Statistical methods for genetic and epigenetic studies

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Yun Bai

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

September, 2016

Acknowledgements

I would like to thank my thesis advisor, Dr. Weihua Guan, for providing invaluable guidance throughout this research, and my co-advisor, Dr. Haitao Chu, for his help and encouragement.

I would also like to thank Dr. Wei Pan, Dr. Xianghua Luo and the other members of my committee, Dr. James Pankow and Dr. Lin Zhang for their insightful comments.

I owe my loving thanks to my family who gave me their unconditional support and encouragement throughout.

Abstract

A common theme to many current large-scale genetic and epigenetic studies is their high-throughput nature of interrogating hundreds of thousands of genetic markers simultaneously. Inherent to these large-scale measurements are the inevitable technical variations of no biological interest. Typically pre-processing methods are applied to remove these technical variations and various other unwanted variations (e.g., batch effects) so that we can obtain unbiased estimates. Most statistical methods typically treat these processed measures as gold standard without any errors in the downstream analysis.

In this thesis, we aim to develop unified modeling approaches to accommodating these technical variations into downstream statistical analysis. Motivated by the Atherosclerosis Risk In Communities (ARIC) Study, we develop alternative statistical methods to incorporate these technical variations to analyze the epigenome-wide methylation data. Specifically we will study the reproducibility of the methylation measures (Chapter 3) and the epigenome-wide association studies (Chapter 4) incorporating these technical variations.

Similar to the epigenome-wide methylation data, the single nucleotide polymorphism (SNP) data provides another genome-wide measures of genetic markers. In the past decade, the genome-wide association studies (GWAS) have found thousands of SNPs associated with various diseases. Most large-scale GWAS have taken a marginal association test approach: testing the association of each trait and marker individually. The GWAS summary statistics (e.g., association test statistics) are generally publicly posted. However the raw genotype and phenotype data are more difficult to share publicly due to privacy and various logistic reasons. Therefore it is desirable to develop statistical methods that can take and mine these publicly available summary data to gain additional insights. In thesis, we develop a statistical method that just needs the summary data from multiple GWAS conducted on the same cohort (i.e., the same genotype data with multiple traits) to identify additional genetic variants that are associated with the outcomes (Chapter 2).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the last decade, the genome-wide association studies (GWAS) have proven very successful and uncovered many disease associated variants. The dominant approach in GWAS is a single-trait-single-variant based association test, where we test for the association of each genetic variant with each outcome separately. For most large cohort studies that have conducted GWAS, there are typically many correlated phenotypes available on each individual. However there haven't been many reported large-scale association studies that try to simultaneously test for multiple correlated phenotypes. Instead, individual trait based GWAS are typically reported in separate studies. Recently there have been many researches suggesting that we can often have improved power by joint test of multiple correlated traits compared to testing each trait separately. Many multi-trait association

test methods have been proposed in the literature, and they can be broadly classified into two classes. The first class is univariate association test based, where we typically summarize multiple traits into one summary score, which is then linked to the genetic variant through familiar association test approach, or use the minimum p-value across multiple traits to summarize the variant association signal. The second class is multivariate test method, where we try to test the overall association of genetic variant with multiple traits simultaneously. Most of these developed methods have assumed that the raw phenotype and genotype data are available. However, it is often logistically difficult to share these raw data. Given that most GWAS have been published with their summary statistics available, it is practically important to develop multi-trait association test methods that can directly take these summary statistics as input. In this thesis, we will develop one such approach that can mine the publicly available GWAS summary statistics on multiple traits from the same cohort to identify more interesting genetic variants.

Although GWAS have been very successful at identifying thousands of disease associated variants, in total the identified variants have only explained a small proportion of the variation of most phenotypes. Besides genetic variants, it is also worthwhile to look at other epigenetic components that can potentially contribute to this missing heritability. In this thesis, we will study one special epigenetic modification: the DNA methylation. Epigenetics is the study of mitotically heritable

modifications in chromatin structure (i.e., modifications not involving the underlying DNA sequence), and their impact on the transcriptional control of genes and cellular function. Epigenetic variation includes post-translational modifications of histone proteins, non-coding RNAs, and DNA methylation, the latter primarily occurring at cystosine-guanine dinucleotides (CpGs). Recent technological advances have provided multiple platforms for systematically interrogating DNA methylation. The recently released HM450 array includes 485,577 CpG sites and provides coverage of 98.9% of RefSeq genes with a global average of 17.2 probes per gene region. Inherent to the large-scale methylation measurements is the inevitable technical variations of no biological interest. Typically pre-processing methods are applied to remove these technical variations and various other unwanted variations (e.g., batch effects) so that we can obtain unbiased estimates. Most statistical methods typically treat these processed measures as gold standard without any errors in the downstream analysis. In this thesis, motivated by the Atherosclerosis Risk In Communities (ARIC) Study, we develop an alternative statistical model to incorporate these technical variations to assess the reproducibility of epigenome-wide methylation measures. We note that from the HM450 methylation array, part of the variation for the reported methylation measures can actually be quantified. Existing methods have typically chosen to ignore this (quantifiable) variation. We will develop a statistical model that

will explicitly incorporate this known variation and can produce more accurate measure of reproducibility.

In the epigenome-wide association study (EWAS), epigenetic marks can be investigated across the epigenome without pre-specifying the genes or regions in which inter-individual variation in DNA methylation is thought to be important for phenotypic variation. However, unlike inherited changes to the genetic sequence, variation in site-specific methylation varies by tissue, stage of development, disease state, and may be impacted by aging and exposure to environmental factors such as diet or smoking. Because DNA methylation patterns can change over time, EWAS (in contrast to GWAS) are subject to many of the same threats to validity that affect traditional epidemiological investigations, including reverse causality and confounding by non-genetic factors that may affect both methylation and risk of disease. Therefore it is important to appropriately adjust for potential confounding factors when conducting EWAS. In EWAS of methylation data, the methylation levels are typically regressed on clinical phenotype of our interest adjusting for various demographic and confounding variables to detect differentially methylated loci associated with outcome. Normal distribution is commonly assumed to model the methylation data. To remove various batch effects of non-biological interest (e.g., chip), we further incorporate additional random effects terms that are typically assumed normally distributed (partly for the modeling

convenience). This leads to the commonly used linear mixed effects model (LMM) widely used in EWAS. Implicit to these LMM, the technical variations are typically assumed to be homogeneous across samples and thus incorporated as part of the random noise in EWAS. Intuitively these EWAS can benefit from accounting for these technical variation. However incorporating these technical variation brings computational challenges to the LMM. In this thesis, we study methods to incorporate these technical variation into association test, and develop a very easy to implement EM algorithm applicable to general LMM with additional measurement errors. We will empirically show that standard statistical methods failing to properly accounting for these heterogeneous technical variations may lead to inflated type I errors. In contrast, our developed method can appropriately control the type I errors.

Chapter 2

Adaptive SPU test for association test of multiple related phenotypes using GWAS summary statistics

2.1 Introduction

Most current genome-wide association studies (GWAS) are typically conducted analyzing each phenotype independently, even when multiple phenotypes are available on each individual. Most detected common variants for most GWAS have had very small effect sizes. For example, the identified genetic variant odds ratio is often in the range of 1.1-1.3. And thus typical GWAS has assembled a large number of samples in order to achieve a decent detection power. Recently there have been more and more evidence showing that by jointly testing multiple correlated traits can actually increase our power and chance of detecting interesting genetic variants. One drawback of testing multiple traits is the potential power loss as more non-associated genetic variants are incorporated. Interpretation of testing multiple traits is also another potential drawback, since a joint significant association does not indicate which specific phenotype are associated. Nonetheless it has been shown that joint multivariate association testing can increase power, thus it leads us one-step further toward identifying more interesting genetic variants.

Recently there have been some methods proposed to study the multi-trait associations (see Klei *et al.*, 2008; Ferreira and Purcell, 2009; Yang *et al.*, 2010; Stephens, 2013; van der Sluis *et al.*, 2013; He *et al.*, 2013, e.g.). A natural approach is to combine the uni-trait analysis results using, e.g., the minimum p-value

approach (Yang *et al.*, 2010). Another class of methods are based on the intuitive idea of dimension reduction: the multivariate traits are summarized into a univariate score, which is then subject to traditional univariate association test. Commonly used dimension reduction methods include the principal component analysis (PCA) (Wang and Abbott, 2008) and the principal components of heritability (PCH) (Klei *et al.*, 2008). For PCA, the top few PCs are constructed to preserve the multivariate trait variation, which however is not guaranteed to capture the association signals. PCH is based on sample splitting and is not efficient. The canonical correlation analysis (CCA) (Ferreira and Purcell, 2009) finds the linear combination of multivariate traits that has the largest correlation with the genetic variant. Galesloot *et al.* (2014) conducted comprehensive simulation studies comparing several commonly used multi-trait association test methods, and concluded that multivariate testing methods typically performed best under majority of the tested scenarios and resulted in a higher power than univariate analysis. Hence they recommended the use of multivariate GWAS methods, even when genetic correlations between traits are weak. For testing multiple neuroimaging phenotypes, Zhang *et al.* (2014) studied the generalized estimating equation based score test for multi-trait associations and compared the performance of various methods. They have found that different tests could provide complementary and useful results. The multivariate testing methods typically have more power

than univariate testing based methods.

Most existing methods typically require the use of raw genotype and phenotype data. When only the GWAS summary Z-scores are available, Stephens (2013) proposed a Bayesian modeling of summary Z-scores by partitioning traits into three categories: directly associated, indirectly associated, and unassociated. The posterior probabilities of these partitions give direct inference of multi-trait associations. In a recently published article, Zhu *et al.* (2015) proposed a chi-square test to test multiple correlated traits using only GWAS summary statistics. They proposed to estimate the correlation of summary Z-statistics from (ideally independent null) SNPs, similar to the idea studied at Stephens (2013). In this paper, we study alternative frequentist methods for multi-trait association testing with only GWAS summary Z-scores that can be quickly implemented.

2.2 Methods

Consider a GWAS with n independent individuals and we want to test the genetic association of SNP with K related continuous traits while adjusting for a set of p covariates. Denote the outcomes as $\boldsymbol{Y}_k = (y_{1k}, \ldots, y_{nk})^T$, $k = 1, \ldots, K$, covariate $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip})^T$, and genotype g_i as the number of minor alleles, $i = 1, \ldots, n$, Let $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$, and $\boldsymbol{G} = (g_1, \ldots, g_n)^T$.

We assume the linear regression model for $\boldsymbol{Y}_k$

$$\boldsymbol{Y}_k = \alpha_k + \boldsymbol{X}\gamma_k + \boldsymbol{G}\beta_k + \epsilon_k,$$

where γ_k is of length p, ϵ_k is of length n and assumed to independently follow a normal distribution with mean zero and variance σ_k^2. Denote the correlation $\rho_{kl} = corr(y_{ik}, y_{il})$, $k, l = 1, \ldots, K$.

Denote $\boldsymbol{G}_e$ as the residual vector of regressing $\boldsymbol{G}$ on $\boldsymbol{X}$. We then have $\hat{\beta}_k = \boldsymbol{G}_e^T\boldsymbol{Y}_k/(\boldsymbol{G}_e^T\boldsymbol{G}_e)$ and can construct a corresponding Z-score $Z_k = \hat{\beta}_k/se(\hat{\beta}_k)$ for testing the genetic association of the k-th trait. Here $se(\hat{\beta}_k) = \hat{\sigma}_k/\sqrt{\boldsymbol{G}_e^T\boldsymbol{G}_e}$. Therefore we have

$$Cov(\hat{\beta}_k) = \frac{\sigma_k^2}{\boldsymbol{G}_e^T\boldsymbol{G}_e},$$

and

$$Cov(\hat{\beta}_k, \hat{\beta}_j) = \frac{\boldsymbol{G}_e^T Cov(\boldsymbol{Y}_k, \boldsymbol{Y}_j)\boldsymbol{G}_e}{(\boldsymbol{G}_e^T\boldsymbol{G}_e)^2} = \frac{\sigma_k\sigma_j\rho_{kj}}{\boldsymbol{G}_e^T\boldsymbol{G}_e}, \quad \forall k \neq j.$$

And hence $corr(\hat{\beta}_k, \hat{\beta}_j) = \rho_{kj}$, and $corr(Z_k, Z_j) \approx \rho_{kj}$. So in general the correlation of Z-scores reflect the correlation of outcomes and are independent of covariates including the genotype. With only summary data, they can be empirically estimated from standardized $\hat{\beta}_k$ across null SNPs. In practice, we choose those roughly independent null SNPs, and use their sample correlation matrix to estimate $\hat{\rho}_{kj}$. Denote the correlation matrix $R = (\rho_{kj})$. Therefore under null, we can assume that for any given SNP, their standardized test statistics across

traits follow a multivariate normal distribution with covariance matrix R. As shown in Lin and Zeng (2010), meta-analysis using GWAS summary statistics is roughly equivalent to performing association test using individual participant genotype and phenotype data. Therefore we expect the previously derived results also apply to GWAS meta-analysis summary statistics.

Given the Z-scores $Z = (Z_1, \ldots, Z_K)^T$ and the associated estimated correlation matrix $\hat{R}$, naturally we can construct the chi-square statistics $\chi^2 = Z^T \hat{R}^{-1} Z$, which follows a K-DF chi-square distribution under null and hence we can easily compute its significance p-value, $1 - F_k(\chi^2)$, where $F_k()$ is the K-DF chi-square distribution function.

In addition we also propose the sum of powered Z-statistics, $S_\gamma = \sum_{k=1}^{K} Z_k^\tau$, following the SPU approach of Pan *et al.* (2014), and denote its significance p-value as p_τ. Following Pan *et al.* (2014), we choose power $\tau = 1, 2, \ldots, 8, \infty$. Here S_∞ is essentially using the maximum statistic $\max_{k=1}^{K} |Z_k|$. And since all Z-scores are standardized, it is also equivalent to using the minimum p-value across all traits. To adaptively choose an optimal power, we study the aSPU test by using the minimum p-value, $S_m = \min_\tau p_\tau$, as a test statistic.

Except for $\tau = 1, 2$, it is in general very hard to derive and compute the analytical null distribution and significance p-value of S_τ. Instead we use the Monte Carlo simulation to approximate its null distribution. Specifically we simulate

$\{T_b = (t_{b1}, \ldots, t_{bK}) : b = 1, \ldots, B\}$ from the zero mean multivariate normal distribution with covariance R. Denote $S_\tau^b = \sum_{k=1}^K t_{bk}^\tau$. Then the significance p-value of S_τ can be computed as

$$p_\tau = \frac{1 + \sum_{b=1}^B I(|S_\tau^b| > |S_\tau|)}{B+1}.$$

Note that all SNPs have the same covariance matrix R under null. Therefore we just need to simulate one set of Monte Carlo samples and efficiently compute significance values for all SNPs simultaneously.

For the adaptive S_m, we can also efficiently compute its significance p-value based on the same set of Monte Carlo samples as following. Note that for S_τ^b itself, we can compute its null p-value based on its rank

$$p_\tau^b = \frac{1 + \sum_{l \neq b} I(|S_\tau^l| > |S_\tau^b|)}{B}.$$

Denote its minimum as $p_m^b = \min_\tau p_\tau^b$. Then we can compute the significance of aSPU as

$$p_m = \frac{1 + \sum_{b=1}^B I(|p_m^b| < |S_m|)}{B+1}.$$

Next we analyze the global lipids consortium summary GWAS data.

