## Title

Random field modeling of multi-trait multi-locus association for detecting methylation quantitative trait loci.



## Journal

Bioinformatics, 38(16)

## ISSN

1367-4803

## Authors

Lyu, Chen
Huang, Manyan
Liu, Nianjun
et al.

## Publication Date

2022-08-10

## DOI

10.1093/bioinformatics/btac443

Peer reviewed

*Bioinformatics*, 38(16), 2022, 3853–3862
https://doi.org/10.1093/bioinformatics/btac443
Advance Access Publication Date: 4 July 2022
Original Paper

OXFORD

Genome analysis

# Random field modeling of multi-trait multi-locus association for detecting methylation quantitative trait loci

**Chen Lyu[1,2], Manyan Huang[1], Nianjun Liu[1], Zhongxue Chen [1], Philip J. Lupo[3], Benjamin Tycko[4], John S. Witte[5,6], Charlotte A. Hobbs[7] and Ming Li [1,*]**

[1]Department of Epidemiology and Biostatistics, Indiana University Bloomington, Bloomington, IN 47405, USA, [2]Department of Population Health, New York University Grossman School of Medicine, New York, NY 10016, USA, [3]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA, [4]Center for Discovery and Innovation, Nutley, NJ 07110, USA, [5]Department of Epidemiology and Population Health, Stanford University, Stanford, CA 94305, USA, [6]Department of Biomedical Data Sciences, Stanford University, Stanford, CA 94305, USA and [7]Rady Children's Institute for Genomic Medicine, San Diego, CA 92123, USA

*To whom correspondence should be addressed.
Associate Editor: Tobias Marschall

Received on September 7, 2021; revised on June 28, 2022; editorial decision on June 30, 2022; accepted on June 30, 2022

## Abstract

**Motivation:** CpG sites within the same genomic region often share similar methylation patterns and tend to be co-regulated by multiple genetic variants that may interact with one another.

**Results:** We propose a multi-trait methylation random field (multi-MRF) method to evaluate the joint association between a set of CpG sites and a set of genetic variants. The proposed method has several advantages. First, it is a multi-trait method that allows flexible correlation structures between neighboring CpG sites (e.g. distance-based correlation). Second, it is also a multi-locus method that integrates the effect of multiple common and rare genetic variants. Third, it models the methylation traits with a beta distribution to characterize their bimodal and interval properties. Through simulations, we demonstrated that the proposed method had improved power over some existing methods under various disease scenarios. We further illustrated the proposed method via an application to a study of congenital heart defects (CHDs) with 83 cardiac tissue samples. Our results suggested that gene *BACE2*, a methylation quantitative trait locus (QTL) candidate, colocalized with expression QTLs in artery tibial and harbored genetic variants with nominal significant associations in two genome-wide association studies of CHD.

**Availability and implementation:** https://github.com/chenlyu2656/Multi-MRF.

**Contact:** li498@indiana.edu

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

## 1 Introduction

DNA methylation can be influenced by genetic variants in a region (Gaunt *et al.*, 2016), referred to as a methylation quantitative trait locus (mQTL or methQTL; here we utilize the first abbreviation). mQTLs are primarily *cis*-acting and located close to the CpG site. Many *cis*-mQTLs are also found to co-localize with genetic variants associated with complex diseases, such as cardiovascular disease (Huan *et al.*, 2019), respiratory disease (Morrow *et al.*, 2018) and metabolic disease (Volkov *et al.*, 2016). Detecting mQTLs is crucial to understand the functional mechanisms of how genotypic variations may influence the disease risk within specific tissues.

Many existing mQTL studies have adopted a single-locus single-trait strategy (Almli *et al.*, 2015; Dick *et al.*, 2014; Smith *et al.*, 2014), by evaluating the association between a genetic variant and a CpG site one-at-a-time. Despite existing successes (Huan *et al.*, 2019), there are also a few limitations. First, the single-locus strategy may suffer from power loss due to heavy multiple testing burden, ignore potential interactions between multiple genetic variants, and fail to detect rare variants as mQTL SNPs. As an alternative, multi-locus testing or region-based analysis has been proposed to integrate the effect of multiple common and rare genetic variants, including burden tests (Li and Leal, 2008), quadratic tests (Wu *et al.*, 2011) and combined tests (Lee *et al.*, 2013). Second, the single-trait strategy may also not be optimal, because CpG sites that are close to one another tend to be co-methylated and share similar methylation patterns. As an alternative, multi-trait methods have been proposed to account for the correlation between CpG sites and to further reduce the burden of multiple testing for

power improvement (Aschard *et al.*, 2014). Broadly speaking, these methods fall into three groups: (i) combining test statistics or *P*-values from univariate analyses (van der Sluis *et al.*, 2013); (ii) dimension reduction methods, such as principal component analysis and canonical correlation analysis (Aschard *et al.*, 2014; Tang and Ferreira, 2012); and (iii) regression frameworks including linear mixed models, multivariate analysis of variance and reverse regression (O'Reilly *et al.*, 2012). However, these methods were all designed for multi-trait single-locus testing. Relatively few methods are available for multi-trait multi-locus testing.

Recently, we and others developed a methylation random field (MRF) method for mQTL detection by modeling the methylation trait with a beta distribution (Lyu *et al.*, 2021). To address the co-methylation between neighboring CpG sites, we propose to extend the MRF to a multi-trait MRF (multi-MRF) method and test the joint association between multiple CpG sites and multiple variants within a genomic region. Similar to the MRF, multi-MRF uses beta distributions to characterize the bimodal and interval properties of methylation traits. It also uses the multi-locus genotypes as the coordinates of a subject in the high-dimensional space and further integrates the effect of multiple common and rare genetic variants with a conditional autoregressive model. As an extension of MRF, multi-MRF is a multi-trait method, and allows flexible correlation structure between neighboring CpG sites (e.g. distance-based correlation). To evaluate the performance of multi-MRF, we conducted simulation studies and compared it with other existing methods, including the dual kernel association test (DKAT; Zhan *et al.*, 2017) and the multiple-testing-adjusted MRF (mMRF). We further illustrated our multi-MRF with a study of congenital heart defects (CHDs) to identify *cis*-acting mQTLs within cardiac tissues.

## 2 Materials and methods

### 2.1 Multi-MRF framework

Assume we have a study of $n$ unrelated subjects who were profiled for $m$ CpG sites, sequenced for $k$ genetic variants, and measured for $l$ non-genetic covariates. For the $i$th subject, let $Y_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,m})$ be the vector of methylation traits; $G_i = (G_{i,1}, G_{i,2}, \ldots, G_{i,k})$ denotes the genotypes, coded as the minor allele counts; and $X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,l})$ be the potential confounders, such as the top principal components of the genomic and epigenomic data. Our research question for detecting mQTLs can be formulated as testing the association between the methylation traits of $m$ CpG sites and the genotypes of $k$ genetic variants within a genomic region, while adjusting for $l$ covariates.

