Kernel Machine Methods for Integrative Analysis of Genome-Wide Methylation and Genotyping Studies

Ni Zhao[1,*], Xiang Zhan[2], Yen-Tsung Huang[3], Lynn M Almli [4], Alicia Smith [5], Michael P. Epstein[6], Karen Conneely[6], Michael C. Wu[7, *]

[1]Departments of Biostatistics, Johns Hopkins University, Baltimore, MD 21205
[2]Department of Public Health Sciences, Pennsylvania State University, Hershey, PA 17033
[3]Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan
[4]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA 30322
[5]Department of Gynecology and Obstetrics, Emory University, Atlanta, GA 30322
[6]Department of Human Genetics, Emory University, Atlanta, GA 30322
[7]Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

*Address for Correspondence:
Ni Zhao
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street, Baltimore, MD 21205
Email: nzhao10@jhu.edu
and
Michael C. Wu
Public Health Sciences Division
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, Seattle, WA 98109
Email: mcwu@fredhutch.org

Abstract

Many large GWAS consortia are expanding to simultaneously examine the joint role of DNA methylation in addition to genotype in the same subjects. However, integrating information from both data types is challenging. In this paper, we propose a composite kernel machine regression model to test the joint epigenetic and genetic effect. Our approach works at the gene level, which allows for a common unit of analysis across different data types. The model compares the pairwise similarities in the phenotype to the pairwise similarities in the genotype and methylation values; and high correspondence is suggestive of association. A composite kernel is constructed to measure the similarities in the genotype and methylation values between pairs of samples. We demonstrate through simulations and real data applications that the proposed approach can correctly control type I error, and is more robust and powerful than using only the genotype or methylation data in detecting trait-associated genes. We applied our method to investigate the genetic and epigenetic regulation of gene expression in response to stressful life events using data that are collected from the Grady Trauma Project. Within the kernel machine testing framework, our methods allows for heterogeneity in effect sizes, nonlinear and interactive effects, as well as rapid p-value computation.

1 Introduction

The etiology for most common human diseases is believed to be multifactorial, with risk factors including heritable genetic variants as well as environmental, behavioral factors and possible interactions between them (Luzzatto and Pandolfi, 2015; Kirchner et al., 2013). In the past few years, genome-wide association studies (GWAS) have been successful in identifying genetic variants, especially in the form of single nucleotide polymorphisms (SNP), that are associated with a number of human diseases (Visscher et al., 2012). However, SNPs discovered by GWAS can usually account for only a small fraction of the genetic varia-

tion of phenotypes in the human population, leading to the so-called "missing heritability" phenomenon (Maher, 2008; Manolio et al., 2009).

Several reasons have been suggested for the missing heritability, including the possibility that environmental or genetic risk factors impact the disease risk through epigenetic modifications, especially via DNA methylation (Johannes et al., 2008, 2009). Consequently, many large GWAS consortia are expanding to simultaneously examine the joint effect of DNA methylation. For example, Liu et al. (2013) identified two clusters within the major histocompatibility complex region whose methylation level mediates the genetic risk of rheumatoid arthritis using genome-wide methylation variation. The importance of jointly considering the genetic and epigenetic effect is further highlighted by the association between epigenetic perturbations and cancer (Zhang et al., 2012; Kulis and Esteller, 2010). Some large scale genomic studies, such as the Cancer Genome Atlas (The Cancer Genome Atlas Consortium, 2012, 2008), International Cancer Genome Consortium (Hudson et al., 2010), now routinely collect information on methylation for different types of cancers. This "multi-dimensional genomic data" provides unique opportunities to explore the regulation in biological processes underlying the disease of interest.

Current examination of the "multi-dimensional" genetic and methylation data often starts by forming SNP-CpG pairs, and then tests the joint effect of each SNP-CpG pair via parametric models, followed by subsequent multiple comparison adjustment (Liu et al., 2013; Hong et al., 2015). This approach suffers from the same limitations as the single SNP analysis in GWAS. First, the method can be underpowered due to the multiple-comparison burden. When the number of SNPs and CpGs increases, the multiple-comparison burden can very quickly become extremely stringent, prohibiting such methods from genome-wide application. Filtering is necessary so that only CpGs that are differentially methylated and SNPs that are partially associated with the phenotype can be included in the integrative analysis (Liu et al., 2013). Second, because of the imperfect linkage disequilibrium (LD)

between the typed SNPs and the true causal variants, this single SNP analysis suffers from poor reproducibility. Third, the single marker analysis fails to detect joint and possible interactive effects from multiple SNPs and CpGs. To overcome these limitations, we propose a region-based approach that combines SNPs and CpGs into biologically meaningful marker-sets and tests their joint association with the phenotype. We focus on gene-based analysis. SNPs and CpGs that are mapped to the same gene form a SNP-set and a CpG-set. It is also possible to group the SNPs and CpGs based on other genomic features such as biological pathways, functional groups or other important genomic regions, which we will broadly referred to as a "gene". We consider a "gene" as the unit of analysis. By combining the genotype and methylation data together, our approach answers the simple and biologically meaningful question "is the 'gene' associated with the phenotype?" and "if so, what is the possible underlying causal relationship?"

We propose to use the semiparametric kernel machine regression (KMR) framework for testing the joint genetic and epigenetic effect on the phenotype. KMR (Liu et al., 2007, 2008) is a powerful and operationally simple approach that has gained much popularity in the field of GWAS (Kwee et al., 2008; Wu et al., 2010; Lin et al., 2011) and rare variant association studies (Wu et al., 2011; Lee et al., 2012). The KMR measures the genetic similarity with a kernel function, a common tool in the support vector machine literature (Cristianini and Shawe-Taylor, 2000), and compares the pairwise similarity in genotype to the pairwise similarity in the phenotype. High correspondence is suggestive of association. For the joint testing, we propose to construct a composite kernel, a weighted average of two kernels that are constructed from genotype and methylation data seperately, to measure the joint similarity in both data types. This composite kernel is then compared to the phenotype for association testing. The weighting parameter in the composite kernel can be varied for optimal power. Statistical significance of association is evaluated via a perturbation based approach, which is more powerful in some scenarios, or a projection approach, which is

computationally faster.

We applied our method to investigate the genetic and epigenetic regulation of gene expression in response to stressful life events using data that are collected from the Grady Trauma Project, and identified 732 genes that showed significant cis-regulation, i.e., the genotype and methylation regulates the expression of the gene itself. For genes that show significant joint association, we further considered a subsequent mediation model to infer the possible causal relation, with the assumption that the methylation can be a potential mediator of the genetic effect on gene expression. The mediation model extends the classical causal steps model (Baron and Kenny, 1986; Judd and Kenny, 1981; MacKinnon et al., 2007), and investigates the possible causal relationships between genotype, methylation and gene expression in this data set by evaluating whether 1) the genotype is associated with methylation and 2) whether the methylation is associated with gene expression on the genotype.

Our proposed framework provides a unified approach for integrative analysis of genotype and methylation. By testing the genetic and epigenetic effect together, our joint analysis approach can be more robust and usually more powerful than methods that investigate only one data type. The gene-based approach provides a common unit of analysis for different data types, reduces the number of comparisons, and facilitates interpretation of results.

2 Method

2.1 Models and Notations

Suppose that data are collected on n independent subjects with quantitative or dichotomous phenotype $(y_1, y_2, \cdots, y_n)'$. We assume that the SNPs and CpGs are grouped into different genes based on prior biological information. $\mathbf{G}$ is a $n \times p_1$ matrix of the genotype data and $\mathbf{M}$ is a $n \times p_2$ matrix containing the methylation values, with each row denoting data for a single subject. p_1 and p_2 are the total number of SNPs and CpG sites in this gene. Let

$\mathbf{X}$ denote a $n \times q$ matrix of additional variables that we want to adjust for, such as age, gender, smoking status, and principal components (PC) for population structure. We let $\mathbf{Z} = (\mathbf{G}, \mathbf{M})$ represent both the genotype and methylation data.

