STATISTICAL METHODS FOR GENETIC AND EPIGENETIC ASSOCIATION STUDIES

Kuan-Chieh Huang

A dissertation submitted to the faculty at 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 in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:
Mengjie Chen
Ethan M Lange
Yun Li
Kari E. North
Wei Sun

ABSTRACT

Kuan-Chieh Huang: Statistical Methods for Genetic and Epigenetic Association Studies
(Under the direction of Yun Li)

First, in genome-wide association studies, few methods have been developed for rare variants which are one of the natural places to explain some missing heritability left over from common variants. Therefore, we propose EM-LRT that incorporates imputation uncertainty for downstream association analysis, with improved power and/or computational efficiency. We consider two scenarios: I) when posterior probabilities of all possible genotypes are estimated; and II) when only the one-dimensional summary statistic, imputed dosage, is available. Our methods show enhanced statistical power over existing methods and are computationally more efficient than the best existing method for association analysis of variants with low frequency or imputation quality.

Second, although genome-wide association studies have identified a large number of loci associated with complex traits, a substantial proportion of the heritability remains unexplained. Thanks to advanced technology, we may now conduct large-scale epigenome-wide association studies. DNA methylation is of particular interest because it is highly dynamic and has been shown to be associated with many complex human traits, including immune dysfunctions, cardiovascular diseases, multiple cancer, and aging. We propose FunMethyl, a penalized functional regression framework to perform association testing between multiple DNA methylation sites in a region and a quantitative outcome. Our results from both real data based

simulations and real data clearly show that FunMethyl outperforms single-site analysis across a wide spectrum of realistic scenarios.

Finally, large studies may have a mixture of old and new arrays, or a mixture of old and new technologies, on the large number of samples they investigate. These different arrays or technologies usually measure different sets of methylation sites, making data analysis challenging. We propose a method to predict site-specific DNA methylation level from one array to another – a penalized functional regression model that uses functional predictors to capture non-local correlation from non-neighboring sites and covariates to capture local correlation. Application to real data shows promising results: the proposed model can predict methylation level at sites on a new array reasonably well from those on an old array.

To my family and wife, I couldn't have done this without you.
Thank you for all of your support along the way.

ACKNOWLEDGMENTS

I would like to offer countless thanks to my adviser, Yun Li, for her tireless support and encouragement, and each of my committee members Wei Sun, Ethan Lange, Kari North, and Mengjie Chen for their thoughtful contributions to this research. I would also like to thank my collaborators from UNC-Chapel Hill for their invaluable comments and suggestions: Karen Mohlke and Leslie Lange from the Department of Genetics for providing the real data used in the illustrative real data example in chapter 3. Finally, I would like to thank my co-workers, friends, and family for their patience and support of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AML	Acute Myeloid Leukemia
ARIC	Atherosclerosis Risk in Communities Study
BMI	Body Mass Index
bp	Base Pair
CARDIA	Coronary Artery Risk Development in Young Adults
CLHNS	Cebu Longitudinal Health and Nutrition Survey
CNV	Copy Number Variant
EWAS	Epigenome-wide Association Study
GCV	Generalized Cross-Validation
GWAS	Genome-wide Association Study
HM27	HumanMethylation27
HM450	HumanMethylation450
JHS	Jackson Heart Study
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein
MaCH	Markov Chain Haplotyper
MAF	Minor Allele Frequency
OLS	Ordinary Least Square
PC	Principle Component
PLINK	PuTTY Link
REML	Restricted Maximum Likelihood

SKAT	SNP-Set (Sequence) Kernal Association Test
SKAT-O	Optimal SNP-Set (Sequence) Kernal Association Test
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TFBS	Transcription Factor Binding Site
WC	Waist Circumference
WHI	Women's Health Initiative

CHAPTER 1:MOTIVATION AND BIOLOGICAL JUSTIFICATION

In this document, we propose statistical methods for assessing association between a genetic variant or sets of epigenetic marks and complex human traits. Because large studies may have a mixture of old and new arrays, or a mixture of old and new technologies, on the large number of epigenetic marks investigated, we also propose a method to predict site-specific DNA methylation level from one array to another. This section provides an overview of the biological problems we are interested in and some of the statistical strategies employed in an attempt to solve them.

To begin, DNA is a double-stranded molecule consisting of four nucleic acid components: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). DNA is found in the nucleus of the vast majority of plant and animal cells and has been compared to a blueprint for the organism in which it is found. Humans have 22 autosomes, in addition to the sex chromosomes X and Y and mitochondrial DNA, accounting for over 5 billion base pairs in total. We will consider primarily autosomal DNA, for which each individual possesses two copies, one inherited maternally and the other paternally. Over 99% of the DNA sequence is the same across humans (Morris and Zeggini 2010); however, there are a large number of ways in which human DNA sequence can differ from one another in a single region including microsatellites, copy number variations (CNVs), insertions, deletions, inversions and single nucleotide polymorphisms (SNPs). Any one of these can be called a genetic variant, meaning that it contains a sequence of nucleic acids that is different from the consensus sequence or from what is most common.

