. MR3456365 https://doi.org/10.1214/15-AOAS865

Hofmeister, B. T., Lee, K., Rohr, N. A., Hall, D. W. and Schmitz, R. J. (2017). Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol.* **18** 1–16.

Houseman, E. A., Christensen, B. C., Yeh, R.-F., Marsit, C. J., Karagas, M. R., Wrensch, M., Nelson, H. H., Wiemels, J., Zheng, S. et al. (2008). Model-based clustering of DNA methylation array data: A recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinform.* **9** 365.

Johnson, W. E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

Kappeler, L. and Meaney, M. J. (2010). Epigenetics and parental effects. *BioEssays* **32** 818–827.

Lee, S.-H., Park, J.-S. and Park, C.-S. (2011). The search for genetic variants and epigenetics related to asthma. *Allergy, Asthma & Immunology Research* **3** 236–244.

Libby, D. L. and Novick, M. R. (1982). Multivariate generalized beta distributions with applications to utility assessment. *J. Educ. Stat.* **7** 271–294.

Lockett, G. A., Patil, V. K., Soto-Ramírez, N., Ziyab, A. H., Holloway, J. W. and Karmaus, W. (2013). Epigenomics and allergic disease. *Epigenomics* **5** 685–699.

Mou, X., Zhang, H. and Arshad, S. H. (2022a). Supplement to "Identifying intergenerational patterns of correlated methylation sites." https://doi.org/10.1214/21-AOAS1511SUPPA

Mou, X., Zhang, H. and Arshad, S. H. (2022b). Computer codes for "Identifying intergenerational patterns of correlated methylation sites." https://doi.org/10.1214/21-AOAS1511SUPPB

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **6** 457–489. MR1762556 https://doi.org/10.2307/3318671

Padmanabhan, N., Jia, D., Geary-Joo, C., Wu, X., Ferguson-Smith, A. C., Fung, E., Bieda, M. C., Snyder, F. F., Gravel, R. A. et al. (2013). Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development. *Cell* **155** 81–93.

Park, H.-S. and Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36** 3336–3341.

Qin, L.-X. and Self, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* **62** 526–533. MR2236835 https://doi.org/10.1111/j.1541-0420.2005.00498.x

Soubry, A., Hoyo, C., Jirtle, R. L. and Murphy, S. K. (2014). A paternal environmental legacy: Evidence for epigenetic inheritance through the male germ line. *BioEssays* **36** 359–371.

Stenz, L., Schechter, D. S., Serpa, S. R. and Paoloni-Giacobino, A. (2018). Intergenerational transmission of DNA methylation signatures associated with early life stress. *Curr. Genomics* **19** 665–675.

Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J. and Liu, S. (2012). IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28** 729–730.

Yu, F., Xu, C., Deng, H.-W. and Shen, H. (2020). A novel computational strategy for DNA methylation imputation using mixture regression model (MRM). *BMC Bioinform.* **21** 1–17.

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. and Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16** 14. https://doi.org/10.1186/s13059-015-0581-9