A conditional autoregressive model can be used as:

$$\begin{aligned} E(Y_{i,p}|Y_{-i,p}) = \mu_{i,p} + \delta\textstyle\sum_{p\neq q} v(Y_{\cdot,p}, Y_{\cdot,q})(Y_{i,q} - \mu_{i,q}) \\ + \gamma\textstyle\sum_{(i,p)\neq(j,q)} s(G_i, G_j)(Y_{j,q} - \mu_{j,q}); \\ \text{for any } 1 \le i \le n,\ 1 \le p \le m,\ 1 \le q \le m; \end{aligned} \tag{1}$$

where $\mu_{i,\cdot} = f(X_i\beta)$ is the expected contribution of non-genetic covariates, $\beta = (\beta_1, \beta_2, \ldots, \beta_l)'$ are the coefficients of these covariates, $f(\cdot)$ is the link function connecting methylation traits with covariates, and $Y_{-i,p}$ represents all methylation traits other than $Y_{i,p}$. Because DNA methylation arrays estimate the methylation level at a CpG site based on the ratio of intensities between methylated and unmethylated alleles, the methylation trait (i.e. beta values) is an interval variable and bimodally distributed. Hence, we adopt a beta regression with a *logit* link to model the properties. Further, $s(G_i, G_j)$ describes the genotypic similarity between subject $i$ and $j$, and is defined by the genetic relationship (Yang *et al.*, 2011): $s(G_i, G_j) = \sum_{h=1}^{k} w_h(G_{i,h} - 2p_h)(G_{j,h} - 2p_h)$, where $p_h$ is the average minor allele counts for the $h$th variant in the study population, and a weighting scheme $w_h$ is incorporated to allow flexible consideration across variants (e.g. allele frequencies or effect sizes). For example, weighting based on the probability density function of a beta distribution, $w_h = dbeta(MAF_h, 1, 25)^2$ recommended by Wu and Pankow (2016), mimics a scenario that rare variants contribute relatively large effect on methylation traits and the effect size decreases as MAF increases. We further use $v(Y_{\cdot,p}, Y_{\cdot,q})$ to model the within-subject similarity of methylation traits between the $p$th and the $q$th CpG sites, where $Y_{\cdot,p} = (Y_{1,p}, Y_{2,p}, \ldots, Y_{n,p})'$ and $Y_{\cdot,q} = (Y_{1,q}, Y_{2,q}, \ldots, Y_{n,q})'$. The trait similarity accounts for the correlation structure among traits, which can be flexible and defined based on prior knowledge. For example, previous studies indicated that the co-methylation of two neighboring CpG sites decreased as their physical distance on the genome increased (Affinito *et al.*, 2020; Lövkvist *et al.*, 2016). Hence, we could assume that the correlation of methylation traits between neighboring CpG sites decreases exponentially as their physical distance increases (Nautiyal *et al.*, 2010):

$$v(Y_{\cdot,p}, Y_{\cdot,q}) = e^{-\frac{|d_p - d_q|}{c}}, \text{ for any } 1 \le p \ne q \le m;$$

where $d_p$ and $d_q$ are the base pair (BP) locations for the $p$th and the $q$th CpG sites, respectively, and $c$ is a constant as the size of the region being tested. Alternatively, if the correlation among CpG sites is assumed to be exchangeable regardless of physical distance, $v(Y_{\cdot,p}, Y_{\cdot,q}) = 1$ can be used for any $1 \le p \ne q \le m$.

Intuitively from Equation (1), the methylation trait of the $p$th CpG site for subject $i$ can be predicted by the methylation traits of other CpG sites within the region through their correlation structure, and by the methylation traits of other subjects through their genotypic similarities. Thus, the parameter $\delta$ is a nuisance parameter that measures the magnitude of adjustment for the correlation of CpG sites within the region. The joint association between $m$ CpG sites and $k$ genetic variants is measured by a fixed parameter $\gamma$, and can be tested against a null hypothesis of $H_0 : \gamma = 0$.

### 2.2 Statistical inference

Equation (1) can be written in a matrix form as:

$$E(Y|Y_-) = \mu + (\delta V + \gamma S)(Y - \mu); \tag{2}$$

where $\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$ as within a beta regression; $Y = (Y_1', Y_2', \ldots, Y_n')'$ is a $n \times m$ matrix for the methylation traits of $n$ subjects at $m$ CpG sites; $Y_- = (Y_{-i,p})_{n\times m}$; $X$ is a $n \times l$ matrix for covariates ; $V$ is a $m \times m$ block matrix modeling the within-subject trait similarities among $m$ CpG sites, with its element $v(p,q) = v(Y_{\cdot,p}, Y_{\cdot,q}) = e^{-\frac{|d_p - d_q|}{c}}$ for any $1 \le p \ne q \le m$, and $v(p,p) = 0$ for any $1 \le p \le m$; and $S$ is a $n \times n$ matrix modeling the genetic similarities among $n$ subjects, with its element $s(i,j) = s(G_i, G_j) = \sum_{h=1}^{k} w_h(G_{i,h} - 2p_h)(G_{j,h} - 2p_h)$ for any $1 \le i \ne j \le n$, and $s(i,i) = 0$ for any $1 \le i \le n$.

We use the estimating equation to construct a generalized score test for $H_0 : \gamma = 0$.

$$\begin{aligned} U_\gamma(\beta, \delta, \gamma) &= \frac{\partial E(Y|Y_-)}{\partial \gamma}^T \{Y - E(Y|Y_-)\} \\ &= (Y - \mu)^T S\{I - \delta V - \gamma S\}(Y - \mu) = 0; \end{aligned} \tag{3}$$

The test statistics can be obtained as (Boos, 1992):

$$Q = \frac{U_\gamma(\hat\beta, \hat\delta, 0)}{n} = \frac{(Y - \hat\mu)' S(I - \hat\delta V)(Y - \hat\mu)}{n}; \tag{4}$$

where the estimated $\hat\beta$ and $\hat\delta$ can be solved iteratively in the linear equations under the null hypothesis. Previous research has demonstrated that the $Q$ statistic follows an asymptotic weighted sum of chi-square distributions with 1 degree of freedom (He *et al.*, 2015; Li *et al.*, 2018a,b). Although this score test tends to be overly conservative when sample size is small and when rare variants are tested, empirical adjustment can be used (Guo *et al.*, 2005).

### 2.3 Simulation studies

We conducted a series of simulations to evaluate the performance of the proposed multi-MRF, and compared it to other existing methods, including the DKAT that was developed for multi-trait multi-locus association tests, and the MRF that was developed for single-trait multi-locus association tests. Both DKAT and multi-MRF use

kernel functions to integrate multiple genetic variants. However, while DKAT uses kernel functions to integrate multiple traits, multi-MRF directly models the correlation of traits. Benjamini–Hochberg's false discovery rate was applied to MRF for multiple testing adjustment. The methods were compared in terms of Type I errors and statistical power.

The genotype data were based on the sequencing data of 1092 unrelated subjects in the 1000 Genomes Project, and we randomly selected a 1 MB segment on Chromosomes 17:7 344 328–8 344 327 (Li *et al.*, 2018a,b). The segment covered a total of 12 735 single nucleotide polymorphisms (SNPs), including both common and rare variants (82.7% had minor allele frequency [MAF] $< 0.05$). To align with the real data (i.e. 450 K array), we assumed that each subject was profiled for 10 CpG sites (average number of CpGs within a gene; see Table 4) within a 10 kb region (average gene size) to be tested. We first randomly sampled the 10 kb region within the 1 MB segment, and then randomly selected 10 non-overlapping positions for 10 CpG sites within the region. Each of the three methods was applied to test the association between the 10 CpG sites and SNPs within the 10 kb region (i.e. aiming to detect mQTLs with *cis*-acting effect). To capture the interval and bimodal properties, the methylation trait of the $p$th CpG site for subject $i$ was simulated following a beta distribution $Y_{i,p} \sim beta(a_{i,p}, b_{i,p})$. The beta distribution was characterized by two shape parameters $a_{i,p}$ and $b_{i,p}$, which can be determined by a mean parameter $\mu_{i,p}$ and a precision parameter $\phi$, so that $a_{i,p} = \mu_{i,p}\phi$, and $b_{i,p} = (1-\mu_{i,p})\phi$. In this study, we fixed $\phi$ to be 30, as suggested by previous literature (Bayes *et al.*, 2012). The mean parameters of 10 CpG sites were simulated under varying correlation structures and other simulation scenarios described below. To evaluate Type I error, the methylation mean parameters were simulated independently from the genotypes. To evaluate statistical power, the methylation means were simulated with a non-genetic component (i.e. covariates' contribution), a genetic component, and a random error to accommodate within-subject correlation.

