KERNEL MACHINE METHODS FOR ANALYSIS OF GENOMIC DATA FROM DIFFERENT SOURCES

Ni Zhao

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2014

Approved by:

Michael C. Wu

D Neil Hayes

Eric T Tchetgen

Wei Sun

Yun Li

ABSTRACT

NI ZHAO: Kernel Machine Methods for Analysis of Genomic Data from Different Sources
(Under the direction of Michael C. Wu)

Comprehensive understanding of complex trait etiology requires examination of multiple sources of genomic variability. Recent advances in high-throughput biotechnology, especially sequencing technology, have enabled multiple platform genomic profiling of biological samples. In this dissertation, we consider using the kernel machine regression (KMR) framework to analyze data from different genetic data sources.

In the first part of this dissertation, we develop a new strategy for identification of large scale, global changes in methylation that are associated with environmental variables or clinical outcomes via a functional regression approach. The density or the cumulative distribution function of the methylation values for each individual can be approximated using B-spline basis functions with the spline coefficients to summarize the individual's overall methylation profile. A variance component score test is proposed to test for association between the overall distribution and a continuous or dichotomous outcome and applied to two real studies.

In the second part, we construct a microbiome regression-based kernel association test (MiRKAT) for testing the association between microbial community profiles and a continuous or dichotomous variable of interest such as an environmental exposure or disease status. This method regresses the outcome on the covariates (including potential confounders) and the microbiome compositional profiles through kernel functions. We demonstrate the improved control of type I error and superior power of MiRKAT compared to existing methods through simulations and real studies.

In the final part, we focus on integrative analysis of genome wide association studies (GWAS) and methylation studies. We propose to use the KMR for first testing the cumulative genetic/epigenetic effect on a trait and for subsequent mediation analysis to understand the mechanisms by which the genomic data influence the trait. In particular, we develop an approach that works at the gene level (to allow for a common analysis unit across data types). We compare pair-wise similarity in trait values between individuals to pair-wise similarity in methylation and genotype values, with correspondence suggestive of association. For a significant gene, we develop a causal steps approach to mediation analysis which enables elucidation of the manner in which the different data types work, or do not work, together.

ACKNOWLEDGMENTS

My most sincere gratitude goes to my dissertation advisor, Dr Michael Wu, for his guidance, understanding, patience, and most importantly, friendship during my graduate studies. He encouraged and taught me to not only grow as a statistician but also as an independent researcher. I should be thankful not only for his close guidance and careful training through all of my dissertation topics , but also for his high spirit and enthusiasm in academic studies. He managed to make my dissertation experience both inspiring and enjoyable that I just feel that I have not learnt enough from him yet.

Dr. Neil Hayes, among all my committee members, I had the longest working relationship with him. I started working with him even before I joined the Department of Biostatistics. I am grateful to work with him and get exposed to the cutting-edge techniques and involved in some pioneering studies in cancer genomics. By setting an great example as a world-class researcher, I learned enormously from him, not only about cancer or genomics, but also about being a scientific researcher in the first place.

I was extremely delighted to work with Dr. Wei Sun for my master thesis, who has been a motivating and encouraging mentor. The work with him has been an unique experience which introduced me to different ideas and methods, both in the field of genetics and statistics. I owe great gratitude to him for all the help he can possibly offer.

I am greatly thankful to Dr. Yun Li for her encouragement and help when I encounter problems in my research. She has the warmest and most welcoming smile. I personally consider her as a perfect combination of statistician and geneticist, from whom I can learn a lot.

I'd like to show my gratitude to Dr. Eric Tchetgen for reading a my dissertation draft, offering special comments and suggestions, especially with respect to the causal mediation analysis. I remembered the time when he kindly went through the effort the derivation and proofs to help me evaluate one of our hypothesis when I went to Boston to ask for help. I sincerely hope that I had more opportunities to work with him.

I would also like to thank my fellow students in the Department of Biostatistics, for the friendship and the fun time we spend together, and all the members in the Neil Hayes research group, especially Michele and Ashley, for being a welcoming academic family and offering help whenever asked.

Finally and most importantly, I would like to thank my husband Xin Huang for all these years of understanding, encouragement and unwavering love. He has been instrumental in instilling confidence in me, constitutes my source of comfort and courage, even in the toughest time. I thank my parents for their faith in me in all of my endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Complex diseases, such as cancer, cardiovascular disease, diabetes and Alzheimer's disease, which constitute the greatest public health burden both nationwide and globally, are considered to be caused by modest effects of multiple genes, interacting with environmental and lifestyle factors. Comprehensive understanding of these complex trait etiology requires examination of multiple sources of genomic variability. Recent advances in high-throughput biotechnology, especially sequencing technology, have enabled multiple platform genomic profiling of biological samples, which can facilitate the characterization of biological systems at multiple levels. For example, the Cancer Genome Atlas (TCGA) project aims to generate comprehensive catalog of the genomic changes of different cancers, including single nucleotide polymorphism (SNP), DNA methylation, gene expression, microRNA expression, for the same set of tumor samples (147, 148, 150). Similarly, the NCI60 project has profiled 60 human cancer cell lines with respect to gene expression, protein expression, microRNA expression and drug responses (221, 20, 178, 195, 182). Integrative analysis of these data sources promises elucidation of the biological processes underlying particular phenotypes. Integrative analysis of "multi-dimensional genomic data" has proven especially challenging. Typical analyses of large scale genomic data that examine each feature individually with subsequent correction for multiple comparisons were problematic(198, 232). First, individual feature analysis is usually underpowered due to the large number of multiple

corrections and the relatively smaller effect size in individual features. Further, difficulty aries in interpretation and formation of biological hypothesis when too many features are called significant. Finally, this method fails to capture the multi-feature/interative effect and usually have poor reproducibility (223, 142).

To overcome many of these limitations, analysis that associate grouped features with outcome has gained popularity during the recent years. For example, in Genome Wide Association Studies (GWAS), multiple-SNP based analyses, in which multiple related SNPs (by proximity to a gene, pathway or functional groups) are combined into SNP set and jointly analyzed for association with outcomes of interest, have emerged as a powerful alternative for identifying associations between multiple gene variants and complex disease. Investigating cumulative effect of multiple related features (e.g genes in a pathway, SNPs in a region or CpGs in a gene) across different platforms has also become a ubiquitous strategy in different complex diseases(218, 94, 176, 240). One particularly popular strategy in the multiple-feature association study is the kernel machine regression (KMR) test, which was initially proposed for gene expression data with continuous or binary phenotypes(112, 111), but has also been extended to candidate gene studies (96),case-control GWAS studies to test for SNP-set effect (229, 187, 139), rare variants studies(230).

This approach is built upon a semi-parametric model within the kernel machine regression framework (40) where the genomic effect can be modeled nonparametrically with simple confounding factors modeled parametrically. Intuitively, this approach constructs a pairwise similarity matrix between genetic measurement through the use of a kernel function, which can then be compared to the similarity between the phenotype of interest with high correspondence suggestive of association. Inference can be conducted through the variance score test, which is operationally simple and fast as it requires model estimation only under the null hypothesis.

The dissertation is organized as follows. In Chapter 2, we review current literature on analysis of multi-dimensional genetic data, with a focus on the KMR framework and identify unsolved problems. In Chapter 3, we develop two related methodologies under the KMR framework for identification of large scale, global methylation changes that are associated with environmental variables, clinical outcomes or other experimental condition. In Chapter 4, we extend this framework to the field of metagenomic studies and develop methods for association between microbial composition and outcomes of interest. In Chapter 5, we develop method to use the powerful kernel machine framework for first testing the cumulative effect of both epigenetic and genetic variability on a trait, and for subsequent mediation analysis to understand the mechanisms by which the genomic data types influence the trait.

CHAPTER 2

Literature Review

2.1 Multidimensional Genetic Data

Genetic research has undergone a dramatic transformation in the past decade because improved technology and reduced cost enabled collection of genetic data at multiple levels. Several large scale studies have collected multidimensional genomic data, including but not limited to whole genome gene expression, genotyping, copy numbers and rare variants; and have demonstrated the great potential of integrative analysis in discovering the complex and interrelated biological foundation underlying disease phenotypes. While multidimensional genomic studies are gaining increasing popularity, the methodology to perform the analysis has not kept pace with the collection of the data.

In this section, we will review the commonly used methods in analyzing different types of genomic studies and defer the KMR framework to Section 2.2

2.1.1 Genome Wide Association Study and Missing Heritability

Proposed almost 20 years ago as a potentially powerful approach to unravel the genetic basis of complex diseases (163), genome-wide association studies (GWAS) have become one of the most common tools for investigating the genetic architecture of human diseases. The rationale underlying GWAS is the "common disease, common

variant", hypothesizing that complex diseases are at least partially attributable to common variants present in more than $1-5\%$ of human population (35, 161, 157). Facilitated by the commercially available "SNP chips" which capture most, although not all, common variants in the genome, GWAS aim to detect association between common variants (especially in SNPs) and disease phenotypes. GWAS have reported hundreds of SNPs that are robustly associated with common phenotypes (133), some of which the biological basis have successfully been elucidated (67, 48, 92) or have shown clinical importance towards personalized medicine (37).

Typically, SNPs discovered by GWAS confer relatively small increments in risk and can together account for only a small fraction of the genetic variation of complex traits in human populations, leading to the perceived problem of missing heritability (130, 134). A number of explanations have been suggested for this missing heritability, including the existence of unmodeled epistatic interactions, the effect of rare variants (157) and inherited epigenetic factors (83, 84).

Current GWAS usually test association between SNPs with a phenotypic trait, one at a time, with stringent genome-wide adjustment for multiple testing. This procedure can be underpowered due to a number of reasons. First, the effect size of individual SNPs can too small to reach the genome-wide significance in GWAS(101). Secondly, the true causal variants are rarely genotyped in practice and detection of association rely on the linkage disequilibrium (LD) between genotyped SNPs and the causal variant. If the LD was not sufficient between the causal variant and SNPs that are genotyped on the GWAS platform, the power of detecting the true association will be reduced (101). This can also cause the problem of poor reproducibility: many of the highly ranked SNPs in the discovery phase of GWAS are false positives and cannot be validated because the estimated association is not for the true causal SNP but the genotyped surrogate markers.

Methods that consider joint effect of multiple SNPs simultaneously can be advantageous because it reduces the total number of tests (hence the number of multiple comparisons) and approximates the causal effect more effectively than could single SNP analysis (175). Moreover, the individual SNP approach considers only the marginal effect of each SNP and fails to accommodate epistatic interaction effect between SNPs (57), which have been shown to be ubiquitous to a number of common human diseases (142), including type I/II diabetes (202, 39), inflammatory bowel disease (33) and Alzheimer's diseases (23, 36, 36). Testing for the epistatic effect e.g. gene-gene interaction, is generally challenging because of the large number of potential interactions (78). Alternative approaches that use prior biological information to form SNP set and test for association between the SNP set and phenotypic traits are successful in improving power and increase the heritability estimates (235, 63).

Rare genetic variants, alleles with a frequency less than $1-5\%$ but potentially higher penetrance, can be essential in influencing complex disease and constitute another source of the missing heritability (179). Because of the relatively lower frequency, rare variants are less likely to be captured by the conventional genotype platforms used in GWAS. The advent of new sequencing technique (135) offers unprecedented opportunities for assessing the contribution of rare genetic variation to complex diseases (50). Standard methods that test for association with single markers are no longer applicable unless the sample size and/or the effect size are extremly large (102, 128). Methods analyzing rare variants involve testing the grouped/cumulative effect for a set of markers across a genomic region, including the burden test and its derivatives (128, 102, 143, 144) and the KMR which will be reviewed in Section 2.2.

2.1.2 Gene Expression Data

Gene expression determines a variety of cellular phenotypes. Gene expression profiling, which measures the activity/expression of thousand of genes simultaneously, has been a routine practice in genetic studies since the microarray technology (212). More recent technologies such as high-throughput RNA sequencing enables not only more accurate determination of gene expression level (145), but structural variations, such as allele-specific expression (167).

The primary goal of many gene expression studies is to identify genes that are differentially expressed under two or more treatment conditions. Traditionally, differential expression was assessed through fold change, t-test or ANOVA one gene at a time, with adjustment for the effects of multiple comparisons using criterions such as Bonferroni correction, false discovery rate or family wise error rate (49, 205). These studies, however successful, have major limitations, including poor reproducibility across studies and lack of interpretability because of the long list of single significant genes. Studies on gene sets, such as genetic pathways (64, 198) have also been very popular. Pathway analysis relies on existing functional annotation and looks for over-representation of functional classes in gene expression, which can be more biologically interpretable and reproducible.

Pattern discovery and class prediction (189) are another two important aspects of gene expression analysis, both of which provide a high-level overview of the data set and aim at forming related subgroups which can capture the biological difference. These two methods approach the phenotype classification differently. Pattern discovery is an unsupervised learning process. It searches for a biologically relevant unknown classes based on gene expression signature using dimension reduction tools, such as singular value decomposition, as well as various clustering techniques. Class prediction (supervised learning), on the other hand, are designed to classify subjects into known

groups, which usually involves a training phase on samples with known class labels and a testing phase, in which the algorithm applies criterions obtained from the training data to predict class labels for the testing samples.

2.1.3 Metagenomics: Genomic Analysis of Microbial Communities

Metagenomics concerns with the genomic study of uncultured microbial community. In metagenomics studies, DNA are collectively sampled from the microorganisms from environment of interest (e.g. agricultural soil, ocean water, or the human gut). The extracted DNA are then sequenced and used to investigate different aspect of the microbial community, such as bio-diversity, dominant microbial classes, biological functions and its effect on human health.

Metagenomic analysis has many distinct features from other genomic analysis. First, the research questions on the metagnomic field are often at the level of microbial communities, within which the organisms and evolutionary relationships are not known. Data preprocessing is usually required before analysis. Different sampling and filtering approaches exist that aim to get the DNA of microorganisms that are of interest while leaving out contaminations that are not of interest. Assembly, binning and annotation are needed to construct operational taxonomic unit (OTU), i.e., species distinction in microbiology (228, 81). Secondly, the sampled sequence data is usually zero-inflated, fragmented and pooled, which are statistically challenging in analysis.

One of the most important aspects of metagenomic studies is to study how a bacterial community be affected by or affects its habitat or host, including the microenvironment within human body. In human studies, microbial composition has been associated with age, gender, BMI, diet and a number of clinical symptoms (204). Distance based analysis is one popular strategy in evaluating the association between microbiota composition and outcomes of interest, in which the phylogenetic distance

based on OTUs is computed between each pair of samples in the study. Multivariate analysis or the top principal coordinates (PCo) of the matrix of pairwise distances are used to test for associations via permutation. Commonly used pairwise distance metrics include weighted and unweighted UniFrac (28, 118, 25) as well as many other important metrics such as the Bray-Curtis (17) metric.

2.1.4 Epigenetics

The term "epigenetics" refers to the heritable and reversible changes in phenotypes that are not coded in the DNA sequence, including DNA methylation, histone modifications and nucleosome positioning. Variation in the epigenome plays a key role in cell differentiation (32, 138) and is considered the main reason of the specialized functions to different cells with the same genome. In multicellular organisms, the ability that epigenetic modifications can be transmitted to offsprings is essential to generate individuals with the same genotype but different phenotypes, such as in cloning or in identical twins (162, 56). Increasing evidence showed that epigenetic modifications are transgenerational, constituting another source of the missing heritability (84).

DNA methylation is the most studied epigenetic event, which occurs almost exclusively on the cytosine at position C5 in CpG dinucleotides. CpG dinucleotides tend to cluster in CpG islands, which are usually defined as regions of at least 200 base pairs with more than 50% G+C content and observed-to-expected CpG ratio of at least 0.6. CpG dinucleotides are pretty rare in human genome, constituting only $\sim 1\%$ of the genome. However, up to 70% of annotated gene promoters are associated with a CpG island, making this the most common promoter type in the vertebrate genome (173, 196). DNA methylation is essential in establishing and maintaining the normal cellular functions, including embryonic development, X-chromosome inactivation and allele-specific methylation related to imprinting (206). Aberrant DNA methylation has

also been related to a variety of human diseases ranging from neurological and autoimmune disorders to cancers(155, 219).

Currently, DNA methylation levels are usually evaluated through two methods: bisulphite sequencing and array based approaches (11, 220, 151, 97), which involve converting unmethylated cytosines to uracil while leaving 5-methylcytosines intact. Advances in next-generation sequencing and array technology has enabled the global assessment of DNA methylation at a high resolution and affordable prices in a large number of samples (159, 105) and hence epigenome wide association studies (EWAS). Similar to GWAS, EWAS aim at identifying differentially methylated CpGs associated with disease states, clinical outcomes, environmental exposures or other experimental conditions (85, 184, 73, 74).

Analysis of methylation data has been shown to be challenging (14). In addition to the problems relating to data preprocessing and normalization (44, 203, 132, 201), associating methylation levels with outcomes is also difficult. A lot of the initial analysis of DNA methylation have utilized statistical methods that were developed for gene expression data, such as differences in abundance levels, cluster analysis and class prediction(88). For example, methods such as t tests, non-parametric tests and generalized linear regression with a quasi-binomial logit link were used to assess the differential DNA methylation in subgroups of samples, in which proper transformation was conducted to make the methylation data from zero to one scale to normally distributed (9, 188). Alternatively, beta regression has also been used to model DNA methylation proportions (54).

2.1.5 Group Based Analysis

Comprehensive understanding of complex trait etiology requires examination of multiple sources of genomic variability. Integrative analysis of these data sources promises

elucidation of the biological processes underlying particular phenotypes. Multi-feature testing, in which the cumulative effect of multiple related features is tested for association with outcome, has gain considerable popularity (198, 222, 208, 176, 240).

In GWAS, a number of SNP set based analysis methods have been developed. SNP set based analysis is a two step procedure with the first step to form SNP sets based on prior biological knowledge and the second step to test for association between the SNP sets with outcome. SNP sets are usually formed based on their physical proximity to a known genomic features (229, 222); e.g, genes or pathways. Then the SNP sets are tested for association against the phenotype as a group via different dimension reduction approaches. Intuitively, SNP set based analysis borrows information across different SNPs that are grouped on the basis of prior biological knowledge and hence provides results with improved reproducibility and increased power, especially when individual-SNP effects are moderate.

Methods that test for cumulative effect of multiple markers/features can be classified into two groups: competitive and self-contained tests (64). In GWAS, the competitive test compares test statistics for SNP set to all the SNPs that are not in the set and test for over representation of the SNPs in the SNP set, such as the Fisher's exact test (27) for pathway effect and the gene set enrichment (198).

Different from competitive tests, a self-contained test compares the a test statistic to a fixed standard and doesn't depend on the effect of background features, which includes the principle component tests, the distance based approach and the kernel machine regression approach.

Principle component is a widely used tool in statistics for dimension reduction. This method seeks to represent the data by a linear combination of a small number of orthogonal principle components and then applies the standard univariate or multivariate analysis to test for association. Gauderman et al. (60) proposed a principal components

analysis based approach (PCA), by which principal components (PCs) are computed from the SNP set and then tested for the association with phenotype of interest. Gao et al.(59) proposed to use kernel function to represent the SNP data and subsequently compute PCs based on the kernel function to test for association and showed superior power compared to the original PCA analysis in case-control GWAS, especially under lower relative risks and lower significance levels. Chen et al. (30) developed a pathway-based analysis using supervised principal components, in which only a selected subset of SNPs most associated with disease outcome is used for construction of PCs and test for association. Adjusting for confounding variables in the PCA methods amounts to adding covariate in to the standard linear or logistic regression model.

Another school of self contained methods involve regression models to relate variation in genomic dissimilarity (or distance) measurement to variation in their phenotype values (222). This genomic distance based regression(GDBR) captures the genotype/haplotype information across multiple loci through the similarity between any two subjects. P-values can be obtained through permutation using a pseudo-F statistic. GDBR has been demonstrated the higher power than several commonly used tests across a wide range of realistic scenarios (106). In addition, a close relation has been established between the GDBR and a class of haplotype similarity tests (236, 207, 181, 91). Unlike PC based approach, adjusting for covariates in GDBR is not straightforward because permutation approach tends to break the correlation structure between the genotype and the confounding variables. The permutation test can also be computationally expensive.

Kernel machine regression(KMR) framework also belongs to self contained global tests. Instead of extracting the PCs from the SNP data, KMR assumes a potentially nonlinear functional relationship between genotypes and the outcome, which can be modeled through the kernel function. The method is fast and efficient as it uses

the asymptotical distribution of variance score test and avoids the time consuming permutation procedure. As a regression based approach, adjusting for covariates are straightforward. Several studies have demonstrated the superior power of KMR compared to other methods under a wide range of practical scenarios (229, 230, 112, 111). Details about the method will be reviewed in Chapter 2.2.

2.2 Semiparametric Model via Kernel Machine Regression

Kernel machine regression (KMR) was proposed in the gene expression framework (112, 111) and extended to test for associations between SNP set and individual complex phenotype (96, 229). Further extension of this approach was applied to censored survival data (21, 108), multivariate outcome (131) and rare variants (185, 8, 230, 99, 100). In this section, we will focus on kernel machine testing with single continuous or dichotomized outcome and defer the extended KMR to Section 2.3.

2.2.1 Specification of $h(Z)$ using Kernel Functions

Suppose the data consist of n subjects. For each subject $i, i = 1, \ldots n$, y_i denotes the phenotype of interest, Z_i is a $1 \times p$ vector of genotypical data, which can be gene expression in a pathway, or genotypes for a set of SNPs or rare variants. X_i denotes a $1 \times q$ vector of confounding variables which we want to adjust for in the model (e.g. demographic or environmental variables). Under the KMR framework, continuous traits y_i depends on X_i and z_i through partial linear model

$$y_i = \beta_0 + X_i\beta + h(Z_i) + \varepsilon_i \tag{2.1}$$

where β is a $q \times 1$ vector of regression coefficients, $h(Z_i)$ is an unknown smooth function and $\varepsilon \sim N(0, \sigma^2)$. Similarly, the model risk of dichotomized trait y_i can be given as:

$$\text{logit}(p(y_i = 1|X_i, Z_i)) = \beta_0 + X_i\beta + h(Z_i) \tag{2.2}$$

Generally, models (2.1) and (2.2) allow for nonparametric modeling of multi-dimensional genomic effect with parametric adjustment of confounding effect. When $h(\cdot) = 0$, the models reduce to standard linear regression or logistic regression model.

The KMR model makes the assumption that $h(\cdot)$ lies in a function space $\mathcal{H}_k$ generated by a positive semidefinite kernel function $K(\cdot, \cdot)$ and this kernel function maps complex and potentially infinite dimensional features into a finite dimensional space. Mercer's theorem (40) states that under some minor regularity conditions, $K(\cdot, \cdot)$ implicitly specifies a unique function space $\mathcal{H}_k$ which can be spanned by a set of orthogonal basis functions such that $h(z) = \sum_{j=1}^{J} \omega_j \phi_j(z) = \phi(z)'\omega$, in which ω is a vector of coefficients. This is the primal representation. Alternatively, $h(z)$ can be represented using a kernel function $K(\cdot, \cdot)$ so that $h(z) = \sum_{l=1}^{L} \alpha_l K(z_l^*, z)$, where $\alpha_1, ...\alpha_L$ be a vector of constants, L being an integer and $z_1^*, ..., z_L^* \in R^p$ (dual representation). In practice, the dual representation is more convenient as it avoids the explicit specification of the basis function and instead only needs to define the kernel function.

The choice of kernel specifies implicitly a complex and nonparametric function space that can capture signals from possibly high-order interaction effects. A wide range of kernels have been described in literatures, with some popular ones listed as follows:

(1) *Linear kernel*: $K(z_1, z_2) = z_1 z_2'$. Linear kernel generates the usual inner product space with basis function $\phi(z) = \{z_1, ..., z_p\}$ and essentially assumes that $h(z) = z'\beta$. Similarly, the weighted linear kernel $K(z_1, z_2) = z_1 W W' z_2'$ generates also a linear function space while allowing different variables to have different relative weights, controlled by the weight matrix $W = \text{diag}[w_1, ..., w_p]$, a $p \times p$ diagonal matrix.

(2) *The d^{th} order polynomial kernel*: $K(z_1, z_2) = (z_1 z_2' + c)^d$ where c is a constant and d determines the order of the polynomial. This kernel implies that $f(z)$ is a d^{th} order polynomial function. When $d = 1$, this first polynomial kernel reduces to the linear kernel with basis function $\phi(z) = \{z_1, ..., z_p\}$. When $d = 2$, the quadratic kernel corresponds to function space with basis function $\phi(z) = \{z_k, z_k z_k'\}$, which is the main effect of each variables in z and their squared and two way interactions.

(3) *Gaussian kernel*: $K(z_1, z_2) = \exp\{-\|z_1 - z_2\|^2/\rho\}$ where $\|z_1 - z_2\|^2 = \sum_{k=1}^{p}(z_{1k} - z_{2k})^2$. The Gaussian kernel corresponds to infinite dimensional function space spanned by radial basis functions. ρ is an extra tuning parameter which controls the degree of linearity, with larger ρ forcing $h(z)$ to be more linear while smaller ρ allows more complex effects to be modeled.

(4) *Weighted identity by state (IBS) kernel.* $K(z_1, z_2) = \sum_{k=1}^{p} w_k\{2I(z_{1k} = z_{2k}) + I(|z_{1k} - z_{2k}| = 1)\}/2p$. The weighted IBS kernel evaluates the genetic distance between a pair of individuals by the fraction of alleles that are shared purely by state (222, 96), subject to proper weighting (230). The weighted IBS kernel has been used in a number of method to measure the similarity using SNPs data or rare variants (229, 230). Because the number of alleles that are identical between subjects is a physical property, this kernel assumes no specific form of the genetic effect, such as the linear effect or polynomial effect with specific order, and allows for epistatic and interaction effects between the SNPs or rare variants.

(5) *Other positive definite kernels.* Many other kernels have been described and tailored to particular data structures. Examples of other choices of kernel functions include the spline kernel, the exponential kernel, the neural network kernel and the sigmoid kernels (177). In fact, any positive semi-definite matrix that measures the similarity between subjects can be used as kernel matrix. Pan et al (152) has established correspondence between genomic distance based approach(GDBR) and KMR in that if

the same positive semi-definite matrix is used as the similarity matrix in distance based approach and the kernel matrix in KMR, the two tests are equivalent up to ignorable constants.