We relate the phenotype to genotype and methylation via a semiparametric KMR model. Specifically, we consider that:

$$g(E(\mathbf{y})) = \beta_0 + \mathbf{X}\boldsymbol{\beta} + h(\mathbf{Z}), \tag{1}$$

where $g(\cdot)$ is a link function that can be set to the identity function or the logistic function when the $\mathbf{y}$ is continuous or dichotomous. $\boldsymbol{\beta}$ is the vector of regression coefficients for the additional covariates $\mathbf{X}$, and $h(\cdot)$ is a possibly nonlinear function of the joint genotype and methylation effect on the phenotype of interest.

2.2 Kernel Machine Regression Model

Under the KMR framework, h is assumed to be a function in a reproducing Kernel Hilbert space $\mathcal{H}$ generated by some kernel function $K(\cdot, \cdot)$. The kernel matrix $\mathbf{K}$ is a matrix with the $(i, i')^{th}$ element being $K(Z_i, Z_{i'})$, which measures the similarity between individual i and i' based on the genotype and methylation values.

We are interested in the joint genotype and methylation effect. The null hypothesis is:

$$H_0 : h(\cdot) = 0.$$

To test this hypothesis, we use the connection with linear mixed model. Specifically, it has been shown that the KMR model in (1) is equivalent to the (generalized) linear mixed model (Liu et al., 2007; Pearce and Wand, 2006)

$$g(E(\mathbf{y})) = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \mathbf{h},$$

in which $\mathbf{h}$ is a random effect with mean 0 and variance $\tau\mathbf{K}$. Under this mixed model specification, testing the null hypothesis $\mathbf{h} = 0$ is equivalent to testing whether $\tau = 0$.

Under the KMR framework, hypothesis testing is conducted via the variance component score test (Kwee et al., 2008; Wu et al., 2010, 2011; Liu et al., 2007). The test statistic is constructed as

$$Q = (\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \hat{\mathbf{y}}_0)/\phi,$$

where $\hat{\mathbf{y}}_0$ are the estimates of $\mathbf{y}$ under the null hypothesis, and ϕ is the estimated dispersion parameter. $\phi = \hat{\sigma}_0^2$, the estimated residual variance under the null hypothesis when $\mathbf{y}$ is continuous, and $\phi = 1$ when $\mathbf{y}$ is dichotomous. Under the null hypothesis, Q asymptotically follows a mixture of χ^2 distribution, which can be obtained easily via moment matching (Liu et al., 2007, 2009) or exact methods (Davies, 1980).

2.3 Composite Kernel for Combined Genotype and Methylation Effect

In principle, any positive semi-definite matrix that satisfies Mercer's theorem (Cristianini and Shawe-Taylor, 2000) can be used as a valid kernel. However, good choices of kernel that fully incorporate the properties of the data can lead to statistical tests with higher power. Constructing a kernel using either genotype or methylation data is a common practice, with many kernels designed for each data type. For example, popular kernels in SNP-set analysis include the linear kernel, the quadratic kernel, the identical by state (IBS) kernel and their weighted counterparts.

- Linear Kernel: $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \mathbf{G}_i'\mathbf{G}_{i'}$
- Quadratic Kernel: $K(\mathbf{G}_i, \mathbf{G}_{i'}) = (\mathbf{G}_i'\mathbf{G}_{i'} + 1)^2$
- IBS Kernel: $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \frac{1}{2p}\sum_{j=1}^{p}(2 - |G_{ij} - G_{i',j}|)$

Suitable kernels for methylation data include the Gaussian kernel, the linear kernel, and

kernels that incorporate both the methylation values and the CpG location (Mayo et al., 2015).

While kernels for individual data type exist, constructing a proper kernel that incorporates both genotype and methylation data is not straightforward. First, it is usually unrealistic to assume that the SNP-set and the CpG-set influence the phenotype in the same manner, and naive construction of kernels using a data-set that concatenates the two data types is often inappropriate. Secondly, the units and scales are inherently different for genotype and methylation data. Genotype data counts the number of minor alleles at each locus and is scaled to 0, 1 and 2. However, methylation data measures the proportion of CpGs that are methylated at each position and is inherently quantitative. Proper weighting is necessary in constructing a kernel that incorporates both data types.

We propose to construct a composite kernel as follows:

$$K(\mathbf{Z}, \mathbf{Z}) = wK_1(\mathbf{G}, \mathbf{G}) + (1 - w)K_2(\mathbf{M}, \mathbf{M}),$$

where K_1 and K_2 are kernel functions for genotype and methylation respectively, and $w \in [0, 1]$ is a weighting parameter. Any valid kernel based on genotype or methylation data can be used for K_1 or K_2, allowing for distinct genotype and methylation effect to be modeled additively. Per the connection with generalized linear mixed model (Liu et al., 2007; Pearce and Wand, 2006), the function h in the kernel machine regression model (1) with such a composite kernel can be considered as a random effect with mean 0 and variance $\tau w\mathbf{K}_1 + \tau(1 - w)\mathbf{K}_2$.

Because of the possibility that $\mathbf{K}_1$ and $\mathbf{K}_2$ are of different scales, and that $Q_G = \phi^{-1}(\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}_1(\mathbf{y} - \hat{\mathbf{y}}_0)$ and $Q_M = \phi^{-1}(\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}_2(\mathbf{y} - \hat{\mathbf{y}}_0)$ are of orders of magnitude different, we standardize $\mathbf{K}_1$ and $\mathbf{K}_2$ before constructing an composite kernel. In specific, we calculate $\eta = \mathrm{SD}(Q_G)/\left(\mathrm{SD}(Q_G) + \mathrm{SD}(Q_M)\right)$, take $\tilde{\mathbf{K}}_1 = (1 - \eta)\mathbf{K}_1$ and $\tilde{\mathbf{K}}_2 = \eta\mathbf{K}_2$, and use $\tilde{\mathbf{K}}_1$ and

$\tilde{\mathbf{K}}_2$ to construct the composite kernel. With some abuse of notation, we designate $\tilde{\mathbf{K}}_1$ and $\tilde{\mathbf{K}}_2$ as $\mathbf{K}_1$ and $\mathbf{K}_2$. By doing this standardization, the genetic and epigenetic data contribute equally to the test statistic when w is selected as 0.5. By doing so, we implicitly scales the genotype and methylation data by weighting at the kernel level instead of at the original data level. w is constrained to ensure the positive semi-definiteness of the composite kernel. w controls the relative contribution of the genotype and methylation to the phenotype. When $w = 1$, the implicit assumption is that the phenotype does not depend on methylation; and $w = 0$ implies that the genotype effect is zero.

When w is fixed, hypothesis testing is straightforward through the usual variance component test by considering the composite kernel as merely a single kernel. In reality, the optimal choice of w depends on the true nature of the genotype and methylation effects and is unknown prior to analysis. In the prediction-based statistical learning literature, w is usually estimated from the data. Unfortunately, this supervised estimation of w can lead to inflated type I error. In this manuscript, we propose two approaches that both consider multiple choices of w in constructing the composite kernel for association testing. The first method utilizes a perturbation based approach for the evaluation of statistical significance. The second method uses the idea of Kernel PCA, and enables analytical computation of the p-value.

2.3.1 Perturbation based Inference Based on Composite Kernel

The objective is to test the joint genotype and methylation effect using a composite kernel, without a prior specification of w. The perturbation based approach starts by considering a grid of $\mathbf{w}$: $(w_1, \ldots, w_L)$ between 0 and 1. This corresponds to L different kernels

$$K_d(\mathbf{Z}, \mathbf{Z}) = w_d K_1(\mathbf{G}, \mathbf{G}) + (1 - w_d) K_2(\mathbf{M}, \mathbf{M}).$$

For each K_d, it is easy to construct the score statistic Q_d and obtain the corresponding p-value p_d. We use the minimum p-value across different choices of ds as the test statistic, which is then compared to its null distribution for the final p-value. The null distribution is obtained through a perturbation-based approach that takes into account of the correlation between the score statistics. The technical details of the perturbation procedure are outlined in Wu et al. (2013), but we provide an overview of the procedure below.