Overall, we considered disease scenarios of varying correlation structures (i.e. exchangeable, autoregressive and distance-based), causal structures between mQTLs and methylation traits (i.e. 'unique', 'half-shared' and 'all-shared'), sample sizes ($n = 50$, 100, 250 and 500), directions of mQTL effects on methylation traits (i.e. one-or bi-directional), and strategies to model the distributions of the methylation traits (i.e. normal, beta or logit). The detailed explanations for each scenario are summarized in Supplementary Table S1. To model the methylation traits, the normal strategy used a linear regression with an *identity* link for methylation traits assuming a normal distribution; the beta strategy used a beta regression with a *logit* link for methylation traits assuming a beta distribution; and the logit strategy used a linear regression with an *identity* link for *logit*-transformed traits assuming a normal distribution after *logit*-transformation. In practice, the normal and beta strategy represent testing beta values as methylation traits, whereas the logit strategy represents testing M values since M values are proportional to the logit transformation of beta values. Because beta regression was not implemented in DKAT, we only considered the normal and logit strategies for DKAT. When the methylation traits were tested for joint association with genotypes, we also considered varying genetic frequencies of the variants within the region (a mixture of common and rare, and rare variants only) and the misspecification of traits correlation. All combinations of these disease scenarios were considered while evaluating the performances of multi-MRF, DKAT and mMRF (MRF with multiple testing adjustment).

### 2.3.1 Type I error

To evaluate Type I errors, we simulated the mean parameters of the methylation traits of 10 CpG sites for the $i$th subject independently from the genotypes:

$$logit(\mu_i) = logit(\mu_0) + \varepsilon_i;\ with\ \varepsilon_i \sim N(0, \Sigma),$$

where $\mu_i = (\mu_{i,1}, \ldots, \mu_{i,10})'$ was the vector of mean parameters for 10 correlated CpG sites; $\mu_0 = (\mu_{0,1}, \ldots, \mu_{0,10})'$ was the baseline methylation trait from non-genetic contribution (i.e. effect from covariates), and was set to 0.1 for all $\mu_{0,p}$, $1 \le p \le 10$, based on real data distribution (Lyu *et al.*, 2021); and $\varepsilon_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,10})'$ represented a multivariate random error to accommodate the within-subject correlation among CpG sites. We assumed a multivariate Gaussian distribution with mean of zero and a variance-covariance matrix ($\Sigma$) describing varying correlation structures.

We considered three correlation structures: (i) exchangeable correlation ($\Sigma_1$), or compound symmetry:

$$\Sigma_1 = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix},$$

assuming that all CpG sites within a candidate region had the same correlation coefficient; (ii) first-order autoregressive correlation ($\Sigma_2$):

$$\Sigma_2 = \sigma^2 \begin{bmatrix} 1 & \rho^1 & \rho^2 & \ldots & \rho^9 \\ \rho^1 & 1 & \rho^1 & \ldots & \rho^8 \\ \rho^2 & \rho^1 & 1 & \ldots & \rho^7 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^9 & \rho^8 & \rho^7 & \ldots & 1 \end{bmatrix},$$

assuming that the correlation between CpG sites decayed in order; and (iii) distance-based correlation ($\Sigma_3$):

$$\Sigma_3 = \sigma^2 \begin{bmatrix} 1 & e^{-\frac{|d_1-d_2|}{c}} & e^{-\frac{|d_1-d_3|}{c}} & \ldots & e^{-\frac{|d_1-d_{10}|}{c}} \\ e^{-\frac{|d_2-d_1|}{c}} & 1 & e^{-\frac{|d_2-d_3|}{c}} & \ldots & e^{-\frac{|d_2-d_{10}|}{c}} \\ e^{-\frac{|d_3-d_1|}{c}} & e^{-\frac{|d_3-d_2|}{c}} & 1 & \ldots & e^{-\frac{|d_3-d_{10}|}{c}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e^{-\frac{|d_{10}-d_1|}{c}} & e^{-\frac{|d_{10}-d_2|}{c}} & e^{-\frac{|d_{10}-d_3|}{c}} & \ldots & 1 \end{bmatrix},$$

assuming that the correlation between CpG sites exponentially decreased as their BP distance on genome increased (Nautiyal *et al.*, 2010); where $d_p$, $1 \le p \le 10$, represented the genomic location of the $p$th CpG site.

The methylation trait at the $p$th CpG site of subject $i$, $Y_{i,p}$, $1 \le p \le 10$, was then simulated following a beta distribution of $Y_{i,p} \sim beta(\mu_{i,p}\phi, (1-\mu_{i,p})\phi)$, where $\phi$ was fixed to 30 as suggested in the literature (Bayes *et al.*, 2012). The Type I errors were also evaluated under varying correlation structures of CpG sites, sample sizes ($n = 50$, 100, 250 and 500), strategies to model the distribution of methylation traits (normal, beta and logit) and genetic frequencies of the SNPs in the region (a mixture of common and rare, and rare variants only). A total of 100 000 replicates were simulated.

### 2.3.2 Statistical power

To evaluate the statistical power, we simulated the mean parameters of the methylation traits of 10 CpG sites for the $i$th subject based on the following model:

$$logit(\mu_i) = logit(\mu_0) + \sum_{h=1}^{k} \begin{bmatrix} G_{i,h}\beta_{h,1} \\ \vdots \\ G_{i,h}\beta_{h,10} \end{bmatrix} + \varepsilon_i;\ with\ \varepsilon_i \sim N(0, \Sigma);$$

where $G_{i,h}$ was the minor allele count for the $h$th SNP of subject $i$ and $(\beta_{h,1}, \ldots., \beta_{h,10})'$ denoted its effects on the 10 CpG sites. We assumed that the effects of mQTL SNPs were inversely associated with MAF as follows:

$$\beta_{h,p} = \begin{cases} \dfrac{1}{MAF_h(1-MAF_h)}, & SNP\ h\ is\ a\ mQTL\ SNP\ for\ the\ p\text{th}\ CpG\ site \\ 0, & SNP\ h\ is\ not\ a\ mQTL\ SNP\ for\ the\ p\text{th}\ CpG\ site \end{cases}$$

For any given CpG site, we assumed 10% of the variants in the region were mQTL SNPs that were causal to each methylation trait.

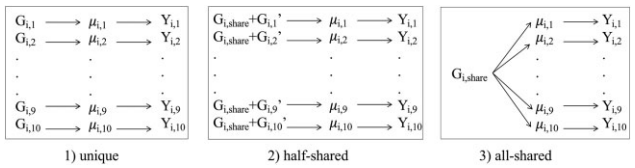


**Fig. 1.** Three causal structures between mQTL SNPs and methylation traits: (1) mutually exclusive mQTL SNPs; (2) half of the mQTL SNPs were shared by all CpG sites and the other were mutually exclusive across CpG sites; and (3) all CpG sites shared the same mQTL SNPs

The mQTL SNPs were randomly selected in each simulation scenarios to represent varying causal structure of methylation traits.

We conducted three sets of simulations representing varying causal structures between mQTL SNPs and methylation traits as illustrated in Figure 1.

*Simulation I:* 'unique' causal structure. In this simulation scenario, each of the 10 CpG sites was influenced by 10% of the SNPs within the region, and these mQTL SNPs were mutually exclusive across CpG sites.

*Simulation II*: 'half-shared' causal structure. In this simulation scenario, each of the 10 CpG sites was influenced by 10% of the SNPs within the region. However, half of these mQTL SNPs were shared by all CpG sites, but the other half of the mQTL SNPs were mutually exclusive across CpG sites.

*Simulation III*: 'all-shared' causal structure. In this simulation scenario, all of the 10 CpG sites shared the same mQTL SNPs which were 10% of all SNPs within the region.

In Simulations I–III, we also simulated the methylation traits under three correlation structures, including exchangeable correlation ($\Sigma_1$), autoregressive correlation ($\Sigma_2$) or distance-based correlation ($\Sigma_3$). The robustness of multi-MRF, DKAT and mMRF was evaluated when the correlation structure between traits was correctly specified or misspecified in the analysis. In addition, we assumed that the mQTL SNPs affecting methylation traits might have either one- or bi-directional effect. The bi-directional effect was simulated by assigning a negative sign to $\beta_b$ for half of the randomly selected mQTL SNPs.