2.2.2 Estimation through Mixed Model Framework

In KMR model, the kernel matrix $K(\cdot,\cdot)$ generates function space $\mathcal{H}_k$ such that $f(z) \in \mathcal{H}_k$. Following the general approach in functional data analysis and additive models (226), Liu et al (112, 111) propose to estimate β and $h(z)$ in models (2.1) and (2.2) by maximizing the penalized likelihood function.

$$J(h,\beta) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - X_i\beta - h(Z_i))^2 - \frac{1}{2}\lambda\|h\|^2_{\mathcal{H}_k} \tag{2.3}$$

and

$$\begin{aligned} J(h,\beta) &= \sum_{i=1}^{n}\{y_i \log(\frac{\mu_i}{1-\mu_i}) + \log(1-\mu_i)\} - \frac{1}{2}\lambda\|h\|^2_{\mathcal{H}_k} \\ &= \sum_{i=1}^{n}(y_i\{\beta_0 + X_i\beta + h(Z_i)\} - \log\{1+\exp(\beta_0 + X_i\beta + h(Z_i))\}) - \frac{1}{2}\lambda\|h\|^2_{\mathcal{H}_k} \end{aligned} \tag{2.4}$$

where λ is the tuning parameter controlling the balance between the goodness of fit and the complexity of the model. When $\lambda = 0$, the model represents a saturated model that interpolates all data points. When $\lambda = \infty$ the model forces $h(Z) = 0$ and reduces to the simple linear or logistic model.

By the Representer Theorem (90), the nonparametric function $h(z)$ in (2.1) and 2.2 can be expressed as

$$h(Z) = \sum_{i=1}^{n}\alpha_i K(.,Z_i) = K\boldsymbol{\alpha} \tag{2.5}$$

Substituting (2.5) into (2.3) and (2.4), the objective function becomes

$$J(h,\beta) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - X_i\beta - \sum_{j=1}^{n}\alpha_j K(Z_i, Z_j))^2 - \frac{1}{2}\lambda\alpha' K\alpha \tag{2.6}$$

and

$$J(h,\beta) = \sum_{i=1}^{n}\{y_i(\beta_0 + X_i\beta + k_i'\alpha) + \log(1+\exp(\beta_0 + X_i\beta + k_i'\alpha)\} - \frac{1}{2}\lambda\alpha' K\alpha \tag{2.7}$$

respectively, in which K is $n \times n$ matrix whose $(i,j)^{th}$ elements is $K(Z_i, Z_j)$ and α is a vector of constant that needs to be estimated. With predefined λ, estimation of β and α can be easily carried out by equating the first derivative of the penalized likelihood function to zero. In reality, the optimal value of λ needs to be estimated through cross validation or by minimizing the generalized cross validation (GCV) score (215), which can be computationally expensive.

In the original paper in 2007, Liu et al. (112) showed that the linear KMR model in equation (2.1) share the same normal equation as in the following linear mixed model:

$$y = \beta_0 + X\beta + h + \varepsilon \tag{2.8}$$

where h is a $n \times 1$ vector of random effects distributed as $N(0, \tau K)$ with $\tau = \lambda^{-1}\sigma^2$, β as a vector of regression coefficients for fixed effects and $\varepsilon \sim N(0, \sigma^2 I)$. Therefore, the estimation of h in model (2.1) corresponds to the best linear unbiased predictor (BLUP) from the linear mixed model which can be obtained through the restricted maximum likelihood method (REML)(69), with simultaneous estimation of the variance component τ.

Parallel to the result from linear KMR, the logistic KMR model in (2.2) were shown

to correspond to the logistic mixed model (111)

$$\text{logit}(\mu) = \beta_0 + X\boldsymbol{\beta} + h \tag{2.9}$$

with h being a $n \times 1$ vector of random effects distributed as $N(0, \tau K)$. Within the logistic mixed model framework, the coefficients β and h can be obtained by fitting the penalized quasi-likelihood (PQL) (146), in which τ is treated as variance component as well as in the linear case.

2.2.3 Variance Component Score Test

In KMR models, it is of great interest to test the overall effect of the genomic features on the outcome, i.e, whether $h(Z) = 0$, with linear adjustment for confounding variables. From the correspondence between KMR models and linear/logistic mixed model, $h(Z)$ is distributed with mean 0 and variance τK. Therefore, $h(Z) = 0$ is equivalent to $\tau = 0$ and the hypothesis can be restated as

$$H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau > 0 \tag{2.10}$$

Test of variance component is nonstandard as the null hypothesis put $\tau = 0$ at the boundary of the parameter space; the likelihood ratio statistic doesn't follow the usual χ^2 distribution (180). Moreover, because the kernel matrix is not block diagonal, the standard approach in mixed models (180) does not apply either and the likelihood ratio doesn't follow a mixture of χ^2_0 and χ^2_1 distribution. Instead, a variance score test was proposed (112, 111, 229) for both quantitative and binary outcomes. The score statistic has the form of

$$Q_\tau = (y - \hat{\mu}_0)' K (y - \hat{\mu}_0) \tag{2.11}$$

in which $\hat{\mu}_0$ is the estimate of y based on the simple linear/logistic regression model

in which no genomic effect is present. Under the null hypothesis, the Q_τ follows a mixture of χ^2 distribution, which can be approximated by a number of approaches (45, 41).

The variance component score test avoids the estimation under the alternative hypothesis and only requires fitting the linear/logistic regression model, which is computationally efficient. For kernels such as the Gaussian kernels which involve additional parameters ρ, the unknown parameter vanishes under the null hypothesis and become inestimable. The variance component score test is valid for any value of ρ, with better choice of ρ merely improves the power.

2.3 Further Work on KMR in Genomic Studies

2.3.1 KMR for Survival Outcomes

The ultimate goal of most genetic studies is to uncover the biological mechanisms underlying human disease, which can subsequently lead to better understanding of the disease process and improved disease prevention and management (75). GWAS with survival outcomes have also been conducted in a number of diseases (5, 77, 53). Traditional approaches that fit individual Cox proportional hazard models to each SNP with subsequent multiple testing adjustment suffers from the same limitations as in the studies with continuous or binary outcomes. In the recent two papers, Lin et al. (108, 21) proposed to use the KMR framework to test for association between a set of genetic markers with censored survival outcome. The model assumes that the survival time T is related to genotype Z and additional covariant X through the Cox proportional hazard model (38) that

$$\lambda(t) = \lambda_0(t) \exp(\beta_0 + X_i\beta + h(Z_i)) \tag{2.12}$$

Through the dual representation, $h(Z) = \sum_{i=1}^{n} \alpha_i K(Z_i, Z)$ where α_i are unknown parameters. Testing the null hypothesis that $H_0 : h(Z) = 0$ is equivalent to testing $H_0 : h(Z) = \sum_{i=1}^{n} \alpha_i K(Z_i, Z) = 0$. The KM score test for censored survival data assumes that $\alpha = \{\alpha_0, \alpha_1, ..., \alpha_n\}'$ follows an arbitrary distribution with mean 0 and variance $\tau \mathbf{K}^{-}$, where K is the $n \times n$ kernel matrix with the $(i, j)^{th}$ element as $K(Z_i, Z_j)$, and K^{-} being the generalized inverse. Then H_0 is equivalent to testing the variance component $H_0 : \tau = 0$, with a score statistic as

$$Q = \hat{\mathbf{M}}' \mathbf{K} \hat{\mathbf{M}} - \hat{q} \tag{2.13}$$

where $\hat{\mathbf{M}} = (\hat{M}_1, \hat{M}_2, ..., \hat{M}_n)'$, where $\hat{M}_i$ being the martingale residual for individual i under the null hypothesis that

$$\hat{M}_i(t) = \Delta_i(t) - \int_0^s Y_i(t) e^{(\hat{\beta}_0 + X_i' \hat{\beta})} d\hat{\Lambda}_0(t)$$

$$\hat{q} = \sum_{i=1}^{n} \int K(Z_i, Z_i) Y_i(t) e^{(\hat{\beta}_0 + X_i \hat{\beta})} d\hat{\Lambda}_0(t) - \sum_{i=1}^{n} \sum_{j=1}^{n} \int \frac{Y_i(t) Y_j(t) e^{(\hat{\beta}_0 + X_i' \hat{\beta})} K(Z_i, Z_j)}{\hat{S}^0(t)} d\hat{\Lambda 0}_0(t)$$

with $Y_i(t) = I(U_i > t)$, the at risk indicator, $\hat{\beta}$ the partial likelihood estimator of β under the null,and $\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \Delta_i I(T_i \leq t) / \hat{S}^0(T_i)$, the Breslow's estimator of $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ under the null model.

Under the null, the score statistic Q asymptotically follows a mixture of χ^2 distributions which can be approximated through resampling approach (21). Specifically, Cai et al.(21) showed that Q converges in distribution to double-integrated martingale processes and p-value can be obtained by approximating the distribution of the martingale processes via resampling to generate realization of the score statistic under the

null.

2.3.2 Multivariate Phenotype Association by KMR

Although most GWAS are analyzed for one phenotype at a time, data for multiple related phenotypes are often collected. Joint analysis of multiple disease-related phenotypes have the potential to revel genes with pleiotropic effect and increase statistical power for association (239, 114). Several papers have developed methods for multivariate association analysis of multiple phenotypes (239, 114, 237, 213). However, most of these multivariate analysis focus on the effect of a single marker.

Similar with the case of univariate outcome, Maity et al. suggested that multivariate analysis can also benefit from marker set analysis via KMR and proposed a multivariate kernel machine regression framework (MVKMR) (131). Specifically, assume that for each individual $i = 1, 2, ..., n$, $\mathbf{Y}_i = (Y_{1i}, ..., Y_{pi})$ be a response vector of phenotypes of interest, X_i be the confounding variables that need to be adjusted, and $\mathbf{Z}_i = (Z_{i1}, ...Z_{iM})'$ being the group of SNPs that are of interest. The model can be constructed as

$$Y_{ki} = X_i\beta_k + h_k(Z_i) + \varepsilon_{ki} \tag{2.14}$$

for $k = 1, ..., p$ and $i = 1, ..., n$,with $\{\varepsilon_{1i}, ...\varepsilon_{pi}\} \sim N(0, \Sigma)$ with $\Sigma = \{\sigma_{kl}\}$ where σ_{kl} reflects the correlation between different phenotypes within the same individual and $h_k(Z)$ represents the genetic effect on the k^{th} phenotypes which can be specified through a kernel function $K_l(\cdot, \cdot)$. The null hypothesis that the SNP sets have no effect on the outcome can be written as

$$H_0 : h_1(\cdot) = h_2(\cdot) = ... = h_p(\cdot) = 0$$

Take $\mathbf{Y} = (Y_{1i}, ..., Y_{1,n}, ..., Y_{P1}, ..., Y_{pn})'$, $\mathbf{h} = (h_1(Z_1), ..., h_1(Z_n), ..., h_p(Z_1), ..., h_p(Z_n))'$

and $\boldsymbol{\varepsilon} = (\varepsilon_1{}', ..., \varepsilon_p')'$, stacked vectors of all the outcome variables, genetic effects and their corresponding residuals. Also define $\mathbf{X} = \text{diag}(X_1, ..., X_p)$ and $\boldsymbol{\beta} = (\beta_1', ..., \beta_p')'$. The model can be rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon} \tag{2.15}$$

where$\boldsymbol{\varepsilon} \sim N(0, \widetilde{\boldsymbol{\Sigma}})$ where $\widetilde{\boldsymbol{\Sigma}} = \Sigma \bigotimes I_n$, i.e, $\widetilde{\boldsymbol{\Sigma}}$ is a $p \times p$ block matrix with each block being a diagonal matrix of σ_{kl} for $k = 1, ..., p$ and $l = 1, ..., p$.

Following the same argument as in univariate KMR, estimation of $\boldsymbol{\beta}$ and $\mathbf{h}$ can be carried out through the linear mixed model framework.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon} \tag{2.16}$$

where $h \sim N(0, \mathbf{K}\boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \text{diag}(\tau_1, ..., \tau_p) \otimes I_n$, and $\boldsymbol{\varepsilon} \sim N(0, \widetilde{\boldsymbol{\Sigma}})$. Testing the null hypothesis that $H_0 : h_1(.) = ... = h_p(.) = 0$ is equivalent to testing $H_0 : \tau_1 = ... = \tau_p = 0$. The corresponding variance component score statistic

$$Q = (\mathbf{Y} - \mathbf{X}'\hat{\boldsymbol{\beta}})'\mathbf{V}_0^{-1}\mathbf{K}\mathbf{V}_0^{-1}(\mathbf{Y} - \mathbf{X}'\hat{\boldsymbol{\beta}}) \tag{2.17}$$

in which $\mathbf{V} = \mathbf{K}\boldsymbol{\Lambda} + \widetilde{\boldsymbol{\Sigma}}$ represents the total variance of y, and $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}_0$ are the estimates of $\boldsymbol{\beta}$ and $\mathbf{V}$ under the null model. P-value can be calculated by comparing Q to a mixture of χ^2 distribution which can be approximated through moment matching or empirical approach.

2.3.3 Kernel Machine Test Under Multiple Candidate Kernels

One major advantage of the KMR is its flexibility: choosing different kernels assumes different functional form of the genetic effect h. As a score test, the test is valid regardless of the choice of kernel, i.e. with well controlled type I error. However, good

choice of kernel generates test with improved power. For example, when epistasis is present, kernels that accommodate nonlinear effects, such as the IBS kernel, can provide improved power for association between SNP set and the phenotype(222). However, if there is no epistatic effect, using the linear kernel can usually be more powerful (108, 229). In practice, information on the underlying genetic effect is seldom known. A number of methods have been proposed that consider multiple candidate kernels simultaneously and choose the optimal test to maximize power, with adjustment of choosing the best kernel.

In the association tests between rare variants and phenotype, Lee et al. (100) extended the sequence kernel association test (SKAT) to allow for correlations among different markers. Specifically, they proposed an optimal test that uses as test statistic a linear combination of burden (156, 144) and SKAT test statistic (230), and showed that it is equivalent to the SKAT statistic with a new family of kernel that includes a correlation parameter ρ.

$$Q_\rho = (y - \hat{\mu})' \hat{\Delta} \hat{V}^{-1} K_\rho \hat{V}^{-1} \hat{\Delta} (y - \hat{\mu})$$

Using notation from generalized linear models, $\hat{\Delta}$ is the estimated link adjustment $\Delta = \text{diag}(g'(\mu))$, $\hat{V} = \text{diag}\{\hat{\phi} v(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2\}$. When canonical link function is used, the test statistic can be simplified to be

$$Q_\rho = (y - \hat{\mu})' K_\rho (y - \hat{\mu}) / \hat{\phi}^2$$

where $K_\rho = GWR_\rho WG'$ is the new kernel function which involves correlation structure, $R_\rho = (1 - \rho)I + \rho 11'$ represents an exchangeable correlation matrix, W is the weight matrix and G the genotype of the rare variants. For each fixed ρ, Q_ρ follows a mixture of χ^2 distribution, which can be easily approximately through the variance inversion or

moment matching approaches.

The minimum of the p-values across different values of ρ was used as test statistic

$$T = \inf_{0 \leq \rho \leq 1} p_\rho$$

Lee et al. derived the theoretical distribution of this SKAT-O statistic by combining the two kernel matrices via projection and approximate Q_ρ as a sum of two independent χ^2 mixture distribution. Sample size and power calculation formula was also derived. By avoiding the computationally expensive resampling approach, this method is fast to implement and applicable to whole genome studies. However, this method can not be extended to kernels beyond the linear kernel.

Several extensions of this approach exist in the literature. A later paper from the same group (99) extended the multiple kernel testing for rare variants and especially studied the problem that the asymptotical p-values from logistic KMR can be too conservative when the sample size is small, leading to incorrect type I error and power loss. Specifically,this paper proposed a method to adjust the asymptotical null distribution of Q and obtain p-values by matching through higher moments, especially kurtosis, via parametric bootstrapping method. A very recent paper (79) extended a similar multiple kernel approach to test for the combined effect of rare and common variants, in which the test statistic is a weighted sum of the KM score statistic constructed by using the rare and common variants separately. All these methods uses linear projection of the kernel matrices to derive the asymptotical distribution of the test statistics, and thus can only allow for linear kernels to be tested.

Other approaches have been proposed that allow for multiple arbitrary kernels to be involved. Wu et al. (233) developed an efficient perturbation procedure that preserves the correlation structure between the genotypes and confounding variables and allows for multiple candidate kernels to be considered simultaneously.

2.4 Causal Mediation Analysis

Establishing causal relationships is one of the central tasks in all aspect of scientific and social research. In the field of medicine and public health, it is a fundamental step in elucidating disease mechanism, designing the best prevention strategy and choosing personalized treatment. It has been heavily debated in philosophy, statistics and epidemiology. Simply put, mediation analysis is a causal model that investigates the role of intermediate variables on the causal path between an independent variable and an outcome variable (71). Mediation models have been extensively used in psychological studies to establish the causal chain between a randomized treatment and outcome variables (123). In the recent years, with the ability to gather sufficient information in human genome, considerable work has been done using mediation models to decipher the genetic causal network (13, 66, 80, 216, 210, 46, 200, 140, 109, 174).

2.4.1 Regression Approach to Mediation

The idea of mediation concerns the extent to which the effect of one variable on another is mediated by some possible intermediate variable. Mediation analysis is an application of associational causal modeling, i.e., it models causality using measures of association (61). A mediation hypothesis is usually represented by a diagram of a causal model (Figure 2.1).

In a mediation model (Figure 2.1), an independent variable X is assumed to cause a set of mediators M, which, in turn, causes the dependent variable Y, so that the effect of X on Y is at least partially through the effect of the mediators. The easiest case of mediation models is the situation when there is only a single mediator M, the effect of which can be modeled through linear regression (Figure 2.1 Panel A & B). In 1986, Baron and Kenny (6) presented a pioneer yet simple mediation model, called causal steps model, for demonstrating that a data set is consistent with the hypothesized

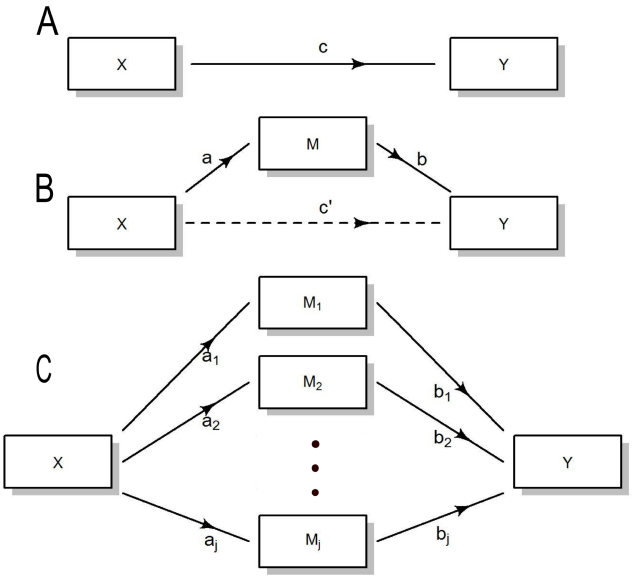

Figure 2.1: Mediation Diagram

causal relationship if the following four steps are satisfied:

1) the independent variable X is correlated with the outcome Y.

$$Y = \alpha_1 + cX + \varepsilon_1 \tag{2.18}$$

2) X is correlated with potential mediator M

$$M = \alpha_2 + aX + \varepsilon_2 \tag{2.19}$$

3) M is associated with the outcome variable Y conditional on X

$$Y = \alpha_3 + c'X + bM + \varepsilon_3 \tag{2.20}$$

4) To establish that M completely mediate the relationship between X and Y, the residual effect of X on Y should be zero after controlling for M (path c' Figure 2.1 Panel B).

The original Baron and Kenny approach did not include additional covariates in the model to adjust for confounding effect. Further studies allowed confounding effect to be corrected by adding extra covariates in the linear models. In this review, we will present the models without considering additional covariates, however, it should be noted that cases with additional covariates is analogous to what is presented.

The first three steps were proposed to be tested through the ordinary least square (OLS) regression models, with the fourth step only required for conclusion of complete mediation, which can be assessed through a bootstrap approach (86). Other methods of estimation, such as logistic regression, multilevel modeling, and structural equal model were proposed in later studies to work with data with non-Gaussian distribution. However, the necessary steps are the same regardless of the analytic methods (58). More contemporary researchers believe that only step 2 and 3 are essential in establishing mediation, especially in the situation of inconsistent mediations (87). Although alternative models have been proposed, causal steps models still constitute a large proportion of all mediation tests in psychology and epidemiology studies (123, 168).

The amount of mediation is considered as indirect effect. From equations (2.18),(2.19) and (2.20),

$$c = c' + ab \tag{2.21}$$

where c is the total effect of X on Y, c' is the direct effect and ab is the indirect effect.

Equation (2.21) holds exactly when a) OLS was used in each of the steps, b) the same cases are used in all analysis and c) all equations adjust for the same covariates. The decomposition of the total effect into direct effect and indirect effect provides the philosophical foundation that can be generalized to more complex causal inference models, including structure equation models and models with counterfactual effects (209).

There are several ways to test for the indirect effect under the causal steps framework. A first and intuitive way to test for the null hypothesis of indirect effect $ab = 0$ is to test that both paths a and b are zero. A more highly recommended strategy to test for the indirect effect ab is to have a single test of ab (125). Sobel (191) proposed a method that uses the asymptotical normal distribution of ab and corresponding Z statistic to calculate p-values via Delta method. The approximate standard error of ab is $b^2 s_a^2 + a^2 s_b^2$ where s_a and s_b are the standard errors of a and b. The Sobel test is considered to be very conservative and usually have low power (127) because the test approximates ab by a symmetric normal distribution while the sample distribution is usually highly skewed.

Bootstrapping method is a relatively new and increasingly popular approach in testing the indirect effect (186, 16), which computes the nonsymmetric confidence bounds for ab from the empirical distribution by resampling approach. There are two commonly used Bootstrap methods in literature. Percentile Bootstrap directly construct the empirical distribution and confidence interval by finding the corresponding percentiles from the estimates from resampled data sets. Bias corrected Bootstrap methods (126, 93) calculate confidence interval and p-values by adjusting the bias between the bootstrapped distribution and the indirect effect. Several recent studies have raised concerns that the bias corrected bootstrapping test can have type I error that are too liberal. In a recent paper (70), Hayes and Scharkow recommended using the bias corrected bootstrap if the power is the main concern while use the percentile bootstrap when the major concern is type I error.

In the cases when outcome Y are categorical variables, the OLS regression approach is no longer applicable. However, the conceptual decomposition of the total effect into direct effect and indirect effect still applies. Mackinnon and Dwyer (122) proposed to

use the modified models for estimation of the standard errors of ab:

$$Y_0^* = \alpha_1 + cX + \delta_1 \tag{2.22}$$

and

$$Y_0^* = \alpha_1 + c'X + bM + \delta_2 \tag{2.23}$$

where Y_0^* is the the unobserved probit of the probability of being in one of the two categories of the outcome variable, c reflects the effect of the program on the probit of the outcome probability in the first equation, c' is the direct effect of the program on the probit of outcome probability adjusted for the effects of the mediator, δ_1 and δ_2 are the residuals in the probit models. The same Sobel tests and Bootstrap approach can be used to test for the indirect effect.

Mediation with Multiple Mediators

Scientific and social researches are replete with situations when multiple mediators exist between an independent variable X and an outcome variable Y. Multiple mediation models that incorporate simultaneous mediation by multiple variables have received less attention in methodological and applied studies than the single mediation models. However, there are a considerable number of methods proposed that aim to test for the overall effect of multiple mediation (192, 19, 31, 124, 93).

Figure 2.1 Panel C represents the situation in which there are j possible mediators between independent variable X and outcome Y. A specific indirect effect through one mediator can be defined as the product of the two paths linking X to Y via one of the mediators (19) and the total indirect effect are defined as the summation of all the specific indirect effect $\sum_{i=1}^{j} a_i b_i$ where $i = 1, ..., j$. Similar to the single mediation case, the total effect can be written as $c = c' + \sum_{i=1}^{j} a_i b_i$. In a more recent paper, Preacher

and Hayes (93) emphasized the difference between the indirect effect through one of the mediators (e.g.,M_1) in a multiple mediation model and the indirect effect in a single mediation model with only M_1 as a mediator. They proposed three testing approaches for the total indirect effect $\sum_{i=1}^{j} a_i b_i$ that mimic the three testing approaches in single mediation analysis: 1) the causal steps approach, which tests for each specific indirect effect and concludes mediation if any of the indirect effect is not zero. 2) Product-of-Coefficients approach which derives the asymptotical variance of the total indirect effect $\sum_{i=1}^{j} a_i b_i$ using multivariate delta method and calculates p-values and confidence interval through the usual Z test. Similar as the Soblel test, this method relies on large sample approximation and can result in a lower power if the sample size and/or effect size are not sufficiently large. 3) Bootstrapping approach, which uses resampling procedure to establish the empirical distribution of the total indirect effect. Preacher and Hayes (93) recommended the use to biased corrected Bootstrapping method for testing of multiple mediation effect.

2.4.2 Counterfactual Approach to Mediation Analysis

While the concept of mediation, defined as indirect effect through the classical regression framework, is appealing theoretically, it is difficult to extend this definition to situations when the effect of exposure and mediator on the outcome have interactions or is non-linear (164, 153). The approach of decomposing the total effect to direct and indirect effect is not readily applicable because holding the mediator at different levels would generate different direct effect at the presence of interaction effect.

Recent progress in mediation analysis has extended the concept of direct and indirect effect to situations when non-linearities and interactions are present and considers the identifiability conditions for a causal relationship. In a paper in 2001, Pearl et al(153) uses the counterfactual notation and formulated new definition of path-specific

effect: the controlled direct effect (CDE) and natural direct effect (NDE). Under the counterfactual framework, the CDE of the exposure on the outcome is defined as $E(Y(x,m) - Y(x^*,m)|C)$, which is the change in the outcome if the treatment was changed from $x^* = 0$ to $x = 1$ while the mediator M is fixed at level m across the population, where C is the confounders that need to be adjusted for. The NDE is defined as $E(Y(x, M(x^*)) - Y(x^*, M(x^*)))$, which is the exposure effect that would be obtained in the outcome if the exposure were changed from x to x^* while the mediator was kept at the level that would be observed as if the independent variable were kept at x^*. Natural and controlled indirect effects (NID and CID) are defined as the difference between the total effect and the corresponding direct effect. VanderWeele and Vansteelandt (209, 211) show that the counterfactual framework can extend the Baron and Kenny formulae for direct and indirect effect to situations when there is interaction effect between the exposure and mediator on the outcome.