This perturbation based approach takes advantage of our knowledge about the asymptotic distribution of the quadratic forms for $(Q_1, \ldots, Q_L)$. For quantitative traits, $(\mathbf{y} - \hat{\mathbf{y}}_0)/\hat{\sigma}_0$ is asymptotically distributed as standard normal when the null hypothesis is true. Then each Q_d can be viewed as a quadratic form of standard normal vectors sandwiching different kernel matrices. The vector of normals are the same across $(Q_1, \ldots, Q_L)$, with all the differences in the kernels. Therefore, we can replace $(\mathbf{y} - \hat{\mathbf{y}}_0)/\hat{\sigma}_0$ using newly generated standard normal vectors to construct the empirical null distribution. The simulated standard normal vectors are rotated properly (using an augmented matrix constructed from all the kernels) to capture the correlation between different kernels. Then for each perturbation sample and each kernel, we can obtain a p-value. We compare the smallest p-value from the original dataset with the smallest p-values across all kernels in all the perturbations for final p-value calculation, which guarantees that we have correct type I error control. The procedure is similar when $\mathbf{y}$ is dichotomous except that we need to use the working linear model.

Although this perturbation approach relies on a Monte Carlo generation of p-value, it offers advantages to permutation because it retains all the possible correlation between additional covariates and genotype/methylation effect, while direct permutation fails to do so. Perturbation is also computationally efficient: it requires only generating random normal vectors while permutation requires reconstruction of kernel matrices, and recalculation of p-values through the moment matching or characteristic function inversion method. Therefore, the perturbation based approach is computationally much faster than permutation.

The number of grids L can impact both the computational speed of the algorithm and the power of the test. With the increase of L, the computational time can have a modest increase, due to the need to generate a rotation matrix with dimensionality equal to the sum of number of nonzero eigenvalues from all of the kernels under consideration, followed by an eigen-decomposition of this matrix. If the rank of individual composite kernel is high (i.e. many SNPs and/or methylation markers with low correlation) and if the number of kernels under consideration is large (i.e. L is large), the reduction in computational speed is more profound. The relationship between L and the power is more complicated. The power depends on the number of tests under consideration (L), and the correlations between these tests and the test using the "optimal" kernel (i.e. the composite kernel with the optimal value of w). By taking into account the correlations between tests using different kernels, the perturbation procedure can have very little power loss even when many tests are conducted if the tests are highly correlated. In the most extreme cases that the tests from the different kernels output exactly the same p-value (for example, the same kernel is repeated multiple times), the final p-value will be the same as the individual p-values. However, in cases when the tests using different kernels are independent, including too many tests can lead to substantial power loss. Because the composite kernel is constructed using a weighted average of two kernels, correlation between tests depends on the weighting parameter w . If we set the grid points too crude, we run into the risk that none of the kernels captures the underlying data structure. However, if we set the grid points too dense, we include many tests that output very different p-values from the "optimal" kernel. Both situations will lead to reduced power. In our simulation and real data analysis, we selected to use five grid points (0, 0.25,0.5,0.75, 1), which showed a good practical balance.

2.3.2 Kernel PCA Based on Composite Kernel

Although computationally more efficient than permutation, the perturbation method still relies on Monte Carlo calculation of p-values. Analytical calculation of p-values cannot be obtained easily because of the possible correlation between the genotype and methylation effects.

Here we propose an alternative approach that starts from the same composite kernel, but linearly transform the kernel space using kernel PCA and basis projection, which enables analytical computation of the final p-values. Model (1) with a composite kernel is equivalent to the following model

$$g(E(\mathbf{y})) = \beta_0 + \mathbf{X}\boldsymbol{\beta} + h_1(\mathbf{G}) + h_2(\mathbf{M}), \tag{2}$$

in which h_1 and h_2 are from function spaces generated by kernel functions K_1 and K_2.

We use kernel PCA, a nonlinear version of PCA to transform the unknown nonparametric functions h_1 and h_2 into linear functions. Consider eigendecomposition of $\mathbf{K}_1$ and $\mathbf{K}_2$ such that $\mathbf{K}_1 = \mathbf{V}_G\boldsymbol{\Lambda}_G\mathbf{V}_G'$ and $\mathbf{K}_2 = \mathbf{V}_M\boldsymbol{\Lambda}_M\mathbf{V}_M'$ where $\mathbf{V}_G$ and $\mathbf{V}_M$ are the eigenvectors and $\boldsymbol{\Lambda}_G = diag(\lambda_{G,1} \geq \lambda_{G,2} \geq \cdots \geq \lambda_{G,k})$ and $\boldsymbol{\Lambda}_M = diag(\lambda_{M,1} \geq \lambda_{M,2} \geq \cdots \geq \lambda_{M,\ell})$ are the associated positive eigenvalues. Let $\mathbf{Z}_G = \mathbf{V}_G\boldsymbol{\Lambda}_G^{1/2}$ and $\mathbf{Z}_M = \mathbf{V}_M\boldsymbol{\Lambda}_M^{1/2}$. Model (2) can be rewritten as

$$g(E(\mathbf{y})) = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_G\boldsymbol{\beta}_G + \mathbf{Z}_M\boldsymbol{\beta}_M. \tag{3}$$

The major purpose of this kernel PCA is to transform the unknown nonparametric functions h_1 and h_2 into linear functions, which can then be easily manipulated. Although the dimension of the nonparametric functions h_1 and h_2 are unknown and potentially infinite, the dimensions of $\mathbf{Z}_G$ and $\mathbf{Z}_M$ are bounded by n. Low-rank approximation is also possible that we only use the leading eigenvectors for $\mathbf{Z}_G$ and $\mathbf{Z}_M$ that can explain the majority of

the variability in the genotype and methylation data.

After linearizing, we project the methylation data to the genotype and construct the following model:

$$g(E(\mathbf{y})) = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_G\boldsymbol{\gamma}_G + \mathbf{Z}_M^*\boldsymbol{\gamma}_M, \tag{4}$$

where $\mathbf{Z}_M^* = (I - \mathbf{P}_G)\mathbf{Z}_M$ with $\mathbf{P}_G = \mathbf{Z}_G(\mathbf{Z}_G'\mathbf{Z}_G)^{-1}\mathbf{Z}_G'$. $\mathbf{Z}_M^*$ is the residuals obtained by performing linear regressions of each component of $\mathbf{Z}_M$ on $\mathbf{Z}_G$. It corresponds to a subspace that is orthogonal to the column space of $\mathbf{Z}_G$. $\boldsymbol{\gamma}_G$ and $\boldsymbol{\gamma}_M$ are the regression coefficients under the transformed model.

We now assume $\boldsymbol{\gamma}_G$ and $\boldsymbol{\gamma}_M$ are random variables with mean 0 and variance τw and $\tau(1-w)$ respectively — this corresponds to that the genetic and epigenetic effect are random effects with mean 0 and variance $\tau w\mathbf{K}_1$ and $\tau(1-w)\mathbf{K}_2^*$ respectively, in which $\mathbf{K}_2^* = \mathbf{Z}_M^*\mathbf{Z}_M^{*\prime} = \mathbf{Z}_M(I - \mathbf{P}_G)\mathbf{Z}_M'$. The null hypothesis can be written as $H_0 : \tau = 0$ under this transformed model. A variance component score statistic under the transformed model can be constructed as follows:

$$\begin{aligned} Q^* &= \phi^{-1}[w(\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}_1(\mathbf{y} - \hat{\mathbf{y}}_0) + (1-w)(\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}_2^*(\mathbf{y} - \hat{\mathbf{y}}_0)] \\ &= wQ_G + (1-w)Q_M^*, \end{aligned} \tag{5}$$

Because Z_G and Z_M^* are orthogonal to each other, Q_G and Q_M^* are asymptotically independent mixtures of χ^2 distributions. Therefore, Q^* also follows a mixture of χ^2 distributions, which can be approximated using the moment matching approach as developed in Ionita-Laza et al. (2013). The details of choosing w and obtaining the final p-value can be found in the supplemental file section S1.