*Simulation IV*: testing only rare variants for mQTL effects. In this simulation, we evaluated the power of all methods when only variants with MAF $< 0.05$ were tested. The scenario was illustrated under 'all-shared' causal structure described above. Three correlation structures and directions of mQTL effects were evaluated in the simulation as well.

In summary, when evaluating the statistical power of each method, the methylation traits were simulated under varying correlation structures (i.e. exchangeable, autoregressive and distance-based), causal structures, sample sizes (i.e. $n = 50$, 100, 250 and 500), effect directions (i.e. one- and bi-direction) and strategies to model the distribution of methylation traits (i.e. normal, beta and logit). We also considered the frequencies of genetic variants being tested (i.e. a mixture of common and rare variants, and rare variants only). For each simulation scenario, a total of 1000 replicates were generated for empirical statistical power.

## 2.4 Application studies

We further applied the proposed multi-MRF to the genomic and epigenomic data of 83 cardiac tissue samples. Alternative methods, DKAT and mMRF, were also applied for comparison. The details of the samples can be found elsewhere (Li *et al.*, 2021). Briefly, each sample was genotyped for ~5 million SNPs using Illumina® Infinium HumanOmni5Exome BeadChip, and was profiled for ~450 K CpG sites using Illumina HumanMethylation450 Beadchip or 850 K CpG sites using Illumina MethylationEPIC Beadchip. For epigenomic data, we used the Bioconductor package 'minfi' in R to combine the raw intensity values from all samples at the same time (Aryee *et al.*, 2014; Fortin *et al.*, 2014, 2017). Functional normalization was applied to raw intensities, which used internal control probes on each array to remove between-array technical variations. Beta values were produced to measure the methylation level of CpG sites, and intensities with detection $P$-values $> 0.01$ were set to missing. We further removed CpG sites with more than 5% missing values or with an SNP in the probe. For genomic data, we used PLINK 1.9 for data processing (Purcell *et al.*, 2007) and the website is available: https://www.cog-genomics.org/plink/. We removed variants that deviated from Hardy–Weinberg equilibrium among control samples ($P$-value $< 0.0001$). After the quality control process, a total of 3 055 128 SNPs and 275 357 CpG sites remained for analysis. To conduct region-based association tests, gene units were defined based on the UCSC genome browser under the genome assembly of GRCh37/hg19. A candidate genomic region was defined as a gene unit along with its 7.5 kb upstream and downstream sequences. To detect multi-locus multi-trait associations, we tested all regions with at least two SNPs and at least two CpG sites.

Within each region, both multi-MRF and mMRF were applied to test the joint association between all SNPs and all CpG sites adjusting for the covariates, including gender, case-control status and top five principal components for the genomic and epigenomic profiles. In addition, both multi-MRF and mMRF were applied with various modeling of trait distributions (i.e. normal, beta or logit) and correlation structures (i.e. exchangeable, autoregressive or distance-based). On the other hand, because the R package of DKAT did not allow modeling of methylation traits with a beta distribution or covariate adjustment, DKAT was applied assuming either a normal or a logit-transformed normal distribution without adjusting for the covariates. A total of 15 695 regions were tested for multi-CpG multi-SNP associations. Bonferroni adjustment was used to account for the multiple testing based on the total number of regions being tested.

## 2.5 Bayesian colocalization

Multiple previous studies have suggested that causal genetic variants for complex diseases may function through regulating the methylation or expression level of genes. Bayesian colocalization was used as a common strategy to map functional regulatory SNPs underlying disease risk (Battle *et al.*, 2017; Giambartolomei *et al.*, 2014). Upon identification of mQTLs, we further used Bayesian colocalization (Giambartolomei *et al.*, 2014) to prioritize the mQTL findings by leveraging the results of two previous CHD genome-wide association studies (GWASs) and the known heart-tissue expression QTLs. The goal of Bayesian colocalization was to evaluate each genomic region for sharing causal genetic variants to two traits (e.g. methylation trait and CHD status, or methylation trait and expression trait). For example, the colocalization analysis between mQTLs and a CHD GWAS would estimate five posterior probabilities (PP0, PP1, PP2, PP3 and PP4) with each supporting a corresponding hypothesis (H0: no association with either methylation or CHD risk; H1: association with methylation trait, but not with CHD risk; H2: association with CHD risk, but not with methylation trait; H3: association with methylation trait and CHD risk through two independent SNPs; H4: association with methylation trait and CHD risk through at least one shared SNP). Two CHD GWASs had a case-parental trio design with 440 and 225 trios, respectively, who were participants of the National Birth Defects Prevention Study. For the expression QTLs, we searched the Genotype-Tissue Expression database for five types of heart tissues, including artery aorta, artery coronary, artery tibial, heart atrial appendage and heart left ventricle. R package '*coloc*' was used for the analysis (Giambartolomei *et al.*, 2014).

# 3 Results

## 3.1 Data simulation studies

### 3.1.1 Type I errors

The Type I errors were evaluated at the $\alpha$ level of 0.05 (Fig. 2) with 100 000 replicates. We considered scenarios of testing a mixture of common and rare variants (Fig. 2A) and testing rare variants only (Fig. 2B). When the methylation traits were modeled with beta distributions (i.e. beta strategy), both multi-MRF and mMRF had reasonably well-controlled Type I errors regardless of the underlying correlation structures among methylation traits. The Type I errors were also robust when the correlation structures were mis-specified.

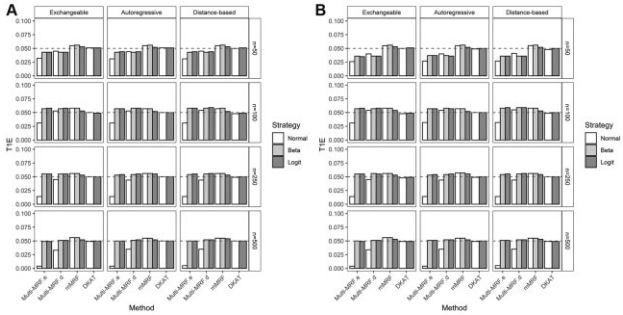


**Fig. 2.** Type I error rates of multi-MRF, mMRF and DKAT at alpha level of 0.05. (**A**) A mixture of common and rare variants was tested for association with traits; (**B**) Only rare variants were tested. Multi-MRF.e and multi-MRF.d represents applying the multi-MRF with exchangeable and distance-based correlation structures, respectively

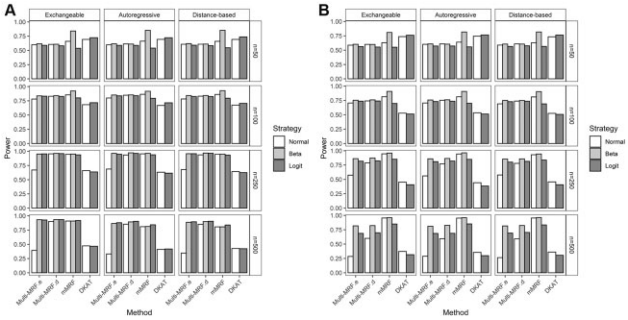


**Fig. 3.** Statistical power for Simulation I, when the causal structure was 'unique'. (**A**) The mQTL SNPs had one-directional effect on methylation traits; (**B**) The mQTL SNPs had bi-directional effect on methylation traits. Multi-MRF.e and multi-MRF.d represent applying the multi-MRF with exchangeable and distance-based correlation structures, respectively

When the methylation traits were modeled with logit-transformed normal distributions (i.e. logit strategy), all methods (multi-MRF, mMRF and DKAT) had well-controlled Type I errors for all actual and assumed correlation structures. However, when the methylation traits were modeled with normal distributions (i.e. normal strategy), multi-MRF showed conservative Type I error rates that were consistently $< 0.05$. The conservativeness is the most severe when an exchangeable correlation was assumed among methylation traits and when the sample size was large. The Type I errors of mMRF and DKAT remained to be well-controlled in such situations.