Specifically, consider a model when there is interactive effect between the exposure and mediator on the outcome,

$$E(M|X = x, C = c) = \beta_0 + \beta_1 x + \beta_2' c \tag{2.24}$$

$$E(Y|X = x, M = m, C = c) = \theta_0 + \theta_1 x + \theta_2 m + \theta_3 xm + \theta_4' c \tag{2.25}$$

Through models (2.24) and (2.25), under some identifiability assumptions, the CDE, NDE and NIE for change of independent variable from level x^* to X is given by under some identification assumptions

$$\begin{aligned} CDE &= (\theta_1 + \theta_3 m)(x - x^*) \\ NDE &= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 x^* + \theta_3\beta_2' c)(x - x^*) \\ NIE &= (\theta_2\beta_1 + \theta_3\beta_2 a)(x - x^*) \end{aligned} \tag{2.26}$$

In cases when there is no interaction effect ($\theta_3 = 0$), the CDE and NDE equal to the direct effect obtained through the Baron-Kenny linear regression approach. However, in this counterfactual framework, the total effect can be decomposed into the sum of natural direct and indirect effect even in models with interactions and non-linearities (153). NDE and NID are useful in evaluating the different mechanism of exposure on outcome, while CDE and CID are often of greater interest in policy evaluation.

Certain identification assumptions are required for expression (2.26) to hold. The direct and indirect effect defined previously are conditional on the levels of covariates C. Let $C = (C_1, C_2)$ where C_1 denotes the confounders of the effect between the exposure and outcome and C_2 denotes the confounders of the mediator-outcome effect. According to VanderWeele et al.(209), two assumptions are required for the identifiability of the controlled direct effect: 1) no unmeasured confounders of the exposure-outcome relationship and 2) no unmeasured confounders of the mediator-outcome relationship. Randomization in the treatment/exposure guarantees the first assumption, but not the second assumption. Thus, in practice, it is required to control for the common causes of treatment/exposure and the outcome to control for the first assumption and control for common causes for mediators and the outcome for the second assumption. For the identification of natural direct and indirect effect, two additional assumptions are required: 3) no unmeasured confounders of the exposure-mediator effect, which will be automatically satisfied if the treatment are randomized and 4) no unmeasured mediator-outcome confounders affected by treatment. It is important to note that randomization can only rule out the confounders with exposure effect but can not rule out confounding effect associated with the mediator effect as mediators are rarely randomized.

2.4.3 Mediation Analysis in Genetic Analysis

Recently, mediation analysis has gained increasing interest in genetic studies to dissect the direct and indirect effect of genetic variants on complex diseases (13, 66, 80, 216, 210, 46, 200, 140, 109, 174). Most of these studies used data from GWAS and applied mediation methods that were developed through social science literatures or epidemiological studies, typically with an assumption that the genomic variation such as SNP, or a quantitative trait locus(QTL), acts as a causal anchor from which all arrows in the corresponding causality diagram are directed outward. However reasonable this assumption appears, cautions should be taken when the sampling scheme is not random, such as in case control studies (104).

Up to now, most of these genetic studies involve analyzing the SNP-trait-trait triads, in which only a single mediator is concerned. For example, Wang et al. used the Sobel's test for binary outcomes to evaluate the mediation effect of smoking and Chronic Obstructive Pulmonary Disease (COPD) on the relations between CHRNA5-A3 genetic locus and lung cancer risk (216), with adjustment for age and other covariates. Chen et al. (29) developed theoretical justification in the form of "causality equivalence theorem", stating the sufficient conditions required for a causal conclusion of the SNP-trait-trait triads. This method has been used, with certain modifications, in several rigorous yet conservative approaches to dissect genetic causal relationship between genotypes and phenotypes, with gene expression or methylation as potential mediators (140, 115).

Multiple mediation models have been explored in genetic studies as well. VanderWeele et al. (210) modeled smoking as a potential mediator between genetic variant on 15q25.1 and lung cancer, while allowing for interaction effect between smoking and COPD, in which mediation through smoking can account for only a small proportion of the total effect. Wang et al (217) constructed a multiple mediation model, in which

smoking and COPD constitute two steps in the causal pathway between CHRNA5-A3 variant and lung cancer. Bootstrapping was used to assess the significance of the indirect effect.

CHAPTER 3

Global Analysis of Methylation via a Functional Regression Approach

3.1 Introduction

Recent advances in high-throughput biotechnology have culminated in the development of large scale epigenome wide association studies (EWAS) (159) in which the DNA methylation at hundreds of thousands of CpGs along the genome can be simultaneously measured across a large number of samples (10, 172). EWAS have resulted in the identification of differentially methylated CpGs associated with differences in disease states, clinical outcomes, environmental exposures, or other experimental conditions (85, 184, 73, 74). These discoveries can provide a breadth of information from fundamental insights into the mechanisms underlying complex disease and to potential biomarkers for diagnosis or prognosis (98, 4).

Despite many successes, analysis of EWAS remains challenging (14). In addition to open questions concerning preprocessing and normalization (44, 203, 132, 201), association analysis with outcome variables is also difficult. Standard analysis proceeds via individual CpG analysis wherein the association between each CpG and an outcome variable (e.g. disease state, environmental exposure, etc.) is assessed one-by-one. After computing a p-value for each CpG, multiple testing criteria such as the false discovery rate (FDR) or Bonferroni corrections are applied. CpGs surviving this correction are

called differentially methylated and followed for validation and interpretation. Recently, alternative approaches based on pathway analysis have also been applied and largely mimic the analyses conducted for gene expression data.

Although individual CpG analysis has been extremely successful in identifying individual CpG sites associated with a variety of outcomes, a question of considerable interest lies in whether there is global differential methylation across the entire epigenome (52). For example, global hypomethylation is believed to occur in cancer (190, 18, 89). Global methylation analysis was traditionally conducted using assays such as examination of repetitive Alu elements and long interspersed nucleotide elements (LINE) (234). Such methods have been widely used for a wide range of diseases and experimental conditions (24, 15, 183, 55). However, these technologies are limited to primarily repeat regions which have only limited coverage on the new individual CpG resolution technologies. Thus, as the field moves towards new technologies, how to assess global methylation within the context of large scale studies with individual CpG level resolution remains unclear. Understanding global methylation is important for improved understanding of the biological systems and mechanisms of disease and also to allow for the continued relevance of traditional models as we transition towards newer technologies.

In this paper, we develop two new, related methodologies for assessing global differential methylation, either epigenome wide or restricted to a large number of CpG sites, using a functional data analysis approach. The intuition behind our approach is that global differences in methylation may be observable through differences in the overall distribution of CpG methylation levels, yet changes in a select, small subset of CpGs (which fails to reflect "global" methylation differences) will not dramatically change the entire distribution. Consequently, for our first strategy, we approximate the density of the methylation distribution for each individual using B-Spline basis functions (160).

For our second strategy, we approximate the cumulative distribution function (CDF) of the methylation distribution for each individual using B-spline basis functions. Then for both approaches, we index the entire distribution of methylation values using the estimated B-spline coefficients. To test for differential global methylation, we employ a variance component test (107) previously used for regression based analysis of gene expression (65, 112, 111) and genetic variants (229, 231).

Our approach offers a number of attractive features. First, since we are using a more robust summary measure rather than the original CpGs, the approach is therefore targeted towards comprehensive, modest changes in methylation globally. Furthermore it is robust to very strong differential methylation in a few CpGs of interest – while interesting this scenario may not reflect true global differential methylation. Second, we will employ a computationally fast variance component test from the kernel machine framework which accommodates the high degree of correlation between spline coefficients while allowing for covariate adjustment.

Finally, our variance component testing approach can be used for a range of outcome types including continuous, dichotomous, survival (21, 108), and multivariate (131) outcomes while adjusting for covariates. The ability to adjust for covariates and confounders is an important feature given recent concerns regarding the need for controlling cell type effects(76)

The remainder of this article is organized as follows. In the next section, we describe our proposed methodology for estimating the density function or the CDF of the methylation distributions for each individual. Then we describe the hypothesis testing procedure using the variance component test in Section 3.3. We assess the performance of our approach via simulations in Section 3.4. In Section 3.5, we apply the proposed work to real data sets to illustrate our approach. We conclude with a brief discussion in Section

3.2 Functional Estimation of Methylation Distributions

The idea behind our approach is that large scale, global differences in methylation will be reflected in differences between individuals in the distribution of their CpG methylation measurements. Thus, our general approach is to approximate the distribution of methylation values for each individual and then test for an association between the distributions and an outcome variable. In this section, we focus on the two proposed approaches for approximating each individual's methylation profile using functional data analytic approaches.

3.2.1 Estimation of the Density for Each Sample

Our first approach for approximating the overall methylation profile for each individual is based on approximation of each individual's density function. In short, we will compute the profile of the histogram for each individual by first creating a fine histogram of the methylation values and then fitting a B-spline to the binned histogram data. The spline coefficients will be used to summarize the profile and will be analyzed in the testing stage.

For the i^{th} sample in the study, $i = 1, \ldots, n$, suppose that the true underlying density function of the methylation values is $H_i(\cdot)$. However in practice, the actual form of this density function is unknown. Instead, we only observe methylation percentages $\{X_{i1}, \ldots, X_{im}\}$, where m is the number of observed probes, and X_{ij} is the methylation level of the j^{th} CpG on the i^{th} sample. To estimate the underlying density, we first generate a fine histogram of the methylation values. In particular, for a pre-specified large number B, we define bins $I_k = [\frac{k-1}{B}, \frac{k}{B})$ for $k = 1, \ldots, B$, and calculate the empirical relative histogram by

$$\widehat{H}_{ik} = \widehat{H}_i(t_k) = \frac{B}{m} \sum_{j=1}^{m} I\left(\frac{k-1}{B} \leq X_{ij} < \frac{k}{B}\right)$$

where t_k denotes the mid-point of the bin I_k for $k = 1, \ldots, B$. Noting that each X_{ij} is the percent methylation (between 0 and 1), then $\widehat{H}_i(t)$ with $t \in I_k$ is the density of probes falling into the k^{th} bin. In principal, B is a constant that can be tuned, and is related to the kernel bandwidth in kernel density estimation area. Larger values of B correspond to more bins and a finer histogram and better capture of small effects, yet greater sensitivity to differences generated by small changes in the overall distribution rather than global changes. Our experience suggests that setting $B = 200$ produces a reasonably fine histogram (Fig. 1a), but in practice, B is also a tuning parameter which can be selected.

Once we have constructed the histogram, we can estimate the smooth methylation profile by fitting a B-spline to the histograms to obtain a smooth curve. In particular, we take a functional data analysis view of the problem and assume that the $\widehat{H}_i(\cdot)$ is simply the observed value from the true functional process $H_i(\cdot)$. The underlying $H_i(\cdot)$ is the profile of the methylation distribution for the i^{th} sample, which we use to summarize the global methylation values for the sample. We can apply standard B-splines to model each $H_i(\cdot)$.

Briefly, B-splines are a sequence of joined polynomial segments between a series of knots which are used to model functional data. Between each pair of knots the curves are modeled as a polynomial of some order greater than 1. For a pre-specified number of interior knots R and order L of the polynomials, the total number of B-spline basis functions is given by $p = R + L$. We model the true methylation profiles $H_i(\cdot)$ by

$$H_i(t) = \sum_{\ell=1}^{p} c_{i\ell}\phi_\ell(t),$$

where $\phi_1(\cdot), \ldots, \phi_p(\cdot)$ are the B-spline basis functions, and $c_{i1}, \ldots, c_{iL}$ are unknown coefficients specific to the i^{th} sample. To estimate the coefficients, we propose to minimize

the penalized least squares criterion

$$\sum_{k=1}^{B} \left(\widehat{H}_{ik} - \mathbf{\Phi}_k' \mathbf{C}_i \right)^2 + \lambda \mathbf{C}_i' S \mathbf{C}_i,$$

where $\mathbf{\Phi}_k = \{\phi_1(t_k), \ldots, \phi_p(t_k)\}'$, $\mathbf{C}_i = (c_{i1}, \ldots, c_{ip})'$ and S is a penalty matrix. Here λ is a penalty parameter that controls the roughness of the fitted function. A larger value of λ results in a smoother estimate while a smaller values of λ produces rougher fit. The resulting estimate of the coefficient vector $\mathbf{c}$ has a closed form and can be computed using standard penalized least squares estimation.

Two important issues in this context are the number and placement of the knots, and the choice of the penalty parameter λ. Since methylation percentages are between 0 and 1 and approximately bimodal, we place more knots in the areas with strong curvature (closer to 0 and 1) and fewer knots in between. In general, we observed that 25-35 knots with polynomial order 4 seems to be a reasonable model for the data. Regarding the choice of λ, there are many available data based methods such as leave one out cross-validation, generalized cross-validation and the restricted maximum likelihood criteria, see e.g., (169). In this article we use generalized cross-validation (GCV) method to select λ.

Although $\widehat{H}_i(\cdot)$ can be thought of as an approximation of the density for the methylation values, strictly speaking, adjustments are needed to ensure that it has the properties of being a probability density function. However, since we are simply using the profile of the histogram as a tool for summarizing the entire profile of methylation values, this is not necessary from the perspective of testing.

3.2.2 Estimation of the Cumulative Distribution Function

Our second approach for approximating the overall methylation profile for each individual is based on approximation of each individual's cumulative distribution function

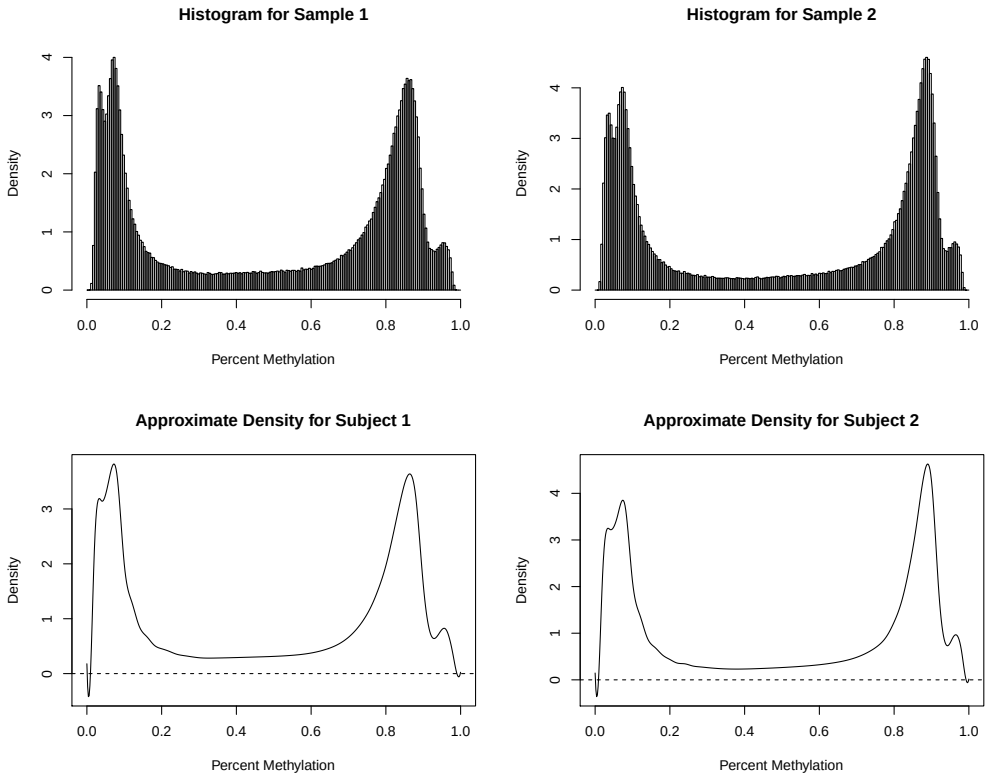

Figure 3.1: Approximating the density for each sample: Example histograms for two samples and the corresponding B-spline approximated densities.

(CDF). Similar to before, we will estimate the empirical CDF (ECDF) and then fit a B-spline to the ECDF. The spline coefficients will again be used to summarize the profile and will be analyzed in the testing stage. The advantage of this approach is two-fold: first, binning to create a histogram is no longer necessary and second, sensitivity of results to knot placement is mitigated.

For the i^{th} sample in the study, $i = 1, \ldots, n$, we assume that the true CDF is $F_i(\cdot)$, and estimate the ECDF as:

$$\widehat{F}_{ik} = \widehat{F}(t_k) = \frac{1}{m} \sum_{j=1}^{m} I(X_{ij} \leq t_k),$$

where $\{t_k, k = 1, \ldots, B\}$ form an equally spaced grid of B points in $[0, 1]$.

In constructing a basis for the CDF, we again use a grid of 35 knots between 0 and 1 due to the nature of methylation data, but in contrast to modeling the density function, we space the knots evenly since the difference in curvature is no longer as apparent. We again assume a B-Spline basis representation for the true CDF with order 4 basis functions and write

$$F_i(t) = \sum_{\ell=1}^{p} c_{i\ell}\phi_\ell(t)$$

where $\phi_1(\cdot), \ldots, \phi_R(\cdot)$ are the B-spline basis functions. As with the estimation for the density functions, the unknown coefficients for the i^{th} subject can be estimated using penalized least squares with $\widehat{F}_{ik}, k = 1, \ldots, B$ as the responses, and the smoothing parameter λ can be estimated using generalized cross validation criterion. Once we have obtained the coefficients for the B-splines for each individual, we can test for association with the outcome variable.

3.3 Variance Component Test in Approximated Distributions

After applying B-splines to approximate either the density or the CDF for each sample in the study, we allow the B-spline coefficients to index the entire distribution. Consequently, to test for global changes in methylation, we need only test whether the spline coefficients are associated with the outcome. To do this while accommodating potential confounding variables and the (typically) high correlation between B-spline coefficients, we propose to use the variance component test used within the SKAT framework for genotype analysis(229, 231, 96).

Here and in the sequel we let $\mathbf{C}_i = [c_{i1}, c_{i2}, \ldots, c_{ip}]'$ denote the vector of B-spline coefficients for the i^{th} individual in the study and $\mathbf{Z}_i$ be a vector of covariates for which we would like to control. We further let y_i denote the outcome of interest. For simplicity, we focus on univariate continuous or dichotomous outcomes, but our framework generalizes naturally to other outcomes such as survival times or multivariate

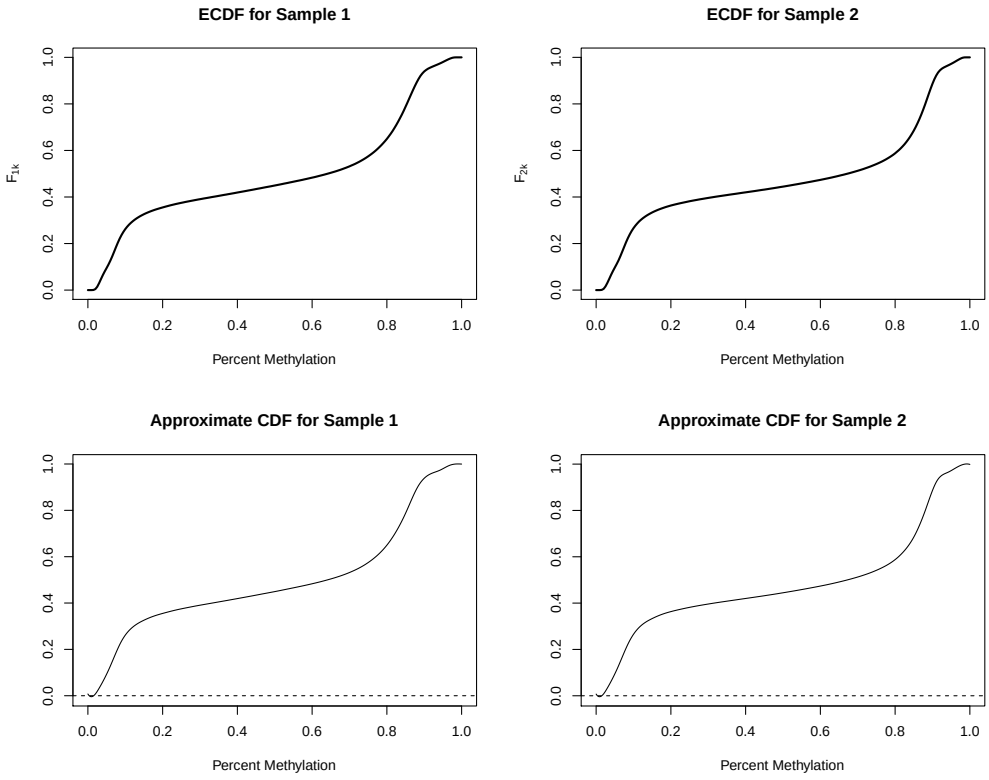

Figure 3.2: Approximating the CDF for each sample: example ECDFs for two samples and the approximated B-spline approximations of the CDFs.

measurements. The objective is to test for association between $\mathbf{C}_i$ and y_i while adjusting for $\mathbf{Z}_i$.

Natural models for relating the variables of interest to the outcome are the linear model

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z}_i + \sum_{j=1}^{p} \beta_j c_{ij} + \varepsilon_i \tag{3.1}$$

for continuous outcomes and the logistic model

$$\text{logit}P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z}_i + \sum_{j=1}^{p} \beta_j c_{ij} \tag{3.2}$$

for dichotomous outcomes, where we define α_0 to be an intercept, $\boldsymbol{\alpha}$ and β_j to be the

regression coefficients corresponding the to covariates and each B-spline coefficient, and ε_i to be a random error term with mean 0 and variance σ^2.

To test for an association between $\mathbf{C}$ and $\mathbf{y}$ corresponds to testing:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0. \tag{3.3}$$

In principle, this can be done using a p-df test, but the $\mathbf{C}$ tend to be highly correlated and p can be large such that power is low. An alternative approach is to assume that the β_j follow some arbitrary distribution $G(\cdot)$ with mean 0 and variance τ. Then τ indexes the significance of the entire group of B-spline coefficients and then testing (3.3) is equivalent to testing

$$H_0 : \tau = 0, \tag{3.4}$$

which can be done using a variance component score test. In particular, for continuous outcomes we can construct the score statistic

$$Q = \frac{(\mathbf{y} - \widehat{\alpha}_0 - \mathbf{Z}\widehat{\boldsymbol{\alpha}})'\mathbf{C}\mathbf{C}'(\mathbf{y} - \widehat{\alpha}_0 - \mathbf{Z}\widehat{\boldsymbol{\alpha}})}{\widehat{\sigma}^2}$$

where $\widehat{\alpha}_0$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\sigma}$ are estimated under (3.4). Similarly, for dichotomous outcomes we can construct the score statistic

$$Q = (\mathbf{y} - \widehat{\mathbf{y}})'\mathbf{C}\mathbf{C}'(\mathbf{y} - \widehat{\mathbf{y}})$$

where $\widehat{\mathbf{y}} = \text{logit}^{-1}(\widehat{\alpha}_0 + \mathbf{Z}\widehat{\boldsymbol{\alpha}})$ and both $\widehat{\alpha}_0$ and $\widehat{\boldsymbol{\alpha}}$ are again estimated under the null, (3.4).

Under the null hypothesis, Q asymptotically follows a mixture of chi-squares distributions. In particular, $Q \sim \sum \lambda_\ell \chi^2_1$ where λ_ℓ are the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{C}\mathbf{C}'\mathbf{P}_0^{1/2}$ and $\mathbf{P}_0 = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for continuous outcomes and $\mathbf{P}_0 = \mathbf{D} - \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$

for dichotomous outcomes with $\mathbf{D} = \text{diag}\{\widehat{y}_i(1-\widehat{y}_i)\}$. This distribution can be approximated use moment matching methods(112, 113) or exact approaches(42, 43, 45) allowing for easy p-value computation.

A key advantage of using the variance component testing framework is that the degrees of freedom of the test adjust naturally to the correlation among the B-spline coefficients. In fact, if the coefficients are perfectly correlated, then the test reduces to a single degree of freedom test since the number of nonzero eigenvalues of $\mathbf{CC}'$ is only one.

3.4 Simulations

We assessed the size and also the power of our proposed approaches for global analysis of methylation profiles.

3.4.1 Type I Error

To examine type I error for both dichotomous and continuous outcomes, we simulated 2,000 data sets containing methylation profiles for n individuals. We generated data under the null by simulating 10,000 methylation values for each individual, from a beta distribution with parameters 0.45 and 0.55. Dichotomous outcomes by assigning $n/2$ of the individuals were to be "controls" ($y_i = 0$) and $n/2$ individuals to be "cases" ($y_i = 1$). For continuous simulations, the outcome variable y_i was simulated as a standard normal independently of the methylation profiles. To each of the simulated data sets, we applied both of the proposed strategies for global analysis methylation profiles. Specifically, we used B-splines to approximate the density of each individual's methylation distribution an to approximate the CDF of each individual's methylation distribution. For the density estimation, we constructed histograms using 200 evenly spaced bins between 0 and 1. The knots for the B-spline were spaced at intervals of

Table 3.1: Type I error simulation results.

	Dichotomous		Continuous	
n	Density	CDF	Density	CDF
40	0.054	0.050	0.046	0.044
60	0.052	0.052	0.049	0.054
80	0.056	0.052	0.046	0.046
100	0.052	0.047	0.044	0.052
500	0.055	0.044	0.045	0.045

0.02 between 0 and 0.3 and between 0.7 and 1. Since the region in the center is less variable, knots were placed at intervals of 0.1 in length. For the CDF estimation, we estimated the ECDF at 1000 evenly spaced points between 0 and 1. A number of 35 knots for B-spline estimation were also evenly spaced between 0 and 1. After estimating the spline coefficients for approximating the density and for approximating the CDF, we applied the variance component score test to test for association between the spline coefficients and the outcome variable. We allowed n to vary as 40, 60, 80, 100, and 500. For each choice of n, we estimated the type I error rate as the proportion of p-values less than $\alpha = 0.05$ across the 2,000 simulations.

The type I error results for both dichotomous and continuous outcomes are presented in Table 3.1 and indicate that the size of the proposed test is correctly controlled at the 0.05 level, even when the sample size is modest.

3.4.2 Power

We examined the power of the proposed approaches for dichotomous outcomes. In particular, we simulated $n/2$ individuals, designated as "controls", each with 10,000 methylation measurements from a beta distribution with parameters 0.45 and 0.55. Then for each of $n/2$ "cases", we simulated $\gamma \times 10,000$ observations from a beta distribution with parameters 0.55 and 0.45 with the remaining $(1-\gamma) \times 10,000$ again sampled from a beta distribution with parameters 0.45 and 0.55. γ indexes the strength of the

difference in the methylation profile between cases and controls and was allowed to take values between 0 and 0.5. We again let n vary as 40, 60, 100, and 500. For each choice of n and γ, we simulated 200 data sets and assessed the power of both the density based and CDF based testing procedures as the proportion of p-values less than $\alpha = 0.05$. Approximation of the density and CDF were done as in the type I error simulations.