Note that we project the methylation data onto the genotype data, similar to a procedure that includes methylation as covariates and test for additional epigenetic effects beyond the genetic effect on the outcome. Projection on the other direction is also possible.

2.4 Simulation Studies

2.4.1 Simulations when the genotype and methylation are independent

We first conducted simulations when genotype and methylation are simulated independently. We simulated gene *ASAH1* (acid ceramidase 1), a gene that encodes enzyme acid ceramidase, which has been associated with a lysosomal storage disorder known as Farber disease (Li et al., 1999). 93 SNPs within *ASAH1* were simulated using HAPGEN2 (Su et al., 2011) to have the same LD structure as the CEU (Utah residents with ancestry from northern and western Europe) samples from international HAPMAP project under release 24 (Altschuler et al., 2005). Supplemental figure S1 shows a heatmap illustrating the LD structure of this simulated gene. We simulated genotypes for 15,000 subjects and randomly selected from this pool to generate data sets with desired sample sizes. 29 of the 93 SNPs are genotyped in the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA), constituting the "typed" SNPs. Only these "typed" SNPs were used for association testing. This mimics a real GWAS in that the causal SNPs may not be genotyped, but are only in LD with the "typed" SNPs. Methylation data (21 CpGs) for the same gene was simulated as multivariate normal $N(0, \Sigma_1)$, with the correlation Σ_1 estimated from a real data (Alisch et al., 2012).

We evaluated the performances of four testing approaches. Specifically, we considered the proposed model (1) that uses the composite kernel to evaluate the joint effect. To construct the composite kernel, we used IBS kernel for genotype because of its ability to model complex genetic effects, and linear kernel for methylation. We evaluated the performance of the kernel PCA approach (KPCA) and perturbation based method (Perturbation) by a grid of w as $0, 0.25, 0.5, 0.75$ and 1. We also evaluated the KMR models (Wu et al., 2011; Lee et al., 2012) that test only the genotype effect (IBS kernel) or the methylation effect (linear kernel). For all methods, 100,000 simulations were conducted to assess the type I error rate at different

α level. 2,000 simulations were conducted to evaluate the statistical power.

We simulated under the following model to evaluate the type I error:

$$y_i = X_i + \varepsilon_i, \tag{6}$$

where $X \sim N(0,1)$, $\varepsilon \sim N(0,1)$. We simulated $n = 500$ and 1000 individuals.

To evaluate the power, we selected 9 SNPs and 2 methylation markers that influence the phenotype. This corresponds to approximately 10% of the SNPs and CpG markers. We simulated two scenarios. In scenario Ia:

$$y_i = X_i + \beta_s \sum_{j \in J_1} G_{i,j} + \beta_m \sum_{j \in J_2} M_{i,j} + \varepsilon_i, \tag{7}$$

in which $G_{i,j}$ is the genotype of the j^{th} SNP in sample i; $M_{i,j}$ is the methylation value of the j^{th} CpG in sample i. $J_1 = \{10, 20, \cdots, 90\}$ and $J_2 = \{10, 20\}$ are the selected SNPs and CpG markers. When $\beta_s = 0$, there is no genetic effect; and when $\beta_m = 0$, there is no methylation effect.

In scenario Ib,

$$y_i = X_i + \beta \times [p \sum_{j \in J_1} G_{i,j} + (1-p) \sum_{j \in J_2} M_{i,j}] + \varepsilon_i, \tag{8}$$

in which $G_{i,j}, M_{i,j}, J_1$ and J_2 are all the same as in scenario Ia. p is a random binary variable that takes value 0 and 1 each with probability of 0.5. β controls the effect size. This simulation mimics real studies when it is unknown whether there is genotype or methylation effect.

2.4.2 Simulations when the genotype and methylation are correlated

We considered two simulation scenarios (scenarios II and III) to evaluate the performance of our method when genotype and methylation are correlated. In scenario II, we simulated data (genotype and methylation together) based on a 50-dimensional multivariate normal with mean 0 and variance Σ_2. Σ_2 has a block-like covariance structure, with three contiguous blocks of sizes 8, 23 and 19. All variables within the same block have a correlation of 0.1, and variables in different blocks have a correlation of 0.01. All simulated variables have variance of 1. Then we randomly selected 21 variables from this multivariate normal to form the methylation data. The remaining 29 variables were used to generate SNP data (0,1, or 2) by comparing the data values to their 36th and 84th percentiles (so that the minor allele frequency is 0.4). This simulation mimics the LD structure in genotype. In this simulation, every SNP is correlated with every CpG, therefore, the correlation between $\mathbf{G}$ and $\mathbf{M}$ is strong.

We simulated in the same way as in (6) to evaluate the type I error. For power, we similarly considered two scenarios:

$$\text{Scenario IIa:} \quad y_i = X_i + \beta_s \sum_{j \in J_1} G_{i,j} + \beta_m \sum_{j \in J_2} M_{i,j} + \varepsilon_i,$$

$$\text{.Scenario IIb:} \quad y_i = X_i + \beta \times [p \sum_{j \in J_1} G_{i,j} + (1-p) \sum_{j \in J_2} M_{i,j}] + \varepsilon_i,$$

in which J_1 and J_2 are randomly selected 3 SNPs and 2 CpG markers, which correspond to approximately 10% of the SNPs and 10% of the CpG markers. β_s, β_m, β and p are all the same as in scenarios Ia and Ib.

Simulation scenario III mimicked a more realistic correlation structure between genotype and methylation, and generated correlated genotype and methylation data via a methylation quantitative trait loci (mQTL). We simulated the genotype data for *ASAH1* the same way

as in scenario I. For methylation data, we first simulated in the same way as in scenario I as well. Then we considered the 29^{th} SNPs as an mQTL that can affect the methylation value of the 19^{th} CpG marker, and added $0.1 \times G_{29}$ to the originally simulated methylation value of the 19^{th} CpG. Type I error and power simulations were conducted the same way as in Scenario I (via equation (6), (7) and (8)).

2.5 Integrative Analysis of Grady Trauma Project

We applied our proposed framework to analyze the genetic and epigenetic regulation of gene expression (GE) using data collected from the Grady Trauma Project (GTP), a large study investigating the role of genetic and environmental factors in predicting response to stressful life events (Gillespie et al., 2009; Davis et al., 2008). Individuals were recruited from the waiting rooms of primary care and obstetrics-gynecology clinics of Grady Memorial Hospital in Atlanta, GA. Genotype, methylation and mRNA expression were measured using whole blood samples. We limited our analysis to 337 samples that have all three types of data (genotype, methylation and GE). The GE and methylation data are available at GEO (Gene Expression Omnibus) under the accessions GSE58137 (Peters et al., 2015) and GSE72680. Details about the genotyping can be found in Almli et al. (2014).

We based all our analysis on the gene level. SNPs that fall between 10kb upstream and 40kb downstream of gene transcription starting site were grouped into SNP-set. CpGs that are located within the same gene were grouped into CpG set. Methylation probes containing SNPs were removed from analysis due to the possibility of biased methylation measurement (Daca-Roszak et al., 2015; Zhi et al., 2013). Because the DNA was extracted from the whole blood, the proportions of different white blood cells may differ among different samples, which may confound the result. We therefore estimated the proportion of white blood cell types using the methylation data via the Houseman's method (Houseman et al., 2012), and adjusted for them in our model. We further adjusted for the age, sex and the top 5 PCs in

our model.