2.3 Application to global lipids GWAS data

We apply the proposed methods to GWAS data from the Global Lipids consortium (Teslovich *et al.*, 2010). The study meta analyzed 46 lipid GWAS studies with a total of more than 100,000 individuals of European ancestry. Around 2.6 million directly genotyped or imputed SNPs were tested for association with each of the four lipid traits in each study: total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG). For each SNP and each trait, evidence of association was combined across studies using a fixed-effects meta-analysis. There were 95 loci that showed genome-wide significant association (p-value $< 5 \times 10^{-8}$) with at least one of the four traits tested. For this meta study, the raw genotype and phenotype data are not available. The univariate summary Z-scores for each SNP and each trait are downloaded from `http://www.sph.umich.edu/csg/abecasis/public/lipids2010`. We note that TC=HDL+LDL+0.2×TG. So TC provides redundant information. In the following we jointly analyze TG,HDL, and LDL to identify those SNPs associated with at least one of the traits.

We first filter out those SNPs that are missing from at least one lipid trait and identified a total of 2,691,421 SNPs with summary Z-scores for all traits. We use 10^9 Monte Carlo simulations to estimate the p-value for SPU and aSPU tests.

There were in total 5395 SNPs from 96 independent loci (at least 1.5M bp away from each other) that passed the genome-wide significance level (p-value $< 5 \times 10^{-8}$) in at least one of the traits. For individual trait analysis, there were 2593 significant SNPs from 52 independent loci for TC, 1808 significant SNPs from 32 independent loci for TG, 2213 significant SNPs from 52 independent loci for HDL, 1769 significant SNPs from 36 independent loci for LDL.

When analyzing the summary statistics for TG,HDL,LDL, we first empirically estimated the correlation matrix of Z-scores using non-null SNPs (excluding those genome-wide significant SNPs in at least one of the traits, i.e., minimum p-value $< 5 \times 10^{-8}$). Although we could potentially only select those roughly independent SNPs, which however could bring in extra randomness of selection.

Applying the 3-DF chi-square test, we identified 4697 significant SNPs from 81 loci. While applying the aSPU test, we identified 4286 significant SNPs from 77 loci.

To compare the results of 3-DF chi-square test and the aSPU test, we only focus on those SNPs with annotations. In common, both methods identified 4130 SNPs, and there were 538 SNPs uniquely identified by the chi-square test. There were in total 128 SNPs from 3 loci uniquely identified by the aSPU test, with the following representative SNPs: rs628751 located at chromosome 6 and basepair 139838419, rs12670798 located at chromosome 7 and basepair 21607352,

rs11700250 located at chromosome 20 and basepair 34132256. The SNP rs628751 is located close to gene LOC645434, rs12670798 is located in gene DNAH11, and rs11700250 is located in gene ERGIC3.

We also look to see wether SPU based tests can identify any interesting SNPs compared to the minimum p-value based approach. For illustration, we pretend we only have meta-analysis results from the three traits, TG,HDL,LDL only. We first identify those SNPs with the minimum meta-analysis p-values among TG,HDL,LDL smaller than 5×10^{-8} (or absolute Z-scores larger than 5.45). The aSPU test identified four independent SNPs located in two loci in addition to these meta-analyses significant SNPs. Three of them, rs11700250, rs2104417, and rs2277862, are located at chromosome 20. For these three SNPs, the absolute values of their reported three meta-analysis Z-scores are all smaller than 4.52, and did not reach genome-wide significance, while all their TC meta-analyses Z-scores are close to 6.23 and thus passed genome-wide significance. Another SNP, rs12751742, is located at chromosome 1, and the reported meta-analysis Z-scores are 3.43, 3.49, -5.08, and 4.48 for the four traits, TC, TG, HDL, and LDL respectively.

Next we conduct simulation studies to investigate the performance of the proposed methods.

2.4 Simulation studies

We simulate a GWAS with n unrelated individuals. The simulation setup mimics the lipids GWAS data (see next section for details). We study three quantitative traits (Y_1, Y_2, Y_3) simulated from a multivariate normal distribution with unit variance and correlation matrix Σ. We consider two covariates: an indicator X_1 following the Bernoulli distribution with probability of 0.5, and a continuous X_2 following the standard normal distribution. We independently simulate m_0 null SNPs with the MAF randomly simulated from $U[0.05, 0.45]$, and $m_1 = 96$ causal SNPs $(G_1, \ldots, G_{96})$ with the MAF of p_0. We set the three pairwise correlations being $\Sigma(1,2) = -0.42$, $\Sigma(1,3) = 0.27$, and $\Sigma(2,3) = -0.08$, based on the computed correlation matrix from the lipids GWAS data. We set $p_0 = 0.25$ and set the causal SNP effects as follows. We first identify 96 independent loci based on the lipids GWAS summary data. Specifically we rank the SNPs based on their minimum p-values and then select the top SNP within each region. From the reported summary Z-statistics Z_{kj} for SNP $j = 1, \ldots, 96$ and trait $k = 1, 2, 3$, we compute their effect size as $\beta_{kj} = Z_{kj}/\sqrt{2Np_0(1-p_0)}$, where $N = 10^5$ is roughly the lipids GWAS sample size. In the appendix, we list the reported lipids GWAS summary Z-statistics for these 96 SNPs.

We then simulate the three outcomes based on

$$\begin{aligned} Y_1 &= 0.5 + 2X_1 + 3X_2 + \sum_{j=1}^{96} G_j\beta_{1j} + \epsilon_1, \\ Y_2 &= 0.5 + 2X_1 + 2X_2 + \sum_{j=1}^{96} G_j\beta_{2j} + \epsilon_2, \\ Y_3 &= 0.5 + 2X_1 + 3X_2 + \sum_{j=1}^{96} G_j\beta_{3j} + \epsilon_3, \end{aligned}$$

where $(\epsilon_1, \epsilon_2, \epsilon_3)$ follow the zero-mean multivariate normal distribution with covariance matrix Σ.

We used 10^7 Monte Carlo simulations to estimate significance p-values for SPU tests. We conducted 100 experiments to estimate Type I errors and the power. The null SNPs are pooled together for Type I error estimation. We also evaluate the accuracy of correlation matrix estimation using summary statistics and their impact on association test with summary statistics. We evaluated the performance of three methods. The first one is based on Bonferroni adjusted minimum p-values across the three outcomes. The last two are the 3-DF chi-square test and the SPU tests based on the estimated empirical correlation matrix using all SNPs. Specifically for each simulated data, we apply the "lm()" R function to compute the Z-statistics for each SNP. We then estimate R using the summary Z-statistics, and simulate random numbers using the "mvtnorm" R package to estimate p-values for SPU statistics. We conducted two sets of

simulations: $n = 10^3, m_0 = 10^5$ and $n = 10^4, m_0 = 10^4$.

2.4.1 Correlation of Z-statistics

As we have shown previously, the correlation of Z-statistics should reflect the correlation of the outcomes. In the simulation study, there are three components that contribute to the variation of outcomes. The first one is the linear component of X_1 and X_2, which will be adjusted out in the linear model. The second one is the component of causal SNP signal, $\eta_k = \sum_{j=1}^{96} G_j \beta_{kj}, k = 1, 2, 3$. And the third one is the random error component ϵ_k. When testing the association of null SNPs, the last two components are not adjusted and hence the correlation matrix of null Z-statistics should roughly be the pairwise correlation of $\eta_k + \epsilon_k, k = 1, 2, 3$. For the previous simulation setup, these two components are independent and we can easily compute the variance due to η_k. Specifically $Var(\eta_k + \epsilon_k) = Var(\epsilon_k) + \sum_{j=1}^{96} 2\beta_{kj}^2 f_j(1 - f_j)$, where f_j is the MAF of G_j. Here $2f_i(1 - f_i) = Var(G_i)$ since we simulate G_i from a Binomial distribution with probability of f_i. $Cov(\eta_k + \epsilon_k, \eta_l + \epsilon_l) = Cov(\epsilon_k, \epsilon_l) + \sum_{j=1}^{96} 2\beta_{kj}\beta_{lj} f_j(1 - f_j)$. We can then compute the correlation matrix, which leads to (-0.47,0.31,-0.10) compared to (-0.42,0.27,-0.08), which are the correlation of ϵ_k.

We estimated three correlation matrices, based on the correlation of summary statistics of just null SNPs, all SNPs, and excluding those significant SNPs with

minimum p-value less than $0.05/(m_0 + m_1)/3$. Table 2.1 summarizes the results. We can see that using null SNPs and excluding those genome-wide significant SNPs lead to very similar results. And all methods have similar and nearly unbiased estimators. Using more samples leads to reduced mean squared error (MSE).

Table 2.1: Bias and root MSE of estimating correlations of (-0.47,0.31,-0.10). There are m_0 null SNPs and 96 causal SNPs.

method	(n, m_0)	Bias			root MSE		
all SNPs	$(10^3, 10^5)$	0.003	-0.004	0.003	0.022	0.025	0.032
	$(10^4, 10^4)$	-0.017	0.013	-0.008	0.020	0.018	0.017
null SNPs	$(10^3, 10^5)$	0.004	-0.004	0.003	0.022	0.025	0.032
	$(10^4, 10^4)$	-0.001	-0.001	-0.001	0.011	0.013	0.013
data-null SNPs	$(10^3, 10^5)$	0.004	-0.004	0.003	0.022	0.025	0.032
	$(10^4, 10^4)$	-0.001	-0.002	-0.002	0.011	0.013	0.014

2.4.2 Type I error summary

We use those data-null SNPs to estimate the correlation matrix. Table 2.2 summarizes the empirical type I errors. We can see that all methods appropriately control the type I errors.

Table 2.2: Type I errors based on 100 simulations: there are $n = 10^3, 10^4$ unrelated individuals, $m_0 = 10^5, 10^4$ null SNPs and $m_1 = 96$ causal SNPs. The type I errors have been scaled by the significance level.

$n = 10^4, m_0 = 10^4$												
α	minP	3-DF	SPU(1,...,8,∞)									aSPU
10^{-4}	1.22	1.15	1.01	1.07	1.22	1.14	1.13	1.21	1.16	1.16	1.20	1.14
10^{-3}	1.03	1.04	0.96	1.07	1.02	1.06	1.03	1.06	1.03	1.05	1.04	1.03
$n = 10^3, m_0 = 10^5$												
α	minP	3-DF	SPU(1,...,8,∞)									aSPU
10^{-4}	1.13	1.10	1.04	1.09	1.12	1.12	1.11	1.14	1.13	1.12	1.14	1.12
10^{-3}	1.06	1.03	1.00	1.02	1.04	1.05	1.06	1.06	1.06	1.07	1.07	1.04

2.4.3 Power comparison

Figure 2.1 and 2.2 summarize the power under significance level 10^{-3} for $n = 10^3$ and $n = 10^4$. For each causal SNP, we first compute its "effect size": the ideal non-centrality parameter for the 3-DF chi-square test based on the true value of β_{kj} and correlation matrix Σ of three outcomes. We then plot the power of different methods against the SNP effect size. In the plot, we compare the power of 3-DF chi-square test, the minimum P-value based approach (denoted as minP), and the aSPU test. We plot the 96 causal SNPs in four sub-figures to better visualize the power of different methods. Overall we can see that under larger sample size $n = 10^4$, the aSPU test has competitive performance.