A SNP is one such genetic variant that occupies only one base pair. As previously stated, much of the genome is shared across humans; however, some of these variants, SNPs included, are quite common with variant or minor allele frequency (MAF) near 0.5. In the 1990's microarray technologies from companies like Affymetrix and Illumina began to capitalize on these common SNPs in the form of genome-wide SNP platforms. Today these technologies can accurately assess up as many as 1 million pre-selected SNPs [e.g. the Affy Axiom or Illumina 1M], however these technologies are limited in that they cannot discover new variants.

As SNP genotyping technology advances, it is now possible to genotype hundreds of thousands of alleles in parallel. This has made it possible to rapidly scan markers across the complete genomes of many people; therefore, the association between the traits of interest and millions of markers could be tested. Recently, genome-wide association studies (GWASs) have identified SNPs related to several complex diseases. The completion of The International HapMap Project has provided a possibility to impute missing genotypes that were not directly genotyped from a cohort or case-control study but were genotyped in the reference samples. Genotype imputation relies on the fact that even two unrelated individuals can share short stretches of haplotype inherited from distant common ancestors. Several methods have been proposed to take imputation into account in genome-wide association studies (GWASs). These existing methods, however, have focused primarily on common variants, which have been the focus of the past wave of GWASs examining either only directly genotyped (or typed) markers, or typed and untyped markers imputed with the aid of smaller scale, lower density reference panels such as those from the International HapMap Project. However, few (if any) methods have been developed for rare variants which have been receiving intensive attention in the past half decade as one of the natural places to explain some missing heritability left over from

common variants, for almost all complex traits studied in the genetics community. In the third chapter, an expectation-maximization likelihood-ratio test (EM-LRT) is developed. This method can accommodate either posterior genotype probabilities (EM-LRT-Prob) or imputed dosages (EM-LRT-Dose). We evaluated our methods and compared them with existing methods through extensive simulations. Our methods clearly show enhanced statistical power over existing methods and computationally more efficient alternative to the best existing method for association analysis of variants with low MAF or imputation quality. We also applied our methods to two data sets: the Cebu Longitudinal Health and Nutrition Survey (CLHNS) and Women's Initiative Study (WHI) of Blood Cell Traits. Consequently, all methods have proper control of type I error and our methods generated more significant *p*-values (and better approached truth in all cases), suggesting power enhancement using our methods.

Although GWASs have identified a large number of loci associated with complex traits, a substantial proportion of the heritability remains unexplained. For example, the >200 loci identified for height can only explain ~20% out of the ~80% total estimated heritability. Recent technological advances have allowed us to conduct large-scale epigenome-wide association studies (EWASs). DNA methylation is of particular interest because it is highly dynamic (Rakyan et al. 2011) and has been shown to be associated with many complex human traits, including immune dysfunctions, cardiovascular diseases, multiple cancer, and aging. Typically, methylation level at hundreds of thousands of sites is measured and each of these sites is examined separately (i.e., single-site analysis). However, because of the correlation structure among the sites and because many of them fall in naturally defined regions (e.g., belonging to the same gene; belonging to the same regulatory region such as an enhancer or DNAse hypersensitivity site), it is conceptually straightforward to imagine achieving enhanced statistical

power by performing region-based test (that is, simultaneously testing multiple sites together) especially when there are multiple low or moderate signals in that region. In the fourth chapter, a penalized functional region-based model is proposed to perform the association testing between DNA methylation marks in a region (explanatory variable) and quantitative trait (response variable). We evaluated our methods and compared them with the benchmark single-site analysis through extensive real data simulations. All the methods have proper control of type I error; however, our methods have enhanced statistical power over the single-site analysis across various settings. Moreover, our methods have much higher statistical power than the existing region-based tests SKAT and SKAT-O. This work is close to finish and we expect to submit the manuscript soon.

Lack of high-throughput profiling technologies used to hinder our understanding of the dynamic state of DNA methylation. Fortunately, geneticists are embraced nowadays by technological advances. For the study of DNA methylation, for example, technological advances constantly provide us with more choices to measure DNA methylation patterns across the genome, including multiple commercial arrays, multiple sequencing-based technologies or protocols (Laird 2010). However, large studies may have a mixture of old and new arrays, or a mixture of old and new technologies, on the large number of samples they investigate. These different arrays or technologies usually measure different sets of methylation sites, making data analysis challenging, if not even impossible. For example, Illumina HumanMethylation27 (HM27) and HumanMethylation450 (HM450) BeadChip are two common microarrays used by the Cancer Genome Atlas (TCGA) project. In several TCGA studies, the DNA methylation profiles of samples collected more recently were measured by HM450, while the others were still measured by HM27. Then when researchers want to utilize data of all samples for downstream

analysis, they can only focus on probes shared between two platforms for simplicity, since re-evaluating all samples using HM450 is both costly and time-consuming. In the fifth chapter, a penalized function regression model is proposed for DNA methylation prediction. We applied the proposed model to a large-scale methylation data set from acute myeloid leukemia patients. As a result, the proposed model can produce accurate imputations when the reference panel (training set) and the target panel (testing set) characterize the same tissue under similar conditions.