We applied the proposed composite kernel test to evaluate the joint genotype and methylation effect in cis-regulation of the GE, i.e., regulation of the expression of the gene itself. The composite kernel was constructed by a weighted average of an IBS kernel from genotype data and a linear kernel from methylation data, with the grid of the weighting parameter $w = (0, 0.25, 0.5, 0.75, 1)$. We used the kernel PCA and projected the methylation data to the genotype data for fast p-value calculation. For comparison, we also used the KMR to evaluate the association between genotype and GE and the association between methylation and GE.

For genes that showed significant cis-regulation, we further conducted a mediation model to infer the causal relationship between the genotype, methylation and GE. For simplicity, we adopt the causal steps model (Baron and Kenny, 1986; MacKinnon, 2008) for mediation analysis, and extend it to multi-dimensional "omic" data by incorporating multivariate KMR framework. This causal steps model assumed that methylation is a potential mediator of the genetic effect on the GE. Then we tested 1) whether genotype and methylation are associated and 2) whether methylation is associated with the phenotype conditional on the genotype data. When genotype, methylation and the phenotype are all univariate, these two conditions can be evaluated by testing whether a and b in Figure 1 are zero using (generalized) linear regression models. In our gene-based analysis framework, $\mathbf{G}$ and $\mathbf{M}$ are multi-dimensional, and the linear model approaches are no longer applicable. Instead, we utilize the KMR framework in which the effect of SNP/CpG-set are modeled nonparametrically. We used the multivariate kernel machine regression model (Maity et al., 2012) and the additive kernel machine regression model (Clark, 2013) to evaluate the associations in the two steps. Details about the statistical methods for evaluating each steps can be found in supplemental file section S2.

3 Results

3.1 Simulation results

The type I errors of the simulations when the data is simulated from the complete null (6) are summarized in table i and supplemental tables S1-S2 . Table i shows the type I error results when **G** and **M** are independent, and supplemental tables S1-S2 show the type I error results when **G** and **M** are correlated. All the methods showed correct type I error control at all the α levels that are tested on when **G** and **M** are independent. When **G** and **M** are correlated via an mQTL (scenario III), the type I errors are also well-controlled at the nominal α level. However, when **G** and **M** are strongly correlated (scenario II), the type I error can be conservative, especially when the sample size is not large enough.

Table ii shows the power result for simulation scenario Ia. Unsurprisingly, in the partial null that $\beta_m = 0, \beta_s \neq 0$, the model that only tests the SNP effect is the most powerful. The method that tests only the methylation effect has correctly controlled type I error, although it does not adjust for the genetic effect. Similar conclusions can be made when $\beta_m \neq 0, \beta_s = 0$. When there is both genetic and epigenetic effect, the proposed composite kernel approaches are more powerful than methods that test only **G** or **M**. In reality, information on the underlying genetic architecture is never known prior to analysis. The proposed composite kernel approaches are more robust than the methods based on only one type of data.

Table iii shows the power result for simulation scenario IIa. Similarly, when there is only genetic effect ($\beta_m = 0, \beta_s \neq 0$), methods that test only the genetic effect is the most powerful. However, unlike in simulation scenario Ia, method that tests only methylation effect has inflated type I error, even when there is no methylation effect. This is due to the correlation between the genotype and the methylation. Similar results hold for the simulation when $\beta_m \neq 0, \beta_s = 0$.

The power result for simulation scenarios Ib, IIb and IIIb was summarized in tables iv,

v and vii. Essentially, when it is unknown whether there is genetic or epigenetic effect, the proposed joint analysis approach is usually more powerful than methods that evaluate only one data type.

We then compared the power result between kernel PCA and the perturbation based approach. When $\mathbf{G}$ and $\mathbf{M}$ are independent, kernel PCA and the perturbation based procedure have similar power. However, when $\mathbf{G}$ and $\mathbf{M}$ are correlated, the perturbation approach showed consistent power gain compared to the kernel PCA approach. The power gain is more profound when the correlation between $\mathbf{G}$ and $\mathbf{M}$ is strong and when the methylation effect is large (Scenario II, table iii). The power loss is mild when the correlation between $\mathbf{G}$ and $\mathbf{M}$ is modest, such as in simulation scenario III (table vi).

The weighting parameter w determines the relative contribution of $\mathbf{G}$ and $\mathbf{M}$ into the score statistic. Our proposed approaches do not aim to directly estimate w. From the mixed model point of view, estimating the w is equivalent to estimating a variance component that disappears when the null hypothesis is true. Instead, our grid search selects the w based on which the composite kernel provides the minimum p-value. Note that the selected w is not the maximum likelihood estimate of the weighting parameter.

Table S3 lists the means and standard deviations (SD) of w that are selected in simulation scenario Ia using the perturbation based approach. When the null hypothesis ($\beta_s = \beta_m = 0$) is true, w is non-identifiable. Empirically, our method selects w to have a mean close to 0.5 and large SD in such situation. When there is no genetic ($\beta_s = 0$) or epigenetic effect ($\beta_m = 0$), our method selects w close to 0 or 1 respectively with small SD. When there are both genetic and epigenetic effects, the selected w is determined by the relative size of genetic and epigenetic effects.

3.2 GTP Integrative Analysis Results

We analyzed the data from the GTP study using our proposed framework. We focused on the cis-regulation of GE, i.e., whether the genotype and the methylation within the gene can regulate the expression of the same gene. We further limited our analysis to 8563 genes with 2 or more SNPs and CpG markers in the gene.

We used the kernel PCA and projected methylation to genotype for fast computation. We used Bonferroni corrected family error rate based on the total number of genes (8563) as cutoff for statistical significance. After adjusting for age, gender, top 5 PCs, and estimated white blood cell proportions, 732 genes showed significant genetic and epigenetic regulation at Bonferroni 0.05 level. We also applied the methods that test only the genetic or epigenetic effect on GE. 652 and 228 genes showed significant genetic and epigenetic effect on GE respectively. The proposed joint analysis provides higher power than methods that use only one data type. A venn-diagram (Figure 2) shows the relationship between these genes that showed significant GE regulation.

Table S4 lists the 11 genes that show significant genetic and epigenetic effect on GE regulation using our integrative analysis approach, but not by testing only **G** or **M** (after Bonferroni correction). Among these genes are *ALOX15* and *GSTM2*, both of which play important roles in response to stress. Genes *ALOX15* and *ALOX12* (a structurally and functionally similar gene to *ALOX15*) encode enzyme 12/15-lipoxygenase (12/15LOX), which is considered as "the central executioner in an oxidative stress-related neuronal death program" (Pallast et al., 2009). 12/15LOX sits in the major pathway through which post-traumatic stress disorder can lead to oxidative stress, which in turn causes neural damage (Miller et al., 2015). *GSTM2* belongs to a superfamily of genes that encodes glutathione S-transferase, a key element in detoxification of oxidative stress. *GSTM2* has been shown to play a prominent role in the etiology of many diseases, including hypertension (Zhou et al., 2008) and many cancers (Gorrini et al., 2013).

We conducted the casual steps mediation model to the 732 genes that showed significant joint genetic and epigenetic effect on GE. Of them, 277 genes showed significant genotype-methylation association (at nominal $\alpha = 0.05$ level), among which 94 genes showed significant methylation-GE association after controlling for the genetic effect, suggesting that methylation possibly mediates the genetic effect in regulating the GE of these genes.

Our causal mediation model assumes the causal diagram $G \to M \to GE$. Because our model is based on association testing for causal interpretation, we can not distinguish between this traditional causal relationship and the reverse causality $G \to GE \to M$. However, our assumption of the causal relationship is reasonable because the traditional causal relationship has much stronger evidence and applies to many more genes than the reverse causality (van Eijk et al., 2012). In addition, it is biologically unlikely that the GE can change the DNA methylation level of the same gene.