Figure 2.1: Power comparison: $n = 10^3, m_0 = 10^5$

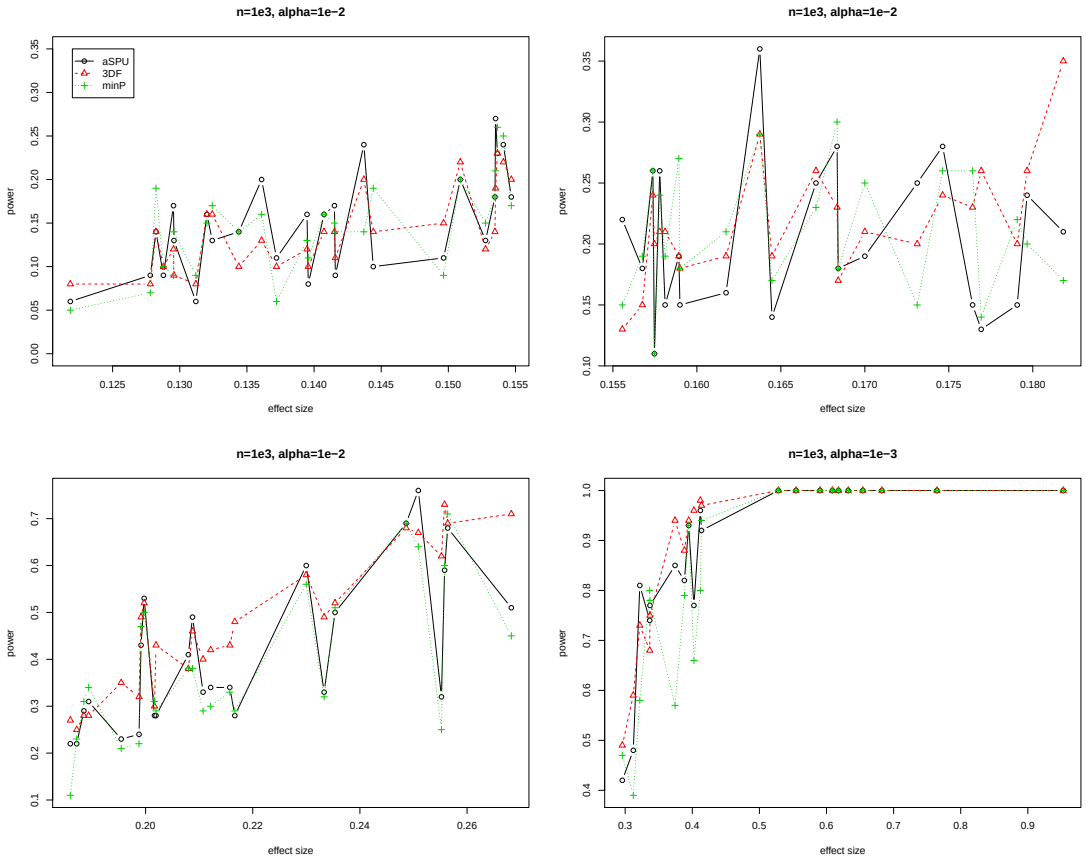


Figure 2.2: Power comparison: $n = 10^4, m_0 = 10^4$

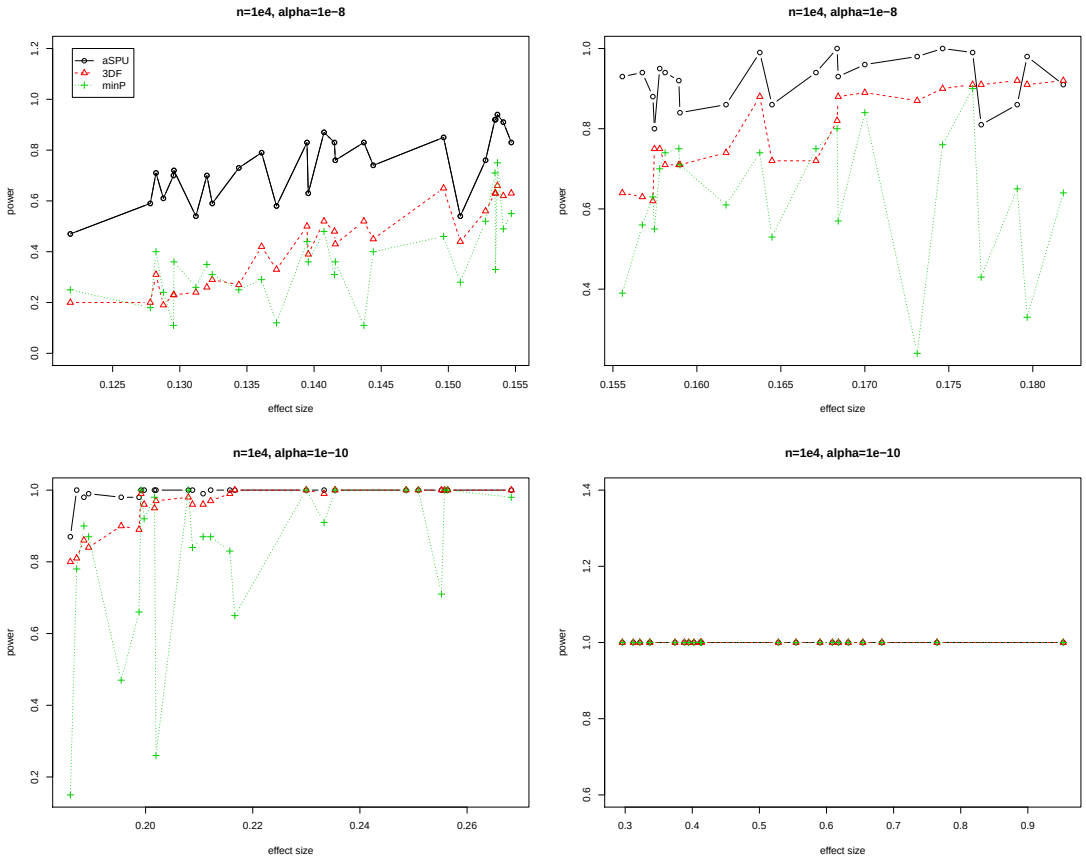


Figure 2.3 and 2.4 summarize the power for all SPU tests.

Figure 2.3: Power of SPU tests under 10^{-3} significance level: $n = 10^3, m_0 = 10^5$

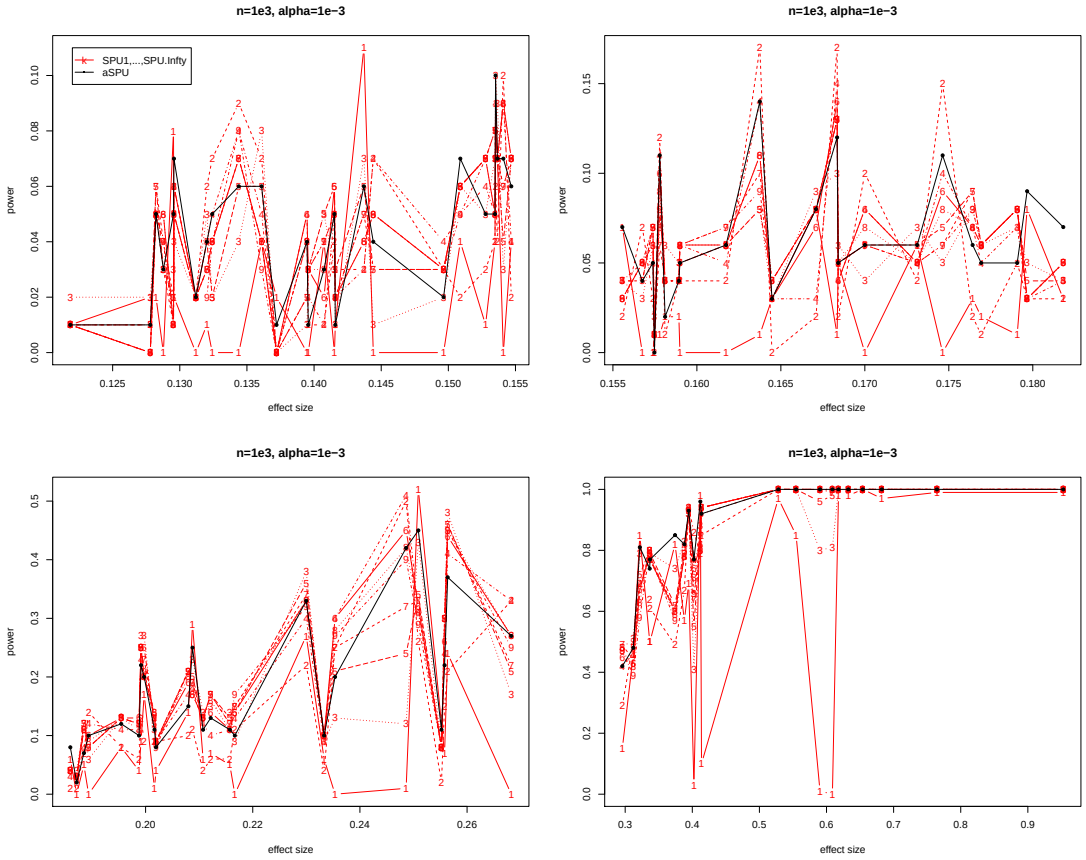
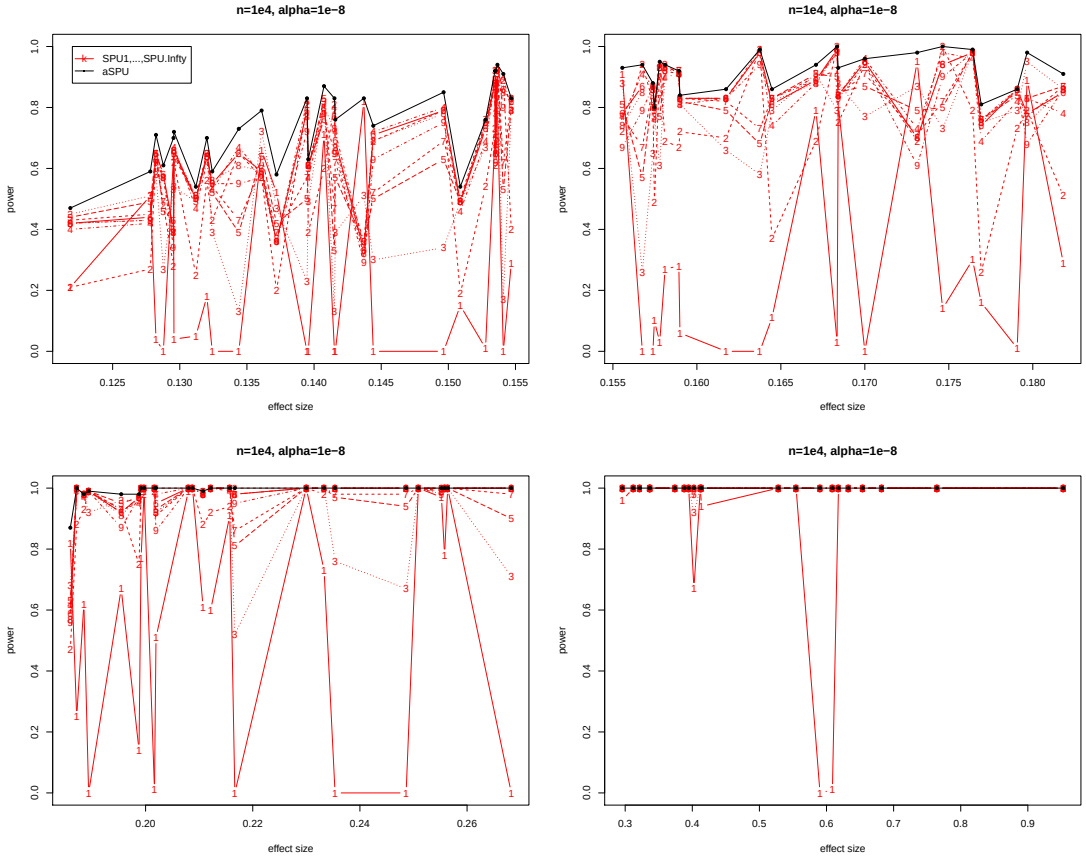


Figure 2.4: Power of SPU tests under 10^{-8} significance level: $n = 10^4, m_0 = 10^4$



2.5 Discussion

In this chapter we have extended the SPU and aSPU tests to test association using only GWAS summary statistics. The SPU and aSPU tests compute their significance p-values based on the Monte Carlo simulation. We also discussed a multi-degrees of freedom chi-square test that can be quickly implemented and offers decent detection power. The SPU tests can adapt to the number of null traits by adaptively weighting the univariate trait association statistics. In the post-GWAS era, many summary GWAS results have been reported. The proposed methods provide an economical way to analyze these summary data without requiring access to the raw genotype and phenotype data, and thus can quickly provide additional results and insights.

2.6 Appendix

2.6.1 96 loci selected for simulation

	SNP	TC-Zscore	TG-Zscore	HDL-Zscore	LDL-Zscore
1	rs1532624	6.718	-7.098	38.678	-6.458
2	rs964184	-15.901	-33.075	14.400	-10.666
3	rs629301	24.350	1.865	-5.413	27.854

4	rs4420638	-22.388	-9.640	9.423	-25.801
5	rs1260326	10.731	24.539	-1.765	3.680
6	rs6511720	-20.890	-1.440	2.445	-23.004
7	rs12678919	0.109	22.849	-20.981	0.567
8	rs1367117	20.803	6.250	-4.232	22.700
9	rs11984636	-0.200	22.160	-20.099	0.114
10	rs1532085	9.102	6.723	20.819	-0.186
11	rs7811265	1.780	16.164	-4.855	-1.472
12	rs2954029	12.464	15.651	-8.657	11.176
13	rs7241918	8.886	-0.394	14.758	3.599
14	rs4299376	-14.096	-3.020	1.247	-14.476
15	rs12916	-14.363	-1.029	-1.495	-14.079
16	rs2131925	13.290	13.710	2.984	8.517
17	rs10401969	12.934	11.282	-0.555	9.618
18	rs1883025	-10.801	-3.286	-12.058	-5.307
19	rs16942887	4.477	-1.573	11.929	0.784
20	rs2479409	-10.136	-2.604	0.982	-11.061
21	rs6882076	-10.940	-6.465	-0.132	-9.748
22	rs9987289	-9.847	2.307	-10.309	-7.650
23	rs2000999	10.153	4.539	-0.622	9.755

24	rs174546	-9.706	10.102	-9.714	-9.518
25	rs6065906	0.037	-8.464	9.747	-1.042
26	rs649129	9.225	-2.248	1.571	9.602
27	rs4846914	1.389	-7.468	9.442	-0.308
28	rs909802	8.376	2.829	-0.137	8.962
29	rs3177928	8.938	2.546	3.000	7.919
30	rs3136441	-1.664	3.421	-8.695	0.463
31	rs10504739	-0.693	-0.696	-8.537	-0.185
32	rs1564348	-8.308	-2.258	1.821	-8.512
33	rs9592961	-3.639	8.199	-6.227	-4.965
34	rs386000	3.860	-0.075	8.130	0.519
35	rs1169288	-7.689	-0.743	-2.410	-8.012
36	rs11220462	6.555	2.395	-2.157	8.004
37	rs4731702	-1.246	-4.879	8.004	-2.390
38	rs1720618	-7.946	-1.098	-0.885	0.251
39	rs41453844	1.572	0.550	7.859	2.022
40	rs7134594	3.827	0.510	7.787	0.579
41	rs838880	-2.153	0.248	-7.618	0.814
42	rs881844	-3.420	1.327	-7.605	-1.047
43	rs514230	-7.523	-0.975	-1.726	-6.816

44	rs643531	4.089	3.598	7.406	0.208
45	rs11136341	-6.127	-0.030	0.203	-7.241
46	rs2807834	-7.228	1.952	-4.209	-6.553
47	rs1030431	7.148	3.937	0.500	5.890
48	rs10761731	-3.225	6.957	-5.156	-3.562
49	rs11065987	6.862	-0.167	3.440	6.044
50	rs442177	3.538	6.827	-5.140	3.296
51	rs3757354	-5.944	-1.246	0.866	-6.785
52	rs2929282	-0.466	-6.736	4.681	0.297
53	rs2925979	-0.684	3.917	-6.700	0.068
54	rs2072183	6.636	2.925	-0.361	6.527
55	rs12027135	-6.600	-2.194	-0.739	-6.435
56	rs2332328	5.302	0.370	-0.312	6.590
57	rs2814982	-6.581	-0.346	-4.841	-5.039
58	rs2293889	-0.545	0.503	-6.550	2.552
59	rs13107325	-2.984	2.425	-6.516	-1.553
60	rs9686661	2.739	6.425	-4.947	2.687
61	rs7941030	-6.403	-0.019	-5.562	-4.642
62	rs10195252	4.298	6.393	-5.340	4.763
63	rs9488822	6.387	3.728	1.147	5.516

64	rs17142153	-0.242	-0.971	-6.384	0.098
65	rs4148008	-0.692	-2.044	6.378	-2.849
66	rs492602	-6.360	-3.609	0.343	-5.377
67	rs2255141	6.359	-2.989	5.124	5.984
68	rs1689800	1.152	-1.626	6.290	-1.468
69	rs2277862	-6.261	-3.041	-2.202	-4.511
70	rs4660293	-2.059	-4.729	6.254	-2.853
71	rs6585689	0.339	-0.056	-6.247	-0.220
72	rs11613352	-2.924	-6.238	5.474	-3.479
73	rs1800562	-5.574	0.322	-1.328	-6.189
74	rs2285942	6.177	2.770	0.133	5.861
75	rs11006644	-0.349	1.025	6.102	-0.547
76	rs1515100	0.472	5.408	-5.997	1.654
77	rs7225700	-4.838	1.367	-1.560	-5.888
78	rs2290159	-5.876	-3.049	-0.777	-4.859
79	rs4082919	2.591	-2.497	5.848	0.919
80	rs12967135	-0.819	4.352	-5.801	-0.246
81	rs2652834	-2.027	3.596	-5.753	-0.896
82	rs181362	-4.189	-2.268	-5.712	-1.732
83	rs6759321	5.675	0.487	2.818	4.894

84	rs41389747	1.333	-0.106	5.657	1.710
85	rs2412710	-0.053	5.624	-4.006	-1.012
86	rs2068888	-3.486	-5.582	2.385	-2.895
87	rs645040	0.931	5.572	-4.653	1.218
88	rs10832963	-5.572	-2.151	-1.069	-5.081
89	rs605066	-1.387	-4.655	5.570	-2.316
90	rs7515577	5.555	1.342	-0.024	5.251
91	rs7255436	2.171	-1.849	5.528	0.634
92	rs11649653	2.318	5.522	-3.111	2.465
93	rs5756931	2.638	5.499	-2.858	2.375
94	rs7134375	0.067	-3.341	5.498	-0.430
95	rs2923084	-0.007	-2.237	5.465	-2.036
96	rs6450176	1.913	4.295	-5.452	1.690

Chapter 3

Incorporate technical variation to assess reproducibility of genome-wide methylation data

3.1 Introduction

Epigenetics is the study of mitotically heritable modifications in chromatin structure (i.e., modifications not involving the underlying DNA sequence), and their impact on the transcriptional control of genes and cellular function. Recent technological advances have provided multiple platforms for systematically interrogating DNA methylation variation across the genome (Laird, 2010; Rauch and Pfeifer, 2010; Bibikova *et al.*, 2011). Among them, the Illumina HumanMethylation450 BeadChip (HM450) (Illumina, Inc.) is a commonly used array constituting a major extension of the previous Infinium HumanMethylation27 BeadChip (HM27) (Illumina, Inc.) and can be used to assess the methylation condition of more than 480,000 cytosines distributed over the entire genome. The HM450 array provides coverage of 98.9% of RefSeq genes with a global average of 17.2 probes per gene region (Dedeurwaerder *et al.*, 2011). A new-generation array, the Infinium MethylationEPIC (EPIC) chip (Moran *et al.*, 2016) is recently introduced with improved coverage in enhancer regions, but large scale data using the EPIC array is not readily available yet.