The power as a function of $\log_{10} \gamma$ is plotted in Figure 3.3 for both the density and CDF based approaches. As anticipated, power grows as a function of both n and γ for both approaches. Importantly, when γ is small, neither approach has much power, which is not necessarily undesirable since the objective is to identify scenarios in which there is an observable change in the overall distribution rather than a few significantly different probes. Overall, using the CDF tended to yield higher power than the density based approach. This is, in part, due to the fact that using the density based approach requires an additional layer of smoothing which can reduce modest effects. However, lack of power for small values of γ may actually be more meaningful in terms of the biological objective.

3.5 Data Applications

We illustrate our proposed methods for global analysis of methylation profiles via application to some real data sets.

3.5.1 Epigenetic Comparison of Newborns and Nonagenarians

Methylation levels at key genes and genome wide are believed to vary with age (224, 95, 165, 129, 82, 72). Recently, Heyn et al. (74) conducted a study to examine differences methylation between blood from newborn infants and nonagenarians. We illustrate our proposed methods for global analysis via application to this data set which was available from the Gene Expression Omnibus (GEO) (47) (`http://www.`

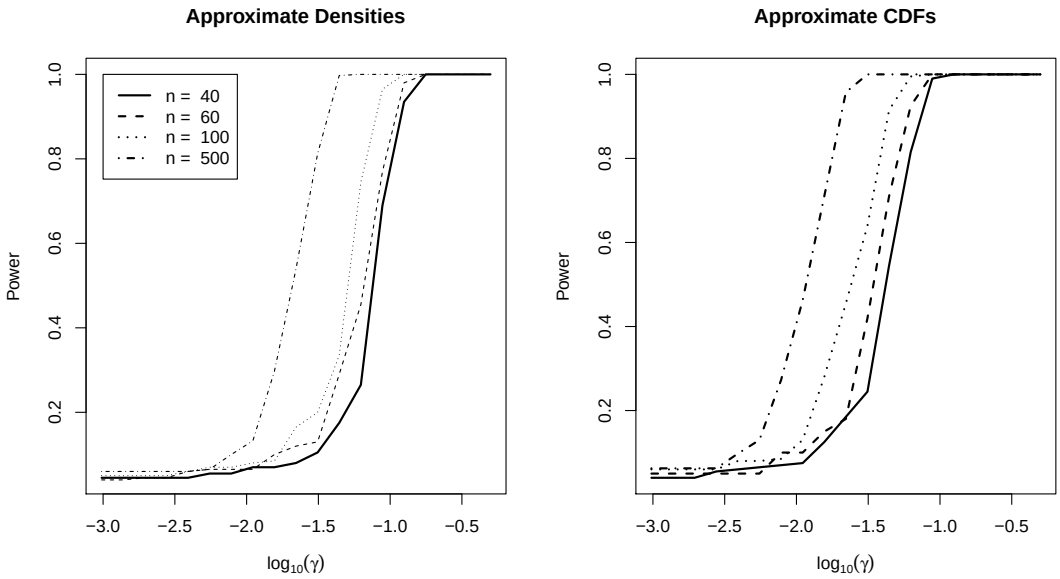

Figure 3.3: Power simulations results.

ncbi.nlm.nih.gov/geo/) under accession number GSE30870. Briefly, the study used the Illumina HumanMethylation450 array to measure methylation at approximately 485,000 CpGs genome wide in blood from 20 nonagenarians and 20 newborn infants. For our analysis, we restricted attention to the approximately 470,000 autosomal CpGs.

We applied both the density and CDF based analysis procedures to the data set to test for global differential expression between newborns and nonagenerians. As in the simulations, for the density based approach, we again computed the relative histogram using 200 evenly spaced breaks between 0 and 1, and then we approximated the density using B-splines with knots placed at intervals of 0.02 between 0 and 0.3 and between 0.7 and 1.0. Between 0.3 and 0.7, knots were evenly spaced at intervals of 0.10. For the CDF based approach, we computed the ECDF for each sample at a grid of 1000 values

between 0 and 1 and then approximated the CDF using B-splines with knots placed at 35 evenly spaced intervals between 0 and 1. For both approaches, GCV was used to estimate the B-spline smoothing parameter λ. The approximate densities and CDFs are shown in Figure 3.4. We then used the variance component test under a logistic model to regress a binary indicator for whether each subject was a nonagenerian on the B-spline coefficients from the the approximate density or from the approximate CDF. No additional covariates were considered.

Using the density based approach, we obtain a p-value for association of 0.265 which fails to meet significance, but on the other hand, if we use the CDF based approach, we obtain a p-value of 0.024. The significant CDF based approach appears to better reflect the authors' observations that the newborn infants tended to have greater methylation genome wide and the nonagenarians tended to have hypomethylation at key genes.

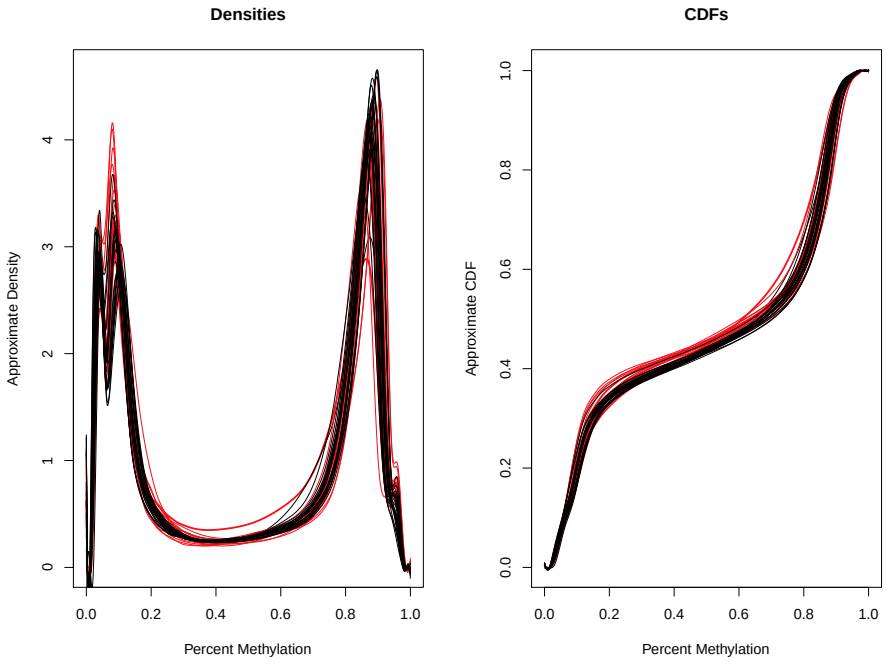

Figure 3.4: Approximate densities and CDFs from the nonagenarian study. Red curves are the nonagenarian methylation profiles and black curves are the infant methylation profiles.

3.5.2 Head and Neck Squamous Cell Carcinoma Methylation Study

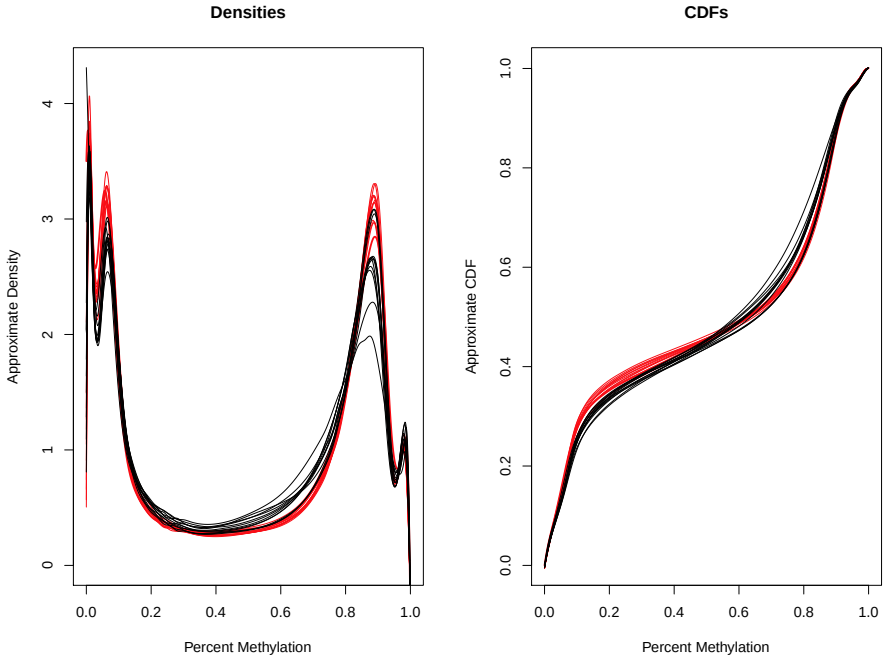

Figure 3.5: Approximate densities and CDFs from the head and neck squamous cell carcinoma study. Red curves are the methylation profiles for the cancer cases and black curves are the methylation profiles for the healthy controls.

Methylation is known to play an important role in a wide variety of cancers including head and neck squamous cell carcinoma (HNSCC) (51, 171, 166). Thus, in a second data example, we applied the proposed methodology to determine if global methylation differences are observed between blood from subjects with HNSCC and healthy controls using a data set available from GEO (accension number GSE40005). The data included genome wide methylation measurements at approximately 485,000 CpGs in blood from 12 cases with HNSCC and 12 controls.

We applied the proposed density and CDF based approaches as in the the study of newborns vs nonagenarians to test for association between global methylation and case-control status while adjusting for age and gender. The approxmate densities and

CDFs for each sample are shown in Figure 3.5. Using the density based approach, we obtain a p-value of 0.0016 while using the CDF based approach we obtain a p-value of 0.00015, both of which are highly significant suggesting that large scale differences in the overall methylation distribution are associated with cancer. This is again reflective of prior knowledge indicating that cancer is associated with large scale differences in methylation (136, 154).

3.6 Discussion

In this article, we propose two new strategies for global analysis of methylation profiles which is based on approximation of either the density of the methylation values for each individual or the CDF of the methylation values for each individual using a functional regression approach. Specifically, by indexing each individual's methylation distribution using B-spline basis coefficients, we can test for association between the overall methylation distribution and an outcome variable, while adjusting for additional covariates, by simply testing the spline coefficients. Although the proposed method tests the global null hypothesis, a key advantage of the proposed method is that we are essentially applying smoothing when we approximate the density or the CDF using B-splines. Therefore, this reduces the influence of single (or a few) probes strongly associated with the outcome.

Overall, of the two proposed methods, the CDF based approach tends to have higher power, as reflected by greater statistical significance in the data applications and in the simulations, due to the additional layer of smoothing done when approximating the density which arises from computing the relative histograms. However, the lower power of the density based approach also means that it is sensitive only to true large scale, global methylation differences.

For hypothesis testing, we focus on testing the spline coefficients using a variance

component test in which the outcome is regressed on the spline coefficients. This allows for natural accommodation of the high correlation among the spline coefficients since the degrees of freedom of the test adapt to the correlation while adjusting for covariates. However, alternative testing procedures are also possible. For example, one could also treat global methylation as the outcome and use a Hotelling's T^2 test or MANOVA to asses significance. While our variance component testing approach and other tests could all protect type I error, alternative methods may yield improved power if the underlying models better reflects the true state of nature.

Our proposed methodology opens doors to new areas of research. First, although we focus on testing global methylation across all CpGs, the approach can be restricted to specific subsets of CpGs such as CpGs falling within specific epigenetically relevant features (e.g. CpG islands, promoters, repeats, etc.) or the CpGs within a particular gene pathway thereby enabling a set or pathway based analysis that tests the global null hypothesis but is more geared towards a true pathway effect. Second, while we have explored the relationship between global methylation and a single dichotomous or continuous outcome, alternative outcome types are possible and warrant further exploration. Finally, while our work focuses on testing the overall methylation distributions, the idea of using a functional regression approach to summarize the overall distribution can also allow for understanding the relationship between outcome variables and other covariates while in the pressence of global methylation differences, i.e. adjusting for the effect of methylation. This is important since methylation can serve as a potential confounder in biological models and adjustment for this can be important. Such explorations remain for future research.

CHAPTER 4

Microbiome Kernel Machine Profiling

4.1 Introduction

The advent of massively parallel sequencing has transformed the field of metagenomics and enabled high-throughput profiling of microbiota in a large number of samples via targeted sequencing of the 16S rDNA gene. Knowledge on how microbial communities differ across individuals can provide key information on the role of communities in relation to variation in biological and clinical variables and is essential for gaining a broader understanding of biological mechanisms underlying disease and response to exposures. Although considerable resources have been devoted to sequencing technologies and to quantifying individual taxa, successful application of microbial profiling to studying biomedical conditions requires novel statistical methods to efficiently test for associations with microbial diversity.

Current strategies for evaluating the association between microbiota composition and outcomes of interest has focused largely on distance based analysis. Using standard analyses, the phylogenetic distance based on the quantified taxa, organized into operational taxonomic units (OTUs), is computed between each pair of samples in the study. Multivariate analysis or the top principal coordinates (PCo) of the matrix of pairwise distances are used to test for associations via permutation. Commonly used pairwise distance metrics include weighted and unweighted UniFrac (118, 117, 119, 25, 68) as

well as many other important metrics such as the Bray-Curtis (17) metric. While there are similarities between metrics, their performances vary and are targeted to specific scenarios.

Although distance based analyses have been successful in identifying many outcomes related to community variability, a key limitation of the existing distance based methods is that they fail to allow for easy covariate adjustment which is essential in order to control for confounders. Failure to control for confounders can easily lead to spurious results through false positives or through attenuated evidence for association. Existing methods are not easily modified to accommodate covariates due to model formulations and due to the reliance on permutation. Further, PCo based analyses implicitly assume the top PCo's capture the variability that is attributable to the outcome of interest. However, since the PCo's are unsupervised and computed without regard to the outcome variable, there is no guarantee that they capture the appropriate signal, possibly resulting in considerable power loss. Finally, given the wide range of distance metrics available, each with differing performance under different scenarios, it can be difficult to choose a particular metric to use. The best metric for any particular data set depends on the underlying true state of nature, which is unknown *a priori* - knowledge on this would generally preclude need for analysis. Using multiple metrics and cherry picking the best result will result in inflated type I error rates and lead to large numbers of spurious results. New methods are needed.

We propose in this paper the microbiota regression-based kernel association test (MiRKAT), a flexible, computationally efficient regression approach for testing the association between microbial community profiles and a continuous or dichotomous variable of interest such as an environmental exposure or disease status, while adjusting for confounders. Our test uses the kernel machine regression framework, previously developed for genotyping data (96, 229, 231), to directly regress the variable of interest

on the covariates (including potential confounders) and the microbiome compositional profiles. An analytical p-value for the association between community profiles and the outcome is rapidly computed via a variance component score test. Intuitively, the kernel machine framework will compares pairwise distance/similarity in the outcome variable to pairwise distance/similarity in the microbiota community profiles such as measured through UniFrac metric or another valid metrics. Consequently, MiRKAT allows for fast, supervised, distance-based association testing under a regression framework that permits controls for potential confounding.

Because the best distance metric to use for testing the association between a particular outcome and microbial diversity is generally unknown, we also develop an "optimal" MiRKAT that simultaneously examines multiple distance metrics, selects the best distance metric to compute a p-value for association, and adjusts for having taken the optimal distance metric. The power of the optimal test is generally close to the power from using the best distance metric and is always better than using poor choices of distances. Thus, this allows for good power in the omnibus while still protecting the type I error rate.

We demonstrate through simulations and analysis of metagenomic studies of irritable bowel syndrome (IBS) and smoking that MiRKAT is often more powerful than existing tests with improved control for type I error across a broad range of models for both continuous and dichotomous variables.

4.2 Methods

4.2.1 Notation

Assume that n samples have been sequenced and microbial communities profiled. For the i^{th} subject, let y_i denote outcome variable of interest, $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})$ corresponds to the abundances of individual OTUs, and $X_i = (X_{i1}, X_{i2}, ..., X_{im})$ are

additional covariates that we want to control for, such as age, gender, and other clinical and environmental variables which are known to influence microbial community diversity and be related to a range of outcomes. The goal is to test for association between the outcome and microbial profiles while adjusting for the covariates X. Note that we will refer to y as an "outcome" that depends on the microbiome composition while in some situations it may be a variable that is thought to influence microbial diversity; however, since our goal is association testing rather than causal modeling, the distinction is unimportant given the duality (65).

4.2.2 Distance Based Association Test for Microbiome Composition

Distance based approach is the most commonly used strategy for association test between microbiota composition and outcomes of interest, which relies on Mental-Carlo generation of p-value using permutation. For the case when only a single distance metric is considered, the method PERMANOVA (Permutational Multivariate Analysis of Variance Using Distance Matrices)(137) can be summarized as follows:

1) Construct an $n \times n$ distance matrix D for all pairs of samples based on the microbiome composition data Z.

2) Obtain centered similarity matric $G = (I - 11'/n)A(I - 11'/n)$, with $A = (-D_{ij}^2/2)$

3) Calculate the projection matrix $H = y(y'y)^{-1}y'$

4) Construct the psedo-F statistic $F = \frac{tr(HGH)/(m-1)}{tr((I-H)G(I-H))/(n-m)}$, where $tr(A)$ is the trace of matrix A.

5) Calculate p-value using permutation by shuffling y.

For microbiome composition data, the OTUs are related by a phylogenetic tree. Phylogenetic distance measures that exploit the degree of divergence between different

sequences are usually much more powerful compared to distance measures that ignores the phylogenetic tree information. Microbiologists have proposed many distance metric to efficiently incorporate the phylogenetic relationship. The most widely used distance metrics are the UniFrac distance (118, 119, 68), which "measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both"(118). The unweighted UniFrac distance (118), denoted as D^U, uses only OTU absence or presence information and measures the fraction of branch length of the phylogenetic tree that are unique to any microbial community. The weighted UniFrac distance (D^W)(120) uses OTU abundance information and weights the branch length by abundance difference between the two communities. The generalized UniFrac distance (28) can be considered as an extension of the weighted and unweighted UniFrc distance with one additional parameter ξ, which can be represented as $D^{(\xi)}$. When $\xi = 1$, the generalized UniFrac distance $D^{(1)}$ corresponds to the weighted UniFrac distance D^W and when $\xi = 0$, the $D^{(0)}$ is reduced to the unweighted UniFrac distance D^U. Beyond the UniFrac distance family, Bray-Curtis (17) is another popular distance for quantifying the compositional dissimilarity between two different samples, which is equivalent to the Sorensen similarity index (194).

Each distance metric focuses on different aspect of microbiome changes and can be most powerful in detecting only a certain scenario. In practice, the optimal choice of distance is unknown in prior in metagenomic studies. Chen et al.(28) developed a method that combines different distance matrices in a single test by taking the maximum of pseudo-F statistics for each distance matrix. Significance is assessed by permutation.

4.2.3 Microbiota Regression-based Kernel Association Test

MiRKAT exploits the kernel machine regression framework to relate the covariates and the microbiota profiles to the outcomes. Specifically, for a continuous outcome variable we use the linear kernel machine model:

$$y_i = \alpha_0 + \boldsymbol{\alpha}' X_i + f(Z_i) + \varepsilon_i \tag{4.1}$$

and for a dichotomous outcome variable (e.g. y = 0/1 for case/control) we use the logistic kernel machine model

$$\text{logit}(P(y_i = 1)) = \alpha_0 + \alpha' X_i + f(Z_i) \tag{4.2}$$

where α_0 is the intercept, $\alpha = [\alpha_1, ..., \alpha_m]'$ is the vector of regression coefficients for the m covariates, and for continuous phenotypes ε_i is an error term with mean zero and variance σ^2.

The relationship between the metagenomic profile and the outcome variable is fully characterized by $f(\cdot)$. Technical mathematical details are omitted, but under the kernel machine framework, $f(Z_i) = \sum_{i'=1}^{n} \gamma_{i'} K(Z_i, Z_{i'})$ for some $\gamma_1, \gamma_2, ..., \gamma_n$. The $K(Z_i, Z_{i'})$ is some positive definite metric of similarity between the i-th and i'-th samples in the study based on their metagenomic profiles. Different choices of $K(Z_i, Z_{i'})$ result in different models. For example, setting $K(Z_i, Z_{i'}) = \sum_{j=1}^{p} Z_{ij} Z_{i'j}$ implies $f(Z_i) = \sum_{j=1}^{p} Z_{ij}\beta_j$, i.e. the model is linear. Replacing the kernel function with more sophisticated similarity metrics results in more complex models.

Within the context of metagenomic studies, a wide range of similarity metrics can be defined. Although existing work focuses primarily on distance metrics for compositional dissimilarity, these can be trivially transformed to similarity metrics. For example, if we compute the UniFrac distance between sample i and the sample i' as $D^U(Z_i, Z_{i'})$, then a

corresponding similarity can be computed as $K_U(\mathbf{Z}_i, \mathbf{Z}_{i'}) = 1 - D^U(\mathbf{Z}_i, Z_{i'})$. Similarly, if we define $D^{BC}(Z_i, Z_{i'})$ to be the Bray-Curtis distance between sample i and the sample i', then the corresponding similarity metric is $K_{BC}(Z_i, Z_{i'}) = 1 - D^{BC}(Z_i, Z_{i'})$. Other similarity metrics can be analogously defined based on commonly used distance metrics.

The goal is to test for association between microbiome composition and the variable of interest. Since the relationship between the y_i and the Z_i is fully determined by the function $f(Z_i)$, then this is equivalent to testing the null hypothesis that $H_0 : f(Z) = 0$.

Through an important relationship between kernel machine regression and mixed models (112, 111, 62), it has been shown that $f(Z)$ can be viewed as a subject specific random effect which follows a distribution with mean 0 and variance τK where τ is a constant and K is the $n \times n$ "kernel matrix" with $(i, i')^{th}$ term equal to $K(Z_i, Z_{i'})$. Then testing for an association between the microbiome composition and the outcome is equivalent to testing the null hypothesis that $H_0 : \tau = 0$. Under the connection with mixed models, this can be done using a standard variance component score test (107). A key advantage of the score test is that it only requires fitting the null model $y_i = \alpha_0 + \alpha' X_i + \varepsilon_i$ for continuous traits and $\text{logit}(P(y_i = 1)) = \alpha_0 + \alpha' X_i$ for dichotomous traits.

In particular, the score statistic is given as

$$Q = (y - \hat{\mu})' K (y - \hat{\mu}) \tag{4.3}$$

$\hat{\mu}$ is the predicted mean of y under H_0, i.e. $\hat{\mu} = \hat{\alpha}_0 + \hat{\alpha}' X$ for continuous traits and $\hat{\mu} = \text{logit}^{-1}(\hat{\alpha}_0 + \hat{\alpha}' X)$ for dichotomous traits; and $\hat{\alpha}_0$ and $\hat{\alpha}$ are estimated under the null model by regressing y on only the covariates X.

Under the null hypothesis, Q asymptotically follows a mixture of chi-square distributions. In regular KMR, p-value can be analytically obtained through moment

matching (113) or exact methods (41, 45). However, in the cases of microbiome composition data, this asymptotical approximation is shown to be too conservative because of the high sparsity of OTU tables. Instead, we propose a residual permutation approach in which we constructed empirical null distribution of Q by generating a large number N of $Q^* = (y^* - \hat{\mu}^*)' K (y^* - \hat{\mu}^*)$ where $(y^* - \hat{\mu}^*)$ is a permutation of the residuals $(y - \hat{\mu})$ in score statistic Q. Then $p = \sum I(Q < Q^*)/N$.

Note that the proposed test is a score test, such that all parameters are estimated under the null hypothesis, i.e. $f(Z)$ does not need to be estimated. Although the permutation requires Monte-Carlo calculation of p-value, it requires fitting a simple linear or logistic regression model once and randomly resample from the corresponding residuals, which can still be efficient. As all the parameters are estimated under the null model, this means that even if a poor similarity metric is chosen, the test is still statistically valid. Better choices of similarity metrics simply improve power. From the perspective of testing, a metric that better captures similarity and that better reflects the true relationship between the metagenomic compositional profiles and the outcome will result in higher power.

4.2.4 Optimal MiRKAT under Multiple Distance Metrics

A problem of significant practical interest in metagenomic analysis is selecting an appropriate distance metric. Different distance metrics are targeted towards different scenarios and consequently yield differential power. However, the best distance metric depends on the underlying true state of nature which is not known prior to analysis. MiRKAT can also lose power if poorly chosen distance metrics are used, though the type I error is protected. Therefore, we develop the optimal MiRKAT which simultaneously examines multiple distance metrics and optimally creates an omnibus test across all possible metrics. The intuition behind the approach is that the optimal MiRKAT will

consider testing using each distance metric, select minimum p-value from all of the distance metrics, and then adjust for having taken the minimum using rapid perturbation methods tailored towards the kernel machine testing set up. Perturbation is similar to permutation except the distribution of the score statistic is used. Consequently, perturbation is often faster than permutation can allow for covariate adjustment even under correlation whereas permutation does not. Technical and algorithmic details are presented in Appendix I.

4.2.5 Numerical Experiments and Simulations

We simulated data based on the strategy of Chen et al (28). In particular, we use a phylogenetic tree that are obtained from a real throat microbiome data set (26) for OTU data simulation. We estimate the mean OTU proportions using the dirichlet distribution from the real data, the parameter of which are subsequently used to simulate OTU count data using multinomial distribution.

We considered several different simulation scenarios to examine the type I error and power of our proposed MiRKAT method, compared with distance based regression. For all scenarios, we considered 2000 simulations with sample size $n = 50, 100$ and 200. We also considered different situations when the highly abundant OTUs or when only the less abundant OTUs have effect on the outcome. For simplicity, we would only show the simulation result for the case when the highly abundant OTUs are of effect with $n = 50$. For simulations when the less abundant OTUs are of effect, the type I error was similar, the power was smaller to the case when the highly abundant OTUs are of effect.

Data were simulated as $y_i = x_{1i} + x_{2i} + \beta h(z_{i1}, z_{i2}, ..., z_{ip}) + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$, x_{1i} and x_{2i} be covariates to be adjusted for and $(z_{i1}, z_{i2}, ..., z_{ip})$ be the abundance of the 10 most abundant OTUs.