CHAPTER 2:LITERATURE REVIEW

This section presents a partial review of many of the papers previously published on the topic of association studies. It is by no means complete since the number of these papers is quite large; however, it is an attempt to show the development of several methods used for these studies.

2.1 Early Methods for Association Studies

Genotype imputation has become standard practice in modern genetic studies (Browning and Browning 2008; Li et al. 2009; Li et al. 2010; Marchini and Howie 2010). For each untyped variant imputed, standard imputation methods estimate posterior probabilities of all possible genotypes. For example, when the untyped variant is bi-allelic with alleles A and B, we obtain posterior probabilities for A/A, A/B, and B/B with the constraint of summation being one. Such probability information can be further summarized into degenerate one-dimensional summary statistics including the mode (the best-guess genotype, or the genotype with the highest posterior probability), or the mean (the imputed dosage).

Since association analysis with phenotypes of interest rather than genotype imputation per se is usually of the ultimate interest, development and evaluation of post-imputation association strategies have therefore attracted considerable attention from the research community (Chen and Abecasis 2007; Lin et al. 2008; Aulchenko et al. 2010; Pei et al. 2010; Jiao et al. 2011; Kutalik et al. 2011; Zheng et al. 2011; Acar and Sun 2013; Liu et al. 2013b). Among them, imputation dosage based methods provide an attractive compromise between

modeling complexity, computational efficiency and statistical power, have been shown analytically to be optimal among methods based on one-dimensional summary statistics (Liu et al. 2013b), and thus have been most commonly adopted in recent imputation-aided genome-wide association studies (Chambers et al. 2011; Auer et al. 2012; Dastani et al. 2012; Berndt et al. 2013). On the other hand, explicitly modeling the probabilities of all possible genotypes using the mixture of regression models (abbreviated Mixture hereafter and detailed below) has the best performance in terms of statistical efficiency, particularly with low imputation quality, but at the cost of increased computational complexity (Zheng et al. 2011).

GWASs have successfully identified sites associated with common diseases but still, a substantial proportion of the causality remains unexplained. In fact, GWASs only study the association between trait and genetic variants at the DNA level, and also some single nucleotide polymorphisms (SNPs) associated with common diseases are not localized near any gene in the pathways involved. Therefore, the unexplained causality could be found in epigenetic variation. Based on the experiences from GWASs, it is inevitable to perform large-scale and systematic studies to detect the epigenetic variation. Epigenome-wide association studies (EWASs) allow us to identify genome-wide epigenetic variants associated with common diseases. DNA methylation is of particular interest because it is highly dynamic and, also the profiling technology for both array- and sequencing-based methods has been well developed. Among these technologies, the whole-genome bisulphite sequencing provides the highest coverage and resolution.

2.1.1 GWAS

We will introduce several existing methods that are widely used in GWAS. First, let $F_i = (f_{i0}, f_{i1}, f_{i2})$ represent the genotype probability vector and X_i represent a particular feature of the imputation procedure

$$X_i = \begin{cases} \underset{j\in\{0,1,2\}}{\operatorname{argmax}}\{f_{ij}\} & , \quad \text{Best-guess} \\ f_{i1} + 2f_{i2} & , \quad \text{Dosage} \end{cases}$$

Best-guess and Dosage methods directly regress the trait Y_i on a particular feature X_i adjusting for the covariates Z_i

$$Y_i = \beta_0 + \beta_1 X_i + \gamma Z_i + \varepsilon_i$$

where $\varepsilon_i \sim N\left(0, \sigma^2\right)$ and i = 1, 2, …, N with N being the sample size. Next, the Mixture method (Zheng et al. 2011) fits the following mixture of regression model

$$Y_i = \sum_{j=0}^{2} f_{ij} \cdot g_j\left(\beta_0, \beta_1, \gamma, \varepsilon_i\right)$$

where $g_j\left(\beta_0, \beta_1, \gamma, \varepsilon_i\right) = \beta_0 + j \cdot \beta_1 + \gamma Z_i + \varepsilon_i$. To estimate the parameters $\left(\beta_0, \beta_1, \gamma\right)$ in the Mixture model, the log-likelihood function is maximized using the Nelder-Mead Simplex Method (Nelder and Mead 1965), implemented in R package *optim*.

2.1.2 EWAS

Typically, DNA methylation level at hundreds of thousands of sites is measured and each of these sites is examined separately (i.e., single-site analysis). For a given CpG site *j*, the methylated (M_j) and unmethylated (U_j) signal intensities are combined to methylation β-values:

$$\beta - \text{value}_j = \frac{\max\left(M_j, 0\right)}{\max\left(M_j, 0\right) + \max\left(U_j, 0\right) + \alpha_\beta}$$

where the inclusion of an offset α_β= 100 is recommended as a stabilization in the situation when both methylated and unmethylated signal intensities are small. By definition, the β-values are bounded to [0,1] interval and can be interpreted as an approximation of the percentage of

methylation. Most regression models used in EWAS treats DNA methylation level as response variable and disease-related phenotypes as explanatory variables. To directly use β-values without transformation, Beta regression (Ferrari and Cribari-Neto 2004) designed for modeling proportions bounded to [0,1] is often used.