### 3.1.2 Statistical power

We evaluated the statistical power of multi-MRF, mMRF and DKAT under four sets of simulations considering varying causal structures between mQTL SNPs and methylation traits. In each simulation, we also considered varying correlation structures of methylation traits, effect directions, sample sizes and the modeling strategies for the distributions of methylation traits. The results are summarized in Figures 3–6.

*Simulation I*: 'unique' causal structure (Fig. 3).

This simulation assumed that the CpG sites had mutually exclusive mQTL SNPs. Under such a scenario, the single-trait testing method with multiple testing adjustment (i.e. mMRF) was often the most powerful. In particular, the power of mMRF assuming a beta distribution (i.e. beta strategy) was consistently higher than that of DKAT regardless of the sample sizes and effect directions of the mQTL SNPs. mMRF also outperformed multi-MRF when the sample size was small (i.e. $n = 50$ or 100) or the mQTL SNPs had bi-directional effect (Fig. 3B). On the other hand, multi-MRF showed improved performance as the sample size increased, and achieved comparable or slightly higher power than mMRF when the sample size was relatively large ($n = 250$ or 500) and the mQTL SNPs had one-directional effect (Fig. 3A). We found the observation reasonable because the methylation traits at multiple CpG sites did not have overlapping causal components, which made the single-trait analysis (i.e. mMRF) a powerful test.

For the comparisons between the multi-trait methods (multi-MRF and DKAT), the power of multi-MRF was lower than that of DKAT when the sample size was very small ($n = 50$) but increased with the sample size and was higher than that of DKAT for all other sample sizes ($n = 100$, 250 and 500). The pattern was consistent regardless of the effect directions of mQTL SNPs.

The performance of each method was also influenced by the strategies to model the methylation trait's distributions. Assuming a beta distribution of methylation traits (i.e. beta strategy) was the most advantageous when the sample size was relatively small (i.e. $n \leq 250$), while assuming a log-transformed normal distribution (i.e. logit strategy) may achieve comparable or slightly higher power when the sample size was relatively large (i.e. $n = 500$). In particular, assuming a normal distribution of methylation traits (i.e. normal strategy) showed substantial power loss when multi-MRF was

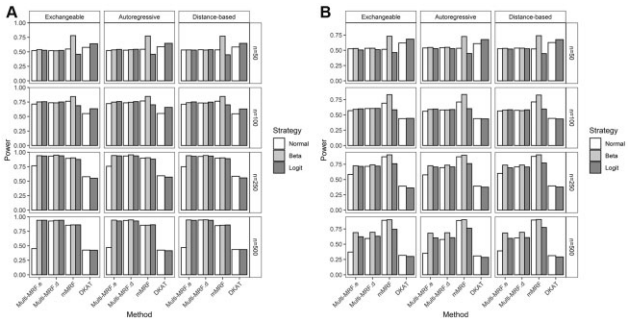


**Fig. 4.** Statistical power for Simulation II, when the causal structure was 'half-shared'. (**A**) The mQTL SNPs had one-directional effect on methylation traits; (**B**) The mQTL SNPs had bi-directional effect on methylation traits. Multi-MRF.e and multi-MRF.d represents applying the multi-MRF with exchangeable and distance-based correlation structures, respectively

applied with an exchangeable correlation structure, which was also consistent with the conservative Type I error observed above. The performance of DKAT was relatively robust assuming either a normal distribution or logit-transformed normal distribution.

*Simulation II*: 'half-shared' causal structure (Fig. 4).

This simulation assumed that half of the mQTL SNPs were shared by all CpG sites and the other half were mutually exclusive across all CpG sites. The results showed similar trend to those of *Simulation I*, and the single-trait method (mMRF) achieved the highest power in most of the scenarios, especially when the sample size was relatively small or the mQTL SNPs had bi-directional effect. When compared with *Simulation I*, we also observed improved performance of multi-trait testing methods. In particular, the power of multi-MRF increased much faster with the sample size, and outperformed mMRF when sample size was relatively large ($n = 250$ or 500) and the mQTL SNPs had one-directional effect (Fig. 4A). In such a situation, multi-MRF gained strength through modeling the correlation of methylation traits due to the shared mQTL SNPs. On the other hand, because 50% of the mQTL SNPs remained mutually exclusive for all methylation traits, mMRF was the best method in various scenarios. In terms of the strategies to model trait distributions, both mMRF and multi-MRF attained the highest power when modeling the methylation trait with a beta distribution.

*Simulation III*: 'all-shared' causal structure (Fig. 5).

This simulation assumed that all CpG sites shared the same mQTL SNPs. In such a scenario, multi-MRF showed the most advantages, and achieved either comparable or substantially higher power in all simulations (Fig. 5). The results were consistent across sample sizes, correlation structures and effect directions. The other multi-trait testing method, DKAT, was also able to outperform mMRF when the sample size was small ($n = 50$), but had lower

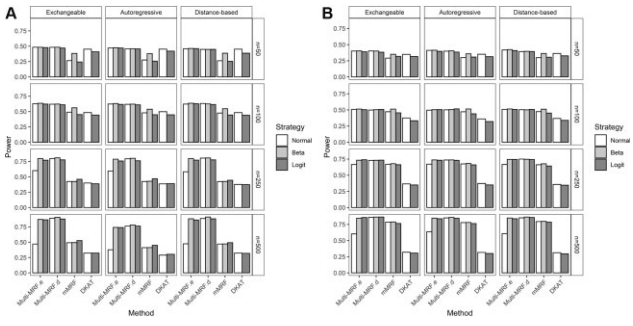


**Fig. 5.** Statistical power for Simulation III, when the causal structure was 'all-shared'. (**A**) The mQTL SNPs had one-directional effect on methylation traits; (**B**) The mQTL SNPs had bi-directional effect on methylation traits. Multi-MRF.e and multi-MRF.d represents applying the multi-MRF with exchangeable and distance-based correlation structures, respectively

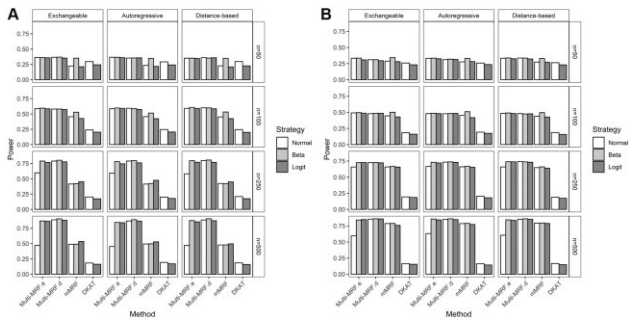


**Fig. 6.** Statistical power for Simulation IV, when only rare variants were tested for association with methylation traits. (**A**) The mQTL SNPs had one-directional effect on methylation traits; (**B**) The mQTL SNPs had bi-directional effect on methylation traits. The causal structure between mQTL SNPs and methylation traits were assumed to be 'all-shared'. Multi-MRF.e and multi-MRF.d represent applying the multi-MRF with exchangeable and distance-based correlation structures, respectively

power when the sample size was larger. In terms of strategies to model trait distributions, multi-MRF also showed robust performance across simulations, and modeling methylation traits with beta distributions had slightly higher power than other strategies. For mMRF, beta strategy worked better with relatively small sample size ($n = 50$ or 100) or bi-directional genetic effect, while logit transformation performed better when the sample size was relatively large ($n = 250$ or 500) under one-directional scenario.

*Simulation IV*: testing rare variants only (Fig. 6).

This simulation aimed to evaluate the performance of methods for detecting rare variants as underlying mQTLs. The scenario was illustrated under 'all-shared' causal structure. The results showed similar patterns with those of *Simulation III*. Multi-MRF models achieved higher power than mMRF and DKAT in all simulation scenarios and the improvement over other methods was most evident when the mQTL SNPs affected the methylation traits in one direction. Modeling the methylation traits with beta distributions also showed improved power than other strategies (i.e. normal distributions or logit-transformed distributions).