We considered situations when X and Z are correlated and when X and Z are independent. In both cases, x_{1i} are simulated as Bernoulli random variable with success probability 0.3. In the case when X and Z are independent,$x_{2i} = \mu_i$, where $\mu_i \sim N(0,1)$. For the case when X and Z are correlated, we let $x_{2i} = \mu_i + cos(\sum z_{ij}), j = 1, ..., p$, where $\sum z_{ij}$ is the total OTUs measurement for the 10 most abundant OTUs.

$h(z_{i1}, z_{i2}, ..., z_{ip}) = \boldsymbol{\beta}' Z$ where $\beta_j = 1/(\bar{z}_j)$ where $\bar{z}_j$ is the average abundance for the j^{th} OTU across all samples, so that the OTUs with higher abundance would have smaller effect size. We tested under the case when $\beta = 0$ for type I error, and assessed power of the tests by changing the values of the coefficient β.

A wide range of distance matrices were constructed using the Unifrac packages from the OTU data $\mathbf{Z}$: $D^W, D^U, D^{(0.5)}, D^{(0)}$ and D^{BC}, which represent the weighted and unweighted UniFrac distance, Generalized UniFrac distance with $\xi = 0.5$ and $\xi = 0$ and the Bray-Curtis distance. We tested for associations between microbial OTUs Z and the trait y using each of the distance metric, with or without adjustment for additional covariates X. Specifically, we applied generalized Unifrac method and our proposed MiKRAT tests using each of these five distance matrices. We also applied omnibus tests which consider all these matrices using PermanovaG (Permutational multivariate analysis of variance using multiple distance matrices) from GUniFrac packages (28) for the distance based approach and our proposed MiKRAT methods, adjusting or not adjusting for confounders X.

4.3 Results

4.3.1 Type I Error Control for MiRKAT and Competing Methods

The type I error of these different simulation situations were shown in Figure 4.3.1. For cases when X and Z are independent, both distance based methods and our proposed MiRKAT are valid, with or without adjusting for covariates X. However, when

X and Z are correlated, all distance based methods produced seriously inflated type I error, even after adjusting for X covariates in the model. The only model that had correctly controlled type I error when X and Z are correlated is the MiRKAT method with covariates X adjustment.

4.3.2 Statistical Power for MiRKAT and Competing Methods

When X and Z are independent, both MiKRAT and distance based methods have valid type I error and can be compared with respect to power. Figure 4.2 shows the power of all constructed tests under this simulation scenario. The power increases for all methods with the increase of association strength. When the same distance matrices were used, the MiKRAT and corresponding distance based method have similar power without adjusting for covariates. In fact, the two methods were shown to be equivalent when the same positive definite matrix was used as similarity matrix in distance based methods and kernel machine regression (152).

Adjusting for X would increase power in MiRKAT, possibly due to reduced variance of estimated residuals. However, adjusting for X in distance based method didn't increase power. For this specific case of our simulation when only the highly abundant bacterial have effect in outcome, the Bray-Curtis distance matrix produced highest power as opposed to other distance matrices, and the proposed omnibus test MiKRAT was the second most powerful test. As have been shown, each distance matrix can perform best under certain scenarios; none of them can have the optimal performance under all conditions. For example, for the simulations under which the less abundant OTUs are of effect, the unweighted Unifrac distance D^U and D^0 had better power than using D^{BC}. The Omnibus MiKRAT method can unify all possible distance matrices and thus be the most robust test for all situations.

For situations when X and Z are correlated, all methods except the proposed

MiRKAT method with X adjustment had seriously inflated type I error, suggesting that adjusting for all potential confounder is an important aspect of this proposed method.

4.3.3 Application to IBS and Smoking Data Sets

We illustrated our methods by application to two data sets. The first data set was the same as analyzed in Chen et al (28),which investigated the smoking effect on the oropharyngeal and nasopharyngeal bacterial community using 454 pyrosequencing of 16S sequence tags (26). The details of the data set can be obtained elsewhere (26, 28). Generally, swab samples were collected from right and left nasopharynx and oropharynx of 29 smoking and 33 nonsmoking adults. The variable region 1-2 (V1-V2) of the bacterial 16S rRNA gene was PCR amplified and subject to multiplexed pryosequencing. OTUs were constructed using QIIME pipeline. For this specific data set, only left oropharyngneal samples were included. Samples with read number less than 500 and OTU with only one read were removed, resulting in an OTU table of 60 samples (28 smokers vs 32 nonsmokers) and 856 OTUs. Covariates in this data included gender, antibiotic use within 3 months and respiratory disease. Chen et al tested association between smoking status and microbial community composition using distance based approach by applying PERMANOVA method, without adjusting for covariates in the model. However, other covariates can be associated with smoking status as well as microbial community composition, thus becoming a confounder. For example, the odds ratio of smoking between males and females is 2.33 in this data set, making gender a potential confonder. We applied our MiKRAT method to analyze the association between smoking and microbial community composition, using the same set of Unifrac distance matrices $D^W, D^U, D^{(0.5)}, D^{(0)}$,and D^{VAW} as in Chen et al as similarity matrix, but adjusted for gender, antibiotic use and respiratory disease. Similar to the findings

from Chen et al, all of the distance matrices achieved statistical significance at 0.05 level, and test using $D^{(0.5)}$ produced the smallest p-value 0.0014, followed by 0.0017 using $D^{(0)}$ .The p-values from D^W, D^U and D^{VAW} are 0.0031 0.0086 and 0.0149 respectively. The Omnibus tests generated a p-value of 0.0042, indicating that smoking can affect the microbial community composition significantly.

Irritable bowel syndrome (IBS) is a functional bowel disorder characterized by abdominal pain and disturbed bowel habits. Substance P(SP) is an important excitable neurotransmitter associated with inflammation and pain in IBS. The second data set we evaluated included samples from 23 IBS patients and 23 healthy adults. We tested the association between microbial community composition and disease status/SP level by applying the MiRKAT method unifying three different kernels: D^W, D^U and D^{BC}, after adjusting for age, gender, BMI and race. There was no significant association between the microbial composition and IBS disease status. SP level is significantly associated with microbial community composition with a p-value of 0.0375 after covariates adjustment. Test using D^U was the most significant with p-values of 0.017 and tests with D^W and D^{BC} failed to generate significant result.

4.3.4 Relationship between MiRKAT and Competing Methods

MiRKAT is closely related to several existing methods for microbiome analysis. In particular, with large sample size, the PERMANOVA method (137) can be shown to be a special case of the kernel machine test under the scenario in which there are no confounding variables (152). Consequently, the MiRKAT with single kernel can be viewed as a generalization of PERMANOVA that accommodates additional covariates. In numerical simulations, the correlation between p-values obtained from single kernel MiRKAT and the corresponding distance based method is usually more than 0.99 at cases when there are no covariates to adjust for.

Similarly, optimal MiRKAT can be viewed as a generalization of the approach of Chen et al (28), which is similar in unifying different types of distance matrices to detect a wide range of biologically relevant changes. However, there is a subtle but important distinction between the permutation approaches of these two methods. The optimal MiRKAT perturbs the minimum p-value across all single-kernel tests to obtain the final p-value while the generalized Unifrac method permutes the maximum F statistics across all PERMANOVA tests. This is potentially problematic as F statistics constructed using different distance matrices can have different degrees of freedom and direct comparison will yield reduced power. On the contrary, p-values are generally scale free and directly comparable. The correlation between p-values obtained from optimal MiRKAT methods and generalized Unifrac method is usually around 0.80.

4.4 Discussion

In this paper, we proposed a kernel machine regression based method (MiRKAT) to test for associations between microbial community composition and outcome of interest, in which covariate effects are modeled parametrically and the microbiome effect is modeled non-parametrically. An extension of this method allows for multiple candidate kernels to be used simultaneously and inherently adjust for multiple testing using perturbation. We showed via simulation that the unifying omnibus test is robust in that it suffered little power loss compared with when the optimal kernel was used, while had substantial power gain compared to when an improper kernel was used.

Microbiome data are statistically challenging to analyze due to their high dimensionality, phylogenetic constraints among OTUs, excessive zeros and over dispersion, which are circumvented by using distance matrices as a summary. Since the best distance matrix was unknown prior to analysis, the omnibus test enables us to detect a wider range of biological relevance. The sparsity of the data can cause problems for

calculating p-values using asymptotical distribution, making the normal approximation too conservative when the sample size is not large enough. Nevertheless, the residual permutation and corresponding perturbation approach are valid under small sample scenario. We recommend using the asymptotic result when sample size is greater than 1000 for a continuous variable and use residual permutation method when sample size is smaller.

As MiRKAT is a regression based method, incorporating additional covariates into the model is very natural. We showed via simulation that when covariates exist which are associated with both outcome of interests and microbial composition, failing to adjust for these confounders would cause seriously inflated type I error. MiRKAT is the only valid (with respect to type I error) method that can properly adjust for confounders in our simulation setting. Meanwhile, the regression framework enables it easily to be extended to other types of outcome variables such as survival, longitudinal and multivariate outcomes, and the ability to effortlessly accommodate a wide range of different distance metrics allows it to serve as a comprehensive framework for a wide variety of metagenomic studies.

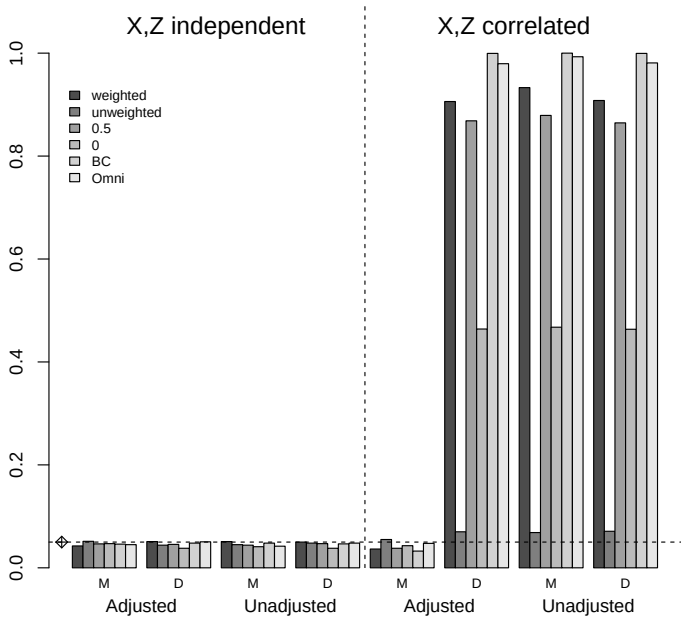

Figure 4.1: Type I error of different methods at $\alpha = 0.05$ level: Data was simulated for $n = 50$ and only the 10 most abundant bacteria have any effect on the outcome. M: MiKRAT D: distance based method. $\diamondsuit$: nominal $\alpha = 0.05$.

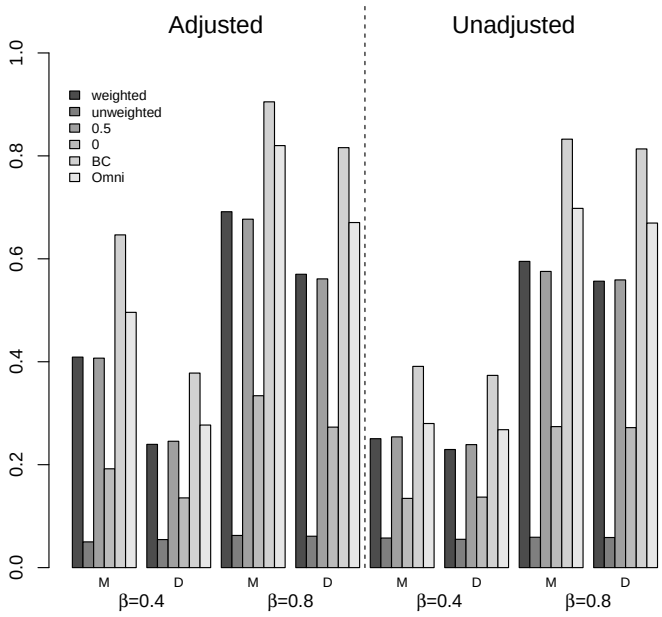

Figure 4.2: Power comparison of different methods: Data was simulated for $n = 50$ and only the 10 most abundant bacteria have any effect on the outcome. Additional covariates X and bacterial effect Z were simulated independently. M: MiKRAT D: distance based method.

CHAPTER 5

Integrative Analysis of Methylation and Genotyping Studies

5.1 Introduction

The etiology for most common human diseases are believed to be multifactorial, with risk factors including heritable genetic variants as well as environmental, behavioral factors and possible interactions between them. In the past few years, genome-wide association studies (GWAS) have been successful in identifying many genetic variants, especially in the form of single nucleotide polymorphism (SNPs), that are associated with a number of common human diseases (214). However, SNPs discovered by GWAS can usually account for only a small fraction of the genetic variation of traits in the human population, leading to the so-called "missing heritability" phenomenon (130, 134). Several reasons have been suggested for the missing heritability, including the lack of power to detect common variants with small effects as well as the complex gene-gene interaction and inherited epigenetic factors, especially methylation (83, 84). Comprehensive understanding of these complex trait etiology requires examination of multiple source of genetic variants. Consequently, many large GWAS consortia are expending to simultaneously examine the joint effect of DNA methylation (115). Integrative analysis of this "multi-dimensional genomic data" were shown to be promising in elucidating the biological processes underlying the disease of interests, for example, in certain cancer (141, 240, 149).

Integrative analysis of genotype and methylation data are challenging in two aspects. First, genotype and DNA methylation are measured in different scales and units, it is unclear how to utilize both data types to determine whether a particular genetic region is associated with the trait of interest. Secondly, it is of great interest in genomic studies to understand the relative roles of different sources of genomic variability in complex trait etiology, for example, whether methylation mediates the genetic effect. In this paper, we propose a two-step framework for first testing the joint effect of both genetic and epigenetic variability on a trait with subsequent mediation analysis to elucidate the possible mechanisms by which these genomic features influence the phenotype.

The standard approach in analyzing GWAS involves testing the effect of each SNP individually on the phenotype of interest, using usually parametric regressions or the Cochran-Armitage trend test. Subsequent adjustment needs to be conducted for multiple comparisons because thousands or even millions of SNPs have been tested. Albeit successful in a lot of applications, this approach has been shown to be underpowered, especially in cases when a large number of genetic variants have only small effect sizes, mainly due to the difficulty in achieving genome-wide significance level. Additionally, individual-SNP analysis suffers from poor reproducibility, because of the imperfect linkage disequilibrium (LD) between the typed SNPs and the true causal variants, which are seldom typed. Instead, we develop an approach that works at the gene level to allow for the common unit of analysis across genotypes and methylation data.

In particular, in our first step, we propose to use the powerful kernel machine framework for testing the cumulative effect of both epigenetic and genetic variants. Kernel machine testing (112, 111) is an operationally simple method for SNP set testing and have been applied in a variety of settings to identify SNP-sets that are associated with a number of diseases (110, 116, 139, 187). We extend this approach to test the joint genetic and epigenetic effect on the phenotype of interest. Pairwise similarities in the

trait values between individuals are compared with the similarity matric constructed from the genotype and methylation data for a particular gene, with high correspondence suggestive of association. Similarity metric in genotype and methylation is constructed by an optimally weighted averages of the similarities using each single data type with the weights determined either by a grid search or adaptive projection method.

In the second step, we adopt classic methods from causal inference literature to investigate the mechanism by which the different genetic variants can affect the phenotype. Specifically, we extend the popular causal steps approach toward mediation analysis to answer the question whether the effect of genetic variants on the phenotype is at least partially mediated through the methylation effect.

Originally developed in psychological and behavior science literatures (6, 121) and extended to many other statistical applications(164, 193, 209) including genetic studies (140, 174, 115), causal steps approach infers causal mediation relationship through a series of linear regression models. In order to establish that the effect of genotypes on phenotypic traits are mediated at least partially through methylation, causal steps model requires the following three conditions to be satisfied(174, 115): 1) genotype is associated with the phenotype 2) genotype is associated with methylation, 3) methylation is associated with the phenotype after controlling for the genotype. Standard mediation analysis that relies on linear regressions are no longer applicable for set-based analysis. In this paper, we develop multidimensional casual steps model counterpart to incorporate gene based genotype and methylation data by using the multivariate version of the kernel machine regression.

Overall, we propose a novel framework that tests the joint genetic and epigenetic effect on the phenotypic trait with subsequent mediation analysis to establish the causal relationship between these variables. The method works at gene level, which unifies the unit in analyzing methylation and genotype data. The kernel machine framework

allows for flexible and robust modeling of the genetic effect, including heterogeneity in effect sizes, non-linear and interactive effects. We demonstrate through simulations and real data analysis that our proposed approach often improves power to detect trait associated genes and often correctly specify the mechanism through which the genetic and epigenetic variability influences power. Further more, extension of this approach to rare variants and sequencing studies are straightforward.

The rest of the article is organized as follows. Section 5.2 defines our notations and elaborate our two-step statistical framework, including the joint genetic/epigenetic effect test and the subsequent causal steps model for mediation. Section 5.3 presents simulation studies of this framework. The manuscript concludes with summaries and discussion in Section 5.5.

5.2 Material and Methods

Current technology has enabled multi-platform genomic profiling of the same biological samples, including genome wide genetic and epigenetic measurement, providing the possibility to analyze their joint effect and to evaluate their relative roles in influencing phenotypic traits. We propose a two step framework that first tests the cumulative effect with subsequent mediation analysis to infer the causal relationship between these variants. We consider set-based test in which SNPs and CpG markers are assigned to SNP or CpG sets based on meaningful biological criteria, such as proximity to predefined genes. This approach unifies the unit of genotype and methylation data and enables joint testing of both effects. For the cumulative effect, we extend the powerful and flexible kernel machine regression framework to allow for possible correlations between genetic and epigenetic variants by constructing optimally weighted kernels from both data types. For the subsequent mediation analysis, we borrow the idea from the classic causal steps model and extend it to multi-dimensional genetic data by incorporating

multivariate kernel machine regression.

5.2.1 Notation

Suppose the data constitute n samples with continuous phenotypic traits $(y_1, y_2, ...y_n)'$. M and G are $n \times p_1$ and $n \times p_2$ matrices of methylation and genotype data with each row denoting methylation or genotype values for a single individual where p_1 and p_2 are the total number of CpG sites or SNP markers within a gene that are considered for the analysis. Let X denote a $n \times q$ matrix of additional variables we want to adjust for in the model, such as age, gender, smoking status, and principal component for population structure.

5.2.2 Cumulative Test of Genetic and Epigenetic effects

For the first step to test for the cumulative effect of genetic and epigenetic variants, we relate the continuous (quantitative) traits through the semiparametric model in which the genetic/epigenetic effect are modeled non-parametrically through the kernel function while the additional covariates are adjusted for parametrically. The model can be represented as:

$$y_i = \beta_0 + X_i\beta + h(G_i, M_i) + \varepsilon_i \tag{5.1}$$

where y_i denotes the phenotypic value for the i^{th} person in the sample, X_i is the set of additional covariates that we wish to control, G_i and M_i are the vectors of genotype and methylation data for p_1 SNPs and p_2 CpG markers in the gene of interest. ε_i is assumed to be independent measurement errors with mean 0 and variance σ^2, β_0 is the intercept and β are the regression coefficients for additional covariates. Let $Z = (G, M)$ represent both the genetic and epigenetic data with $h(Z_i) = h(G_i, M_i)$. For testing the

null hypothesis that the gene has no effect on phenotype,

$$H_0 : h(Z) = 0 \tag{5.2}$$

Under the kernel machine regression framework, the function $h(G_i, M_i)$ are defined only through a positive definite kernel function $K(\cdot, \cdot)$, which measures the similarities between different individuals in genetic and epigenetic variants. In principle, any positive definite matrix that satisfies the conditions of Mercer's theorem (40) can be used as a valid kernel. However, good choice of kernels can lead to statistical tests with substantial higher power.

The kernel machine regression framework has been widely applied to SNP-set analysis with several kernels specifically designed for genotype data under different genetic effect models (96, 229, 106). Popular kernels that in SNP-set analysis include the linear, the identical by state (IBS) and the weighted IBS kernels. Specifically, the linear kernel is the usual inner product between the covariate vectors for different individuals and corresponds to an underlying model which assumes a linear relationship between the phenotype and the SNPs in the SNP set. The IBS kernel and the weighted IBS kernels evaluate the similarity between different subjects by counting the number of alleles that are shared identical, which accommodate the non-linearity between the SNP effect. Thus, the (weighted) IBS kernel can sometimes offer improved power to linear kernel when epistasis is present (222), but when no epistasis is present, using the linear kernels is usually more powerful. Kernel machine regression has also recently gained popularity in analyzing methylation data, with suitable kernels including the gaussian kernel, linear kernel, quadratic kernel.

Constructing a proper joint kernel that incorporates both genetic and epigenetic effects is not straightforward. First, it is usually unrealistic to assume that the SNP-set and the methylation markers influence the phenotype in the same manner and naive

construction of kernels using data set that concatenates the two data types are usually improper. Secondly, the units and scales are inherently different for methylation and genotype data. Genotype data counts the number of minor alleles at each loci and are scaled to 0, 1 and 2 while methylation data measures the proportion of methylation levels at each position and are inherently quantitative. Proper weighting scheme should be performed in constructing the composite kernel. Methylation data can be easily standardized to have mean 0 and unite variance. However, standardization of genotype data are nevertheless non-intuitive.

In this paper, we consider a composite kernel approach to model the joint effect of genetic and epigenetic variants. In particular, we assume

$$K_c(Z_i, Z_j) = wK_1(G_i, G_j) + (1 - w)K_1(M_i, M_j) \tag{5.3}$$

where $Z = (G, M)$ include both the genetic and epigenetic variants, $w \in (0, 1)$,and K_1 and K_2 are proper kernel functions for genotype and methylation data respectively.

The composite kernel can be viewed as a weighted average of two kernels which correspond to genotype and methylation effect respectively. The weights are constrained to ensure the positive definiteness of K. One key advantage of this composite kernel is that it allows for distinct mechanism for methylation and genotype effects. For example, we can construct the composite kernel K_c by choosing K_1 to be the IBS kernel which allows for epistatic effect on the SNP set and choosing K_2 to be the linear kernel, which models the effect of methylation markers linearly. Weighting of genotype and methylation data are carried out at the kernel level instead of at the original data level.

Within the prediction-based statistical learning literature, considerable studies have been devoted to estimation and prediction based on composite kernels (22, 199). Testing under the composite kernel is rarely investigated. For fixed weight w one may directly apply the kernel machine test with K_c considered as any single kernel. Briefly, by using

the connection between the kernel machine regression model and the linear mixed model $y_i = \beta_0 + X_i\beta + h_i + \varepsilon_i$ where h is the random effect distributed as $N(0, \tau K_c)$, the test of no genetic/epigenetic effect $h = 0$ is equivalent to the variance component test $H_0 : \tau = 0$. One can use the variance component score test $Q_c = (y - \hat{y})' K_c (y - \hat{y}))/\hat{\sigma}^2$, where $\hat{y}$ and $\hat{\sigma}^2$ are the estimates under the null model. Asymptotically, Q_c is distributed as an unknown mixture of χ^2 distribution, with the mixture weights determined by the eigen values of $P_0^{1/2} K_c P_0^{1/2}$ where $P_0 = I - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}$, $\tilde{X} = (1, X)$. Many methods can be used to approximate the distribution of Q_c under H_0, including the moment matching (112, 113) and exact methods based on inversion of the characteristic function (43, 45).

In reality, optimal w is unknown as it depends on the true nature of the genetic and epigenetic effects. It is tempting to estimate w from the data and then resort to the same variance component score test as if w is fixed. Unfortunately, this supervised estimation and testing approach would inflate the type I error just as in the multiple testing problem. In this section, we propose two approaches to test the joint genetic and epigenetic effect using this composite kernel which avoid the direct estimation of w. The first method involves constructing multiple candidate composite kernels using a grid search of w, calculating p-values under each kernel, taking the minimum of these p-values and adjusting for taking the minimum by a computationally efficient perturbation test. The second approach calculates the asymptotical p-value through kernel PCA via projection. We will present these two methods in detail in the subsequent sections 5.2.2 and 5.2.2.

Perturbation based Inference

The basic idea underlying perturbation based approach is to generate a number of composite kernels across a range of w, compute p-values under each candidate kernel, take the minimum p-value and adjust for this multiple comparison via perturbation. This approach follows the same perturbation scheme as reported in previous research,which focuses on testing under multiple kernels constructed from the same data type, and extends to incorporate composite kernels constructed from different data types. This method starts with constructing L composite kernels using different choices of weights $w = (w_1, ..., w_d, ..., w_L)$,

$$K_{cd} = w_d K_1 + (1 - w_d) K_2$$

where $(w_1, ..., w_d, ..., w_L)$ are selected between 0 and 1.

For quantitative traits, under H_0, $(y-\hat{y})/\hat{\sigma}$ is asymptotically distributed as standard normal when sample size n is reasonably large. Then each $Q_d = (y - \hat{y})' K_{cd} (y - \hat{y})/\hat{\sigma}^2$ can be viewed as a quadratic form of two standard normal vectors flanking different kernel matrices. Under H_0, each $Q_1, ..., Q_L$ share the same vector of standard normals, with all the differences lie in the kernel matrices which measure the similarities of the genetic and epigenetic data between individuals. This method perturbs each Q_d by replacing $(y - \hat{y})/\hat{\sigma}$ using newly generated standard normal vectors to construct the empirical null distribution. For completeness, we include in this paper the detailed steps in this perturbation approach following (233)

1. For each K_{cd}, construct the variance component score test and obtain corresponding p-values.

2. Obtain the minimum p-values $p^o = \min_{1 \le d \le L} p_d$

3. For each $d \in (1, ..., L)$, obtain the $\Lambda_d = \text{diag}(\lambda_{d,1}, ..., \lambda_{d,m_d})$ and $V_d = [\text{v}_{d,1}, ..., \text{v}_{d,m_d}]$ with $\lambda_{d,1}, ..., \lambda_{d,m_d}$ being the positive eigenvalues of $P_0^{1/2} K_{cd} P_0^{1/2}$ and $\text{v}_{d,1}, ..., \text{v}_{d,m_d}$ being their corresponding eigenvectors.

4. Construct
$$\Sigma = \begin{bmatrix} I & V_1'V_2, & \dots & V_1'V_L \\ V_2'V_1 & I, & \dots & V_2'V_L \\ \vdots & \vdots & \dots & \vdots \\ V_L'V_1 & V_L'V_2 & \dots & I \end{bmatrix}$$
and construct Cholesky decomposition of $\Sigma = RR'$.