Among the 94 genes that are discovered in the GTP data are *BTN3A2*, *CTSW*, *CDC16* and *NAPRT1*, all of which have been shown to convey both eQTL (expression QTL) and mQTL (van Eijk et al., 2012; Wagner et al., 2014), i.e., the genetic variations in these genes have a cis-effect on the methylation and GE level. Previous research has shown strong evidence of the causal relationship $G \to M \to GE$ for genes *BTN3A2*, *CTSW*, *CDC16* and *NAPRT1* (van Eijk et al., 2012). In fact, we have replicated most of the previous findings (4 out of 6) that showed a strong evidence of this causal relationship (van Eijk et al., 2012).

4 Discussion

In this paper, we propose a statistical framework for integrative analysis of genome-wide methylation and genotype data. We propose to test the joint genetic/epigenetic effect via a flexible, semiparametric KMR by constructing composite kernels. This approach unifies the units in analyzing genotype and methylation data, and allows for easy interpretation of results. The composite kernel, which is a weighted average of genotype and methylation

specific kernels, is flexible to incorporate SNP-specific and methylation-specific kernels.

This composite kernel approach, albeit very flexible, assumes additive genetic and epigenetic effect, and does not allow for genotype-methylation interaction. Incorporating additional G-M interaction is possible by constructing a composite kernel as $\mathbf{K} = w_1\mathbf{K}_1 + w_2\mathbf{K}_2 + (1 - w_1 - w_2)\mathbf{K}_1 \circ \mathbf{K}_2$, where $\circ$ represents element-wise product and $\mathbf{K}_1 \circ \mathbf{K}_2$ models the **G**-**M** interaction, similar to the method in testing gene-gene interaction (Larson and Schaid, 2013). Statistical significance may be evaluated by choosing different values of w_1 and w_2. This, however, can be very computationally expensive for perturbation-based approach due to the large number of combinations that w_1 and w_2 can be. For the kernel PCA approach, this implies an additional orthogonalization and a more-complex two-degree integration to obtain the p-value. In addition to the computational difficulties, understanding the causal relationship when **G** and **M** interact is very challenging, or even impossible. In causal inference literature, there are methods that use the counterfactual definition in evaluating the mediation effect with interactive effect. The approach, however, relies on assumptions that are not easy to evaluate and can not be easily adapted to our situation when the mediator (methylation) is multidimensional. The causal steps model can only evaluate the causal relationship assuming there is no interaction.

Of the two methods that are proposed in this paper, the perturbation based approach is more powerful than the kernel PCA, especially when the **G** and **M** have high correlation. This is due to the fact that the kernel PCA loses information when one data type is projected on the other one. Nevertheless, we show, via simulations and theoretical justification, that kernel PCA can have comparable power to the perturbation based approach in many realistic situations, such as when **G** and **M** are independent or have only modest correlation, or when the methylation effect is small (if we project the methylation data onto the genotype data). In addition, kernel PCA is much faster in computation compared to the perturbation based approach, especially when people are interested in very small α levels, such as in genome

wide association studies, in which a large number of perturbation is required. We show in Figure S2 the comparison of computation time needed for the kernel PCA and perturbation approach at different sample sizes.

We considered a multivariate causal steps model for mediation in our analysis of the GTP data set. Our multivariate causal steps model extends the classical univariate causal steps model (Baron and Kenny, 1986; Judd and Kenny, 1981) in social psychological studies to incorporate multidimensional mediator and multidimensional independent variables. Critical to the mediation analysis is the correct specification of the causal diagram prior to analysis. Further assumptions include 1) no unmeasured confounding of the genotype-methylation relationship, 2) no unmeasured confounding of the genotype-phenotype relationship, 3) no unmeasured confounding of the methylation-phenotype relationship, and 4) no genotype-induced confounder for the methylation-phenotype association. In usual genomic studies, because of the random assortment of genetic alleles, the first two assumptions hold automatically. Potential confounding of the methylation-phenotype should be carefully studied and adjusted in the model.

Although the specific steps in our causal mediation model, i.e., the multivariate kernel machine regression model for testing the genetic effect on methylation and the additive kernel machine regression model for testing the epigenetic effect conditional on the genetic data, are adopted from previous research, we bring these association testing procedures into a novel multivariate causal mediation framework in which $\mathbf{M}$ is considered as a multivariate mediator. Our research extends the classical causal steps model (Baron and Kenny, 1986; Judd and Kenny, 1981) to incorporate multi-dimensional mediator.

It is noteworthy that the proposed mediation analysis result should be taken with caution, especially in the context of disjoint effect or inconsistent mediation (Huang and Pan, 2016). For example, with two exposures (G_1 and G_2) and two mediators (M_1 and M_2), it is possible that $G_1 \to M_1$ and $M_2 \to Y$. In this case of disjoint effect, the proposed analysis approach

will capture this as "mediation effect", while in fact there is no real causal effect between the genotype (the exposures) and the phenotype. Further, the proposed mediation model only tests for marginal mediation effect, i.e., the overall effect mediated through the group of methylation markers regardless through which element of $\mathbf{M}$. In reality, element-wise mediation effect may cancel each other out, resulting in a non-significant marginal mediation effect. For example, it is possible that G_1 increases the phenotype through M_1 but decreases the phenotype through M_2. In this case of inconsistent mediation, we may observe no marginal effect despite true mediation effect. Therefore, causal interpretation for such high-dimensional mediator model needs to be cautious (Huang and Pan, 2016).

Our proposed approaches are motivated by the need to integrate genotype and methylation data from the same set of samples, which are increasingly collected in GWAS. Although the composite kernel and the perturbation based approach have been studied previously (Wu et al., 2013), they provide a natural way for joint analysis of such data, and lead to powerful tests with interpretable result. The specific steps in our causal mediation model are adopted from previous research, however, we bring them into a novel framework in which the methylation data is tested as a multivariate mediator.

In summary, our proposed gene-based analysis provides a general framework for assessing the relationship between genotype, methylation and a phenotype of interest, which can be easily extended to analyze other types of data. The composite kernel approach is directly applicable to analyze the joint effect of different data types, such as common and rare variants, GE, miRNA, gene environment interactions, etc. The mediation model may be applied to studies when the potential mediator is multi-dimensional, such as the gene expressions for multiple genes in a pathway. The method, albeit preliminary, provides some insights to understanding the potential causal relationships and can be helpful in our better understanding the biological processes, and identifying novel targets for clinical practice and prevention.

5 Figures and Tables

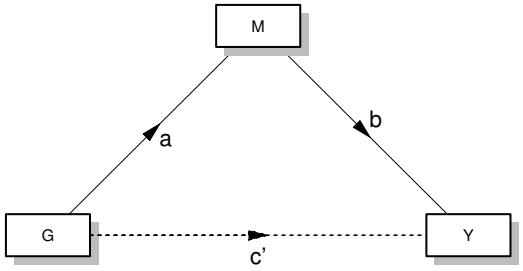

Figure 1: Mediation diagram. The causal steps model considers the genotype (**G**) as the independent exposure, whose effect on the phenotype Y may be mediated through methylation (**M**). The path a and b in the figure represent the indirect path, and c' represents the direct effect.

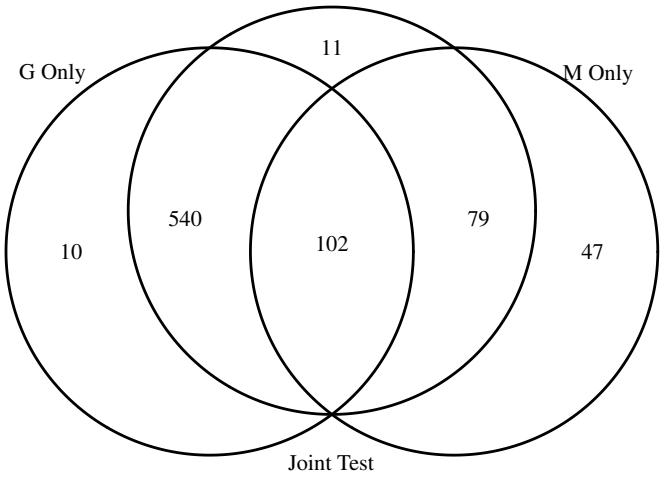

Figure 2: Venn Diagram shows the number of genes that show significant association with GE using 1) joint genotype-methylation test 2) genotype only test and 3) methylation only test.