Using the HM450 or other Illumina array, DNA methylation level is estimated using fluorescence intensity from the methylated (M) and unmethylated (U) alleles. In methylation-trait association tests, two methylation measures have been

commonly used, the β-value, calculated as M/(M+U), and the M-value, calculated as log2(M/U) (Du *et al.*, 2010). For each probe sequence, a median of 14 beads are randomly distributed on the array, and the average intensities across the beads are used to compute the M and U values. However, the bead-to-bead variation of the intensities, which is reported in the GenomeStudio software (as standard error of M and U value), is often ignored in the subsequent analysis.

Kuan *et al.* (2010*b*) proposed a weighted clustering approach to select CpG sites and cluster samples accordingly with distinct characteristics. This algorithm weights each β-value by the corresponding detection p-value, which can be viewed as an indirect measure of the bead-to-bead variability. Also, an alternative methylation measure, the N-value, has been proposed, which explicitly uses the intensity variability between beads to weight the M and U values (Ryu, 2013). The authors showed that the N-value has desirable normal distribution and can improve statistical power in association studies compared to the β and M values. To derive this weighted measure, a linear relationship was assumed between log-transformed M (or U) values and the log of standard deviation of intensities across beads, which is difficult to validate. The interpretation of N-value and the corresponding effect size are also not as straightforward as the β-value. In this study, we will consider a direct likelihood-based approach to incorporating the bead-to-bead variability in analysis of methylation data.

Sandoval *et al.* (2011) validated the HM450 chip by showing that methylation patterns measured in colorectal cancer cell lines and normal mucosa were consistent with those found in bisulfite genomic sequencing. However, similar to other microarray experiments, it is still important to evaluate the impact of technical variation on the measurement. A well-known source of bias for epigenome-wide association studies (EWASs) of DNA methylation is the so-called "batch effect", which is largely caused by technical differences from one chip to the next. In our previous study (Bose *et al.*, 2014), we used technical replicates to evaluate consistency of methylation measurement at each CpG site on the HM450 array. Specifically, we calculated the intraclass correlation coefficient (ICC) to compare the within- and between-replicate variations, where the within-replicate variation represents technical variation, and between-replicate variation represents biological variation. The ICC value can serve as a measure for the impact of batch effects on each probe, with high ICC value indicating good measurement accuracy at the corresponding CpG site. Using a mixture model approach, we identified three clusters of CpG sites on the HM450 array based on the ICC values (=0, 0-0.37, and >0.37). We further demonstrated that the significant association results in EWAS are more likely to be identified at CpG sites with high ICC than those with low ICC. Another study (Shvetsov *et al.*, 2015) calculated the ICCs using the M-value instead of β-value, and observed similar distribution for the ICCs.

Dugua *et al.* (2015) used a simulation study to demonstrate the impact of ICC on the power of EWAS, and Chen *et al.* (2016) proposed a filtering algorithm prior to EWASs to remove CpG sites with low ICC values.

In this study, we will define a modified ICC measure, taking into account of measurement variability of multiple beads for each probe on the HM450 array. We will calculate the new measure to the DNA methylation data in the Atherosclerosis Risk in Communities (ARIC) study. The modified ICC eliminates one known technical variation ("bead-to-bead" variation), and we believe that it will be useful to guide subsequent association tests in EWASs. Our approach can also be directly applied to other Illumina methylation arrays including the new EPIC array.

3.2 Methods

Using technical replicates, we previously calculated intraclass correlation coefficient (ICC) for each probe on the HM450 array to quantify the impact of technical errors (Bose *et al.*, 2014). We demonstrated that probes can be potentially classified according to different levels of reliability, and that the significant association results are more likely to be identified at CpG sites with high ICC than those with low ICC. In this study, we will re-calculate ICC, taking into account of measurement variability of multiple beads for each probe, and investigate whether the

updated ICC values can measure the reliability of probes more accurately.

On the methylation array, such as HM450, the fluorescence intensities of the methylated and unmethylated alleles are measured at each targeted cytosine position. For each probe sequence, multiple beads are used. Let M denote the average signal from methylated alleles and U the average signal from unmethylated alleles, at a given CpG site. The methylation level is estimated as

$$\beta = \frac{M}{M + U + 100}.$$

Here the denominator has been added an constant of 100 to stabilize the methylation level measures.

The standard errors of the averaged signals, M and U, denoted as τ_M and τ_U respectively, are available in the Illumina GenomeStudio output. We approximate the variability of β-value (denoted as τ_β^2) using the delta method, assuming the two signals are independent.

$$\tau_\beta^2 \approx (\partial\beta/\partial M)^2\tau_M^2 + (\partial\beta/\partial U)^2\tau_U^2 = \frac{(U+100)^2\tau_M^2 + M^2\tau_U^2}{(M+U+100)^4}. \tag{3.1}$$

3.2.1 Assessment of methylation measurement reproducibility

Given a set of samples with technical replicates, denote Y_{ij} as the measured methylation level at a specific CpG site for the i-th replicate of the j-th biological sample. A simple model describing variation of methylation is the following one-way ANOVA with random effects:

$$Y_{ij} = \mu + u_j + \epsilon_{ij}, \tag{3.2}$$

where μ is the overall mean methylation level. The random effect term u_j follows the zero-mean normal distribution with variance σ_b^2 and is shared by all replicates for sample j reflecting sample characteristics; and ϵ_{ij} follows the zero-mean normal distribution with variance σ_w^2 and is a random noise term including other technical errors for multiple measures of the same biological sample, such as chip effects and experiment conditions. It is generally assumed that ϵ_{ij} and u_j are independent. The ratio $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$ is known as the intraclass correlation (ICC) since $\rho = corr(Y_{ij}, Y_{kj})$ for $i \neq k$, the correlation between two technical replicates of the same biological sample. The ICC parameter ρ directly measures the reproducibility of methylation levels. This modeling approach has been adopted by Bose *et al.* (2014) to assess the technical reproducibility of methylation levels in the Atherosclerosis Risk In Communities (ARIC) Study.

To take into account the measurement error, we propose the following model

$$Y_{ij} = \mu + u_j + \nu_{ij} + \epsilon_{ij}, \tag{3.3}$$

where the additional term $\nu_{ij} \sim N(0, \tau_{ij}^2)$ models the measurement variability across beads and similarly as before $u_j \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma_w^2)$. τ_{ij}^2 can be calculated using standard output from the GenomeStudio software, specifically the standard error of M and U, based on the delta method (3.1). Compared to model (3.2), the measurement error term ν_{ij} will capture part of the variation in both u_j and ϵ_{ij}. For the ARIC data, we will show that this model better captures the data variation and will lead to more accurate measure of technical reproducibility. For model (3.3) we can check that

$$E(Y_{ij}) = \mu, \quad Var(Y_{ij}) = \tau_{ij}^2 + \sigma_b^2 + \sigma_w^2, \quad Cov(Y_{ij}, Y_{kj}) = \sigma_b^2.$$

Therefore the covariance among technical replicates is not changed, but the marginal variation of methylation levels has an extra term τ_{ij}^2 from the measurement error.

A modified ICC, after subtracting the bead-to-bead variation, can be calculated as:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}.$$

Compared to the original ICC values, which was defined as

$$ICC_{original} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2 + \tau_\beta^2}.$$

The modified ICC eliminates the known technical variation, and can better reflect the chip-to-chip variation and other technical variation. In the following, we re-analyze the ARIC methylation data and classify probes on the HM450 array based on these modified ICC, and identify the set of probes that were impacted the most by bead-to-bead variation.

For sample j, denote $Y_j = (Y_{1j}, \ldots, Y_{n_j j})^T$, where n_j is the number of replicates for the j-th sample, $\mathbf{1} = (1, \ldots, 1)^T$, $V_j = Cov(Y_j)$. The log likelihood can be written as

$$L = \sum_{j=1}^{n} -\frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|V_j| - \frac{1}{2}(Y_j - \mu\mathbf{1})^T V_j^{-1}(Y_j - \mu\mathbf{1}).$$

The maximum likelihood estimate (MLE) of σ_b^2 and σ_w^2 can be easily obtained through EM algorithms (Dempster *et al.*, 1977) as follows.

3.2.2 MLE with EM algorithm

For the jth sample, we treat u_j as missing data. The complete data log likelihood is

$$\sum_{j=1}^{n}[\log \Pr(u_j) + \sum_i \log \Pr(y_{ij}|u_j)],$$

where $\Pr(y_{ij}|u_j) \sim N(\mu + u_j, \sigma_w^2 + \tau_{ij}^2)$. We can easily check that

$$V_j = Cov(Y_j) = \sigma_b^2 \mathbf{1}\mathbf{1}^T + \mathrm{diag}(\sigma_w^2 + \tau_{1j}^2, \ldots, \sigma_w^2 + \tau_{n_j j}^2).$$

In the E-step, we have

$$E(u_j|Y_j) = \sigma_b^2 \mathbf{1}^T V_j^{-1}(Y_j - \mu\mathbf{1}), \quad Var(u_j|Y_j) = \sigma_b^4 \mathbf{1}^T V_j^{-1}\mathbf{1}.$$

In the M-step, the parameter σ_b^2 can be estimated by

$$\min_{\sigma_b^2} \sum_{j=1}^{n} [\frac{E(u_j^2|Y_j)}{\sigma_b^2} + \log(\sigma_b^2)].$$

So

$$\hat{\sigma}_b^2 = \frac{\sum_{j=1}^{n} E(u_j^2|Y_j)}{\sum_j n_j}.$$

The parameter σ_w^2 and μ can be estimated by

$$\min_{\mu,\sigma_w^2} \sum_{i,j} \frac{E[(y_{ij} - \mu - u_j)^2|Y_j]}{\sigma_w^2 + \tau_{ij}^2} + \log(\sigma_w^2 + \tau_{ij}^2).$$

So

$$\hat{\mu} = \frac{\sum_{i,j}(\sigma_w^2 + \tau_{ij}^2)^{-1}[y_{ij} - E(u_j|Y_j)]}{\sum_{i,j}(\sigma_w^2 + \tau_{ij}^2)^{-1}},$$

and $\hat{\sigma}_w^2$ can be solved with a constrained numerical optimization ($\sigma_w^2 \geq 0$). We can also directly maximize the marginal likelihood of Y_j to estimate the fixed effects parameter μ, which will lead to the following generalized least squares (GLS) estimate

$$\hat{\mu} = \frac{\sum_{j=1}^{n} \mathbf{1}^T V_j^{-1} Y_j}{\sum_{j=1}^{n} \mathbf{1}^T V_j^{-1}\mathbf{1}}.$$

In each iteration of the previous EM algorithm, we need to numerically solve a constrained optimization problem to estimate $\hat{\sigma}_w^2$, which could be computing

intensive. In the following we develop an alternative and very easy to program EM algorithm.

We treat both u_j and ν_{ij} as missing data. The complete data log likelihood is now

$$\sum_j [\log \Pr(u_j) + \sum_i \log \Pr(\nu_{ij}) + \log \Pr(y_{ij}|u_j, \nu_{ij})].$$

Here each term is an one-dimensional normal distribution density with

$$\Pr(y_{ij}|u_j, \nu_{ij}) \sim N(\mu + u_j + \nu_{ij}, \sigma_w^2).$$

Recall $V_j = \sigma_b^2 \mathbf{1}\mathbf{1}^T + \text{diag}(\sigma_w^2 + \tau_{1j}^2, \ldots, \sigma_w^2 + \tau_{n_j j}^2)$. In the E-step we compute

$$E(u_j|Y_j) = \sigma_b^2 \mathbf{1}^T V_j^{-1}(Y_j - \mu\mathbf{1}), \quad Var(u_j|Y_j) = \sigma_b^4 \mathbf{1}^T V_j^{-1}\mathbf{1},$$

$$E(\nu_{ij}|Y_j) = \tau_{ij}^2 e_i^T V_j^{-1}(Y_j - \mu\mathbf{1}), \quad Var(\nu_{ij}|Y_j) = \tau_{ij}^4 e_i^T V_j^{-1} e_i,$$

where e_i is a column vector of zeros except the ith element equal to 1. And

$$Cov[(u_j, \{\nu_{ij}\})|Y_j] = \text{diag}(\sigma_b^2, \{\tau_{ij}^2\}) - Var[E[(u_j, \{\nu_{ij}\})|Y_j]].$$

So

$$E(u_j \nu_{ij}|Y_j) = Cov(u_j, \nu_{ij}|Y_j) + E(u_j|Y_j)E(\nu_{ij}|Y_j).$$

In the M-step, for parameter σ_b^2, we have

$$\min_{\sigma_b^2} \sum_j [\frac{E(u_j^2|Y_j)}{\sigma_b^2} + \log(\sigma_b^2)],$$

so

$$\hat{\sigma}_b^2 = \frac{\sum_j E(u_j^2|Y_j)}{\sum_j n_j}.$$

For parameter σ_w^2 and μ, we have

$$\min_{\mu,\sigma_w^2} \sum_{i,j} \frac{E[(y_{ij} - \mu - u_j - \nu_{ij})^2|Y_j]}{\sigma_w^2} + \log(\sigma_w^2),$$

so

$$\hat{\mu} = \frac{\sum_j n_j[y_{ij} - E(u_j + \nu_{ij}|Y_j)]}{\sum_j n_j},$$

and

$$\hat{\sigma}_w^2 = \frac{\sum_j n_j E[(y_{ij} - u_j - \nu_{ij})^2|Y_j]}{\sum_j n_j}.$$

We can also use the previously derived GLS estimate

$$\hat{\mu} = \frac{\sum_{j=1}^n \mathbf{1}^T V_j^{-1} Y_j}{\sum_{j=1}^n \mathbf{1}^T V_j^{-1} \mathbf{1}}.$$

Now we have closed-form updates at each M-step. The proposed EM algorithm is very easy to program and it generally has very stable performance in that each iteration always increase the log likelihood.