It was later proposed to use the log_2 ratio of the methylated to unmethylated signal intensities (Allison et al. 2006):

$$M-\text{value}_j = \log_2\left(\frac{\max\left(M_j,0\right)+\alpha_M}{\max\left(U_j,0\right)+\alpha_M}\right)$$

where $\alpha_M = 1$ is usually specified and the *M*-values are therefore defined on $\left(-\infty,\infty\right)$. It has also been shown that after ignoring α_β and α_M, there is a logit (with base 2) relationship between β-values and *M*-values:

$$M-\text{value}_j \approx \log_2\left(\frac{\beta-\text{value}_j}{1-\beta-\text{value}_j}\right)$$

As a result, after transforming β-values, the linear regression model for CpG site *j* is fit as $M_i = \alpha+\gamma Z_i+\varepsilon_i$, where M_i is *M*-value$_i$ at site *j*, Z_i is the covariate vector, $\varepsilon_i \sim N\left(0,\sigma^2\right)$, and i = 1, 2, …, *N* with *N* being the sample size. Moreover, if one wants to adjust for batch effect, a mixed model can be easily adopted by specifying a batch-specific random effect in the model. In addition, the feature selection using test statistics is similar for *M*- and β-values for relatively large sample sizes but *M*-values allow more reliable identification of true positives for small sample sizes (Zhuang et al. 2012).

2.2 Early Methods for DNA Methylation Prediction

DNA methylation is an important epigenetic modification involved not only in normal development (Reik 2007; Smith and Meissner 2013) but also in risk and progression to many diseases (Bergman and Cedar 2013). It has been shown to play a key role in the regulation of gene transcription, X-inactivation, cellular differentiation, as well as other critical processes such as aging (Bird 2002; Gonzalo 2010). Recently, the emergence of powerful technologies such as microarray-based DNA methylation studies (Bibikova et al. 2011) and whole-genome bisulfite sequencing (Harris et al. 2010) has enabled the profiling of DNA methylation levels at high resolution. Numerous studies employed these high-throughput approaches to characterize changes of DNA methylation patterns and their corresponding tissue and disease-specific differentially methylated regions on a genome-wide scale (Irizarry et al. 2009; Berman et al. 2011; Varley et al. 2013).

As new technology emerge, researchers tend to replace old methylation profiling platforms with new ones. However, different platforms usually target CpG sites at different locations and resolutions, which hinder joint analysis of data from different platforms. For instance, Illumina HumanMethylation27 (HM27) and HumanMethylation450 (HM450) BeadChip are two common microarrays used by The Cancer Genome Atlas (TCGA) project. While HM27 investigates 27,578 CpG sites predominantly located near CpG islands, HM450 provides broader coverage with 485,577 probes spanning 96% of CpG islands and 92% of CpG shores across a larger number of genes. Several TCGA studies have used HM450 to gather methylation profile data for more recently collected samples while still using HM27 to measure DNA methylation in the older test subjects. These mixed profiles compel researchers to focus on those probes shared between the two platforms when using the data for downstream analysis, as

re-evaluating all samples using HM450 is not only expensive but also time-consuming.

Most existing methods for DNA methylation prediction assume the DNA methylation is binary, the DNA methylation status is 0 for unmethylated and 1 for methylated. They also have limited predictions to specific regions of the genome. Moreover, most existing methods attempt to predict the DNA methylation level or status using HM450 probes as well as some features, such as DNA composition, predicted DNA structure, repeat elements, transcription factor binding sites (TFBSs), evolutionary conservation, and etc.

CHAPTER 3:EM-LRT

3.1 Introduction

Limited evaluations of existing methods (including methods that explicitly model posterior probabilities) on variants with low imputation quality suggest much reduced power compared with accurately imputed variants, for instance, as demonstrated in Figure 2 and 3 of (Zheng et al. 2011) and Figure 2 of (Liu et al. 2013b) Analysis of variants with low imputation quality is not surprisingly a challenging problem due to the low correlation between imputed and true genotypes. It is nevertheless an increasingly important problem because as sequencing-based reference panels continue to grow (Altshuler et al. 2012; Fu et al. 2013)we have increasingly more well imputed markers but also even more markers with relatively low imputation quality, particularly at markers with lower allele frequencies (Duan et al. 2013a)(Altshuler et al. 2010b; Liu et al. 2012; Zhang et al. 2013; Duan et al. 2013b). It is thus highly warranted to seek alternative and potentially more efficient methods to model imputation uncertainty for these markers. In this chapter, we develop expectation-maximization likelihood-ratio tests (EM-LRT) that can accommodate either posterior genotype probabilities, when available (EM-LRT-Prob), or imputed dosages (EM-LRT-Dose). Simulations and real data application demonstrate the validity of the proposed methods and suggest them as a computationally more efficient alternative to the best existing method (Mixture) for association analysis of variants with low MAF or imputation quality.