Table i: Empirical Type I Error Rate at different α levels when $\mathbf{G}$ and $\mathbf{M}$ are independent.

α	n	KPCA	Perturbation	G	M
0.05	500	0.0482	0.0491	0.0495	0.0496
	1000	0.0481	0.0498	0.0495	0.0503
0.01	500	0.0096	0.0098	0.0096	0.0099
	1000	0.0093	0.0096	0.0097	0.0095
0.001	500	0.00093	0.00099	0.00095	0.00097
	1000	0.00094	0.00099	0.00099	0.00096

Table ii: Empirical Power for Simulation Scenario Ia

n	β_s	β_m	KPCA	Perturbation	G	M
500	0	0.1	0.3235	0.3585	0.0485	0.4440
	0.1	0	0.4245	0.4390	0.5585	0.0480
	0.1	0.1	0.6070	0.6245	0.5235	0.4160
	0.2	0.2	1.0000	0.9995	0.9935	0.9635
1000	0	0.1	0.6305	0.6435	0.0465	0.7535
	0.1	0	0.7915	0.8000	0.8770	0.0435
	0.05	0.05	0.3185	0.3280	0.2680	0.2400
	0.1	0.1	0.9615	0.9620	0.8845	0.7605

Table iii: Empirical Power for Simulation Scenario IIa

n	β_s	β_m	KPCA	Perturbation	G	M
500	0	0.1	0.2505	0.3205	0.0605	0.4150
	0.2	0	0.7930	0.8125	0.8710	0.1390
	0.05	0.05	0.1125	0.1405	0.0975	0.1510
	0.1	0.1	0.4395	0.5705	0.3405	0.5535
1000	0	0.1	0.6490	0.7100	0.1010	0.7970
	0.1	0	0.4155	0.4465	0.5155	0.0955
	0.05	0.05	0.2270	0.2865	0.1760	0.2885
	0.1	0.1	0.8855	0.9430	0.7235	0.8990

Table iv: Empirical Power for Simulation Scenario Ib

n	β	KPCA	Perturbation	G	M
500	0.1	0.6300	0.6535	0.3950	0.3650
	0.3	0.9725	0.9755	0.5080	0.5225
1000	0.1	0.8125	0.8210	0.4805	0.4285
	0.2	0.9695	0.9725	0.5360	0.4910

Table v: Empirical Power for Simulation Scenario IIb

n	β	KPCA	Perturbation	G	M
500	0.2	0.8600	0.8845	0.4960	0.5645
	0.3	0.9980	0.9980	0.6525	0.6140
1000	0.1	0.5145	0.5685	0.3105	0.4235
	0.2	0.9995	0.9995	0.6505	0.6215

Table vi: Empirical Power for Simulation Scenario IIIa

n	β_s	β_m	KPCA	Perturbation	G	M
500	0	0.1	0.3265	0.3520	0.0435	0.4545
	0.1	0	0.4185	0.4230	0.5390	0.0440
	0.1	0.1	0.6345	0.6540	0.5405	0.4680
	0.2	0.2	0.9990	0.9995	0.9920	0.9685
1000	0	0.1	0.6530	0.6690	0.0565	0.7780
	0.1	0	0.8315	0.8345	0.9110	0.0405
	0.05	0.05	0.3285	0.3370	0.2755	0.2370
	0.1	0.1	0.9555	0.9540	0.8670	0.7690

Table vii: Empirical Power for Simulation Scenario IIIb

n	β	KPCA	Perturbation	G	M
500	0.1	0.3725	0.3940	0.2840	0.2630
	0.2	0.9530	0.9545	0.5255	0.4915
1000	0.1	0.7455	0.7560	0.4680	0.4165
	0.2	1.0000	1.0000	0.5315	0.5200

References

R. S. Alisch, B. G. Barwick, P. Chopra, L. K. Myrick, G. A. Satten, K. N. Conneely, and S. T. Warren. Age-associated dna methylation in pediatric populations. *Genome Res*, 22 (4):623–32, 2012.

L. M. Almli, R. Duncan, H. Feng, D. Ghosh, E. B. Binder, B. Bradley, K. J. Ressler, K. N. Conneely, and M. P. Epstein. Correcting systematic inflation in genetic association tests that consider interaction effects: application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiatry*, 71(12):1392–1399, Dec 2014.

D. Altschuler, L. Brooks, A. Chakravarti, F. Collins, M. Daly, P Donnelly, and International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063): 1299–320, 2005.

R. M. Baron and D. A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 51(6):1173–82, 1986.

J. J. Clark. *Estimation and Hypothesis Testing with Additive Kernel Machines for High-Dimensional Data*. PhD thesis, University of North Carolina at Chapel Hill, 2013.

N Cristianini and J Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel Based Learning Method*. Cambridge Univerisity Press, Cambridge, UK, 2000.

P. Daca-Roszak, A. Pfeifer, J. Zebracka-Gala, D. Rusinek, A. Szybinska, B. Jarzab, M. Witt, and E. Zietkiewicz. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies. *BMC Genomics*, 16:1003, Nov 2015.

R. B. Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3): 323–333, 1980.

R. G. Davis, K. J. Ressler, A. C. Schwartz, K. J. Stephens, and R. G. Bradley. Treatment barriers for low-income, urban African Americans with undiagnosed posttraumatic stress disorder. *J Trauma Stress*, 21(2):218–222, Apr 2008.

C. F. Gillespie, B. Bradley, K. Mercer, A. K. Smith, K. Conneely, M. Gapen, T. Weiss, A. C. Schwartz, J. F. Cubells, and K. J. Ressler. Trauma exposure and stress-related disorders in inner city primary care patients. *Gen Hosp Psychiatry*, 31(6):505–514, 2009.

C. Gorrini, I. S. Harris, and T. W. Mak. Modulation of oxidative stress as an anticancer strategy. *Nat Rev Drug Discov*, 12(12):931–947, Dec 2013.

X. Hong, K. Hao, C. Ladd-Acosta, K. D. Hansen, H. J. Tsai, X. Liu, X. Xu, T. A. Thornton, D. Caruso, C. A. Keet, Y. Sun, G. Wang, W. Luo, R. Kumar, R. Fuleihan, A. M. Singh, J. S. Kim, R. E. Story, R. S. Gupta, P. Gao, Z. Chen, S. O. Walker, T. R. Bartell, T. H. Beaty, M. D. Fallin, R. Schleimer, P. G. Holt, K. C. Nadeau, R. A. Wood, J. A. Pongracic, D. E. Weeks, and X. Wang. Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nat Commun*, 6: 6304, 2015.

E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012.

Y. T. Huang and W. C. Pan. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413, Jun 2016.

T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, and M. et al. Guyer. International network of cancer genome projects. *Nature*, 464(7291):993–998, Apr 2010.

I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin. Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *Am. J. Hum. Genet.*, May 2013.

F. Johannes, V. Colot, and R. C. Jansen. Epigenome dynamics: a quantitative genetics perspective. *Nat. Rev. Genet.*, 9(11):883–890, Nov 2008.

F. Johannes, E. Porcher, F. K. Teixeira, V. Saliba-Colombani, M. Simon, N. Agier, A. Bulski, J. Albuisson, F. Heredia, P. Audigier, D. Bouchez, C. Dillmann, P. Guerche, F. Hospital, and V. Colot. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.*, 5(6):e1000530, Jun 2009.

C. M. Judd and D. A. Kenny. Process Analysis. *Evaluation Review*, 5(5):602–619, October 1981. doi: 10.1177/0193841x8100500502. URL `http://dx.doi.org/10.1177/0193841x8100500502`.