#### 3.1.3 Simulation summary

In summary, multi-MRF, mMRF and DKAT were all able to detect multi-locus mQTL with reasonably controlled Type I errors. However, the current version of DKAT did not allow covariate adjustment. Both multi-MRF and mMRF allowed the adjustment of confounding factors, and outperformed DKAT in most of the simulation scenarios. For the comparison between multi-MRF and mMRF, multi-MRF showed more advantages when multiple methylation traits share the same mQTL SNPs, the sample size was relatively large ($n \geq 250$), and the mQTL SNPs had one-directional effect on methylation traits. On the other hand, mMRF was more appropriate when the methylation traits had largely mutually exclusive mQTL SNPs. The conclusion was consistent across varying correlation structures.

The choice of correlation structures and modeling strategies of the methylation traits may also affect the testing power of each method. For multi-MRF, assuming distance-based correlation and beta distribution of methylation traits showed the most robust performance across simulation scenarios. We also found that applying multi-MRF assuming exchangeable correlation and normal distributions of methylation traits would yield conservative Type I errors and reduced statistical power, which should be avoided in practice. For mMRF, modeling methylation traits with beta distributions (i.e. beta strategy) performed best when the sample size was relatively small ($n = 50$ or 100), while modeling with logit-transformed normal distribution (i.e. logit strategy) may be a better choice for studies of larger sample sizes ($n \geq 250$). For DKAT, modeling methylation traits with logit-transformed normal distributions showed higher power with relatively small sample size, while modeling with normal distributions (i.e. normal strategy) may be more appropriate for larger sample sizes.

### 3.2 Application study

To identify *cis*-acting mQTLs, we tested a total of 15 695 genomic regions within 83 cardiac tissue samples. For each region, we applied multi-MRF, mMRF and DKAT to evaluate the joint association between all SNPs and all CpG sites. Based on the aim of our study and the simulation results, we prioritized the findings by applying multi-MRF assuming that the methylation traits followed beta distributions with a distance-based correlation structure. The application of multi-MRF identified a total of 162 significant mQTL regions after multiple testing adjustment. The complete results are summarized in Supplementary Table S2.

We further leveraged the findings from two previous CHD GWAS studies by comparing with the GWAS testing *P*-values of the SNPs within those 162 genomic regions. Most of these regions (107 out of 162) harbored SNPs with nominally significant associations in both GWAS studies. On the other hand, among the 15 695 regions tested, a total of 5406 contained nominally significant SNPs in both GWAS phases. A Fisher's exact test for enrichment ($p_{val} = 1.8 \times 10^{-16}$) showed that the GWAS-associated SNPs were significantly over-represented or enriched in the identified mQTL regions (107 out of 162) compared with all the regions tested (5406 out of 15 695).

In Table 1, we summarized the top 10 mQTL candidates among these 107 regions. The 10 mQTL regions were located on Chromosomes 2, 7, 8, 11, 16, 18, 20 and 21.

### 3.3 Bayesian colocalization

We further conducted Bayesian colocalization to prioritize mQTL findings leveraging the findings from two CHD GWASs and the expression QTL studies. The full colocalization results of 162 regions identified by multi-MRF are summarized in Supplementary Table S3. Among which, gene *BACE2*, as an mQTL candidate, had a high posterior probability to colocalize with expression QTLs in artery tibial ($PP4 > 0.8$). The gene region also harbored SNPs with nominal significance association with CHD risk in both phases of GWASs (Table 2). For the other identified mQTL regions, we did not observe strong evidence for colocalization, which may partly be due to insufficient power. It should also be noted that the existing GWASs and eQTL studies commonly utilized the single-trait single-locus testing strategy with a focus to detect common variants. In contrast, multi-MRF is a multi-locus multi-trait method aiming to detect both common and rare variants.

### 3.4 Comparison with the results of DKAT and mMRF

A total of 74 302 and 344 significant mQTL regions were identified by mMRF assuming methylation traits following a beta distribution, and by DKAT assuming either a normal or logit-transformed normal

**Table 1.** Top 10 significant genetic regions that were identified by multi-MRF and overlapped with nominal significant SNPs in two phases of CHD GWAS[a]

| Chr | Region | Gene | nCpGs | nSNPs | N1[b] | N2[c] | Method | Normal | Beta | Logit |
|---|---|---|---|---|---|---|---|---|---|---|
| chr18 | 34 815 507–35 153 500 | CELF4 | 64 | 463 | 10 | 15 | multi-MRF | 8.80e−10 | 7.85e−16 | 5.14e−09 |
| | | | | | | | mMRF | 0.13 | 0.13 | 0.05 |
| | | | | | | | DKAT | 0.16 | — | 0.21 |
| chr8 | 97 956 591–98 466 225 | TSPYL5 and LOC101927066 | 36 | 630 | 4 | 5 | multi-MRF | 5.28e−08 | 1.82e−11 | 1.46e−09 |
| | | | | | | | mMRF | 0.40 | 0.32 | 0.43 |
| | | | | | | | DKAT | 1.37e−04 | — | 4.32e−05 |
| chr7 | 18 119 064–19 049 537 | HDAC9 | 41 | 1314 | 20 | 16 | multi-MRF | 1.11e−12 | 8.08e−11 | 5.31e−15 |
| | | | | | | | mMRF | 0.06 | 0.02 | 0.01 |
| | | | | | | | DKAT | NA | — | NA |
| chr8 | 72 102 167–72 467 392 | EYA1 | 19 | 438 | 40 | 11 | multi-MRF | 4.74e−08 | 1.16e−10 | 1.59e−11 |
| | | | | | | | mMRF | 0.14 | 0.14 | 0.23 |
| | | | | | | | DKAT | 0.20 | — | 0.08 |
| chr20 | 13 968 645–16 041 341 | MACROD2 | 24 | 2807 | 86 | 56 | multi-MRF | 4.71e−10 | 1.47e−10 | 3.25e−10 |
| | | | | | | | mMRF | 0.04 | 0.05 | 0.03 |
| | | | | | | | DKAT | 0.03 | — | 0.02 |
| chr2 | 236 395 232–237 047 944 | AGAP1 | 270 | 844 | 23 | 33 | multi-MRF | 1.30e−11 | 4.99e−10 | 1.12e−08 |
| | | | | | | | mMRF | 5.43e−03 | 6.91e−03 | 2.38e−03 |
| | | | | | | | DKAT | 5.22e−04 | — | 0.04 |
| chr11 | 118 861 342–118 894 002 | CCDC84 | 8 | 51 | 1 | 8 | multi-MRF | 4.92e−07 | 6.27e−10 | 1 |
| | | | | | | | mMRF | 0.34 | 0.36 | 0.35 |
| | | | | | | | DKAT | 0.33 | — | 0.25 |
| chr18 | 7 559 813–8 414 359 | PTPRM | 25 | 1211 | 35 | 40 | multi-MRF | 6.23e−09 | 6.81e−10 | 1.30e−07 |
| | | | | | | | mMRF | 0.79 | 0.86 | 0.84 |
| | | | | | | | DKAT | 0.01 | — | 1.12e−03 |
| chr16 | 88 774 245–88 859 128 | PIEZO1 and MIR4722 | 96 | 188 | 5 | 1 | multi-MRF | 1.76e−07 | 1.96e−09 | 7.65e−07 |
| | | | | | | | mMRF | 2.63e−03 | 5.12e−03 | 2.66e−03 |
| | | | | | | | DKAT | 7.47e−04 | — | 5.06e−04 |
| chr21 | 38 113 425–38 370 194 | HLCS | 19 | 376 | 1 | 3 | multi-MRF | 1.41e−11 | 2.16e−09 | 3.55e−11 |
| | | | | | | | mMRF | 0.02 | 0.01 | 0.03 |
| | | | | | | | DKAT | 0.30 | — | 0.44 |

[a]Logit transform represents M values here because M values are proportional to the logit transformation of beta values.
[b]Number of nominal significant SNPs in CHD GWAS1.
[c]Number of nominal significant SNPs in CHD GWAS2.

distribution, respectively. However, we found relatively few overlapping findings between methods, and no overlaps for all methods (Supplementary Fig. S1). We think this is mainly because each method has unique advantages and disadvantages under various underlying causal scenarios. Based on the simulation results, multi-MRF was the most powerful in detecting rare variants which contributed to all CpG sites within the region, while mMRF was the most appropriate for situations when methylation traits were largely determined by mutually exclusive mQTL SNPs. On the other hand, as an extension of SKAT, we hypothesized that DKAT might be more advantageous to detect common variants regulating methylation traits.