3.3 Application to technical replicates of ARIC methylation data

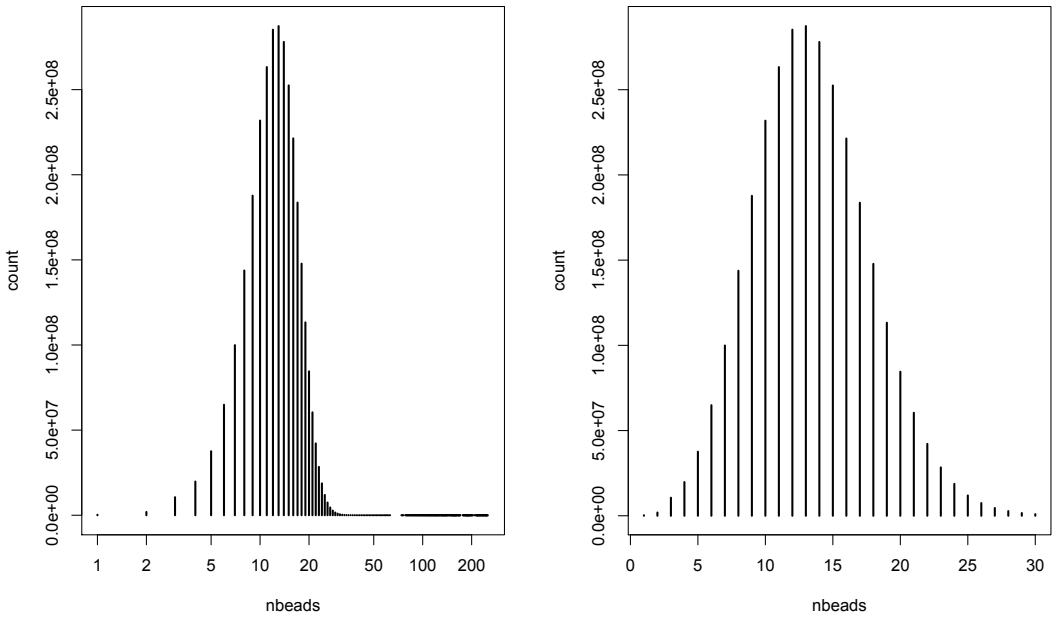
The ARIC Study is a prospective cohort study of cardiovascular disease risk in White and Black adults from four US communities (The ARIC Investigators,

1989). Subjects were seen at baseline (Visit 1) in 1987-1989, with four follow-up visits (Visits 2-5) thereafter. Among the ARIC samples, genome-wide leukocyte DNA methylation levels have been measured for 2097 African American (AA) participants using the HM450 chip. Bisulfite-converted DNA was used for hybridization on the HM450 BeadChip, following the Illumina Infinium HD Methylation protocol (`www.illumina.com`). This consisted of a whole genome amplification step followed by enzymatic end-point fragmentation, precipitation and re-suspension. The re-suspended samples were hybridized to the complete set of bead-bound probes, followed by ligation and single-base extension during which a fluorescently labeled nucleotide is incorporated and scanned. The degree of methylation is determined for each CpG cytosine by measuring the amount of incorporated label for each probe. The intensities of the images were extracted using Illumina GenomeStudio 2011.1, Methylation module 1.9.0 software. The methylation score for each CpG was represented as a β value according to the fluorescent intensity ratio. Beta values may take any value between 0 (unmethylated) and 1 (completely methylated). Background subtraction was conducted with the GenomeStudio software using built-in negative control bead types on the array.

There are in total 130 individuals with technical replicates in the ARIC samples. Among them 125 individuals have two technical measurements, and 5 individuals have three technical measurements. For each CpG site assayed on the HM450 chip, we calculate their ICC values using the proposed methods accounting for the bead-to-bead measurement error. Note that the set of technical replicates analyzed here are slightly different from those used in Bose *et al.* (2014), due to additional QC procedures implemented. Therefore for comparison, the ICC values using the original approach of Bose *et al.* (2014) are re-calculated.

We analyze a total of 485,577 CpG sites included on the HM450 array. Figure 3.1 summarizes the distribution of beads across all sites. The left plot shows the whole range and is plotted on a log scale due to the skewness of distribution. The right plot shows the distribution in a narrow range, i.e., number of beads ≤ 30. The mode is around 13 beads per CpG site. In practice, CpG sites with number of beads < 3 are often excluded from subsequence analysis.

Figure 3.1: Distribution of number of beads



The distributions of the two ICC estimates are shown in Figure 3.2. The median ICC values are 0.27 and 0.38 respectively for the old and new definitions. For the old ICC values, we observe one cluster of sites with relatively high ICC values (mode ~ 0.71), one cluster with relatively low ICC values (mode ~ 0.09), and an third cluster of 40,688 sites having zero ICC. For the new ICC values, the cluster of sites with relatively high ICC values has mode ~ 0.77, the cluster with relatively low ICC values has mode ~ 0.18, the cluster of zero ICC has 34,161 sites.

Figure 3.3 summarizes the change of ICC values. For majority of the CpG sites, the newly estimated ICC values are larger than the old approach. Specifically in 429,509 CpG sites, we obtain improved ICC values, and for 54,463 CpG sites, the newly estimated ICC values are smaller. Notably the new ICC values have an additional cluster around $\rho = 1$ (5,928 CpG sites), which implies that most of the measurement variation at these CpG sites are possibly due to bead-to-bead variation.

Figure 3.2: Comparison of estimated ICC values between the proposed method and the approach of Bose *et al.* (2014).

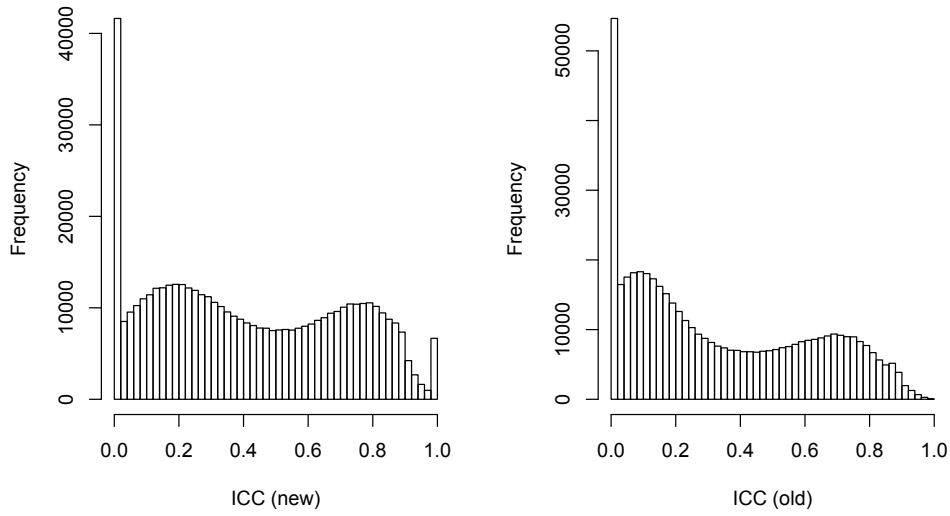


Figure 3.3: Difference of estimated ICC values between the proposed method and the approach of Bose *et al.* (2014).

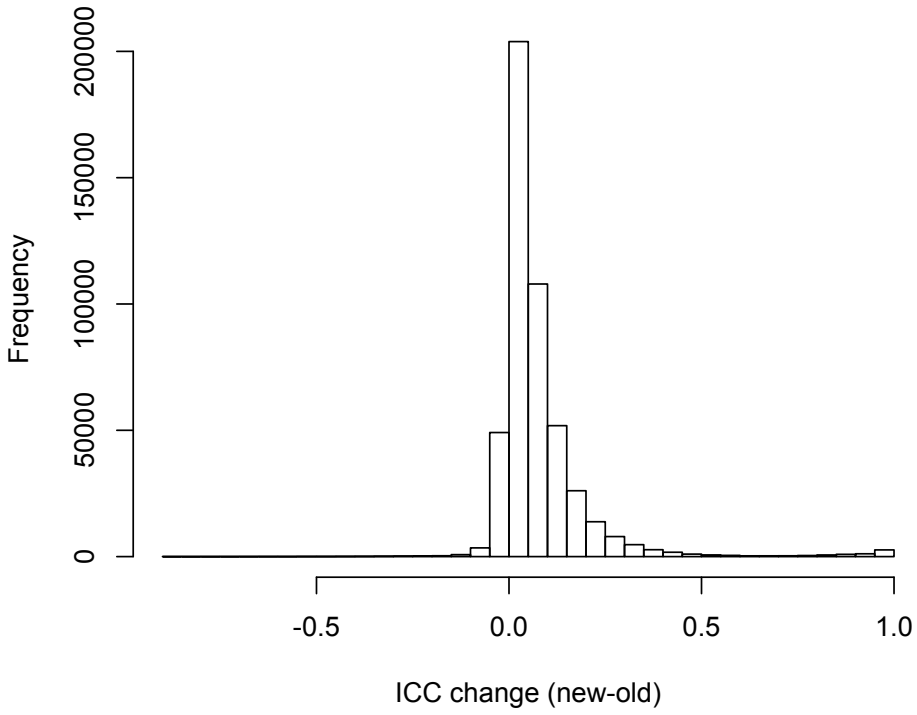


Figure 3.4 shows the ICC change (new-old) by probe type. Overall the ICC improve slightly more for type II probes (median of change is 0.047 for type II vs 0.037 for type I). But at the tail of the distribution, type I probes seem to have larger changes than type II (the 95th percentile for type II is .225, but .515 for type I).

Figure 3.4: ICC change for different probe types.

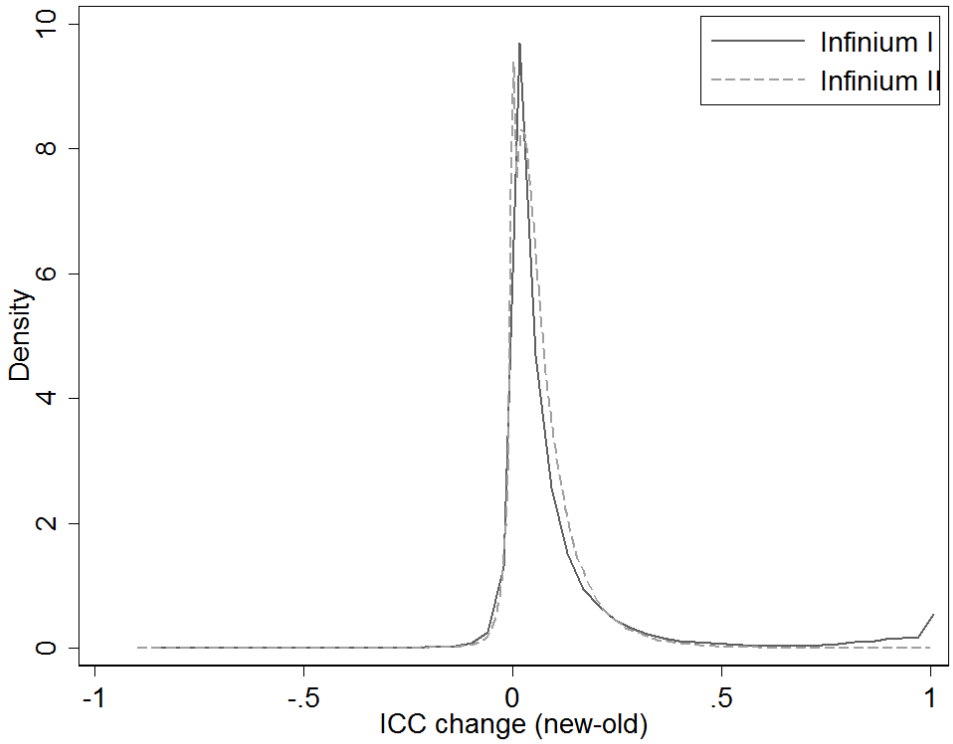


Figure 3.5 compares the new ICC values between different probe types. Figure 3.6 compares the new ICC values between different CpG sites. The observation from Figure 3.5 and 3.6 does not change from that in Bose *et al.* (2014).

Figure 3.5: ICC distribution for different probe types.

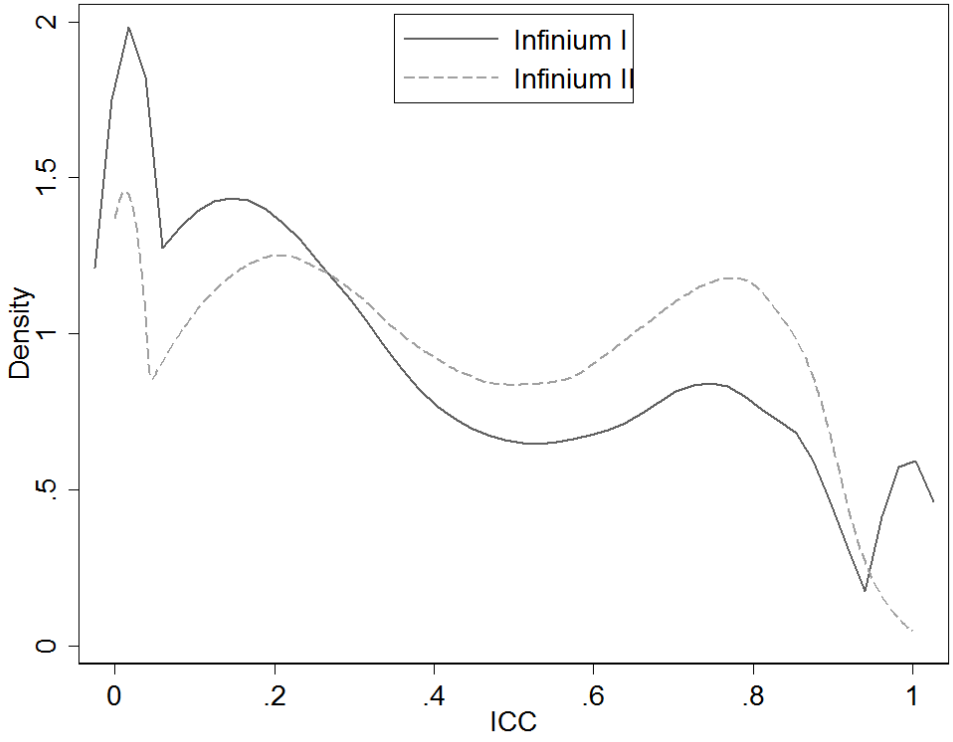
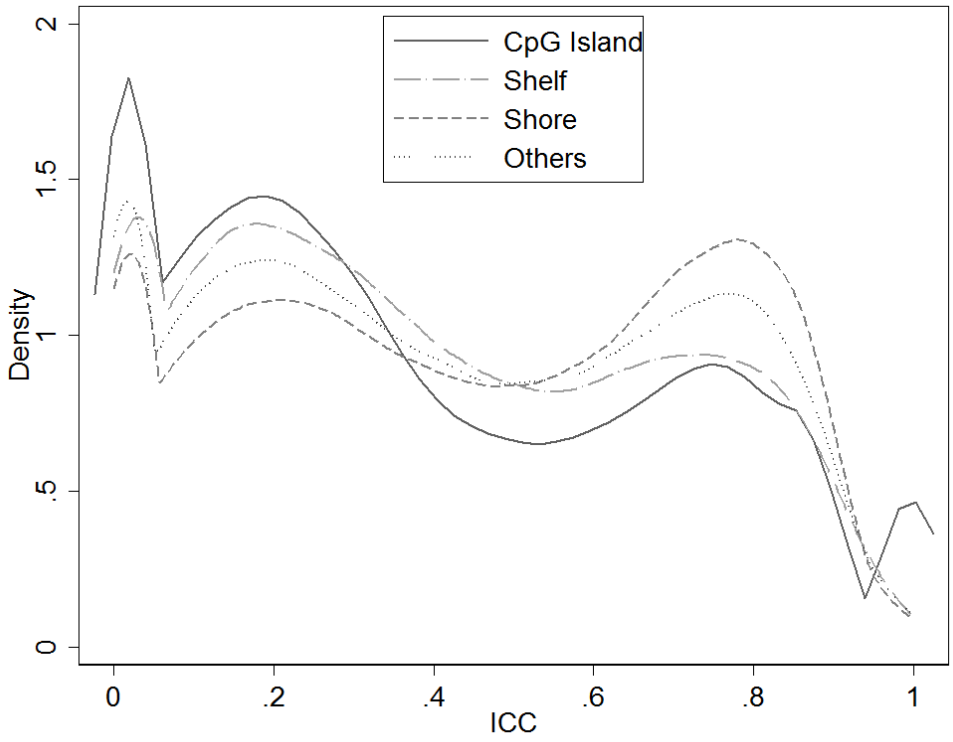