3.2 Methods

3.2.1 A Hierarchical Modeling Framework to Simulate Data

We adopt a hierarchical model that generates posterior probabilities, imputed dosages, and true genotypes using marker-specific information including minor allele frequency (MAF) and imputation quality measure (R^2), as well as a quantitative trait with which we test for genetic association. The model has three stages: the first stage generates genotype probabilities based on marker-specific information (*genotype probability* stage); the second stage employs a multinomial distribution with probabilities from the first stage to generate allele counts (*allele count* stage); and the final stage fits a linear regression model to generate quantitative trait values (*trait* stage).

Genotype Probability Stage. For a specific marker with MAF q and imputation quality R^2, the genotype probability vector $F_i = (f_{i0}, f_{i1}, f_{i2})$ for the i-th sample is drawn from a Dirichlet distribution with parameters $\alpha = (\alpha_0, \alpha_1, \alpha_2)$, where f_{ij} is the probability of having j copies of the minor allele for the i-th sample and $\sum_{j=0}^{2} f_{ij} = 1$. The parameters in the Dirichlet distribution are: $\alpha_0 = (1-q)^2/c$, $\alpha_1 = 2q(1-q)/c$, $\alpha_2 = q^2/c$ with $c = R^2/(1-R^2)$. Here we give some brief explanations. First, this distribution gives reasonable expected values for $F_i = (f_{i0}, f_{i1}, f_{i2})$ such that $E(f_{i0}) = (1-q)^2$, $E(f_{i1}) = 2q(1-q)$, and $E(f_{i2}) = q^2$, which are the expected probabilities of having 0, 1, or 2 copies of minor alleles assuming Hardy-Weinberg Equilibrium. Next, when R^2 approaches to 1, $F_i = (f_{i0}, f_{i1}, f_{i2})$ approaches to a distribution that takes three possible values, (1,0,0), (0,1,0), and (0,0,1) (i.e., the probability of having a particular genotype is either 0 or 1), which is the expected situation when there is no imputation ambiguity. Given the genotype

probability vector, the imputed dosage is $D_i = f_{i1} + 2f_{i2}$.

Allele Count Stage. The allele count vector $X_i = (x_{i0}, x_{i1}, x_{i2})$ for the *i*-th sample is drawn from a multinomial distribution with genotype probabilities specified in the previous stage, where $x_{ij} = 1$ if the *i*-th sample has *j* copies of the minor allele; and 0 otherwise, with the constraint of $\sum_{j=0}^{2} x_{ij} = 1$. Additionally, the genotype G_i for the *i*-th sample is generated using this allele count vector, specifically $G_i = x_{i1} + 2x_{i2}$. Our simulation framework, taking imputation quality R^2 into account using *c* above, renders $corr^2(G_i, D_i) = R^2$.

Trait Stage. In the final stage, a linear regression model is used to generate quantitative trait Y_i using genotype G_i and covariates Z_i, $Y_i = \beta_0 + \beta_1 G_i + \gamma Z_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ and i = 1, 2, …, N.

3.2.2 Expectation-Maximization Likelihood-Ratio Test

Our primary goal is to test for marker-trait association when marker genotypes G are not directly observed but rather imputed. We propose the following expectation-maximization likelihood ratio tests (EM-LRT). We consider two common scenarios after genotype imputation: 1) when posterior probabilities of genotypes are available and 2) when only dosages are available.

Scenario I: When Posterior Probabilities Are Available [EM-LRT-Prob]. Under this scenario, the true genotype G_i is missing but genotype probability vector $F_i = (f_{i0}, f_{i1}, f_{i2})$ is estimated, i = 1, 2, …, N with N being the sample size. Given the observations (y_i, G_i, z_i, f_i) where f_i is the observed value for F_i, the complete data likelihood is

$$L^*\left(\beta,\sigma,\gamma|y,G,z,f\right)=\prod_{i=1}^{N}f\left(y_i|G_i,z_i,f_i\right)\cdot P\left(G_i|z_i,f_i\right)=\prod_{i=1}^{N}f\left(y_i|G_i,z_i\right)\cdot P\left(G_i|f_i\right)\propto\prod_{i=1}^{N}f\left(y_i|G_i,z_i\right)$$

where the second equality holds because trait y_i is independent of genotype probability vector f_i conditional on true genotype G_i and true genotype is independent of covariates z_i conditional on genotype probability vector. Therefore, with Gaussian distribution, the corresponding complete-data log-likelihood is

$$l^*\left(\beta,\sigma,\gamma|y,G,z,f\right)\propto\sum_{i=1}^{N}-\log\sigma-\left[y_i-\left(\beta_0+\beta_1G_i+\gamma z_i\right)\right]^2\Big/2\sigma^2$$