H. Kirchner, M. E. Osler, A. Krook, and J. R. Zierath. Epigenetic flexibility in metabolic regulation: disease cause and prevention? *Trends Cell Biol.*, 23(5):203–209, May 2013.

M. Kulis and M. Esteller. DNA methylation and cancer. *Adv. Genet.*, 70:27–56, 2010.

L. C. Kwee, D. Liu, X. Lin, D. Ghosh, and M. P. Epstein. A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, 82(2):386–397, Feb 2008.

N. B. Larson and D. J. Schaid. A kernel regression approach to gene-gene interaction detection for case-control studies. *Genet. Epidemiol.*, 37(7):695–703, Nov 2013.

S. Lee, M. C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, Sep 2012.

C. M. Li, J. H. Park, X. He, B. Levy, F. Chen, K. Arai, D. A. Adler, C. M. Disteche, J. Koch, K. Sandhoff, and E. H. Schuchman. The human acid ceramidase gene (ASAH): structure, chromosomal location, mutation analysis, and expression. *Genomics*, 62(2):223–31, 1999.

X. Lin, T. Cai, M. C. Wu, Q. Zhou, G. Liu, D. C. Christiani, and X. Lin. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genet Epidemiol*, 35(7):620–31, 2011.

D. Liu, X. Lin, and D. Ghosh. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63(4): 1079–88, 2007.

D. Liu, D. Ghosh, and X. Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9:292, 2008.

H. Liu, Y. Tang, and H H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

Y. Liu, M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, K. Shchetynsky, A. Scheynius, J. Kere, L. Alfredsson, L. Klareskog, T. J. Ekstrom, and A. P. Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, 31(2):142–7, 2013.

L. Luzzatto and P. P. Pandolfi. Causality and Chance in the Development of Cancer. *N. Engl. J. Med.*, 373(1):84–88, Jul 2015.

D. P. MacKinnon. *Introduction to statistical mediation analysis*. Erlbaum, Mahwah, NJ, 2008.

D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz. Mediation analysis. *Annu Rev Psychol*, 58:593–614, 2007.

B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, Nov 2008.

A. Maity, P. F. Sullivan, and J. Y. Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol*, 36(7):686–95, 2012.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009.

T. R. Mayo, G. Schweikert, and G. Sanguinetti. M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, 31(6):809–816, Mar 2015.

M. W. Miller, E. J. Wolf, N. Sadeh, M. Logue, J. M. Spielberg, J. P. Hayes, E. Sperbeck, S. A. Schichman, A. Stone, W. C. Carter, D. E. Humphries, W. Milberg, and R. McGlinchey. A novel locus in the oxidative stress-related gene ALOX12 moderates the association between PTSD and thickness of the prefrontal cortex. *Psychoneuroendocrinology*, 62:359–365, Dec 2015.

S. Pallast, K. Arai, X. Wang, E. H. Lo, and K. van Leyen. 12/15-Lipoxygenase targets neuronal mitochondria under oxidative stress. *J. Neurochem.*, 111(3):882–889, Nov 2009.

N.D. Pearce and M.P. Wand. Penalized splines and reproducing kernel methods. *The American Statistician*, 60:233–240, 2006.

M. J. Peters, R. Joehanes, L. C. Pilling, C. Schurmann, K. N. Conneely, J. Powell, E. Reinmaa, G. L. Sutphin, A. Zhernakova, K. Schramm, Y. A. Wilson, S. Kobes, T. Tukiainen, Y. F. Ramos, H. H. Goring, M. Fornage, Y. Liu, S. A. Gharib, B. E. Stranger, P. L. De Jager, A. Aviv, D. Levy, J. M. Murabito, P. J. Munson, T. Huan, A. Hofman, A. G. Uitterlinden, F. Rivadeneira, J. van Rooij, L. Stolk, L. Broer, M. M. Verbiest, M. Jhamai, P. Arp, A. Metspalu, L. Tserel, L. Milani, N. J. Samani, P. Peterson, S. Kasela, V. Codd, A. Peters, C. K. Ward-Caviness, C. Herder, M. Waldenberger, M. Roden, P. Singmann, S. Zeilinger, T. Illig, G. Homuth, H. J. Grabe, H. Volzke, L. Steil, T. Kocher, A. Murray, D. Melzer, H. Yaghootkar, S. Bandinelli, E. K. Moses, J. W. Kent, J. E. Curran, M. P. Johnson, S. Williams-Blangero, H. J. Westra, A. F. McRae, J. A. Smith, S. L. Kardia, I. Hovatta, M. Perola, S. Ripatti, V. Salomaa, A. K. Henders, N. G. Martin, A. K. Smith, D. Mehta, E. B. Binder, K. M. Nylocks, E. M. Kennedy, T. Klengel, J. Ding, A. M. Suchy-Dicey, D. A. Enquobahrie, J. Brody, J. I. Rotter, Y. D. Chen, J. Houwing-Duistermaat, M. Kloppenburg, P. E. Slagboom, Q. Helmer, W. den Hollander, S. Bean, T. Raj, N. Bakhshi, Q. P. Wang, L. J. Oyston, B. M. Psaty, R. P. Tracy, G. W. Montgomery, S. T. Turner, J. Blangero, I. Meulenbelt, K. J. Ressler, J. Yang, L. Franke, J. Kettunen, P. M. Visscher, G. G. Neely, R. Korstanje, R. L. Hanson, H. Prokisch, L. Ferrucci, T. Esko, A. Teumer, J. B. van Meurs, A. D. Johnson, M. A. Nalls, D. G. Hernandez, M. R. Cookson, R. J. Gibbs, J. Hardy, A. Ramasamy, A. B. Zonderman, A. Dillman, B. Traynor, C. Smith, D. L. Longo, D. Trabzuni, J. Troncoso, M. van der Brug, M. E. Weale, R. O'Brien, R. Johnson, R. Walker, R. H. Zielke, S. Arepalli, M. Ryten, and A. B. Singleton. The transcriptional landscape of age in human peripheral blood. *Nat Commun*, 6:8570, 2015.

Z. Su, J. Marchini, and P. Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–5, 2011.

The Cancer Genome Atlas Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, 2008.

The Cancer Genome Atlas Consortium. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–25, 2012.

K. R. van Eijk, S. de Jong, M. P. Boks, T. Langeveld, F. Colas, J. H. Veldink, C. G. de Kovel, E. Janson, E. Strengman, P. Langfelder, R. S. Kahn, L. H. van den Berg, S. Horvath, and R. A. Ophoff. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, 13:636, Nov 2012.

P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *Am. J. Hum. Genet.*, 90(1):7–24, Jan 2012.

J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.*, 15(2):R37, Feb 2014.

M. C. Wu, P. Kraft, M. P. Epstein, D M. Taylor, S J. Chanock, D J. Hunter, and X. Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89(1): 82–93, Jul 2011.

M. C. Wu, A. Maity, S. Lee, E. M. Simmons, Q. E. Harmon, X. Lin, S. M. Engel, J. J.

Molldrem, and P. M. Armistead. Kernel machine SNP-set testing under multiple candidate kernels. *Genet. Epidemiol.*, 37(3):267–275, Apr 2013.

S. Zhang, C. C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*, 40(19):9379–91, 2012.

D. Zhi, S. Aslibekyan, M. R. Irvin, S. A. Claas, I. B. Borecki, J. M. Ordovas, D. M. Absher, and D. K. Arnett. SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics*, 8(8):802–806, Aug 2013.

S. G. Zhou, P. Wang, R. B. Pi, J. Gao, J. J. Fu, J. Fang, J. Qin, H. J. Zhang, R. F. Li, S. R. Chen, F. T. Tang, and P. Q. Liu. Reduced expression of GSTM2 and increased oxidative stress in spontaneously hypertensive rat. *Mol. Cell. Biochem.*, 309(1-2):99–107, Feb 2008.