5. Generate $r = [r_1, \dots, r_m]' \sim N(0, I)$ with $m = \sum_{d=1}^{d=L} m_d$ and obtain $r^* = Rr$. For the d^{th} kernel, assign $r_d^* = [r_a^*, \dots, r_{a+m_d}^*]$ where $a = \sum_{j=1}^{j=d-1}(m_j + 1)$. This corresponds to that $r^* \sim N(0, \Sigma)$.

6. Compute $Q_d^* = r_d^{*\prime} \Lambda_d r_d^*$ for each d and corresponding p-values p_d^*. Take the minimum $p^* = \min(p_1^*, \dots, p_d^*, \dots, p_L^*)$.

7. Repeat the steps 5 and 6 for a large number of B times to obtain $p_1^*, \dots, p_B^*$.

8. Obtain the final p-value by $p = B^{-1} \sum_{b=1}^{B} I(p_b^* \leq p^o)$.

The principle behind perturbation is similar to that behind permutation in that they both generate a large number of test statistics under H_0 to construct the empirical null distribution. However, perturbation is advantageous because it retains all the possible correlation between additional covariates and genetic/epigenetic effect through the unchanged kernel matrix. On the other hand, permutation requires all the covariates to be uncorrelated to the genetic/epigenetic data for covariates adjustment, which is usually unsatisfied in many situations. Secondly, perturbation is much more computationally efficient: it requires only generating a number of random normal vectors while

permutation requires reconstruction of kernel matrices, and recalculation of p-values through the moment matching or characteristic function inversion method. Finally, it is necessary to directly use the minimum p-value as test statistic instead of the maximum score statistics because Q values constructed from different kernels are dramatically different with respect to degrees of freedom, making Qs not directly comparable. Instead, p-values are scale free.

Kernel PCA approach

One major advantage of the aforementioned composite kernel approach is that it adaptively model the genetic effect and epigenetic effect by varying the weight w. Therefore, it can maintain high power across a wide range of scenarios, including cases when only the genetic (or epigenetic) variation has an effect on the phenotypic trait and the cases when both of them influence the trait value with different effect size. Although computationally more efficient than the permutation based approach, the perturbation method still relies on Monte Carlo calculation of p-values. Analytical calculation of p-values can not be obtained easily because of the possible correlation between the genetic and epigenetic effects.

In this section, we consider an adaptive approach that also relies on the composite kernel. However, we approximate the composite kernel space via kernel PCA and basis projection, to enable analytical computation of the final p-values. It is apparent that model (5.1) with composite kernel (5.3) is equivalent to the following model

$$y_i = \beta_0 + X_i\beta + h_1(G_i) + h_2(M_i) + \varepsilon_i \tag{5.4}$$

in which $h_1(\cdot)$ and $h_2(\cdot)$ are random effects with $h_1(\cdot) \sim N(0, \tau w K_1)$ and $h_2(\cdot) \sim N(0, \tau(1-w)K_2)$.

Consider $\lambda_{G,1} \leq \lambda_{G,2} \leq \cdots \leq \lambda_{G,k}$ and $\lambda_{M,1} \leq \lambda_{M,2} \leq \cdots \leq \lambda_{M,l}$ are the

positive eigenvalues of kernel matrices K_1 and K_2 with corresponding eigenvectors $[\mathrm{v}_{G,1}, \mathrm{v}_{G,2}, \ldots, \mathrm{v}_{G,k}]$ and $[\mathrm{v}_{M,1}, \mathrm{v}_{M,2}, \ldots, \mathrm{v}_{M,l}]$. Let $Z_G = [\mathrm{v}_{G,1}, \mathrm{v}_{G,2}, \ldots, \mathrm{v}_{G,k}]$ and $Z_M = [\mathrm{v}_{M,1}, \mathrm{v}_{M,2}, \ldots, \mathrm{v}_{M,l}]$. Then Z_G and Z_M can be viewed as the basis for the corresponding kernel K_1 and K_2. Model (5.4) can be rewritten as

$$y_i = \beta_0 + X_i\beta + Z_{G,i}\beta_G + Z_{M,i}\beta_M + \varepsilon_i \tag{5.5}$$

with $\beta_G = [\beta_{G,1}, \ldots, \beta_{G,j}, \ldots, \beta_{G,k}]'$ and $\beta_M = [\beta_{M,1}, \ldots, \beta_{M,j}, \ldots, \beta_{M,l}]'$. Testing $H_0 : h(G, M) = 0$ is equivalent to testing $H_0 : \beta_{G,1} = \cdots = \beta_{G,k} = \beta_{M,1} = \cdots = \beta_{M,j} = 0$.

Further, we linearly transform the model (5.5) via projection and construct the following model (79):

$$y_i = \beta_0 + X_i\beta + Z_G^*\gamma_G + Z_M\gamma_M + \varepsilon_i \tag{5.6}$$

where $Z_G^* = (I - M)Z_G$ with $M = Z_M(Z_M'Z_M)^{-1}Z_M'$ is the projection matrix onto the column space of Z_M. $\gamma_G = [\gamma_{G,1}, \ldots, \gamma_{G,j}, \ldots, \gamma_{G,k}]'$ and $\gamma_M = [\gamma_{M,1}, \ldots, \gamma_{M,j}, \ldots, \gamma_{M,l}]'$ are the regression coefficient under the transformed model.

Note that Z_G^* is the residuals by performing linear regressions of each component of Z_G on Z_M and corresponds to a subspace that is orthogonal to the column space of Z_M. We assume γ_G and γ_M are random variables that $\gamma_G \overset{iid}{\sim} N(0, \tau w)$ and $\gamma_M \overset{iid}{\sim} N(0, \tau(1 - w))$ (79). The null hypothesis $\beta_G = 0, \beta_M = 0$ in the previous model (5.5) is equivalent to $H_0 : \tau = 0$ under this transformed model.

Similar to the perturbation based approach, we construct composite kernels by varying $0 = w_1 < \cdots < w_d < \cdots < w_L = 1$ that $K_{cd}^* = w_d K_1^* + (1 - w_d)K_2$. A variance

component score statistics under the transformed model can be constructed as follows:

$$\begin{aligned} Q^*_{cd} &= w_d(y-\hat{y})'K_1^*(y-\hat{y}) + (1-w_d)(y-\hat{y})'K_2(y-\hat{y}) \\ &= w_d Q^*_G + (1-w_d)Q_M \end{aligned} \tag{5.7}$$

where $K_1^* = Z_G^* Z_G^{*\prime} = (I-M)Z_G Z_G'(I-M)$ and K_2 is the same kernel matrix from epigenetic data as previous. We consider the minimum p-values across different choices of w as test statistic

$$p^{o*} = \min_{1\le d\le L} p_d^*$$

where $p_d^* s$ is the p-value for Q^*_{cd}. It is easy to see that Q^*_G and Q^*_M are asymptotically independent with both of them following a mixture of χ^2 distribution. We can approximate this mixture of χ^2 distribution via moment matching(79).

To be specific, $Q^*_G = k_1 =\sim \sum_{q=1}^{m_G} \lambda_{G,q}\chi_1^2$ and $Q^*_M = k_2 =\sim \sum_{q=1}^{m_M} \lambda_{M,q}\chi_1^2$ with $\lambda_{G,q}$ being the eigenvalues of $P_0^{1/2}K_1^*P_0^{1/2}$ and $\lambda_{M,q}$ being eigenvalues of $P_0^{1/2}K_2P_0^{1/2}$, in which P_0 is defined as before. Consider q_d being the $(1-p^{o*})^{th}$ percentile of Q^*_{cd} for $w = w_d$. Then the final p-value can be calculated as

$$\begin{aligned} p &= 1 - P[Q^*_{c,1} < q_1, \ldots, Q^*_{c,L} < q_L] \\ &= 1 - P[w_1k_1 + (1-w_1)k_2 < q_1, \ldots, w_Lk_1 + (1-w_L)k_2 < q_L] \\ &= 1 - P[k_1 < \min_{1\le d\le L} \frac{q_d - w_dk_2}{1-w_d}] \end{aligned} \tag{5.8}$$

Overall, the key idea behind this approach is to construct two orthogonal kernel spaces via projection, leading to a mixture of independent χ^2 distributions and enabling analytical p-value calculation. In this section, we project the SNP effect onto the kernel space of the methylation effect, however, projection on the other direction can be conducted similarly. Both approaches are considered in our simulation studies.

5.2.3 Subsequent Mediation Analysis

Having established the cumulative effect of a set of SNPs/CpG markers on certain phenotype, a natural subsequent step would be to explore the causal relationship between these variables. In this section, we adopt the classic casual steps model (6, 121) for mediation analysis, and extend it to multi-dimensional genomic data by incorporating multivariate kernel machine regression framework.

The idea of mediation concerns the extent to which the effect of one variable on another is mediated by some possible intermediate variable, which can usually be represented by directed diagrams (figure 5.1) with arrows representing causal relationships. In the setting of genomic study, if the effect of genotype (G) on certain phenotype (y) is at least partially directed through methylation (M) level, then the methylation can be considered as a mediator, represented by (M2) in figure 5.1. Here G is considered as an independent variable, which is reasonable under the assumption that genotypes are randomly segregated and fixed at DNA level. DNA methylation, on the other hand, changes with age, diet, various environmental exposure, etc. M is considered as a potential mediator, assuming that DNA methylation variation is prior to and contributes to the occurrence of the phenotype, either a disease or change in quantitative traits. This assumption is reasonable if the time sequence of events can be established or the reverse causality can be ruled out via scientific reasons. For example, if the DNA methylation level is measured at baseline (such as at infant stage) before the disease occurrence/phenotypic change, the causal model M2 can be assumed with confidence. Similarly, suppose we are interested in the cis-regulation of gene expression. Because it is unlikely that the gene expression level can affect the DNA methylation of the same gene, the mediation model M2 can also be assumed.

Originally developed in the 1980s (6), the causal steps model is one of the most popular mediation tests among psychology and epidemiology studies (123, 168) despite

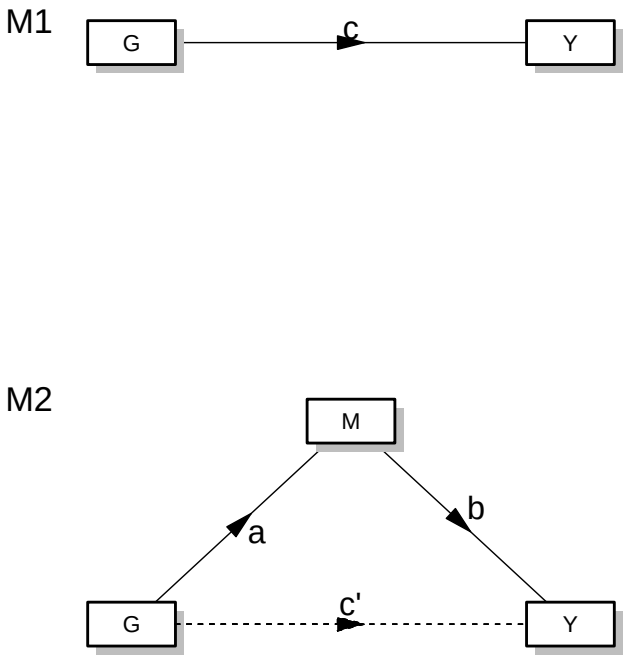

Figure 5.1: Mediation Diagram

many recent alternatives. Under the causal steps model, the mediation effect of M on the effect of G on y can be established if the following conditions are satisfied:

1) G and y are associated,

2) G is associated with M

3) M is associated with y conditional on G

4) G is independent of y conditional on M

In practice, step 1) is usually implicit that causality test is generally conducted only for genes that are significantly associated with phenotype of interest. In our

mediation framework, we conduct subsequent mediation analysis only for genes that show significant cumulative effect on the phenotypic trait based on the aforementioned cumulative test presented in Section 5.2.2.

Conditions 2) and 3) are the most essential steps in establishing a causal chain (186, 168). In the standard situation when G, M and y are all univariate, these two conditions can be tested via the linear regression models $M = \beta_1 + aG + \varepsilon$ and $y = \beta_2 + c'G + bM + \varepsilon$ in which we test $H_0 : a = 0$ and $H_0 : b = 0$ for the two conditions respectively, where the a, b and c' represent the causal effect as in the diagram 5.1. For cases when G and M are multi-dimensional, such as in our situation when G and M corresponds to a set of SNPs and CpG markers respectively, the linear model approaches are no longer applicable. Instead, we utilize the kernel regression framework in which the effect of SNP/methylation set are modeled nonparametrically. We defer the detailed model specification to sections 5.2.3 and 5.2.3.

The final step 4) aims at establishing a complete mediation of M on the $G \to y$ relationship, i.e, all effect of G on y are through M with no other intermediate variables. This is a challenging problem both practically and conceptually. Practically, in the simple case when G, M and y are all univariate, this equates to testing the model $y = \beta_0 + c'G + bM + \varepsilon$ with $H_0 : c' \neq 0$ vs $H_1 : c' = 0$, which is a non-standard equivalence test for which rigorous testing is difficult to accomplish. For this simple case, Chen et al. developed equivalence test procedure via careful specification of the alternative space, a range that are considered sufficiently close to 0 that the difference are practically ignorable, and parametric bootstrapping (29, 140). However, this approach does not work for the case with multi-dimensional G and M because of the difficulty in specifying a "tight"alternative space within the multi-dimensional context. Conceptually, complete mediation implies the process by which G influences y has been completely explained and no other intermediate variable should be investigated. With

issues such as imperfect measurement, sample size and power to detect significant associations, claiming complete mediation can be misleading in complex situations(168). Therefore, we include in our multivariate causal steps model only steps 2) and 3).

Test the association between genotype and methylation

In this step, the multi-dimensional methylation data M is considered as continuous outcome and testing for the association between G and M is required, i.e, a situation when we want to test for the joint effect of multiple markers on multiple outcomes simultaneously. Commonly adopted strategies for kernel machine analysis on multiple outcomes include (1) multiple univariate kernel machine tests with subsequent multiple testing correction, (2) use a summary statistic (e.g, the mean, or methylation from surrogate CpG site (7)) obtained from the multiple outcome as an outcome and fit univariate kernel machine regression, and (3) multivariate regression approach based on kernel machine regression.

In the first framework, each methylation marker was considered as a separate outcome and the similar kernel regression model can be applied to each methylation marker. Then methods such as false discovery rate or family wise error rate can be applied to adjust for multiple comparison. Although intuitive and computationally efficient, this method fails to utilize the correlations within different methylation markers, resulting in reduced power, especially when the number of comparisons is large. The second method makes a strong assumption that the effect of genotype on all methylation markers are the same and can be captured by the chosen summary statistic. By using the summary statistic, this method reduces the dimensionality of the continuous outcome and can lead to potentially improved power if the assumption are satisfied. However, in cases when the assumption are violated, this method can cause serious power loss.

The third method utilizes a multivariate kernel machine regression(MVKMR) framework to evaluate the joint effect of multiple genotype markers on the methylation status (131). Under this framework, the model can be written as

$$M_{ij} = \tilde{X}_i\beta_j + g_j(G_i) + \varepsilon_{ij} \tag{5.9}$$

where $i = 1, ..., n$ with n being the sample size, $j = 1, ..., p_1$ with p_1 being the number of CpG markers, and $\tilde{X}_i$ are the additional covariates to adjust for, including intercept. $\varepsilon_{i1}, ..., \varepsilon_{ip_1} \sim N(0, \Sigma)$ with $\Sigma = \sigma_{kl}$ where σ_{kl} reflects the correlation between M_k and M_l of the same individual. Notice that in this scenario the covariates $\tilde{X}_i$ and Y_i are the same for all methylation markers. Testing the null hypothesis that G is not associated with M would be equivalent to testing

$$H_0 : g_1(G) = g_2(G) = ... = g_{p1}(G) = 0$$

Again, we use kernel machine framework to specify $g_j(G)$ by a kernel function $K(\cdot, \cdot)$ where $g_j(G) = \sum_{l=1}^{L} \alpha_l K_j(G_l^*, G)$ for some integer L, some constants α_l and some $G_1^*, ..., G_L^* \in R^{p_2}$. A score test can be used to test the proposed null hypothesis.

Define

$$\mathbf{M} = (M_{11}, ..., M_{1n}, ..., M_{p_1,1}, ..., M_{p_1,n})'$$

$$\mathbf{g(G)} = \{g_1(G_1), ..., g_1(G_n), ..., g_{p_1}(G_1), ..., g_{p_1}(G_n)\}'$$

$$\mathbf{K} = \text{diag}(K_1, ..., K_{p_1})$$

Also define $\tilde{\mathbf{X}} = \text{diag}(\tilde{X}, ..., \tilde{X})$, and $\boldsymbol{\beta} = \{\beta_1^T, ..., \beta_{p_1}^T\}'$. The model can be rewritten in matrix form that

$$\mathbf{M} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{g(G)} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \sim N(0, \tilde{\boldsymbol{\Sigma}})$ with $\tilde{\boldsymbol{\Sigma}} = \Sigma \otimes I_{p1}$, which is a p_1 by p_1 block matrix with each block as a diagonal matrix $\sigma_{kl}\mathbf{1}_n$ for $k = 1, ..., p_1$ and $l = 1, ..., p_1$. This represents the fact that the correlation is nonzero for methylation markers from the same individual and methylation from different samples are independent.

The score like statistics, as shown by Maity et al(131), can be obtained as $T = (\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}})'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$ can be estimated under the null model. To obtain the p-value, first, the eigen-decomposition of $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}'$ is calculated and $T = \mathbf{r}'\mathbf{D}\mathbf{r}$ where $\mathbf{r} = \mathbf{U}'\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{M} - \mathbf{Z}'\boldsymbol{\beta})$. Under the null hypothesis, $\mathbf{r} \sim N(0, \mathbf{U}'\mathbf{P}_0\mathbf{U})$ where $\mathbf{P}_0 = \tilde{\boldsymbol{\Sigma}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}$. Therefore, we can approximate the distribution of T as a mixture of χ^2 distribution with weights being the eigen values of $\mathbf{U}'\mathbf{P}_0^{1/2}\mathbf{K}\mathbf{P}_0^{1/2}\mathbf{U}$.

Two simplifications of the original MVKMR can be considered to fit our specific setting that the multiple outcome are methylation level from the same individual. The first one is to assume a common kernel function K for all $g_1(G), ..., g_{p_1}(G)$. In other words, we assume $K_1 = ... = K_{p_1} \equiv K$ and $\mathbf{K} = I_{p1} \otimes K$. This approach can greatly facilitate the computation of the weights in the mixture of χ^2 distributions for p-value calculation. Please refer to Appendix II for details about the simplification in calculating p-value.

A second simplification is that instead of considering each methylation marker as distinct outcomes and fitting a multivariate kernel regression model, we consider the methylation values as a Markov Chain of several methylation blocks(12). Within each block, the methylation level at different CpG marker can be considered to have a common mean and be collapsed as a single outcome variable. For example, the CpG markers in the same CpG island can be considered to have a common underlying mean, given the small reported variance for methylation values within the same island (241, 34). Other algorithm that infer the block structure, such as those that infer linkage

disequilibrium (158) or clustering algorithm, may also be used to define methylation blocks.

Association between methylation and phenotype conditional on genotype

We consider the additive least square kernel machine regression framework for this step. The model can be specified as

$$y_i = \beta_0 + X_i\beta + h_1(G_i) + h_2(M_i) + \varepsilon_i$$

in which β is a vector of regression coefficients and $\varepsilon_i \sim N(0, \sigma^2)$. Simlarly, we consider two function space H_1 and H_2 generated from two positive semidefinite kernels K_1 and K_2 corresponding to genotype and methylation data respectively. Notice that y is related to M only through $h_2(M)$ and testing the effect of methylation would be equivalent to testing the null hypothesis that $H_0 : h_2(M) = 0$. By establishing the correspondence between additive least square kernel machine regression and the additive linear mixed model framework, a variance component score test of Zhang and Lin can be employed for this test, in which (238).

$$Q = (y - \hat{\beta}_0 - X\hat{\beta} - \hat{h_1})'K_2(y - \hat{\beta}_0 - X\hat{\beta} - \hat{h_1})$$

where $\hat{\beta}_0, \hat{\beta}$, $\hat{h}_1$ are estimated under the null model $y_i = \beta_0 + X_i\beta + h_1(G_i) + \varepsilon_i$.

Notice that the null model belongs to the kernel machine regression framework and estimation can be obtained by maximizing the following scaled penalized likelihood function

$$J(h, \beta) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - X_i'\beta - h_2(G_i)\}^2 - \frac{1}{2}\lambda\|h_2\|^2_{H_2}$$

where λ is the tuning parameter controlling the tradeoff between goodness of fit and

the complexity of the model. The estimate of λ, β and h_2 can be obtained via restricted maximum likelihood (REML) using the connection between the least square kernel machine and the linear mixed model(112) or by fitting a penalized additive model using generalized cross validation (GCV) criteria (225, 227). Similarly, we can approximate the statistics Q by a mixture of χ^2 distributions with weights being the diagonal elements of D, in which D is the positive eigenvalues of matrix $P_0K_2P_0$, and P_0 is the projection matrix under the null $P_0 = I - (I + \lambda^{-1}K_1)^{-1}\lambda^{-1}K_1 + X(X'(I + \lambda^{-1}K_1)^{-1}X)^{-1}X'(I + \lambda^{-1}K_1)^{-1}$

5.3 Simulations Study

Performance of our proposed cumulative test and subsequent causal steps model was evaluated by simulation studies under a variety of realistic scenarios, in all of which the SNP sets were generated on the basis of the LD structure of a real gene and the methylation sets were simulated to have the same correlation structure as obtained from real data. This allows us to investigate the size and power of our proposed testing framework in real data analysis.

In all our simulations, we considered gene ASAH1, acid ceramidase 1(229), a 28.5-kb-long gene which encodes enzyme acid ceramidase. Expression of this gene has been associated with prostate cancer and mutations are associated with a lysosomal storage disorder known as Farber disease (103, 170, 242).

5.3.1 Simulation for Cumulative Test

We first conducted simulation to investigate the size and power of the proposed cumulative test. We simulated genotype data for 93 SNPs within gene ASAH1 using HAPGEN2 (197) to have the same LD structure as the CEU (CEPH [Utah residents with ancestry from northern and western Europe])samples from international

HAPMAP project under release 24(3). 29 of these 93 HapMap SNPs are genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0 (1), constituting the "typed" SNPs. These were the set of SNPs used for kernel construction in all testing scenarios. Methylation data of the 21 CpG markers within gene ASAH1 is simulated as multivariate normal $N(\mu, \Sigma)$ with Σ simulated to have the same correlation structure as estimated from a real study (2) and μ being a common mean vector. Simulation of methylation data using the symmetric or exchangeable correlation structure generates very similar result with respect to type I error and power to all our proposed methods and thus were omitted from this manuscript. We considered two scenarios in simulating the methylation data: 1) Methylation is independent of genotype data $\mu = -0.1$. 2) Methylation level depends on genotype $\mu = -0.1 + 0.1G_{31}$, where G_{31} is a randomly picked 31^{th} SNP within the SNP set.

We first conducted simulations to verify that the proposed joint test procedure can properly control the genome-wide type I error. We simulated n continuous outcomes via the following null model with $n = 200, 500, 1000$

$$y_i = X_i + \varepsilon_i \tag{5.10}$$

where X is a vector of covariate with $X \sim N(0, 1)$. We considered the a number of testing approaches. Table 5.1 summarizes all the methods that we considered in evaluating the cumulative tests.

Out of the 6 models that we considered, model 1 and 2 are projection based kernel PCA approach and model 3 are perturbation based omnibus test. Both the perturbation based omnibus test and the kernel PCA method start by constructing composite kernel $K_c = wK_1 + (1-w)K_2$ where K_1 and K_2 are based on the genotype and methylation data respectively. For methylation, we considered the linear kernel while for genotype data, we considered the IBS kernels because of its ability to model interactive

Table 5.1: Cumulative Effect Tests Model Specification

Model	Class	$K_1(G)$	$K_2(M)$	Comments
M1	Kernel PCA	IBS	Linear	$K_2(M)$ to $K_1(G)$
M2	Kernel PCA	IBS	Linear	$K_1(G)$ to $K_2(M)$
M3	Perturbation	IBS	Linear	
M4	Naive	K = ZZ'		Z = [G,M]
M5	SNP only	IBS	–	
M6	Methylation only	–	Linear	

effects. For the kernel PCA method, we investigated the performance of both projection directions that one projects the SNP effect onto the kernel space of the methylation effect and the other projects in the opposite direction. For the perturbation based approach, we used a grid of w as $0, 0.25, 0.5, 0.75, 1$. For comparison, we considered several alternative models. The naive model (M4) tests the joint effect of methylation and genotype via data matrix Z which is a naive concatenation of G and M. We also considered models that tests only the SNP effect and the methylation effects. 100,000 simulations were conducted to assess the genome-wide type I error rate.

Simulations were also conducted under the alternative to assess the empirical power of the proposed joint genetic/epigenetic effect tests. Similarly as in the type I error simulation, we considered situations when the genotype and methylation are independent or situations when genotype can affect the average methylation level. Within each of the scenario, we simulated under the models

$$y_i = X_i + \beta_s G_{i,29} + \beta_m M_{i,19} + \varepsilon_i \tag{5.11}$$

and

$$y_i = X_i + \beta_s G_{i,29} G_{i,55} + \beta_m M_{i,19} + \varepsilon_i \tag{5.12}$$

where $G)_{\cdot,29}$, and $M_{\cdot,19}$ are the 29^{th} SNPs and the 19^{th} CpG marker in the set of simulated genotype and methylation data, which constitute the causal SNPs and CpG

markers. In the first simulation scenario, we essentially considered one causal SNP with additive effect and one causal CpG marker while in the second scenario, we considered two potential causal SNPs with an interactive epistatic effect, but no separate main effect, along with one causal CpG markers. By varying the values of β_s and β_m, we control the strength of association between genetic/epigenetic variation and the phenotypic outcome. By choosing different β_s and β_m, we constructed situations when there is only genotype effect or methylation effect or both with different effect sizes. 2000 simulations were conducted to assess the empirical power.

5.3.2 Simulation for Multivariate Causal Steps Model

We also evaluated the performance of our proposed multivariate causal steps model via simulations on gene ASAH1. The multivariate causal steps model constitutes two steps with the first step to establish G and M are associated and the second step to test M and y are associated conditional on G. We aim at evaluating the size and power for the two steps individually.