In this complete data log-likelihood, terms involving true genotype G_i, namely G_i and G_i^2 , are not observed and will be replaced in the *E-step* by their conditional expectations given the observed data. Their conditional expectations are

$$E\left(G_i|y_i,f_i,z_i\right)=\sum_{G_i=0}^{2}G_i\cdot P\left(G_i|y_i,f_i,z_i\right)=C^{-1}\cdot\sum_{G_i=0}^{2}G_i\cdot P\left(y_i|G_i,z_i\right)\cdot P\left(G_i|f_i\right)$$

$$E\left(G_i^2|y_i,f_i,z_i\right)=\sum_{G_i=0}^{2}G_i^2\cdot P\left(G_i|y_i,f_i,z_i\right)=C^{-1}\cdot\sum_{G_i=0}^{2}G_i^2\cdot P\left(y_i|G_i,z_i\right)\cdot P\left(G_i|f_i\right)$$

where $C=\sum_{G_i=0}^{2}f\left(y_i|G_i,z_i\right)\cdot P\left(G_i|f_i\right)$, and $P\left(G_i|f_i\right)=f_{i0}^{I(G_i=0)}f_{i1}^{I(G_i=1)}f_{i2}^{I(G_i=2)}$.

In the *M-step*, the maximum likelihood estimates of the parameter $\theta=\left(\beta_0,\beta_1,\gamma,\sigma\right)$ are obtained as follows:

$$\left(\hat{\beta}_0,\hat{\beta}_1,\hat{\gamma}\right)=\left\{\left[I,G,Z\right]^T\left[I,G,Z\right]\right\}^{-1}\left[I,G,Z\right]^TY=\begin{bmatrix}I^TI & I^TG & I^TZ\\ G^TI & G^TG & G^TZ\\ Z^TI & Z^TG & Z^TZ\end{bmatrix}^{-1}\begin{bmatrix}I^TY\\ G^TY\\ Z^TY\end{bmatrix}$$

$$\hat{\sigma}^2=\left(Y-\hat{\beta}_0-G\hat{\beta}_1-Z\hat{\gamma}\right)^T\left(Y-\hat{\beta}_0-G\hat{\beta}_1-Z\hat{\gamma}\right)\Big/n$$

We repeat the *E-step* and *M-step* until convergence ($\delta<10^{-6}$).

To speed up the EM algorithm, we suggest using the naïve parameter estimates as starting values, that is, the parameter estimates derived by fitting a simple linear regression on

trait *Y* using dosage *D* and covariates *Z* (a.k.a Dosage or standard method). Our EM-LRT-Prob approach shares some similarity with the seminar work by Lander and Botstein (Lander and Botstein 1989) for interval mapping, in which the authors also used mixture model framework, treating genotypes at quantitative trait sites as missing data.

Scenario II: When Only Dosages Are Available [EM-LRT-Dose]. We propose a framework that first uses the conditional (on dosages) distribution to sample genotype probabilities given the imputed dosages, and then apply the EM algorithm detailed above in Scenario I.

First, we derive the probability density function for f_{i1}, the probability of having one copy of the minor allele conditioning on imputed dosage

$$f\left(f_{i1}=p\middle|D_i\right)=\frac{C'}{B(\alpha)}\cdot\left[1-0.5\left(D_i+p\right)\right]^{\alpha_0-1}p^{\alpha_1-1}\left[0.5\left(D_i-p\right)\right]^{\alpha_2-1}$$

where *C'* is the normalizing constant, $p\in\left[0,\min\left(2-D_i,1,D_i\right)\right]$, and $B(\cdot)$ is the beta function [Appendix B]. Second, we select the envelope function $g(p)=\max_p f\left(f_{i1}=p\middle|D_i\right)$ such that $f(p)\le g(p)$ for all *p*. Third, we perform the following steps to sample f_{i1}: 1) generate $p\sim U\left(0,\min\left(2-D_i,1,D_i\right)\right)$; 2) generate $U\sim U(0,1)$; 3) accept p if $U<f(p)/g(p)$. Finally, we calculate f_{i0} and f_{i2} using the relationship $D_i=f_{i1}+2f_{i2}$ and $\sum_{j=0}^{2}f_{ij}=1$.

The drawback of the above rejection sampling approach is that it can be computationally rather expensive especially when the envelope function is large. Fortunately, we can use an approximation approach when MAF is not high. For example, when MAF is low enough, the probability of having two copies of the minor allele is close to zero. In that case, we adopt an approximation approach (referred hereafter as dosage approximation approach) by setting the

probability of having one copy of the minor allele to dosage when MAF is below certain threshold depending on the imputation quality [details are shown below in subsection **Numerical Simulation: MAF Threshold**].