Figure 3.6: ICC distribution for different CpG sites



3.3.1 Classification of CpG sites using truncated normal mixture models

Next we fit a mixture of truncated normal models to the estimated ICC values (Lee and Scott, 2012)

$$\theta_0\delta_0 + \theta_1\delta_1 + \sum_{k=1}^{K} p_k N(\mu_k, \sigma_k^2|(0,1)),$$

where $\sum_{k=1}^{K} p_k = 1 - \theta_0 - \theta_1$, and δ_0 and δ_1 are two point masses at 0 and 1 respectively. We choose $K = 2$ based on the overall ICC distributions in Figure 3.2. For the newly estimated ICC values, the four mixing proportion estimates are $\{0.071, 0.012, 0.638, 0.279\}$, the two μs are $\{0.216, 0.759\}$, and two σs are $\{0.261, 0.109\}$. When fitting the same mixture models to the old ICC values, the four mixing proportions are $\{0.084, 0, 0.589, 0.327\}$, i.e., the $ICC = 1$ cluster has a mixing proportion of 0. The two μs were $\{0.035, 0.681\}$, and two σs are $\{0.240, 0.135\}$ for old ICC.

With a cutoff ICC value of 0.37, the numbers of CpG sites classified into each of the three clusters (for the old ICC) are: 40,688 (ICC = 0), 242,641 (low reliability) and 200,643 (high reliability cluster). For the new ICC values, the size of the three clusters are 34,161 (ICC=0), 201,700 (low reliability) and 248,111 (high reliability), with an additional set of 5,928 CpG sites with $ICC = 1$.

3.4 Discussion

In the analysis of array-based DNA methylation, it is important to consider the technical variation of measures. In this paper, we derived a standard error of methylation β-value, which reflects the bead-to-bead variation on the HM450 array. We further incorporated it into the calculation of the intraclass correlation coefficient (ICC) for each HM450 array probe. Using technical replicates available in the Atherosclerosis Risk in Communities (ARIC) methylation data, we demonstrated that the ICCs were greatly improved after taking into account of this particular measurement error. The results suggest that accounting for the bead-level variation as part of technical error can improve accuracy and efficiency of subsequent epigenome-wide association analyses.

The bead-to-bead variation is a part of the total technical variation on the array-based DNA methylation. Our results show that the modified ICCs, removing the bead-to-bead variation specifically, are improved by a median value of 0.04. For 6,000 CpG sites, the new ICC values approach 1, suggesting that most of the technical variation could be explained by the bead-to-bead variation at corresponding probes. Most of these sites are measured using the type I probes, and are in the regions rich of CpG sites (CpG islands).

The standard error we derived is for the reported β-value from the GenomeStudio software, often referred as the "raw" value. Often, these β-values are normalized to remove potential batch effects (e.g., Johnson *et al.*, 2007), or to remove probe type (type I vs II) bias (e.g., Teschendorff *et al.*, 2013; Pidsley *et al.*, 2013). The derivation of standard error of normalized β-value depends on the exact normalization method, which is not always straight-forward. In our experiences, these normalization methods do not change the distribution of methylation measures greatly. One could still use the same standard errors from the "raw" values to approximate the bead-to-bead variation after normalization, though extensive study will be needed to evaluate the impact of each normalization method.

Although we use the β-value as the methylation measure in this paper, it is easy to extend the calculation to the M-value as well. The standard error of M-value can be similarly computed using the delta method. The HM450 array is being replaced by a newer generation of Illumina's methylation array, the Infinium MethylationEPIC (EPIC) array. Our method can also be directly applied to the EPIC data, because the technology is similar.

In summary, we have proposed a statistical approach to incorporate the bead-to-bead variation in the analysis of array-based DNA methylation measure. We evaluated the improvement of ICC values, which explicitly quantify the ratio of

technical and biological variation, using the proposed method. The newly estimated ICC values for all CpG sites on the HM450 chip are available in supplemental data. The R scripts for the proposed method is freely available upon request. We hope that our findings can lead to powerful statistical tests for future EWASs using the HM450 or other types of methylation arrays.

Chapter 4

Incorporate technical variation in epigenome-wide association study of methylation data with application to the Atherosclerosis Risk In Communities (ARIC) Study

4.1 Introduction

In the past decade, genome-wide association studies (GWAS) have successfully identified thousands of SNPs associated with various disease phenotypes. However in combination these identified SNPs only explained a small proportion of the phenotype variance (Manolio *et al.*, 2009). Various hypotheses have been suggested for this missing heritability phenomenon. The epigenetic change is one possible source that can contribute to the phenotype variability. The epigenetics is the study of mitotically heritable modifications in chromatin structure (i.e., modifications not involving the underlying DNA sequence), and their impact on the transcriptional control of genes and cellular function. Among the many epigenetic mechanisms, the DNA methylation has been widely studied recently (Breitling *et al.*, 2011; Wan *et al.*, 2012; Shenker *et al.*, 2013; Aslibekyan *et al.*, 2015; Demerath *et al.*, 2015). With the rapid advance of biotechnology, DNA methylation level can be measured across the genome at hundreds of thousands of CpG sites, covering most of the known genes. It makes it feasible to carry out epigenome-wide association studies (EWASs) in large cohorts of samples.

The relationship between methylation and trait can be bi-directional. In many biological pathways, environmental or biological factors are responsible to change in DNA methylation level. However, similar to other array-based genetic studies

(e.g., microarray gene expression), methylation data generated from microarray chips is also subject to many sources of non-biological/technical variations, so-called "batch effects". Therefore, the total variation in methylation level consists of both biological and technical variations, which are typically modeled using normal distributions in a linear model approach.

In most analyses of EWASs, the technical variation is assumed to be homogeneous across samples and thus incorporated as part of the random noise in a linear model. But the power of association test will be greatly improved if the technical variation can be explicitly modeled. For example, methylation levels at many CpG sites are found to systematically vary across chips and different chip positions. A linear mixed-effects model (LMM) is often applied to model the chip and chip-position effects. Another source of technical variation inherent to the Illumina Infinium technology for DNA methylation measurement is due to the use of different number of probes for each CpG site and each DNA sample, which haven't been widely modeled in EWASs. However incorporating these technical variation brings computational challenges to EWAS. In this paper, we study methods to incorporate this technical variation into association tests, and develop a very easy to implement EM algorithm for reliable model estimation. Currently there are multiple platforms for systematically interrogating DNA methylation variation across the genome (Bibikova *et al.*, 2011). We will use the

Illumina HumanMethylation450 BeadChip (HM450) (Illumina, Inc.) as an illustration example, but the methods can be directly applied to data from other types of Illumina arrays. The HM450 array can assess the methylation condition of more than 480,000 cytosines distributed over the entire genome, and provides coverage of 98.9% of RefSeq genes with a global average of 17.2 probes per gene region (Dedeurwaerder *et al.*, 2011; Du *et al.*, 2010). We illustrate our proposed methods using an application to the Atherosclerosis Risk In Communities (ARIC) Study.

4.2 Methods

On the HM450 array, the fluorescence intensities of the methylated and unmethylated alleles are measured at each targeted cytosine position. For each probe sequence, multiple beads are used. Let M denote the average signal from methylated alleles and U the average signal from unmethylated alleles, at a given CpG site. The methylation level can be estimated as

$$\beta = \frac{M}{M + U + 100}.$$

The standard errors of the averaged signals, M and U, denoted as τ_M and τ_U respectively, are available in the Illumina GenomeStudio output. Following Bai *et al.* (2016), we approximate the variability of β-value (denoted as τ_β^2) using the

delta method as follows

$$\tau_\beta^2 = \frac{(U+100)^2\tau_M^2 + M^2\tau_U^2}{(M+U+100)^4}. \tag{4.1}$$

4.2.1 Association test accounting for technical variation

In EWASs, it is common to encounter various batch effects that need to be properly accounted for in the association analysis. Following Demerath *et al.* (2015), we adopt the linear mixed effects model (LMM) to account for various cluster effects due to batch differences. Given the jth cluster with n_j correlated samples, $j = 1, \ldots, n$, denote Y_j as the measured methylation levels, X_j the set of covariates (intercept included) to be adjusted, Z_j the random effects covariates, W_j the measurement error (with variance assumed known), and ϵ_j the random errors. We assume

$$Y_j = X_j\beta + Z_jU_j + W_j + \epsilon_j, \tag{4.2}$$

where the fixed effects coefficient β is of length p, and random effects U_j is of length K. Assume that $U_j \sim N(0, V_\theta)$, $W_j \sim N(0, \Gamma_j)$, and $\epsilon_j \sim N(0, \sigma^2\boldsymbol{I})$, where $\Gamma_j = \text{diag}(\tau_{1j}^2, \ldots, \tau_{n_jj}^2)$ with τ_{ij}^2 pre-computed following (4.1). The three random errors are assumed mutually independent.

Specifically for the LMM used for EWAS in the ARIC study (see application section), there are two sources of random effects U_j: (1) different samples on the

same methylation chip shares one random intercept. Therefore the first set of covariates in Z_j code different chips. (2) different samples on the same row in one methylation chip shares another random intercept. The second set of covariates in Z_j code different chip rows. We further assume these two random intercepts are independent.

4.2.2 Model estimation and association test

The maximum likelihood estimates and restricted maximum likelihood estimates (MLE/REML) of the parameters can be readily obtained through the EM algorithm (Dempster *et al.*, 1977) (see appendix for details of technical derivations). Given the estimated covariance parameters, we can check that $\hat{\beta} = (\sum_{j=1}^{n} X_j^T P_j^{-1} X_j)^{-1} (\sum_{j=1}^{n} X_j^T P_j^{-1} Y_j)$, where $P_j = Z_j V_{\hat{\theta}} Z_j^T + \Gamma_j + \hat{\sigma}^2 \boldsymbol{I}$ is the covariance matrix for the jth cluster. The large-sample asymptotic covariance matrix of $\hat{\beta}$ is $(\sum_{j=1}^{n} X_j^T P_j^{-1} X_j)^{-1}$. We apply the Wald statistics to test the epigenome-wide association of methylation levels.

4.2.3 Proposed method versus the standard LMM

Compared to the standard LMM, the proposed method (denoted as LMMt) added an additional term W_j assumed as the known technical variation. When W_j has constant variation across all clusters, it can be absorbed into the random error

component ϵ_j and hence LMMt will reduces to a standard LMM. Otherwise when W_j does not have constant variation, LMMt will be different from a standard LMM. If we do assume the LMMt model is correct, then the standard LMM will have a wrong "working" covariance matrix, though the mean model is still correct. Thus we can treat the standard LMM estimation as a form of generalized estimation equations (GEE). Therefore the model based covariance matrix of standard LMM estimates will under-estimate the true covariance matrix. We note that the robust sandwich covariance matrix will provide a valid estimate, and generally the sandwich covariance matrix will be larger than the LMM covariance matrix. So we expect that under the LMMt model, the standard LMM will under-estimate the true variation and produce smaller p-values compared to LMMt (leading to inflated type I errors). The following numerical studies will provide numerical confirmations.

4.3 Numerical study

4.3.1 Application to ARIC methylation data

The ARIC Study is a prospective cohort study of cardiovascular disease risk in White and Black adults from four US communities (The ARIC Investigators, 1989). Subjects were seen at baseline (Visit 1) in 1987-1989, with four follow-up

visits (Visits 2-5) thereafter. Among the ARIC samples, genome-wide leukocyte DNA methylation levels have been measured for 2097 African American (AA) participants using the HM450 chip. Bisulfite-converted DNA was used for hybridization on the HM450 BeadChip, following the Illumina Infinium HD Methylation protocol (`www.illumina.com`). This consisted of a whole genome amplification step followed by enzymatic end-point fragmentation, precipitation and re-suspension. The re-suspended samples were hybridized to the complete set of bead-bound probes, followed by ligation and single-base extension during which a fluorescently labeled nucleotide is incorporated and scanned. The degree of methylation is determined for each CpG cytosine by measuring the amount of incorporated label for each probe. The intensities of the images were extracted using Illumina GenomeStudio 2011.1, Methylation module 1.9.0 software. The methylation score for each CpG was represented as a β value according to the fluorescent intensity ratio. Beta values may take any value between 0 (unmethylated) and 1 (completely methylated). Background subtraction was conducted with the GenomeStudio software using built-in negative control bead types on the array.

The ARIC methylation levels have been previously shown to have high accuracy and reliability (Bose *et al.*, 2014; Bai *et al.*, 2016). Imputed white blood cell (WBC) counts and cell type differentials have been obtained for all subjects

(Houseman *et al.*, 2012; Demerath *et al.*, 2015).

We study the association of methylation levels with smoking phenotype defined as an 0-1 indicator comparing never smokers versus current smokers (in total of 1932 AA subjects). We fit the LMM with methylation beta values as the dependent variable, and with chip and chip row specified as random effects and the following variables specified as fixed effects: sex, age, field center, total WBC, leukocyte cell type proportions (neutrophils, lymphocytes, monocytes and eosinophils), and 10 PCs from the Illumina Infinium HumanExome Beadchip genotype array (Grove *et al.*, 2013) to account for potential confounding by genetic ancestry.

After removing those samples with missing covariates, we analyzed 1640 AA subjects in total. Following standard practice in association analysis, multiple test corrections were used to control the family-wise error at 0.05. With around 500,000 CpG sites, the Bonferroni corrected epigenome-wide significance level is 10^{-7}.