Type I error for each of the two steps in the causal model was evaluated by conducting 2000 simulations under the null model, in which genotype and methylation was independently simulated as previously described in Section 5.3.1. Outcome data was simulated under the following model that doesn't rely on G or M:

$$y_i = X_i + \varepsilon_i$$

where $X \sim N(0,1)$ and $\varepsilon_i \sim N(0,1)$ were simulated independently.

We considered two causal scenarios in assessing the power of each individual test in the causal steps model. The complete mediation are situations when the genetic effect on the outcome is through and only through the changes in the methylation level, which equates to nonzero a and b effects and $c = 0$ in the causal diagram. The

partial mediation corresponds to scenario that there is both direct and indirect genetic effect that methylation is a partial mediator between genotype and outcome.

In both the complete mediation and partial mediation scenario, genotype data was simulated the same way as previously described in Section 5.3.1. Methylation data was similarly simulated to be multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ, where Σ is the same as in Section 5.3.1. However, we simulated the 21 CpG markers as three contiguous blocks of size 3,15 and 3 respectively. The mean methylation level remains the same within each block, but the average methylation level can vary between blocks. Specifically, we considered the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \mu_3, \ldots, \mu_3)^T$, with μ_1, μ_2 and μ_3 repeated for 3,15 and 3 time respectively. Notice that the simulation in Section 5.3.1 that all the CpG markers have a common mean constitutes a special case of this simulation. We simulated under the case that a randomly chosen SNP 31 in the genotype data can affect the methylation levels by $\mu_{i1} = -0.1 + 0.1G_{i,31}$, $\mu_{i2} = -0.1 + 0.2G_{i,31}$, and $\mu_{i3} = -0.1$.

The continuous outcome was simulated by $y_i = X_i + 0.2G_{i,29} + 0.2\bar{M}_{1-3} + \varepsilon_i$ for the partial mediation and $y_i = X_i + 0.2\bar{M}_{1-3} + \varepsilon_i$ for the case of complete mediation, in both of which $X \sim N(0, 1)$, $G_{i,29}$ is the genotype from the 29th SNP in the gene and $\bar{M}_{1-3}$ represents the mean methylation level across first block of CpG markers.

For the first step, we conducted the MVKMR as in Section 5.2.3 in which M is considered as multivariate outcome and regressed on the genotype G through kernel function $K(G)$. We considered two testing approaches: the first approach (MG block) utilized the contiguous block information and collapses markers within the same block as a single outcome variable, and the second approach (MG original) considered each CpG marker as distinct outcome for MVKMR. For both approaches, the IBS kernel was used because of its robustness. For the second step, we utilized the additive least square kernel machine regression framework presented in Section 5.2.3 with linear kernels for

M and IBS kernels for G

5.4 Results

5.4.1 Empirical Size and Power for Cumulative Test

Table 5.2 summarizes the size result of the simulation studies on testing the cumulative effect of the genetic and epigenetic variation. We used $100,000$ simulations to evaluate the type I error at $\alpha = 0.05, 0.01$ and 0.001. We showed in Table 5.2 the type I error for cases when G and M are independent. For cases when G and M are related, the type I error result are essentially the same. From the simulations, all these methods have well controlled type I error for all the sample sizes we tested on, at all α levels that are tested on.

Table 5.2: Empirical Type I Error Rate at different α levels

α	n	M1	M2	M3	M4	M5	M6
	200	0.948	0.944	1.004	1.000	0.981	1.005
0.05	500	0.954	0.947	1.008	1.001	0.980	1.006
	1000	0.952	0.953	1.007	1.002	0.986	1.004
	200	0.943	0.973	0.962	1.072	0.955	1.034
0.01	500	0.945	0.966	0.959	1.072	0.949	1.038
	1000	0.973	0.950	0.984	1.026	0.978	1.017
	200	1.119	1.141	0.987	0.998	1.119	1.042
0.001	500	1.138	1.129	1.062	1.005	1.100	1.062
	1000	1.037	1.027	1.008	0.978	1.012	1.022

Presented as type I error divided by corresponding α

Figure 5.2 presents the power for all the tests, with the upper panel showing the power result for simulation case (5.11), in which the SNP and CpG have additive effect

and the lower panel showing the result for the case 5.12, in which we observe epistatic SNP effect.

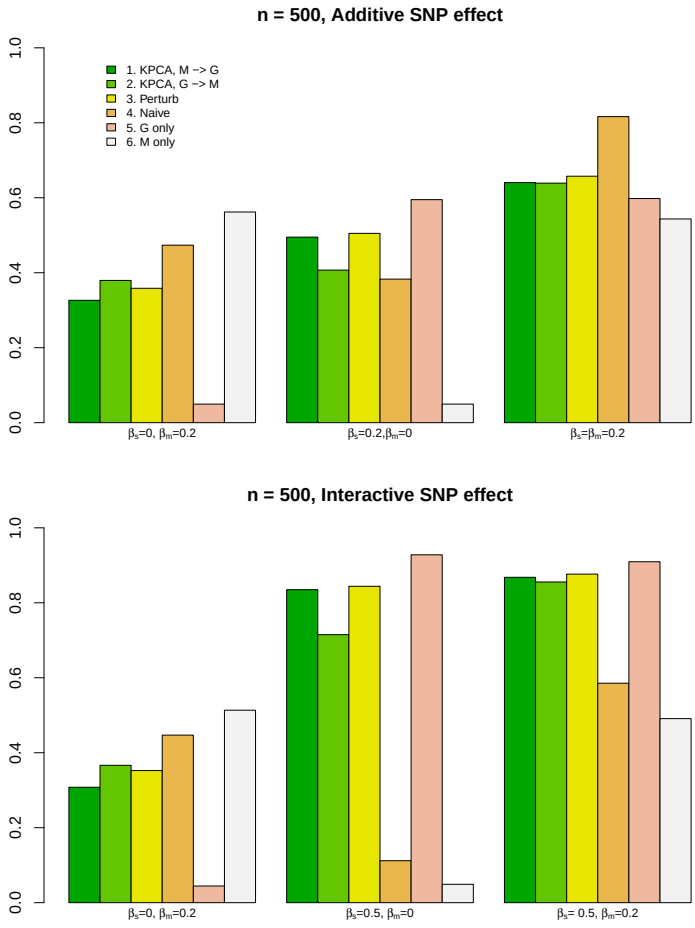

Figure 5.2: Empirical Power for Tests on Cumulative Genetics and Epigenetic Effects

On the basis of the figure 5.2, all the methods have increased power when the effect size increases. We first compared the proposed joint test models (M1-M3) with models (M5-M6) that tested the single genetic/epigenetic effect. Unsurprisingly, in situations when there is only genetic effect ($\beta_m = 0, \beta_s \neq 0$), the model(M5) that only tests the SNP-set effect is more powerful than the joint tests (M1-M3) because of the fewer degrees of freedom. Similar conclusions can be drawn for situations when there is only the epigenetic effect. However, in reality, information on the underlying genetic

architecture is never known in prior. In situations when there is both genetic/epigenetic effect, testing on only one of the effects can cause a considerable power loss compared to joint testing.

Secondly, comparison was made between our proposed composite kernel based approaches M1-M3 with the naive model M4, which constructed linear kernels based on naive concatenation of genetic and epigenetic variations. For the first simulation scenario with additive SNP and methylation effect, the naive model has the highest power when there is both genetic and epigenetic effects. The composite kernel approaches, including the perturbation based approaches and the kernel PCA methods, on the other hand, have only modest power loss in this situation compared to the naive model. For situations when there is interactive SNP effect, the naive model performs poorly as it fails to accommodate the epistatic effect. Across all simulation scenarios, the composite kernel approach can lead to substantially improved power over poor choices of kernels and only modest power loss compared to when using the optimal kernel.

Finally, we compared between the proposed composite kernel based methods. The perturbation approach have superior or similar power compared to the kernel PCA approaches across all simulation scenarios. Among the kernel PCA approaches, projection of the epigenetic variation on the kernel space of genetic variation provides higher power when the overall genetic effect is bigger (or more extremely, methylation do not have an effect at all). Similarly, projecting the genetic effect on the epigenetic effect provides higher power when the overall epigenetic effect is dominating the genetic effect. Overall,the power loss for projecting the epigenetic effect onto the kernel space of genetic effect is modest when the epigenetic effect is dominating. However,the power loss can be substantial for projecting the genetic effect onto the kernel space of epigenetic effect when there is dominating genetic effect.

5.4.2 Multivariate Causal Steps Model Results

Figure 5.3 represents the type I error and empirical power evaluated at $\alpha = 0.05$ level for the two steps in the causal mediation analysis under three different causal scenarios for sample size $n = 500$. The results for $n = 200$ or 1000 are similar.

Under the null model, all the methods, including the two MVKMR models that test the association between M and G and the additive least square kernel regression that tests the association between M and the phenotype y conditional on G, showed valid type I error. Also, all tests have adequate power under both alternative scenarios, including the partial mediation and the complete mediation scenario. Unsurprisingly,the MG block test that utilized the block structure of the methylation data have higher power than the corresponding test that treats each CpG marker as individual outcomes.

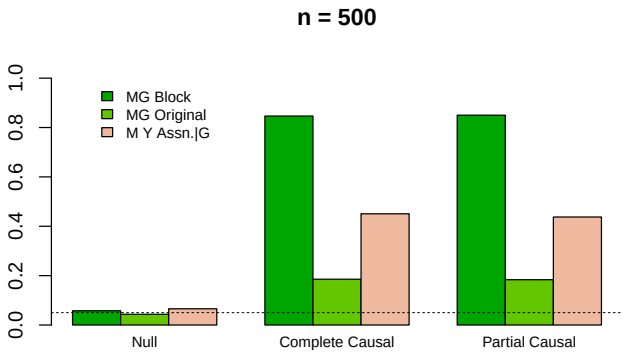

Figure 5.3: Type I error and empirical power for causal mediation tests

5.5 Discussion

In this paper, we propose a statistical framework for integrative analysis of genome wide methylation and genotype studies, in which we first test the cumulative genetic/epigenetic effects with subsequent mediation analysis to decide their relative roles. Our analysis is based on SNP-set and CpG marker set that are constructed using prior biological knowledge, such as proximity to a known gene or other genomic features. This approach unifies the units in analyzing methylation and genotype data. We model the cumulative genetic/epigenetic effect via a flexible, semiparametric kernel machine regression by constructing composite kernels. For genes that show significant association with the phenotype, we construct mediation analysis to understand their relative roles in affecting the phenotype. The subsequent mediation model tests whether methylation is a mediator by first testing the association between genotype and methylation and then testing the association between methylation and the phenotype adjusting for genotypic effect.

We proposed two approaches for the cumulative test, both of which start by constructing a composite kernel as a weighted average of two kernels based on genotype and methylation respectively. In principle, any valid kernel that measures the similarity between genotype or methylation from different individuals can be used in constructing the composite kernel, allowing for flexible modeling with different genotype and methylation effect mechanism. Therefore, the composite kernel approach can be advantageous than the naive model which concatenates the genotype and methylation data to construct a single common kernel (with possible weighting) and tests the joint effect through this common kernel. By using the single common kernel, the naive model assumes a common genetic and epigenetic effect mechanism which can be unreasonable. The composite kernel approach, on the other hand, can incorporate kernels that are specifically designed for SNP data and methylation data, incorporating possible

epistatic SNP effect and other possible interaction effect in methylation data.

Both of our cumulative test approaches rely on the adaptive selection of the optimal weights in constructing the composite kernel. The perturbation based approach selects the optimal weight via a grid search of different w, and adjusts for multiple comparison via computationally efficient perturbation. In our simulations, we constructed composite kernels as a weighted average of IBS kernels for SNP data and linear kernel for methylation data. However, it should be noted that even more flexible model can be conducted by incorporating multiple kernels for genetic and epigenetic data respectively.

The kernel PCA approach relies on the projection of genetic variation onto the epigenetic variation, or reversely. Although the projection enables decomposition of the genetic and epigenetic effect to facilitate analytical p-value computation, it can weaken the association signal. In our simulation experiments, the kernel PCA approach tends to be less powerful than the perturbation approach, especially when the sample size is modest ($n = 200$, data not shown). The direction of projection can also be important, the relative power of which depends on the relative genetic and epigenetic effect sizes. In our simulation, the power loss can be magnificent for projecting the genetic variation to the epigenetic variation when the genetic effect is dominating, especially when the sample size is modest ($n = 200$, data not shown). In practice, we suggest using the perturbation approach than the kernel PCA approach in the context of the joint association testing of genetic and epigenetic variation.

For genes that show significant joint genetic/epigenetic effect, we propose a two step process to explore the possible causal relationship between genotype, methylation and phenotype. The multivariate causal steps model extends the classic univariate causal steps model to incorporate multidimensional mediator and independent variable. In each step, we test association via the kernel machine regression framework for flexible

semiparametric modeling of genetic and epigenetic effect. Critical to this process is the correct specification of the causal model prior to analysis that methylation level change is prior to and possibly contributes to variation in the phenotype: i.e, methylation is a potential mediator between genotype on phenotype. This assumption is reasonable in situations when the time order of events can be established, such as in cases when the methylation level is measured at baseline (such as at birth or before disease) or when the phenotype is unlikely to cause methylation level change. Application of the causal model without consideration of the underlying model assumption can result in spurious conclusions.

In summary, we have proposed a statistical framework for genome wide integrative analysis of methylation and genotype studies, with first testing the cumulative association of genetic and epigenetic variation at each gene with phenotypic trait and subsequent mediation analysis for causal inference. We employed a semiparametric kernel machine regression framework to allow for flexible modeling of SNP-set and methylation set effect. We show via simulation and real data application the well controlled genome wide type I error as well as the superior power compared to competing models.

APPENDIX I

Exact Method for MiRKAT Using Multiple Kernels

Because of the sparsity of the OTU data, the asymptotic approximation from the mixture of χ^2 distribution can be too conservative when sample size are not large enough. Inference based on the exact distribution is difficult because of the correlation of test statistics under different kernels. Instead we can resort to residual permutation to obtain exact p value.

Suppose that d different kernels $K = (K_1, ..., K_k, ..., K_d)$ were considered, the following residual permutation approach can be used to obtain the final p-value as well as kernel specific p-values at the same time.

1. Fit the null linear or logistic regression model by regressing y on X and obtain the residuals $r = y - \hat{\mu}$.

2. For each K_k, with k in $(1, ..., d)$, Q_k can be calculated as $Q_k = r' K_k r$.

3. Permute the residual r for a large number of B times, and with each permutated r_b^*, calculate $Q_{kb}^* = r_b^{*\prime} K_k r_b^*$.

4. The kernel specific p-value for k^{th} kernel is estimated as $p_k = B^{-1} \sum_{b=1}^{B} I(Q_k < Q_{kb}^*)$ and the minimum p-value across all the d kernels specific p-values $p_o = \min_{1 \le k \le d} p_k$.

5. For each permutation b and each kernel k, $p_{kb}^* = (B-1)^{-1} \sum_{b' \neq b, b'=0}^{B} I(Q_{kb}^* < Q_{kb'}^*)$, which equates to getting a p-value for each permutated data set under each kernel. The minimum p-value across all d kernels can be obtained as $p_{ob}^* = \min_{1 \le k \le d} p_{kb}^*$.

6. Calculate the final p-value as $p = B^{-1} \sum_{b=1}^{B} I(p_o > p_{ob}^*)$

Notice that for each permutation b, $p^*_{1b}, ..., p^*_{db}$ are calculated using the same set of reasmpled residuals and thus are correlated. Kernel specific p-values can be obtained through steps 1 to 4.

APPENDIX II

Obtaining χ^2 Mixture Weights for Joint Association Test

Methylation data M is considered as multivariate continuous outcome. The model (5.9) that tests the association between the methylation and genotype is:

$$M_{ij} = \tilde{X}_i \beta_j + g_j(G_i) + \varepsilon_{ij}$$

where $i = 1, ..., n$ with n being the sample size, $j = 1, ..., p_1$ with p_1 being the number of CpG markers, and $\tilde{X}_i$ are the additional covariates to adjust for, including intercept. $\varepsilon_{i1}, ..., \varepsilon_{ip_1} \sim N(0, \Sigma)$ with $\Sigma = \sigma_{kl}$ where σ_{kl} reflects the correlation between M_k and M_l of the same individual. For the first simplification, we consider a common kernel function K for all $g_1(G), ..., g_{p_1}(G)$. In other words, we assume $K_1 = ... = K_{p_1} \equiv K$ and $\mathbf{K} = I_{p1} \otimes K$.

Define

$$\begin{aligned}
\mathbf{M} &= (M_{11}, ..., M_{1n}, ..., M_{p_1,1}, ..., M_{p_1,n})' \\
\mathbf{g(G)} &= \{g_1(G_1), ..., g_1(G_n), ..., g_{p_1}(G_1), ..., g_{p_1}(G_n)\}' \\
\mathbf{K} &= I_{p1} \otimes K \\
\tilde{\mathbf{X}} &= \text{diag}(\tilde{X}, ..., \tilde{X}) \\
\boldsymbol{\beta} &= \{\beta_1^T, ..., \beta_{p_1}^T\}'
\end{aligned} \tag{5.13}$$

The model can be rewritten in matrix form that

$$\mathbf{M} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{g(G)} + \boldsymbol{\varepsilon}$$

Then the test statistics T can be written as:

$$\begin{aligned} T =& (\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}})'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ =& (\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\Sigma^{-2} \otimes K)(\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

Because $\mathbf{M} - \mathbf{X}\hat{\boldsymbol{\beta}} \sim N(0, \tilde{\boldsymbol{\Sigma}}) = N(0, \Sigma \otimes I_{p1})$, T can be approximated by a mixture of χ^2 distribution with weights being eigenvalues of $\tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\Sigma}}^{1/2}$, which can be written as $(\Sigma^{1/2} \otimes I_{p1})(\Sigma^{-2} \otimes K)(\Sigma^{1/2} \otimes I_{p1}) = (\Sigma^{1/2}\Sigma^{-2}\Sigma^{1/2}) \otimes K = \Sigma^{-1} \otimes K$. Suppose $\lambda_1, ..., \lambda_l$ and $\mu_1, ...\mu_m$ are eigenvalues of K and Σ respectively, then the weights for the mixture of χ^2 distribution are $\lambda_i\mu_j$ with $i \in (1, ..., l), j \in (1, ..., m)$. In this way, we reduce the eigen-decomposition of $np_1 \times np_1$ matrices to matrices of sizes $n \times n$ for computation efficiency.

BIBLIOGRAPHY

[1] Affymetrix, inc. http://www.affymetrix.com.

[2] 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.

[3] B. L. C. A. C. F. D. M. Altschuler, D., P. Donnelly, and I. H. Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[4] N. Attar. The allure of the epigenome. *Genome Biology*, 13:419, 2012.

[5] E. M. Azzato, P. D. Pharoah, P. Harrington, D. F. Easton, D. Greenberg, N. E. Caporaso, S. J. Chanock, R. N. Hoover, G. Thomas, D. J. Hunter, and P. Kraft. A genome-wide association study of prognosis in breast cancer. *Cancer Epidemiology Biomarkers and Prevention*, 19(4):1140–1143, 2010.

[6] 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.

[7] V. Barrera and M. A. Peinado. Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale. *Nucleic Acids Res.*, 40(22):11490–11498, Dec 2012.

[8] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, 35(7):606–19, 2011.

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

[10] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

[11] M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E. W. Garcia, B. Wu, D. Doucet, N. J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D. L. Barker, M. S. Chee, J. Floros, and J. B. Fan. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.*, 16(3):383–393, Mar 2006.

[12] A. Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, 2002.

[13] E. H. Blackburn. Walking the walk from genes through telomere maintenance to cancer risk. *Cancer Prev Res (Phila)*, 4(4):473–475, Apr 2011.

[14] C. Bock. Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

[15] V. Bollati, A. Baccarelli, L. Hou, M. Bonzini, S. Fustinoni, D. Cavallo, H.-M. Byun, J. Jiang, B. Marinelli, A. C. Pesatori, et al. Changes in dna methylation patterns in subjects exposed to low-dose benzene. *Cancer research*, 67(3):876–880, 2007.

[16] K. A. Bollen and R. Stine. Direct and Indirect Effects: Classical and Bootstrap Estimates of Variability. *Sociological Methodology*, 20:115–140, 1990.

[17] J. Bray and J. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349, 1957.

[18] A. R. Brothman, G. Swanson, T. M. Maxwell, J. Cui, K. J. Murphy, J. Herrick, V. Speights, J. Isaac, and L. R. Rohr. Global hypomethylation is common in prostate cancer cells: a quantitative predictor for clinical outcome? *Cancer genetics and cytogenetics*, 156(1):31–36, 2005.

[19] R. L. Brown. Assessing specific mediational effects in complex theoretical models. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2):142–156, 1997.

[20] K. J. Bussey, K. Chin, S. Lababidi, M. Reimers, W. C. Reinhold, W. L. Kuo, F. Gwadry, Ajay, H. Kouros-Mehr, J. Fridlyand, A. Jain, C. Collins, S. Nishizuka, G. Tonon, A. Roschke, K. Gehlhaus, I. Kirsch, D. A. Scudiero, J. W. Gray, and J. N. Weinstein. Integrating data on dna copy number with gene expression levels and drug sensitivities in the nci-60 cell line panel. *Mol Cancer Ther*, 5(4):853–67, 2006.

[21] T. Cai, G. Tonini, and X. Lin. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*, 67(3):975–86, 2011.

[22] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla. Composite kernels for hyperspectral image classification. *Geoscience and Remote Sensing Letters, IEEE*, 3(1):93–97, 2006.

[23] O. Carlborg and C. S. Haley. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.*, 5(8):618–625, Aug 2004.

[24] K. Chalitchagorn, S. Shuangshoti, N. Hourpai, N. Kongruttanachok, P. Tangkijvanich, D. Thong-ngam, N. Voravud, V. Sriuranpong, and A. Mutirangura. Distinctive pattern of line-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene*, 23(54):8841–8846, 2004.

[25] Q. Chang, Y. Luan, and F. Sun. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, 12:118, 2011.

[26] E. S. Charlson, J. Chen, R. Custers-Allen, K. Bittinger, H. Li, R. Sinha, J. Hwang, F. D. Bushman, and R. G. Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One*, 5(12):e15216, 2010.

[27] D. I. Chasman. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet. Epidemiol.*, 32(7):658–668, Nov 2008.

[28] J. Chen, K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, R. G. Collman, F. D. Bushman, and H. Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–13, 2012.

[29] L. S. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol*, 8(10):R219, 2007.

[30] X. Chen, L. Wang, B. Hu, M. Guo, J. Barnard, and X. Zhu. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genet. Epidemiol.*, 34(7):716–724, Nov 2010.

[31] M. W. Cheung. Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(2):227–246, 2007.

[32] A. S. Chi and B. E. Bernstein. Developmental biology. Pluripotent chromatin state. *Science*, 323(5911):220–221, Jan 2009.

[33] J. H. Cho, D. L. Nicolae, L. H. Gold, C. T. Fields, M. C. LaBuda, P. M. Rohal, M. R. Pickles, L. Qin, Y. Fu, J. S. Mann, B. S. Kirschner, E. W. Jabs, J. Weber, S. B. Hanauer, T. M. Bayless, and S. R. Brant. Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. *Proc. Natl. Acad. Sci. U.S.A.*, 95(13):7502–7507, Jun 1998.

[34] R. K. Chodavarapu, S. Feng, Y. V. Bernatavichute, P. Y. Chen, H. Stroud, Y. Yu, J. A. Hetzel, F. Kuo, J. Kim, S. J. Cokus, D. Casero, M. Bernal, P. Huijser, A. T. Clark, U. Kramer, S. S. Merchant, X. Zhang, S. E. Jacobsen, and M. Pellegrini. Relationship between nucleosome positioning and dna methylation. *Nature*, 466(7304):388–92, 2010.

[35] F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581, Nov 1997.

[36] O. Combarros, M. Cortina-Borja, A. D. Smith, and D. J. Lehmann. Epistasis in sporadic Alzheimer's disease. *Neurobiol. Aging*, 30(9):1333–1349, Sep 2009.

[37] G. M. Cooper, J. A. Johnson, T. Y. Langaee, H. Feng, I. B. Stanaway, U. I. Schwarz, M. D. Ritchie, C. M. Stein, D. M. Roden, J. D. Smith, D. L. Veenstra, A. E. Rettie, and M. J. Rieder. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*, 112(4):1022–1027, Aug 2008.

[38] D. Cox. Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34:187–220, 1972.

[39] N. J. Cox, M. Frigge, D. L. Nicolae, P. Concannon, C. L. Hanis, G. I. Bell, and A. Kong. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat. Genet.*, 21(2):213–215, Feb 1999.

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

[41] R. Davies. The distribution of a linear combination of chi-2 random variables. 29(3):323–333, 1980.

[42] R. B. Davies. Numerical inversion of a characteristic function. *Biometrika*, 60(2):415–417, 1973.

[43] 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.

[44] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, 2011.

[45] P. Duchesne and P. Lafaye de Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54:858–862, 2010.

[46] A. Dumitriu, J. C. Latourelle, T. C. Hadzi, N. Pankratz, D. Garza, J. P. Miller, J. M. Vance, T. Foroud, T. G. Beach, and R. H. Myers. Gene expression profiles in Parkinson disease prefrontal cortex implicate FOXO1 and genes under its transcriptional regulation. *PLoS Genet.*, 8(6):e1002794, Jun 2012.

[47] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[48] A. O. Edwards, R. Ritter, K. J. Abel, A. Manning, C. Panhuysen, and L. A. Farrer. Complement factor H polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424, Apr 2005.

[49] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[50] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, 11(6):446–450, Jun 2010.

[51] A. K. El-Naggar, S. Lai, G. Clayman, J. Lee, M. A. Luna, H. Goepfert, and J. G. Batsakis. Methylation, a major mechanism of p16/cdkn2 gene inactivation in head and neck squamous carcinoma. *The American journal of pathology*, 151(6):1767, 1997.

[52] C. Eng, J. G. Herman, and S. B. Baylin. A bird's eye view of global methylation. *Nature Genetics*, 24(2):101–102, 2000.

[53] G. Fatemifar, C. J. Hoggart, L. Paternoster, J. P. Kemp, I. Prokopenko, M. Horikoshi, V. J. Wright, J. H. Tobias, S. Richmond, A. I. Zhurov, A. M. Toma, A. Pouta, A. Taanila, K. Sipila, R. Lahdesmaki, D. Pillas, F. Geller, B. Feenstra, M. Melbye, E. A. Nohr, S. M. Ring, B. St Pourcain, N. J. Timpson, G. Davey Smith, M.-R. Jarvelin, and D. M. Evans. Genome-wide association study of primary tooth eruption identifies pleiotropic loci associated with height and craniofacial distances. *Human Molecular Genetics*, 2013.