Hypothesis Testing. To assess whether a variant is associated with phenotypic trait of interest *Y*, we perform the following hypothesis testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Note that the same β_1 is assumed across all three possible genotypes. We propose to use the likelihood ratio test for this purpose. Specifically, hypothesis testing is performed as follows: 1) use the EM algorithm described previously to find the ML estimates $\hat{\theta}$ for $\theta = (\beta_0, \beta_1, \gamma, \sigma)$, and then compute the log-likelihood $l^*(\hat{\theta})$; 2) find the ML estimates $\hat{\theta}$ under H_0; and 3) compute the likelihood-ratio statistics (*LRS*): $LRS = 2\left[l^*(\hat{\theta}) - l^*(\tilde{\theta})\right]$. The LRT will reject the H_0 if $LRS > \chi^2_\alpha$, where χ^2_α is the $(1-\alpha)100^{\text{th}}$ percentile of the χ^2_α-distribution with degree of freedom (d.f.) = 1.

3.2.3 Numerical Simulation

MAF Threshold. To achieve optimal balance between performance and computational efficiency, we use extensive simulations to find the MAF threshold between the choice of rejection sampling and dosage approximation. Given R^2, we calculate two sets of mean squared error (MSE) between sampled genotype probability $\hat{f}_{i1}$ and truth f_{i1} using rejection sampling and dosage approximation, respectively.

Type I Error Evaluation. We assess the validity of EM-LRT-Dose, Dosage, EM-LRT-Prob, Mixture and gold-standard (based on true genotypes) under various combinations of R^2 and MAF. Specifically, we simulate data sets each with 2,000 samples using pre-specified marker-specific information R^2 and MAF, which allows us to generate genotype probabilities, dosages,

and true genotypes. Next, we simulate the trait values Y_i according to the linear model for sample i with a set of pre-specified parameters, where $i = 1, \ldots, 2{,}000$. For simplicity, we set $(\beta_0, \beta_1, \gamma, \sigma) = (1, 0, 1, 1)$.

We repeat the simulation ten million times. For each simulated data set, we perform association testing based on the true genotypes (truth), as well as based on imputed data using the standard Dosage method, Mixture method, and our proposed EM-LRT-Prob and EM-LRT-Dose methods. The empirical type I error rate of each method is calculated as the proportion of observed p-values that fall below the specified significance level. In addition, we calculate the Spearman correlation between the observed and gold-standard (true genotype based) p-values.

Statistical Power Assessment. To evaluate the statistical power of different methods, we again simulate data sets each with 2,000 samples using a combination of marker-specific information R^2 and MAF, and parameters $(\beta_0, \beta_1, \gamma, \sigma) = (1, \beta_1, 1, 1)$ where $\beta_1 \in [0, 1.5]$. We again repeat the simulations one million times. Similarly, for each simulated data set, we performed the same set of tests. The power of each method is calculated as the proportion of observed p-values that fall below the significance threshold $\alpha = 5 \times 10^{-5}$.

3.3 Results

3.3.1 MAF Threshold

We used simulations to determine the MAF threshold specific to each R^2 such that the rejection sampling is advantageous (quantified by lower MSE in estimating f_{i1}, the probability of having one copy of the minor allele) over dosage approximation when exceeding the MAF threshold (Figure 1 and Table 1). We observed the two sampling methods have similar performance (measured by MSE) when MAF is not high (below 20%-30% depending on R^2). In

such cases, we chose the simple dosage approximation method due to computational efficiency (Runtimes for rejection sampling and dosage approximation based on 2,000 samples are also shown). We also observed inferior performance (larger MSE) of both methods for low MAFs with intermediate R^2 values. Both can be explained by a combination of the imputation quality and the variation in f_{i1} as a function of both imputation quality R^2 and MAF q. Specifically, we have Var(f_{i1}) = $R^2 \times 2q(1-q) \times (1 - 2q(1-q))$, which increases with MAF q as well as with R^2. Low imputation quality R^2 coupled with low MAF leads to relatively little variation in f_{i1}, rendering both sampling methods capable of estimating it relatively accurately. On the other hand, high imputation quality implies dosages close to true genotype values 0, 1, and 2, as well as f_{i1} close to 0 or 1, thus allowing accurate inference of f_{i1} despite the larger variation in the values of f_{i1} across individuals. In the intermediate R^2 range, variation in f_{i1} coupled with imputation uncertainty makes inference challenging for both approaches.

3.3.2 Empirical Type I Error Simulation

As shown in Table 2 (at significance level 5E-02) and Table 3 (at significance level 5E-05), all the methods have proper control of type I error across in the range of R^2 and MAF examined: $0.1 \leq R^2 \leq 0.3$ and $1\% \leq \text{MAF} \leq 20\%$. Next, as shown in Figure 2, Spearman correlation with true p-values increases for every method when R^2 increases. The overall correlation is low in the range of MAF and R^2 examined. This low correlation is expected given the high level of imputation uncertainty and consistent with previous results (Zheng et al. 2011), confirming that association inference is challenging with low frequency variants, or with variants imputed with a high level of uncertainty. Although the absolute performance of all methods is not particularly impressive, we observe that EM-LRT methods always show slightly higher

Spearman correlation than Dosage method especially when MAF is low, suggesting that the EM-LRT *p*-values better approach gold-standard *p*-values. When R^2 and MAF are high, all methods perform similarly (results not shown), consistent with results shown in literature (Zheng et al. 2011; Liu et al. 2013b).