## 4 Discussion

We proposed a multi-MRF method for mQTL detection from testing the association between a set of CpG sites and a set of genetic variants within a genomic region. This method leverages co-methylation among neighboring CpG sites, and in doing so achieved improved power over single-trait analysis in various situations. The benefits are most evident when multiple traits share the same genetic mechanisms or when the sample size is relatively large ($n \geq 250$). The multi-MRF shares similar strengths with previous region-based methods (He *et al.*, 2014; Li *et al.*, 2014; Lyu *et al.*, 2021) as a powerful multi-locus test for rare variants, accounting for the linkage disequilibrium or potential interactions among SNPs and was especially tailored for beta-distributed traits.

In the past few years, several computational tools have been developed for mQTL detection (Ongen *et al.*, 2016; Scherer *et al.*, 2021; Shabalin, 2012). These existing methods conduct single-locus, single-trait analysis, and are commonly used for detecting common variants (e.g. MAF > 0.05) as QTLs. A major goal of our study is to consider the effect of rare variants for mQTL detection. The proposed multi-MRF may serve as a complementary method to the existing ones. We also conducted additional simulations to compare the proposed multi-MRF with Matrix-eQTL, a well-established benchmark method (Shabalin, 2012). The simulation results are summarized in Supplementary Figures S2–S4. The results showed Matrix-eQTL had significantly inflated Type I errors for testing a mixture of common and rare variants or rare variants only (Supplementary Figs S2A and B). The power comparison would not be meaningful under such a situation. When only common variants were tested, Matrix-eQTL had well-controlled Type I errors, especially when sample size was relatively large (Supplementary Fig. S3). We further evaluated the statistical power of all methods (Supplementary Fig. S4) for testing common variants. Matrix-eQTL had highest power when the sample size is 50, while multi-MRF outperformed Matrix-eQTL for larger sample sizes (Supplementary Fig. S3). In summary, Matrix-eQTL is not appropriate for detecting rare mQTL SNPs due to the inflated Type I errors. When only common variants are tested, multi-MRF, DKAT and Matrix-eQTL are all viable options.

To evaluate the performance of all methods, our simulations were conducted to evaluate Type I errors (i.e. false positive rate) or statistical power (i.e. 1−false negative rate). It should be noted that the numbers of false positives and false negatives in practice will depend on the number of tests conducted under null hypothesis (H0: genes without mQTLs) and alternative hypothesis (H1: genes with

**Table 2.** mQTL regions colocalized with eQTLs in heart tissues with a threshold of PP4 > 0.8

| Chr | Regions | Gene | Source for coloc | PP0 | PP1 | PP2 | PP3 | PP4 | Multi-MRF *P* value | N1[a] | N2[b] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chr21 | 42 532 227–42 661 961 | *BACE2* | mQTL—Artery Tibial | 2.53e−05 | 4.96e−04 | 8.79e−03 | 0.13 | 0.86 | 1.71e−06 | 20 | 21 |

[a]Number of nominal significant SNPs in CHD GWAS1.
[b]Number of nominal significant SNPs in CHD GWAS2.

mQTLs). For example, we conducted a simulation scenario of 2000 tests (1000 under $H_0$ and 1000 under $H_1$) with 500 samples. We further assumed that the mQTLs had 'all-shared' causal structure as described in *Simulation III*. Applying multi-MRF with distance-based correlation had an estimated Type I error rate of 0.051 and statistical power of 0.87. The full confusion matrix was observed as:

| | Test positive (rejecting H0) | Test negative (not rejecting H0) | Total |
|---|---|---|---|
| H0 is true | 51 | 949 | 1000 |
| H1 is true | 870 | 130 | 1000 |
| Total | 921 | 1079 | 2000 |

For whole-genome analysis, multiple testing adjustment (e.g. Bonferroni correction or Benjamini–Hochberg's false discovery rate) should also be applied to limit the number of false positives.

A major feature of multi-MRF is that flexible assumptions can be made to model the potential correlation between CpG sites. Due to the utilization of Generalized Estimating Equation for statistical inference, the method may be robust to the misspecification of correlation structure. However, for better estimation and statistical power, it is highly recommended to choose the appropriate correlation structure based on existing knowledge. Multi-MRF is especially advantageous to detect mQTLs when the same mQTL SNPs contribute to the methylation traits of all CpG sites within the region. In Table 3, we provide an empirical guideline to choose among methods based on their strengths and limitations. We think other existing tools, such as fastQTL and MAGAR (Ongen *et al.*, 2016; Scherer *et al.*, 2021), share many strengths with Matrix-eQTL.

When applied to a tissue-specific study of CHDs, the multi-MRF identified some *cis*-mQTLs regions with evidence of biological plausibility. One of the mQTL candidates, *BACE2*, was colocalized with expression QTL in artery tibial and overlapped with nominal significant findings in two CHD GWASs, suggesting a potential pathway linking genetic variants, DNA methylation and gene expression to CHD status. Previous literature indicated that *BACE2* was a critical region for Down syndrome, a type of syndromic CHD (Asim *et al.*, 2015). However, its association with non-syndromic CHD has not been reported and may merit further investigation. In addition, several genes identified by multi-MRF have been previously reported in association with CHDs. For example, both an animal study (Guo *et al.*, 2011) and an epidemiological study (Li *et al.*, 2018a) suggested that *EYA1* was associated with conotruncal heart malformations. Meanwhile, *HDAC9* is involved in cardiac development. The Bayesian test for colocalization is underpowered probably due to the small sample size in mQTL study ($n = 83$). Nevertheless, we believe the mQTLs identified by multi-MRF may also reveal some novel signals to represent rare variants of relatively large and pleiotropic effect.

Although applying different methods, very few overlaps were observed among the mQTL regions identified by multi-MRF, mMRF and DKAT. This observation is reasonable because each method has unique advantages and disadvantages under various causal scenarios. They should be viewed as complementary methods for detecting mQTL regions underlying varying biological mechanisms.

Our study must be considered in the light of certain limitations. First, DNA methylation varies by tissue and cell types. We conducted tissue-specific analysis but were not able to quantify the cell types within tissues. A few computational methods, referred to as deconvolution methods, have been developed to estimate cell compositions from DNA methylation data (Li and Wu, 2019; Rowland *et al.*, 2022; Zhang *et al.*, 2021). These methods may be used to adjust for the heterogeneity due to different cell types. Second, the computation time of multi-MRF depends on the sample size, the dimensionality of traits and the number of genetic variants within a genomic region, and the

**Table 3.** Empirical guideline to choose among methods in practice

| | Considerations | Multi-MRF | MRF | DKAT | Matrix-eQTL |
|---|---|---|---|---|---|
| Input | Programming language | R | R | R | R |
| | Adjustment of covariates | √ | √ | — | √ |
| | Genomic locations of CpG sites | √ | — | — | — |
| Working scenario | Multi-trait testing | √ | — | √ | — |
| | Multi-variant testing | √ | √ | √ | — |
| | Detecting rare variants as mQTLs | √ | √ | √ | — |
| | Detecting common variants as mQTLs | √ | √ | √ | √ |
| | Accounting for pleotropic effect on multiple CpG sites | √ | — | √ | — |
| | Top performer for modeling traits with beta distributions | √ | √ | — | — |
| | Top performer for modeling traits with normal distributions after logit-transformation | — | — | √ | √ |
| | Top performer with small sample size (e.g. <100) | √ | √ | — | — |
| | Top performer with large sample size (e.g. 500) | — | — | √ | √ |
| Speed | Fast method for quick results | — | — | — | √ |
| Output | Method gives *P* value | √ | √ | √ | √ |
| | Burden of multiple testing adjustment | Low | Median | Low | High |

**Table 4.** Distributions for the numbers of CpG and variants within a gene

| | Min | First quartile | Median | Mean | Third quartile | Max |
|---|---|---|---|---|---|---|
| No. of CpG sites | 2 | 5 | 8 | 12.84 | 14 | 296 |
| No. of variants | 2 | 32 | 56 | 110 | 111 | 7200 |

**Table 5.** Runtime for a representative gene with varying numbers of CpG sites and variants

| Number of CpGs | Number of variants | Runtime |
|---|---|---|
| 2 | 2 | 0.31 s |
| 13 | 109 | 1.32 s |
| 50 | 3000 | 51.84 s |
| 93 | 7200 | 6.37 min |
| 270 | 844 | 9.84 min |
| 296 | 246 | 14.44 min |

number of bootstrap resampling for empirical *P*-values to avoid conservative Type I errors (Wu *et al.*, 2011). In our application data, the distributions for the number of CpG sites and number of variants within a gene are given in Table 4. The expected runtimes are provided in Table 5.