We fit two LMMs to the data: the standard LMM (implicitly assuming homogeneous technical variations), and the proposed LMM accommodating technical variations (denoted as LMMt). In addition, we also filter out those CpG sites with poor reproducibility. Specifically we only focus on those CpG sites with ICC larger than 0.5 (Bai *et al.*, 2016). Based on the two approaches, we identified 1863

epigenome-wide significant CpG sites. Among them, 1738 CpG sites are identified by both approaches, 65 CpG sites are only identified by LMM, and 60 CpG sites are only identified by LMMt. Figure 4.1 shows the boxplot for these two sets of CpG sites uniquely identified by both methods. Generally we can see that the significant CpG sites uniquely identified by the proposed LMMt have larger ICC values than LMM. Figure 4.2 compares the LMM and LMMt p-values for these CpG sites further divided into two groups based on an ICC cutoff value of 0.75.

Figure 4.1: ICC (reproducibility measure) comparison of epigenome-wide CpG sites identified by LMM and LMMt.

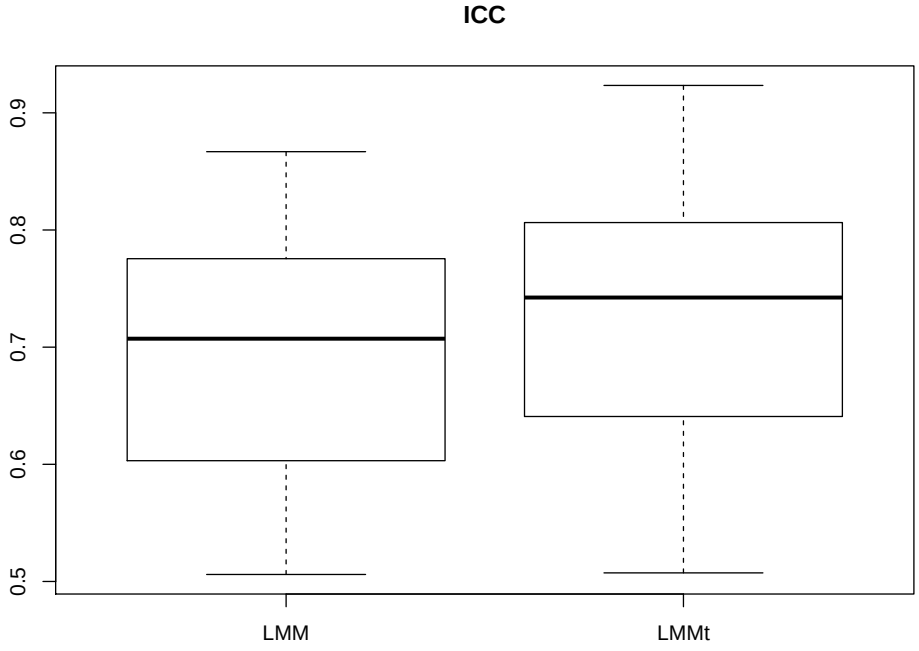
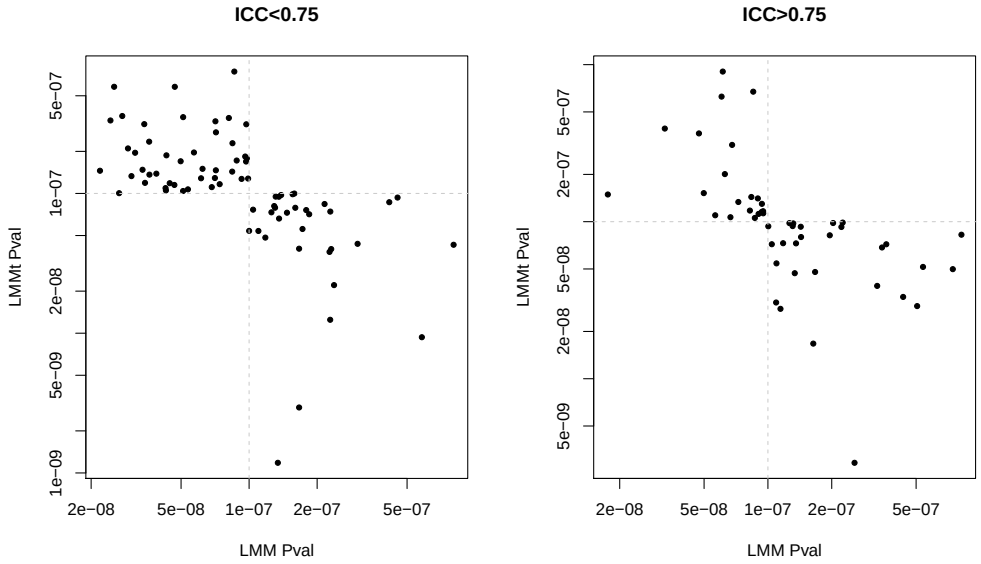
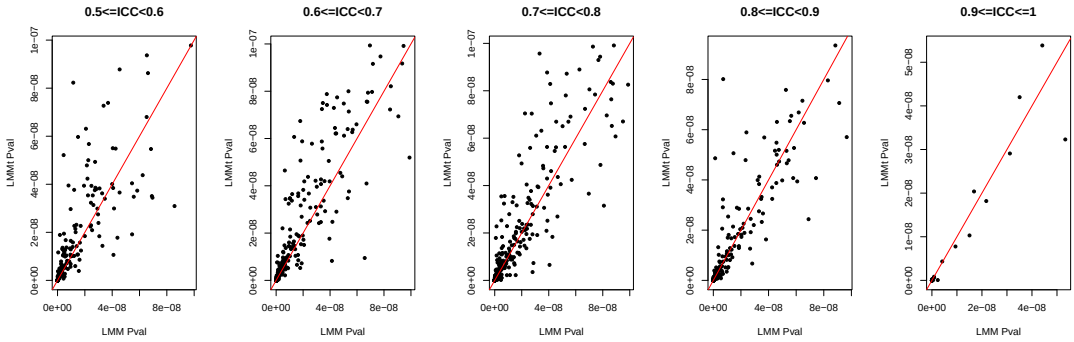


Figure 4.2: Comparison of p-values for epigenome-wide CpG sites uniquely identified by LMM and LMMt.



For the common set of 1738 epigenome-wide significant CpG sites, we divide them into five groups based on the ICC cutoff values (0.5,0.6,0.7,0.8,0.9,1). Figure 4.3 compares the LMM and LMMt p-values. Overall we can observe the trend that LMM and LMMt perform more and more similarly with increasing ICC values. This is intuitive since large ICC value means small technical variation compared to the biological variation, and hence we expect to see similar results of LMMt as LMM.

Figure 4.3: Comparison of p-values for epigenome-wide CpG sites identified by both LMM and LMMt.



4.3.2 Simulation study

We consider $n = 1000$ samples, and simulate a standard normal covariate X, and two independent standard normal random effects from two clusters, each simulated randomly as a factor of 100 and 200 levels respectively. The variances of technical variations are generated uniformly from interval $[0.2, 2]$. The outcomes Y are simulated as a zero mean multivariate normal random vector with covariance matrix $\sigma_0^2 \boldsymbol{I}_n + \sigma_1^2 Z_1 Z_1^T + \sigma_2^2 Z_2 Z_2^T + W$, where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix, Z_1 and Z_2 are the design matrices (of dimension 100×100 and 200×200) for the two random effects factors respectively, and W is the diagonal matrix of simulated technical variations. Here $\sigma_0 = \sigma_1 = \sigma_2 = 1$.

We fit the mean of Y as $\beta_0 + \beta_1 X$ and estimate the three variance parameters $\sigma_j^2, j = 0, 1, 2$. We fit both the LMM and LMMt using the REML, and apply the Wald tests to test $H_1 : \beta_1 = 0$. We conduct 1000 simulations to estimate the type I errors at the 0.05 and 0.01 levels. Table 4.1 summarizes the empirical type I errors. Overall we can see that LMM has inflated type I errors. In contrast LMMt controls the type I errors well.

Table 4.1: Type I errors for LMM and LMMt

α	0.01	0.05
LMM	0.015	0.055
LMMt	0.010	0.051

4.4 Discussion

Technical variation is common in genetic data, especially those generated from microarray platform, which is often ignored in association analysis. In this paper, we have applied a statistical method to quantify a particular type of technical variation ("bead-to-bead" variation) reported in array-based DNA methylation data, and developed novel regression framework that incorporates this variation explicitly.

In many biological pathways, DNA methylation is affected by the change of environmental factors. Explaining the variation in DNA methylation is an essential step to understand this biological process. However, the methylation measures in practice, e.g., those from array-based platform, are subject to technical variations as well. Modeling the effects of technical factors in the equation, in addition to those of biological factors, help to set up appropriate statistical models for variation of methylation. Our new approach targets on a particular type of measurement variation which is due to the use of multiple beads at each probe of the microarray. This approach can be easily extended to model other types of technical variation, if they can be explicitly quantified.

In summary, we have proposed a novel regression framework to incorporate known technical variations in association analysis of DNA methylation. The R

scripts for the proposed method is freely available upon request. We hope that our methods can lead to powerful statistical tests for future EWASs using the HM450 or other types of methylation arrays.

4.5 Appendix

It is generally appreciated that computing the MLE/REML for standard LMMs is challenging due to the non-negative definite constraint on the random effects covariance matrix and dependence among observations. Robust EM algorithms have been developed and well-studied for standard LMMs. The incorporation of additional measurement error components in the proposed LMM brings additional challenges to computing the MLE/REML. In the following, we develop easy-to-implement EM algorithms to solve the MLE/REML for general LMM with additional measurement error components.

4.5.1 Maximum likelihood estimation

Denote $P_j = Z_j V_\theta Z_j^T + \Gamma_j + \sigma^2 \boldsymbol{I}$. We have $Y_j \sim N(X_j\beta, P_j)$. We can easily derive the generalized least squares (GLS) estimates of the fixed effects parameter β

$$\hat{\beta} = (\sum_{j=1}^{n} X_j^T P_j^{-1} X_j)^{-1} (\sum_{j=1}^{n} X_j^T P_j^{-1} Y_j).$$

Note that (U_j, W_j, Y_j) jointly follows a multivariate normal distribution with

$$Cov(Y_j, U_j) = Z_j V_\theta, \quad Cov(Y_j, W_j) = \Gamma_j, \quad Cov(U_j, W_j) = 0.$$

We treat both U_j and W_j as missing data. The complete data log likelihood is

$$\sum_j [\log \Pr(U_j) + \Pr(W_j) + \log \Pr(Y_j | U_j, W_j)],$$

where $\Pr(Y_j|U_j, Wj) \sim N(X_j\beta + U_j + W_j, \sigma^2 \boldsymbol{I}^2)$ is a simple linear regression model.

In the E-step we compute

$$E(U_j|Y_j) = V_\theta Z_j^T P_j^{-1}(Y_j - X_j\beta), \quad Var(U_j|Y_j) = V_\theta Z_j^T P_j^{-1} Z_j V_\theta,$$

$$E(W_j|Y_j) = \Gamma_j P_j^{-1}(Y_j - X_j\beta), \quad Var(W_j|Y_j) = \Gamma_j P_j^{-1} \Gamma_j.$$

And

$$Cov[(U_j^T, W_j^T)^T | Y_j] = \text{diag}(V_\theta, \Gamma_j) - Var[E[(U_j, W_j)|Y_j]].$$

Hence

$$Cov(U_j, W_j|Y_j) = -V_\theta Z_j^T P_j^{-1} \Gamma_j,$$

and

$$E(U_j W_j^T | Y_j) = Cov(U_j, W_j|Y_j) + E(U_j|Y_j) E(W_j|Y_j)^T.$$

In the M-step, we estimate covariance parameter θ by

$$\min_\theta tr[V_\theta^{-1} \sum_j E(U_j U_j^T | Y_j)] + n \log(|V_\theta|),$$

where $E(U_j U_j^T|Y_j) = E(U_j|Y_j)E(U_j|Y_j)^T + Var(U_j|Y_j)$. The variance parameter σ^2 can be estimated by

$$\min_{\sigma^2} \sum_j \frac{E[\|Y_j - X_j\beta - U_j - W_j\|^2|Y_j]}{\sigma^2} + n_j \log(\sigma^2),$$

which leads to

$$\hat{\sigma}^2 = \frac{\sum_j E(\|Y_j - X_j\beta - U_j - W_j\|^2|Y_j)}{\sum_j n_j}.$$

Note that conditional on Y_j, $Y_j - X_j\beta - U_j - W_j$ follows a multivariate normal distribution with mean

$$[\boldsymbol{I} - (V_\theta Z_j^T + \Gamma_j)P_j^{-1}](Y_j - X_j\beta),$$

and covariance

$$V_\theta Z_j^T P_j^{-1} Z_j V_\theta + \Gamma_j P_j^{-1} \Gamma_j - V_\theta Z_j^T P_j^{-1} \Gamma_j - \Gamma_j P_j^{-1} Z_j V_\theta.$$

To reduce the bias of estimating covariance parameters (θ and σ^2), we can use the residual maximum likelihood (REML) estimation described as follows.

4.5.2 REML estimation

For the LMM

$$Y_j = X_j\beta + Z_j U_j + W_j + \epsilon_j,$$

assume that $U_j \sim N(0, V_\theta)$, $W_j \sim N(0, \Gamma_j)$, and $\epsilon_j \sim N(0, \sigma^2 \boldsymbol{I})$, where $\Gamma_j = \text{diag}(\tau_{1j}^2, \ldots, \tau_{n_j j}^2)$ with τ_{ij}^2 given. The three random errors are assumed mutually independent. Denote $P_j = Z_j V_\theta Z_j^T + \Gamma_j + \sigma^2 \boldsymbol{I}$. Assume a constant prior for β, $\Pr(\beta) \approx 1$. We have $(Y_j|\beta) \sim N(X_j\beta, P_j)$. We can easily check that conditional on Y_j, β follows a multivariate normal distribution with mean $\hat{\beta} = (\sum_{j=1}^n X_j^T P_j^{-1} X_j)^{-1}(\sum_{j=1}^n X_j^T P_j^{-1} Y_j)$, and covariance $V_\beta = (\sum_{j=1}^n X_j^T P_j^{-1} X_j)^{-1}$.

To derive the EM algorithm, we just need the conditional distribution of (U_j, W_j) on Y_j. We can easily check that they still follow the multivariate normal distribution. Specifically for U_j, we have mean $E(U_j|Y_j) = V_\theta Z_j^T P_j^{-1}(Y_j - X_j\hat{\beta})$, and variance

$$Var(U_j|Y_j) = V_\theta Z_j^T P_j^{-1} Z_j V_\theta + V_\theta Z_j^T P_j^{-1} X_j V_\beta X_j^T P_j^{-1} Z_j V_\theta.$$

For W_j, we have mean $E(W_j|Y_j) = \Gamma_j P_j^{-1}(Y_j - X_j\hat{\beta})$, and variance

$$Var(W_j|Y_j) = \Gamma_j P_j^{-1} \Gamma_j + \Gamma_j^T P_j^{-1} X_j V_\beta X_j^T P_j^{-1} \Gamma_j.$$

We can further check that

$$Cov(U_j, W_j|Y_j) = E[Cov(U_j, W_j|Y_j, \beta)|Y_j] + Cov[E(U_j|Y,\beta), E(W_j|Y,\beta)|Y_j]$$

$$= -V_\theta Z_j^T P_j^{-1} \Gamma_j + V_\theta Z_j^T P_j^{-1} X_j V_\beta X_j^T P_j^{-1} \Gamma_j.$$

Chapter 5

Conclusion and Discussion

In Chapter 2, we have studied statistical methods to mine the publicly available GWAS summary statistics. In the study, we have considered the standard chi-square statistics and the SPU tests. The SPU tests are based on summing the chi-square statistics across all traits directly (and hence ignored the trait correlation). SPU tests could have potential power loss when traits have heterogeneous effects and correlations. Potentially we could perform an adaptive test based on combining the SPU tests and the standard chi-square test. A similar Monte Carlo sampling from the multivariate normal distribution can be used to compute the p-values for the combined adaptive test. It is worthwhile to empirically investigate its performance.