[54] S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[55] J. C. Figueiredo, M. V. Grau, K. Wallace, A. J. Levine, L. Shen, R. Hamdan, X. Chen, R. S. Bresalier, G. McKeown-Eyssen, R. W. Haile, et al. Global dna hypomethylation (line-1) in the normal colon and lifestyle characteristics and dietary and genetic factors. *Cancer Epidemiology Biomarkers & Prevention*, 18(4):1041–1049, 2009.

[56] M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. Heine-Suner, J. C. Cigudosa, M. Urioste, J. Benitez, M. Boix-Chornet, A. Sanchez-Aguilera, C. Ling, E. Carlsson, P. Poulsen, A. Vaag, Z. Stephan, T. D. Spector, Y. Z. Wu, C. Plass, and M. Esteller. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U.S.A.*, 102(30):10604–10609, Jul 2005.

[57] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, 10(4):241–251, Apr 2009.

[58] P. A. Frazier, A. P. Tix, and K. E. Barron. Testing Moderator and Mediator Effects in Counseling Psychology Research. *Journal of Counseling Psychology*, 51(1):115–134, Jan. 2004.

[59] Q. Gao, Y. He, Z. Yuan, J. Zhao, B. Zhang, and F. Xue. Gene- or region-based association study via kernel principal component analysis. *BMC Genet.*, 12:75, 2011.

[60] W. J. Gauderman. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.*, 155(5):478–484, Mar 2002.

[61] L. A. Gelfand, J. L. Mensinger, and T. Tenhave. Mediation analysis: a retrospective snapshot of practice and more recent directions. *The Journal of general psychology*, 136(2):153–176, Apr. 2009.

[62] D. Gianola and J. B. van Kaam. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178(4):2289–303, 2008.

[63] M. E. Goddard, N. R. Wray, K. Verbyla, and P. M. Visscher. Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, 24(4):pp. 517–529, 2009.

[64] J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007.

[65] J. J. Goeman, S. A. Van De Geer, F. De Kort, and H. C. Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[66] J. Gu, M. Chen, S. Shete, C. I. Amos, A. Kamat, Y. Ye, J. Lin, C. P. Dinney, and X. Wu. A genome-wide association study identifies a locus on chromosome 14q21 as a predictor of leukocyte telomere length and as a marker of susceptibility for bladder cancer. *Cancer Prev Res (Phila)*, 4(4):514–521, Apr 2011.

[67] J. L. Haines, M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. L. Spencer, S. Y. Kwan, M. Noureddine, J. R. Gilbert, N. Schnetz-Boutaud, A. Agarwal, E. A. Postel, and M. A. Pericak-Vance. Complement factor H variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421, Apr 2005.

[68] M. Hamady, C. Lozupone, and R. Knight. Fast unifrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data. *ISME J*, 4(1):17–27, 2010.

[69] D. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[70] S. M. Hayes, A.F. The relative trustworthiness of tests of indirect effects in statistical mediation analysis. *Psychological Science*, 2013.

[71] M. Hernan and J. Robins, editors. *Causal Inference*. CRC Press, 2011.

[72] D. G. Hernandez, M. A. Nalls, J. R. Gibbs, S. Arepalli, M. van der Brug, S. Chong, M. Moore, D. L. Longo, M. R. Cookson, B. J. Traynor, et al. Distinct dna methylation changes highly correlated with chronological age in the human brain. *Human molecular genetics*, 20(6):1164–1172, 2011.

[73] H. Heyn, F. J. Carmona, A. Gomez, H. J. Ferreira, J. T. Bell, S. Sayols, K. Ward, O. A. Stefansson, S. Moran, J. Sandoval, et al. Dna methylation profiling in breast cancer discordant identical twins identifies dok7 as novel epigenetic biomarker. *Carcinogenesis*, 34(1):102–108, 2013.

[74] H. Heyn, N. Li, H. J. Ferreira, S. Moran, D. G. Pisano, A. Gomez, J. Diez, J. V. Sanchez-Mut, F. Setien, F. J. Carmona, et al. Distinct dna methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*, 109(26):10522–10527, 2012.

[75] J. Hirschhorn. Genomewide association studies–illuminating biologic pathways. *The New England Journal of Medicine*, 360(5):1699–701, 2009-04-23 00:00:00.0.

[76] 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.

[77] Y.-T. Huang, R. S. Heist, L. R. Chirieac, X. Lin, V. Skaug, S. Zienolddiny, A. Haugen, M. C. Wu, Z. Wang, L. Su, K. Asomaning, and D. C. Christiani. Genome-wide analysis of survival in early-stage non-small-cell lung cancer. *Journal of Clinical Oncology*, 27(16):2660–2667, 2009.

[78] D. Hunter and P. Kraft. Drinking from the fire hose–statistical issues in genomewide association studies. *N. Engl. J. Med.*, 357(5):436–439, 2007.

[79] 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.

[80] T. Ishii, R. Wakabayashi, H. Kurosaki, A. Gemma, and K. Kida. Association of serotonin transporter gene variation with smoking, chronic obstructive pulmonary disease, and its depressive symptoms. *J. Hum. Genet.*, 56(1):41–46, Jan 2011.

[81] H. Jiang, L. An, S. M. Lin, G. Feng, and Y. Qiu. A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. *PLoS ONE*, 7(10):e46450, 2012.

[82] P. Jintaridth and A. Mutirangura. Distinctive patterns of age-dependent hypomethylation in interspersed repetitive sequences. *Physiological genomics*, 41(2):194–200, 2010.

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

[84] 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.

[85] B. R. Joubert, S. E. Håberg, R. M. Nilsen, X. Wang, S. E. Vollset, S. K. Murphy, Z. Huang, C. Hoyo, Ø. Midttun, L. A. Cupul-Uicab, et al. 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*, 120:1425–31, 2012.

[86] C. M. Judd and D. A. Kenny. Process Analysis. *Evaluation Review*, 5(5):602–619, Oct. 1981.

[87] D. Kenny, D. Kashy, and N. Bolger. Data analysis in social psychology. 1998.

[88] A. Khalili, T. Huang, and S. Lin. A Robust Unified Approach to Analyzing Methylation and Gene Expression Data. *Comput Stat Data Anal*, 53(5):1701–1710, Mar 2009.

[89] Y.-I. Kim, A. Giuliano, K. D. Hatch, A. Schneider, M. A. Nour, G. E. Dallal, J. Selhub, and J. B. Mason. Global dna hypomethylation increases progressively in cervical dysplasia and carcinoma. *Cancer*, 74(3):893–899, 2006.

[90] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applicatoins*, 33(1):82–95, 1970.

[91] L. Klei and K. Roeder. Testing for association based on excess allele sharing in a sample of related cases and controls. *Hum. Genet.*, 121(5):549–557, Jun 2007.

[92] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, Apr 2005.

[93] A. F. H. Kristopher J. Preacher. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3):879–891, 2008.

[94] Z. Kutalik, J. S. Beckmann, and S. Bergmann. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat Biotechnol*, 26(5):531–9, 2008.

[95] B. Kwabi-Addo, W. Chung, L. Shen, M. Ittmann, T. Wheeler, J. Jelinek, and J.-P. J. Issa. Age-related dna methylation changes in normal human prostate tissues. *Clinical cancer research*, 13(13):3796–3802, 2007.

[96] 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.

[97] C. Ladd-Acosta, M. J. Aryee, J. M. Ordway, and A. P. Feinberg. Comprehensive high-throughput arrays for relative methylation (CHARM). *Curr Protoc Hum Genet*, Chapter 20:1–19, Apr 2010.

[98] P. W. Laird et al. The power and the promise of dna methylation markers. *Nature Reviews Cancer*, 3:253–266, 2003.

[99] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*, 91(2):224–37, 2012.

[100] 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.

[101] S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, 88(3):294–305, Mar 2011.

[102] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, 83(3):311–321, Sep 2008.

[103] 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.

[104] Y. Li, B. M. Tesson, G. A. Churchill, and R. C. Jansen. Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in Genetics*, 26(12):493 – 498, 2010.

[105] Y. Li, J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, J. Sun, Y. Huang, H. Zheng, H. Cao, X. Zhou, S. Guo, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck, J. Wang, and X. Zhang. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.*, 8(11):e1000533, 2010.

[106] W. Y. Lin and D. J. Schaid. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet. Epidemiol.*, 33(3):183–197, Apr 2009.

[107] X. Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

[108] 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.

[109] M. A. Lindquist. Functional causal mediation analysis with application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309, 2012.

[110] C. Y. Liu, M. C. Wu, F. Chen, M. Ter-Minassian, K. Asomaning, R. Zhai, Z. Wang, L. Su, R. S. Heist, M. H. Kulke, X. Lin, G. Liu, and D. C. Christiani. A Large-scale genetic association study of esophageal adenocarcinoma risk. *Carcinogenesis*, 31(7):1259–1263, Jul 2010.

[111] 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.

[112] 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.

[113] 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.

[114] J. Liu, Y. Pei, C. J. Papasian, and H. W. Deng. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet. Epidemiol.*, 33(3):217–227, Apr 2009.

[115] 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.

[116] A. E. Locke, K. J. Dooley, S. W. Tinker, S. Y. Cheong, E. Feingold, E. G. Allen, S. B. Freeman, C. P. Torfs, C. L. Cua, M. P. Epstein, M. C. Wu, X. Lin, G. Capone, S. L. Sherman, and L. J. Bean. Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome. *Genet. Epidemiol.*, 34(6):613–623, Sep 2010.

[117] C. Lozupone, M. Hamady, and R. Knight. Unifrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, 7:371, 2006.

[118] C. Lozupone and R. Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235, Dec 2005.

[119] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight. Unifrac: an effective distance metric for microbial community comparison. *ISME J*, 5(2):169–72, 2011.

[120] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73(5):1576–1585, Mar 2007.

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

[122] D. P. Mackinnon and J. H. Dwyer. Estimating Mediated Effects in Prevention Studies. *Evaluation Review*, 17(2):144–158, Apr. 1993.

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

[124] D. P. MacKinnon, J. L. Krull, and C. M. Lockwood. Equivalence of the mediation, confounding and suppression effect. *Prevention science : the official journal of the Society for Prevention Research*, 1(4):173–181, Dec. 2000.

[125] D. P. MacKinnon, C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, 7(1):83–104, Mar. 2002.

[126] D. P. Mackinnon, C. M. Lockwood, and J. Williams. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research*, 39(1):99–128, 2004.

[127] D. P. Mackinnon, G. Warsi, and J. H. Dwyer. A Simulation Study of Mediated Effect Measures. *Multivariate Behav Res*, 30(1):41, Jan 1995.

[128] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, 5(2):e1000384, Feb 2009.

[129] S. Maegawa, G. Hinkal, H. S. Kim, L. Shen, L. Zhang, J. Zhang, N. Zhang, S. Liang, L. A. Donehower, and J.-P. J. Issa. Widespread and tissue specific age-related dna methylation changes in mice. *Genome research*, 20(3):332–340, 2010.

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

[131] 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.

[132] J. Maksimovic, L. Gordon, and A. Oshlack. Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome Biology*, 13(6):R44, 2012.

[133] T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, 363(2):166–176, Jul 2010.

[134] 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.

[135] E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[136] C. J. Marsit, B. C. Christensen, E. A. Houseman, M. R. Karagas, M. R. Wrensch, R.-F. Yeh, H. H. Nelson, J. L. Wiemels, S. Zheng, M. R. Posner, et al. Epigenetic profiling reveals etiologically distinct patterns of dna methylation in head and neck squamous cell carcinoma. *Carcinogenesis*, 30(3):416–422, 2009.

[137] B. McArdle and M. Anderson. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

[138] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, Aug 2008.

[139] N. J. Meyer, Z. J. Daye, M. Rushefski, R. Aplenc, P. N. Lanken, M. G. Shashaty, J. D. Christie, and R. Feng. SNP-set analysis replicates acute lung injury genetic risk factors. *BMC Med. Genet.*, 13:52, 2012.

[140] J. Millstein, B. Zhang, J. Zhu, and E. E. Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genet*, 10:23, 2009.

[141] S. Monti, B. Chapuy, K. Takeyama, S. J. Rodig, Y. Hao, K. T. Yeda, H. Inguilizian, C. Mermel, T. Currie, A. Dogan, J. L. Kutok, R. Beroukhim, D. Neuberg, T. M. Habermann, G. Getz, A. L. Kung, T. R. Golub, and M. A. Shipp. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large b cell lymphoma. *Cancer Cell*, 22(3):359–72, 2012.

[142] J. H. Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.*, 56(1-3):73–82, 2003.

[143] S. Morgenthaler and W. G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, 615(1-2):28–56, Feb 2007.

[144] A. P. Morris and E. Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, 34(2):188–193, Feb 2010.

[145] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, Jul 2008.

[146] B. N and C. D. Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88:9–25, 1993.

[147] C. G. A. R. Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, 2008.

[148] C. G. A. R. Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, Sep 2012.

[149] C. G. A. R. Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–25, 2012.

[150] C. G. A. R. Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct 2012.

[151] M. Oda, J. L. Glass, R. F. Thompson, Y. Mo, E. N. Olivier, M. E. Figueroa, R. R. Selzer, T. A. Richmond, X. Zhang, L. Dannenberg, R. D. Green, A. Melnick, E. Hatchwell, E. E. Bouhassira, A. Verma, M. Suzuki, and J. M. Greally. High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res.*, 37(12):3829–3839, Jul 2009.

[152] W. Pan. Relationship between genomic distance based regression and kernel machine regression for multi-marker association testing. *Genetic epidemiology*, 35(4):211–216, 2011.

[153] J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 411–420, San Francisco, CA, 2001. Morgan Kaufmann.

[154] G. M. Poage, E. A. Houseman, B. C. Christensen, R. A. Butler, M. Avissar-Whiting, M. D. McClean, T. Waterboer, M. Pawlita, C. J. Marsit, and K. T.

Kelsey. Global hypomethylation identifies loci targeted for hypermethylation in head and neck cancer. *Clinical Cancer Research*, 17(11):3579–3589, 2011.

[155] A. Portela and M. Esteller. Epigenetic modifications and human disease. *Nat. Biotechnol.*, 28(10):1057–1068, Oct 2010.

[156] A. L. Price, G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, 86(6):832–838, Jun 2010.

[157] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, 69(1):124–137, Jul 2001.

[158] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, 69(1):1–14, Jul 2001.

[159] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541, 2011.

[160] J. Ramsay and B. Silverman. *Functional data analysis*. Wiley Online Library, 2005.

[161] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17(9):502–510, Sep 2001.

[162] W. M. Rideout, K. Eggan, and R. Jaenisch. Nuclear cloning and epigenetic reprogramming of the genome. *Science*, 293(5532):1093–1098, Aug 2001.

[163] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996.

[164] J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–55, 1992.

[165] T. Rönn, P. Poulsen, O. Hansson, J. Holmkvist, P. Almgren, P. Nilsson, T. Tuomi, B. Isomaa, L. Groop, A. Vaag, et al. Age influences dna methylation and gene expression of cox7a1 in human skeletal muscle. *Diabetologia*, 51(7):1159–1168, 2008.

[166] S. L. B. Rosas, W. Koch, M. d. G. da Costa Carvalho, L. Wu, J. Califano, W. Westra, J. Jen, and D. Sidransky. Promoter hypermethylation patterns of p16, o6-methylguanine-dna-methyltransferase, and death-associated protein kinase in tumors and saliva of head and neck cancer patients. *Cancer Research*, 61(3):939–942, 2001.

[167] J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, N. Bhardwaj, M. Rubin, M. Snyder, and M. Gerstein. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, 7:522, 2011.

[168] D. D. Rucker;, K. J. Preache;, Z. L. Tormala;, and R. E. Petty;. Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, 5(6), 2011.

[169] W. Ruppert and R. Carroll. *Semiparametric regression*. Cambridge University Press, 2003.

[170] A. F. Saad, W. D. Meacham, A. Bai, V. Anelli, S. Elojeimy, A. E. Mahdy, L. S. Turner, J. Cheng, A. Bielawska, J. Bielawski, T. E. Keane, L. M. Obeid, Y. A. Hannun, J. S. Norris, and X. Liu. The functional effects of acid ceramidase overexpression in prostate cancer progression and resistance to chemotherapy. *Cancer Biol Ther*, 6(9):1455–60, 2007.

[171] M. Sanchez-Cespedes, M. Esteller, L. Wu, H. Nawroz-Danish, G. H. Yoo, W. M. Koch, J. Jen, J. G. Herman, and D. Sidransky. Gene promoter hypermethylation in tumors and serum of head and neck cancer patients. *Cancer research*, 60(4):892–895, 2000.

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

[173] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.*, 103(5):1412–1417, Jan 2006.

[174] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 37(7):710–7, 2005.

[175] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, 70(2):425–434, Feb 2002.

[176] D. J. Schaid, J. P. Sinnwell, G. D. Jenkins, S. K. McDonnell, J. N. Ingle, M. Kubo, P. E. Goss, J. P. Costantino, D. L. Wickerham, and R. M. Weinshilboum. Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet Epidemiol*, 36(1):3–16, 2011.

[177] B. Schalkopf and A. Smola, editors. *Learning with Kernels*. MIT Press, Cambridge, Massachusetts, 2002.

[178] U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, 24(3):236–44, 2000.

[179] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, 19(3):212–219, Jun 2009.

[180] S. Self and K. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of American Statistical Association*, 82:605–610, 1987.

[181] Q. Sha, H. S. Chen, and S. Zhang. A new association test using haplotype similarity. *Genet. Epidemiol.*, 31(6):577–593, Sep 2007.

[182] U. T. Shankavaram, W. C. Reinhold, S. Nishizuka, S. Major, D. Morita, K. K. Chary, M. A. Reimers, U. Scherf, A. Kahn, D. Dolginow, J. Cossman, E. P. Kaldjian, D. A. Scudiero, E. Petricoin, L. Liotta, J. K. Lee, and J. N. Weinstein. Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, 6(3):820–832, 2007.

[183] P. Sharma, J. Kumar, G. Garg, A. Kumar, A. Patowary, G. Karthikeyan, L. Ramakrishnan, V. Brahmachari, and S. Sengupta. Detection of altered global dna methylation in coronary artery disease patients. *DNA and cell biology*, 27(7):357–365, 2008.

[184] J. Shen, S. Wang, Y.-J. Zhang, H.-C. Wu, M. G. Kibriya, F. Jasmine, H. Ahsan, D. P. Wu, A. B. Siegel, H. Remotti, et al. Exploring genome-wide dna methylation profiles altered in hepatocellular carcinoma using infinium humanmethylation 450 beadchips. *Epigenetics*, 8(1):0–1, 2013.

[185] D. Shriner and L. K. Vaughan. A unified framework for multi-locus association analysis of both common and rare variants. *BMC Genomics*, 12:89, 2011.

[186] P. E. Shrout and N. Bolger. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods*, 7(4):422–45, 2002.

[187] I. M. Shui, L. A. Mucci, P. Kraft, R. M. Tamimi, S. Lindstrom, K. L. Penney, K. Nimptsch, B. W. Hollis, N. Dupre, E. A. Platz, M. J. Stampfer, and E. Giovannucci. Vitamin d-related genetic variation, plasma vitamin d, and risk of

lethal prostate cancer: a prospective nested case-control study. *J Natl Cancer Inst*, 104(9):690–9, 2012.

[188] K. D. Siegmund. Statistical approaches for the analysis of DNA methylation microarray data. *Hum. Genet.*, 129(6):585–595, Jun 2011.

[189] D. K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.*, 32 Suppl:502–508, Dec 2002.

[190] J. Soares, A. E. Pinto, C. V. Cunha, S. Andre, I. Barão, J. M. Sousa, and M. Cravo. Global dna hypomethylation in breast carcinoma. *Cancer*, 85(1):112–118, 1999.

[191] M. E. Sobel. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13:290–312, 1982.

[192] M. E. Sobel. Direct and indirect effects in linear structural equation models. *Sociological Methods & Research*, 16(1):155–176, 1987.

[193] M. E. Sobel. Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, 33(2):230–251

[194] T. Sorenson. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*, 5(1-34):4.7, 1948.

[195] J. E. Staunton, D. K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, U. Scherf, J. K. Lee, W. O. Reinhold, J. N. Weinstein, J. P. Mesirov, E. S. Lander, and T. R. Golub. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, 98(19):10787–92, 2001.

[196] R. Straussman, D. Nejman, D. Roberts, I. Steinfeld, B. Blum, N. Benvenisty, I. Simon, Z. Yakhini, and H. Cedar. Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.*, 16(5):564–571, May 2009.

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

[198] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 2005.

[199] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. *Machine Learning*, 79:73–103, 2010.

[200] M. S. Teng, L. A. Hsu, S. Wu, H. H. Chou, C. J. Chang, Y. Z. Sun, S. H. Juan, and Y. L. Ko. Mediation analysis reveals a sex-dependent association between ABO gene variants and TG/HDL-C ratio that is suppressed by sE-selectin level. *Atherosclerosis*, 228(2):406–412, Jun 2013.

[201] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–196, 2013.

[202] H. K. Tiwari and R. C. Elston. Restrictions on components of variance for epistatic models. *Theor Popul Biol*, 54(2):161–174, Oct 1998.

[203] N. Touleimat and J. Tost. Complete pipeline for infinium® human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics*, 4(3):325–341, 2012.

[204] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1031, Dec 2006.

[205] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, 98(9):5116–5121, Apr 2001.

[206] B. Tycko. Allele-specific DNA methylation: beyond imprinting. *Hum. Mol. Genet.*, 19(R2):R210–220, Oct 2010.

[207] J. Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.*, 72(4):891–902, Apr 2003.

[208] J. Y. Tzeng and D. Zhang. Haplotype-based association analysis via variance-components score test. *Am. J. Hum. Genet.*, 81(5):927–938, Nov 2007.

[209] T. VanderWeele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468, 2009.

[210] T. J. VanderWeele, K. Asomaning, E. J. Tchetgen Tchetgen, Y. Han, M. R. Spitz, S. Shete, X. Wu, V. Gaborieau, Y. Wang, J. McLaughlin, R. J. Hung, P. Brennan, C. I. Amos, D. C. Christiani, and X. Lin. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. *Am. J. Epidemiol.*, 175(10):1013–1020, May 2012.

[211] T. J. VanderWeele and S. Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 2010.

[212] J. S. Verducci, V. F. Melfi, S. Lin, Z. Wang, S. Roy, and C. K. Sen. Microarray analysis of gene expression: considerations in data mining and statistical treatment. *Physiol. Genomics*, 25(3):355–363, May 2006.

[213] C. J. Verzilli, N. Stallard, and J. C. Whittaker. Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. *Genet. Epidemiol.*, 28(4):313–325, May 2005.

[214] 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.

[215] G. Wahba, editor. *Spline Models for Observational Data*. SIAM, 1990.

[216] J. Wang, M. R. Spitz, C. I. Amos, A. V. Wilkinson, X. Wu, and S. Shete. Mediating effects of smoking and chronic obstructive pulmonary disease on the relation between the CHRNA5-A3 genetic locus and lung cancer risk. *Cancer*, 116(14):3458–3462, Jul 2010.

[217] J. Wang, M. R. Spitz, C. I. Amos, X. Wu, D. W. Wetter, P. M. Cinciripini, and S. Shete. Method for evaluating multiple mediators: mediating effects of smoking and COPD on the association between the CHRNA5-A3 variant and lung cancer risk. *PLoS ONE*, 7(10):e47705, 2012.

[218] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81(6):1278–83, 2007.

[219] X. M. Wang, T. C. Greiner, M. Bibikova, B. L. Pike, K. D. Siegmund, U. K. Sinha, M. Muschen, E. B. Jaeger, D. D. Weisenburger, W. C. Chan, D. Shibata, J. B. Fan, and J. G. Hacia. Identification and functional relevance of de novo DNA methylation in cancerous B-cell populations. *J. Cell. Biochem.*, 109(4):818–827, Mar 2010.

[220] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schubeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, 37(8):853–862, Aug 2005.

[221] J. N. Weinstein, T. G. Myers, P. M. O'Connor, S. H. Friend, J. Fornace, A. J., K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes, and K. D. Paull. An information-intensive approach to the molecular pharmacology of cancer. *Science*, 275(5298):343–9, 1997.

[222] J. Wessel and N. J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet*, 79(5):792–806, 2006.

[223] S. M. Williams, J. L. Haines, and J. H. Moore. The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? *Bioessays*, 26(2):170–9, 2004.

[224] V. L. Wilson, R. Smith, S. Ma, and R. Cutler. Genomic 5-methyldeoxycytidine decreases with age. *Journal of Biological Chemistry*, 262(21):9948–9951, 1987.

[225] S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Ass*, 99(467):673–686, 2004.

[226] S. N. Wood, editor. *Generalized additive models: an introduction with R*, volume 66. CRC press, 2006.

[227] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society*, 73(1):3–36, 2011.

[228] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Comput. Biol.*, 6(2):e1000667, Feb 2010.

[229] 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.

[230] 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, 2011.

[231] 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.

[232] M. C. Wu and X. Lin. Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Stat Methods Med Res*, 18(6):577–93, 2009.

[233] 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.

[234] A. S. Yang, M. R. Estécio, K. Doshi, Y. Kondo, E. H. Tajara, and J.-P. J. Issa. A simple method for estimating global dna methylation using bisulfite pcr of repetitive dna elements. *Nucleic acids research*, 32(3):e38–e38, 2004.

[235] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, Jul 2010.

[236] A. Yuan, Q. Yue, V. Apprey, and G. Bonney. Detecting disease gene in DNA haplotype sequences by nonparametric dissimilarity test. *Hum. Genet.*, 120(2):253–261, Sep 2006.

[237] M. A. Zapala and N. J. Schork. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci. U.S.A.*, 103(51):19430–19435, Dec 2006.

[238] D. Zhang and X. Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.

[239] H. Zhang, C. T. Liu, and X. Wang. An Association Test for Multiple Traits Based on the Generalized Kendall's Tau. *J Am Stat Assoc*, 105(490):473–481, Jun 2010.

[240] 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.

[241] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements, 2013.

[242] J. Zhou, M. Tawk, F. D. Tiziano, J. Veillet, M. Bayes, F. Nolent, V. Garcia, S. Servidei, E. Bertini, F. Castro-Giner, Y. Renda, S. Carpentier, N. Andrieu-Abadie, I. Gut, T. Levade, H. Topaloglu, and J. Melki. Spinal muscular atrophy associated with progressive myoclonic epilepsy is caused by mutations in asah1. *Am J Hum Genet*, 91(1):5–14, 2012.