Empirically, a median-sized gene with 13 CpG sites and 109 variants can be tested in 1.32 s on a local MacBook (Catalina 10.15.7; x86_64-apple-darwin17.0). Over 75% and 95% of genes in our dataset (450 K array) have <15 and 50 CpG sites, respectively. We also expect that there could be more genes with a larger number of CpG sites for studies with 800 K arrays and bisulfite sequencing. We recommend partitioning large genes into smaller regions each with <300 CpG sites to ensure computational speed. Testing a large number of genetic markers has been a challenge because of the computational burden to estimate the full set of eigenvalues of a high-dimensional matrix. Approximation strategies have been proposed and successfully applied to region-based testing methods, such as SKAT (Lumley *et al.*, 2018). We are evaluating similar strategies to improve the computation speed of our method. Third, within the current multi-MRF framework we only considered unrelated subjects. Future work may evaluate multi-trait analysis incorporating family data and consider the potential interaction between genetics and epigenetics. Fourth, due to the small sample size of the cardiac tissues, we chose MAF < 0.05 as an operational cutoff for rare variants. A lower threshold (e.g. 0.01) may be used for studies with a larger sample size.

## Acknowledgements

We wish to thank the anonymous reviewers and the Associate Editor for their valuable comments.

## Funding

This work was supported in part, by the National Heart, Lung and Blood Institute under award number [K01HL140333 to M.L.], the Eunice Kennedy Shriver National Institute of Child Health and Human Development under award number [R03HD092854 to M.L.] and [R01HD039054 to C.A.H.], the National Institute of Dental and Craniofacial Research under award number [R03DE024198 to N.L.] and [R03DE025646 to N.L.] and the National Science Foundation under award number [DMS 2002865 to N.L.].

*Conflict of Interest*: none declared.

## Data availability

The genetic and epigenetic data supporting the current study will be deposited to the database of Genotypes and Phenotypes (dbGaP) following the data sharing guideline of NHLBI and NICHD, and are available from the corresponding author on reasonable request.

## References

Affinito,O. *et al.* (2020) Nucleotide distance influences co-methylation between nearby cpg sites. *Genomics*, **112**, 144–150.

Almli,L.M. *et al.* (2015) A genome-wide identified risk variant for PTSD is a methylation quantitative trait locus and confers decreased cortical activation to fearful faces. *Am. J. Med. Genet.*, **168**, 327–336.

Aryee,M.J. *et al.* (2014) Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Aschard,H. *et al.* (2014) Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.*, **94**, 662–676.

Asim,A. *et al.* (2015) Down syndrome: an insight of the disease. *J. Biomed. Sci.*, **22**, 41.

Battle,A. *et al.*; eQTL manuscript working group. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

Bayes,C.L. *et al.* (2012) A new robust regression model for proportions. *Bayesian Anal.*, 7, 841–866.

Boos,D.D. (1992) On generalized score tests. *Am. Stat.*, **46**, 327–333.

Dick,K.J. *et al.* (2014) DNA methylation and body-mass index: a genome-wide analysis. *Lancet*, **383**, 1990–1998.

Fortin,J.-P. *et al.* (2017) Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, **33**, 558–560.

Fortin,J.-P. *et al.* (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.*, **15**, 503.

Gaunt,T.R. *et al.* (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.*, **17**, 61.

Giambartolomei,C. *et al.* (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.

Guo,C. *et al.* (2011) A Tbx1-Six1/Eya1-Fgf8 genetic pathway controls mammalian cardiovascular and craniofacial morphogenesis. *J. Clin. Invest.*, **121**, 1585–1595.

Guo,X. *et al.* (2005) Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Stat. Med.*, **24**, 3479–3495.

He,Z. *et al.* (2015) Set-based tests for genetic association in longitudinal studies. *Biometrics*, **71**, 606–615.

HeZ. *et al.* (2014) Modeling and testing for joint association using a genetic random field model. *Biometrics*, **70**, 471–479.

Huan,T. *et al.* (2019) Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.*, **10**, 4267.

Lee,S. *et al.* (2013) General framework for Meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.*, **93**, 42–53.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Li,B. *et al.* (2018a) In silico analyses reveal the relationship between SIX1/EYA1 mutations and conotruncal heart defects. *Pediatr. Cardiol.*, **39**, 176–182.

Li,M. *et al.* (2018b) Detecting rare mutations with heterogeneous effects using a family-based genetic random field method. *Genetics*, **210**, 463–476.

Li,M. *et al.* (2014) A generalized genetic random field method for the genetic association analysis of sequencing data. *Genet. Epidemiol.*, **38**, 242–253.

Li,M. *et al.* (2021) Mapping methylation quantitative trait loci in cardiac tissues nominates risk loci and biological pathways in congenital heart disease. *BMC Genom. Data*, **22**,1–12.

Li,Z. and Wu,H. (2019) TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.*, **20**, 1–17.

Lövkvist,C. *et al.* (2016) DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res.*, **44**, 5123–5132.

Lumley,T. *et al.* (2018) FastSKAT: sequence kernel association tests for very large sets of markers. *Genet. Epidemiol.*, **42**, 516–527.

Lyu,C. *et al.* (2021) Detecting methylation quantitative trait loci using a methylation random field method. *Brief. Bioinform.*, **22**, bbab323.

Morrow,J.D. *et al.* (2018) Human lung DNA methylation quantitative trait loci colocalize with chronic obstructive pulmonary disease genome-wide association loci. *Am. J. Respir. Crit. Care Med.*, **197**, 1275–1284.

Nautiyal,S. *et al.* (2010) High-throughput method for analyzing methylation of CpGs in targeted genomic regions. *Proc. Natl. Acad. Sci. USA*, **107**, 12587–12592.

O'Reilly,P.F. *et al.* (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, **7**, e34861.

Ongen,H. *et al.* (2016) Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, **32**, 1479–1485.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rowland,B. *et al.* (2022) THUNDER: a reference-free deconvolution method to infer cell type proportions from bulk Hi-C data. *PLoS Genet.*, **18**, e1010102.

Scherer,M. *et al.* (2021) Identification of tissue-specific and common methylation quantitative trait loci in healthy individuals using MAGAR. *Epigenetics Chromatin.*, **14**, 44.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Smith,A.K. *et al.* (2014) Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics.*, **15**, 145.

Tang,C.S. and Ferreira,M.A. (2012) A gene-based test of association using canonical correlation analysis. *Bioinformatics*, **28**, 845–850.

van der Sluis,S. *et al.* (2013) TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet.*, **9**, e1003235.

Volkov,P. *et al.* (2016) A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. *PLoS One*, **11**, e0157776.

Wu,B. and Pankow,J.S. (2016) Sequence kernel association test of multiple continuous phenotypes. *Genet. Epidemiol.*, **40**, 91–100.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.

Zhan,X. *et al.* (2017) Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, **206**, 1779–1790.

Zhang,W. *et al.* (2021) Complete deconvolution of DNA methylation signals from complex tissues: a geometric approach. *Bioinformatics*, **37**, 1052–1059.