It has been argued in the literature that these Monte Carlo sampling approaches may not scale well to the genome-wide association test, since the typical genome-wide significance level of 5×10^{-8} means that we need to perform more than 10^9 Monte Carlo simulations to reliably approximate the p-values of any genome-wide significant variant. However note that in most GWAS, we typically only observe a very small number of genome-wide significant variants. Thus we can use the following adaptive Monte Carlo sampling to efficiently compute p-values for all variants. Specifically we can first perform, e.g., 10^5 Monte Carlo simulations, which can be used to accurately estimate those p-values larger than 10^{-3}. Next we only need to consider those variants with p-values less than 10^{-3}. We can then perform, e.g., 10^7 Monte Carlo simulations, which can be used to accurately estimate those p-values larger than 10^{-5}. If we continue simulations as previously by sequentially increasing the number of Monte Carlo simulations, we just need to perform large-scale simulations (e.g., 10^9) for those genome-wide significant variants. This approach is thus more computationally efficient than naively applying 10^9 Monte Carlo simulations to all variants.

An implicit assumption for the multi-trait association test methods is that the detected association refers to the joint effects of all traits, and does not tell us which one of the traits is actually driving the association. Therefore additional tests are needed to further test whether each trait is associated with the variant.

It is worthwhile to investigate unified approach to testing joint and individual associations simultaneously.

In our proposed mixed effects model for EWAS, we have used the Wald test and assumed the normal distribution to compute the significance p-values, mainly for computational convenience. It is known that for the mixed effects model, the Wald test may not have good finite sample performance, since it does not account for the effects of estimating all the variance parameters. There are potentially two alternative approaches. Firstly, we can use the approximate F-test following the idea of Kenward and Roger (1997), which tried to use the F-test to properly account for the uncertainty of estimated variance parameters. Secondly, we can use the parametric Bootstrap approach to computing the Monte Carlo p-value. This approach is more computing intensive, though generally applicable. From a statistical perspective, it is also worthwhile to investigate the performance of score test and the likelihood ratio test. The score test generally has less inflation compared to the Wald test, though it could be conservative. The likelihood ratio test is generally the most powerful under the correct model assumptions. It is worthwhile to investigate their relative performance in the context of our proposed mixed models that incorporate known technical variation through thorough numerical studies.

We acknowledge that the causal pathway between DNA methylation and biological factors can be in either direction. In the EWAS of ARIC smoking data, we have treated the methylation levels as the outcomes, since we believe it is more natural to interpret the model as smoking is likely to change the methylation levels. In other EWAS, it is possible that we would treat the methylation levels as the covariate, which then leads to a standard measurement error model (Carroll *et al.*, 2006). It's worthwhile to empirically compare the two modeling approaches.

We have used the delta method to approximately compute the technical variation of the methylation β value, which has been computed based on the average of M and U values across multiple beads. Potentially, we could completely recover the technical variation information by simultaneously analyzing the (M, U) values across all beads, which should be modeled appropriately to incorporate their dependence. This may give us the complete information and hence potentially lead to more power, yet it will also bring both computational and modeling challenges as we now need to account for both the dependence across beads within each individual and the dependence across the same chip and same row on the chip (batch effects). It is worthwhile to investigate this modeling approach in the future.

It has been suggested that the methylation measures (in the range of 0 to 1) should be modeled with other (e.g., the Beta) distributions instead of the normal distribution. For these non-continuous outcomes, it is natural to extend

our proposed method to a generalized linear mixed effects model (GLMM) based approach that incorporate the technical variation as part of the random effects. For a standard GLMM, computing the maximum likelihood estimates (MLE) is typically very challenging. Generally numerical methods (e.g., Laplace approximation, or the more computing-intensive Gauss-Hermite quadrature) are used to approximate the MLE. The computing intensive Markov Chain Monte Carlo approaches have also been applied to more accurately compute the MLE for the GLMM. Incorporating the technical variation will bring more complications to computing the MLE. We will investigate the use of various numerical approximation methods and the Monte Carlo EM algorithm along the line of our previously proposed EM algorithm for the LMM.

References

Aslibekyan,S., Demerath,E.W., Mendelson,M., Zhi,D., Guan,W., Liang,L., Sha,J., Pankow,J.S., Liu,C., Irvin,M.R., Fornage,M., Hidalgo,B., Lin,L.A., Stanton Thibeault,K., Bressler,J., Tsai,M.Y., Grove,M.L., Hopkins,P.N., Boerwinkle,E., Borecki,I.B., Ordovas,J.M., Levy,D., Tiwari,H.K., Absher,D.M. and Arnett,D.K. (2015) Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. *Obesity,* **23** (7), 1493–1501.

Bai,Y., Pankow,J.S., Kilaru,V., Demerath,E.W., Bressler,J., Fornage,M., Grove,M.L., Tsai,M.Y., Guan,W. (2016) Incorporate technical variation to assess reproducibility of genome-wide methylation data. tech report.

Barfield,R.T., Almli,L.M., Kilaru,V., Smith,A.K., Mercer,K.B., Duncan,R., Klengel,T., Mehta,D., Binder,E.B., Epstein,M.P., Ressler,K.J. and Conneely,K.N. (2014) Accounting for population stratification in DNA methylation studies. *Genetic Epidemiology,* **38** (3), 231–241.

Bibikova,M., Barnes,B., Tsan,C., Ho,V., Klotzle,B., Le,J.M., Delano,D., Zhang,L., Schroth,G.P., Gunderson,K.L., Fan,J.B. and Shen,R. (2011) High density DNA methylation array with single CpG site resolution. *Genomics,* **98** (4), 288–295.

Bose,M., Wu,C., Pankow,J.S., Demerath,E.W., Bressler,J., Fornage,M., Grove,M.L., Mosley,T.H., Hicks,C., North,K., Kao,W.H., Zhang,Y., Boerwinkle,E. and Guan,W. (2014) Evaluation of microarray-based DNA methylation measurement using technical replicates: the Atherosclerosis Risk In Communities (ARIC) Study. *BMC Bioinformatics,* **15** (1), 312.

Breitling,L.P., Yang,R., Korn,B., Burwinkel,B. and Brenner,H. (2011) Tobacco-smoking-related differential DNA methylation: 27k discovery and replication. *American Journal of Human Genetics,* **88** (4), 450–457.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M. (2006) Measurement error in nonlinear models: a modern perspective. *CRC press.*

Chen,J., Just,A.C., Schwartz,J., Hou,L., Jafari,N., Sun,Z., Kocher,J.P.A., Baccarelli,A. and Lin,X. (2016) CpGFilter: model-based CpG probe filtering with replicates for epigenome-wide association studies. *Bioinformatics (Oxford, England),* **32** (3), 469–471.

Dedeurwaerder,S., Defrance,M., Calonne,E., Denis,H., Sotiriou,C. and Fuks,F. (2011) Evaluation of the Infinium Methylation 450k technology. *Epigenomics,* **3** (6), 771–784.

Demerath,E.W., Guan,W., Grove,M.L., Aslibekyan,S., Mendelson,M., Zhou,Y.H., Hedman,.K., Sandling,J.K., Li,L.A., Irvin,M.R., Zhi,D., Deloukas,P., Liang,L., Liu,C., Bressler,J., Spector,T.D., North,K., Li,Y., Absher,D.M., Levy,D., Arnett,D.K., Fornage,M., Pankow,J.S. and Boerwinkle,E. (2015) Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Human Molecular Genetics,* **24** (15), 4464–4479.

Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological),* **39** (1), 1–38.

Dugua,P.A., English,D.R., MacInnis,R.J., Joo,J.E., Jung,C.H. and Milne,R.L. (2015) The repeatability of DNA methylation measures may also affect the power of epigenome-wide association studies. *International Journal of Epidemiology,* **44** (4), 1460–1461.

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 bioinformatics,* **11**, 587.

Ferreira,M.A.R. and Purcell,S.M. (2009) A multivariate test of association. *Bioinformatics,* **25** (1), 132–133.

Galesloot,T.E., van Steen,K., Kiemeney,L.A.L.M., Janss,L.L. and Vermeulen,S.H. (2014) A comparison of multivariate genome-wide association methods. *PLoS ONE,* **9** (4), e95923.

Grove,M.L., Yu,B., Cochran,B.J., Haritunians,T., Bis,J.C., Taylor,K.D. and others. (2013) Best practices and joint calling of the HumanExome BeadChip: the CHARGE consortium. *PloS One,* **8** (7), e68095.

He,Q., Avery,C.L. and Lin,D.Y. (2013) A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic Epidemiology,* **37** (8), 759–767.

Houseman,E.A., Accomando,W.P., Koestler,D.C., Christensen,B.C., Marsit,C.J., Nelson,H.H., Wiencke,J.K. and Kelsey,K.T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics,* **13**, 86.

Jaffe,A.E. and Irizarry,R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology,* **15** (2), R31.

Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England),* **8** (1), 118–127.

Joubert,B.R., Hberg,S.E., Nilsen,R.M., Wang,X., Vollset,S.E., Murphy,S.K., Huang,Z., Hoyo,C., Midttun,., Cupul-Uicab,L.A., Ueland,P.M., Wu,M.C., Nystad,W., Bell,D.A., Peddada,S.D. and London,S.J. (2012) 450k epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives,* **120** (10), 1425–1431.

Kenward, M. G. and Roger, J. H. (1997) Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics,* **53**, 983–997.

Klei,L., Luca,D., Devlin,B. and Roeder,K. (2008) Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology,* **32** (1), 9–19.

Kuan,P.F., Wang,S., Zhou,X. and Chu,H. (2010*b*) A statistical framework for Illumina DNA methylation arrays. *Bioinformatics (Oxford, England),* **26** (22), 2849–2855.

Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics,* **11** (3), 191–203.

Lee,G. and Scott,C. (2012) EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis,* **56** (9), 2816–2829.

Lin,D.Y. and Zeng,D. (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology,* **34** (1), 60–66.

Marabita,F., Almgren,M., Lindholm,M.E., Ruhrmann,S., Fagerstrm-Billai,F., Jagodic,M., Sundberg,C.J., Ekstrm,T.J., Teschendorff,A.E., Tegnr,J. and Gomez-Cabrero,D. (2013) An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics,* **8** (3), 333–346.

Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A. and others. (2009) Finding the missing heritability of complex diseases. *Nature,* **461** (7265), 747–753.

Meng,H., Joyce,A.R., Adkins,D.E., Basu,P., Jia,Y., Li,G., Sengupta,T.K., Zedler,B.K., Murrelle,E.L. and van den Oord,E.J.C.G. (2010) A statistical

method for excluding non-variable CpG sites in high-throughput DNA methylation profiling. *BMC bioinformatics,* **11**, 227.

Moran,S., Arribas, C., Esteller, M. (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics,* **8** (3), 389–399.

Pan,W., Kim,J., Zhang,Y., Shen,X. and Wei,P. (2014) A powerful and adaptive association test for rare variants. *Genetics,* genetics.114.165035.

Pidsley,R., Y Wong,C.C., Volta,M., Lunnon,K., Mill,J. and Schalkwyk,L.C. (2013) A data-driven approach to preprocessing Illumina 450k methylation array data. *BMC Genomics,* **14**, 293.

Rauch,T.A. and Pfeifer,G.P. (2010) DNA methylation profiling using the methylated-CpG island recovery assay (MIRA). *Methods (San Diego, Calif.),* **52** (3), 213–217.

Ryu,D. (2013) Quantifying and Normalizing Methylation Levels in Illumina Arrays. *Journal of Biometrics & Biostatistics,* **04** (03).

Sandoval,J., Heyn,H., Moran,S., Serra-Musach,J., Pujana,M.A., Bibikova,M. and Esteller,M. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics,* **6** (6), 692–702.

Shenker,N.S., Polidoro,S., van Veldhoven,K., Sacerdote,C., Ricceri,F., Birrell,M.A., Belvisi,M.G., Brown,R., Vineis,P. and Flanagan,J.M. (2013) Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Human Molecular Genetics,* **22** (5), 843–851.

Shvetsov,Y.B., Song,M.A., Cai,Q., Tiirikainen,M., Xiang,Y.B., Shu,X.O. and Yu,H. (2015) Intraindividual Variation and Short-term Temporal Trend in DNA Methylation of Human Blood. *Cancer Epidemiology Biomarkers & Prevention,* **24** (3), 490–497.

Stephens,M. (2013) A unified framework for association analysis with multiple related phenotypes. *PLoS ONE,* **8** (7), e65245.

Teslovich,T.M., Musunuru,K., Smith,A.V., Edmondson,A.C., Stylianou,I.M., and others. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature,* **466** (7307), 707–713.

Teschendorff,A.E., Marabita,F., Lechner,M., Bartlett,T., Tegner,J., Gomez-Cabrero,D. and Beck,S. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics (Oxford, England),* **29** (2), 189–196.

The ARIC Investigators. (1989) The atherosclerosis risk in communities (aric) study: design and objectives. *American Journal of Epidemiology,* **129** (4), 687–702.

Touleimat,N. and Tost,J. (2012) Complete pipeline for Infinium Human Methylation 450k BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics,* **4** (3), 325–341.

van der Sluis,S., Posthuma,D. and Dolan,C.V. (2013) TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet,* **9** (1), e1003235.

Wan,E.S., Qiu,W., Baccarelli,A., Carey,V.J., Bacherman,H., Rennard,S.I., Agusti,A., Anderson,W., Lomas,D.A. and Demeo,D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Human Molecular Genetics,* **21** (13), 3073–3082.

Wang,K. and Abbott,D. (2008) A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology,* **32** (2), 108–118.

Yang,Q., Wu,H., Guo,C.Y. and Fox,C.S. (2010) Analyze multivariate phenotypes

in genetic association studies by combining univariate association tests. *Genetic Epidemiology,* **34** (5), 444–454.

Zhang,Y., Xu,Z., Shen,X. and Pan,W. (2014) Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage,* **96**, 309–325.

Zhu,X., Feng,T., Tayo,B., Liang,J., Young,J.H., Franceschini,N., Smith,J., Yanek,L., Sun,Y., Edwards,T., Chen,W., Nalls,M., Fox,E., Sale,M., Bottinger,E., Rotimi,C., Liu,Y., McKnight,B., Liu,K., Arnett,D., Chakravati,A., Cooper,R. and Redline,S. (2015) Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics,* **96**, 21–36.