3.3.3 Empirical Power Simulations

When R^2 and MAF are high, all methods have similar performance. In this chapter, we focus on scenarios where $0.1 \leq R^2 \leq 0.3$ and $1\% \leq \text{MAF} \leq 20\%$. As shown in Figure 3, EM-LRT-Prob and Mixture methods are consistently the most powerful methods among all methods evaluated. However, these methods are not applicable in scenario II when only imputed dosages are available. It is thus valuable to notice that EM-LRT-Dose method approaches the statistical efficiency of EM-LRT-Prob, outperformng the standard Dosage method especially when R^2 or MAF is low. For example, when $R^2 = 0.1$ and MAF = 0.05, the power for EM-LRT-Prob, Mixture, EM-LRT-Dose and Dosage are 84.5%, 84.5%, 82.1%, and 61.4% under $\beta_1 = 1$.

3.4 Real Data Application

3.4.1 CLHNS

We applied the proposed EM-LRT methods as well as other existing methods to the Cebu Longitudinal Health and Nutrition Survey (CLHNS) study of 1,800 unrelated Filipino women. We performed association analysis across chromosome 16, where we have previously identified the variants near *CDH13* gene associated with plasma adiponectin level (Wu et al. 2010).

We conducted association testing with standardized adiponectin level measured in 2005 on a log scale as the quantitative trait and adjusted for age and BMI also measured in 2005. Additionally, we excluded subjects from the analysis if they met one or more of the following

criteria: 1) subjects with adiponectin level missing or outside of the range mean +/– 4 standard deviations (n=19); 2) subjects carrying the R221S variant (n=53) (Croteau-Chonka et al. 2012); and 3) subjects with missing age or BMI covariate information (n=20). In total, 1,717 subjects were tested for association with adiponectin level.

These 1,717 subjects were genotyped on the Affymetrix Genomewide Human SNP Array 5.0 GWAS chip (Lange et al. 2010) and also on the Illumina HumanExome Beadchip. Specifically, we first established the truth by employing PLINK (Purcell et al. 2007) to perform association on the true genotypes separately, finding 10 true positives (p-value $< 5\times10^{-6}$) on Affymetrix 5.0 and 5 on exome chip (with 2 overlapping). Next, to mimic a setting of low imputation quality, we masked all neighboring GWAS SNPs within 2kb of the 13 true positives before genotype imputation (22 SNPs were masked). Finally, we performed imputation using the MaCH imputation software (Li et al. 2010) using the ASN panel from the Phase I 1000 Genomes Project (March 2012 release, version 3) as reference. To evaluate the performance of the proposed methods along with other alternatives, we used markers overlapping between the ASN reference panel and the exome chip, but not on the Affymetrix 5.0, at which we have both imputed genotypes and true genotypes (from exome chip genotyping). We then conducted association testing on the imputed genotypes (dosages or probabilities) using our proposed EM-LRT methods, Dosage, and Mixture method.

Figure 4 shows the Q-Q plot for the 1,135 SNPs on chromosome 16 with $R^2 \leq 0.3$ and true p-value $> 5\times10^{-6}$. Q-Q plots are used to assess the number and magnitude of observed associations between tested SNPs and the trait under study, by comparing the observed $-\log_{10} p$-values to what is expected under the null hypothesis of no association. Early departure from the identity line suggests either that there is uncontrolled confouding leading to false positives (for

example, due to population stratification) or that a considerable proportion of SNPs are associated with the trait of interest (and thus not under the null distribution). Focusing on variants with p-values > 5×10^{-6} based on experimental genotypes allowed us to examine the type I error empirically. Overall, this Q-Q plot suggests that all methods have proper control of type I error with all points falling within the 95% confidence bands with the exeption of one variant. The single potential false positive, rs8045889 with a true p-value = 0.0148; R^2 = 0.0736; and MAF = 0.4271, was identified by Dosage, EM-LRT-Prob, and Mixture (EM-LRT-Dose has a borderline p-value of 0.0002). In addition, we observe overall deflation in the test statistics (observed larger p-values) of all methods when compared with truth. The median (mean) p-values are 0.6407, 0.5614, 0.5568 and 0.5568 (0.6075, 0.5552, 0.5512, and 0.5543) for Dosage, EM-LRT-Dose, EM-LRT-Prob and Mixture respectively, compared with the true median (mean) of 0.5008 (0.5009). The tendency towards large p-values is expected and driven by the loss of information due to imputation uncertainty.

While establishing the validity is crucial, we are more interested in the power to identify genuine associations. Table 4 tabulates p-values from all four methods together with the truth for variants with $R^2 < 0.3$ and true p-value < 5×10^{-8}. Although all variants reach the genome-wide significance threshold regardless of the method, we observed that EM-LRT-Dose or EM-LRT-Prob generated more significant p-values (and better approached truth in all cases) than the alternatives for five out of the six variants interrogated, suggesting power enhancement using our methods.

3.4.2 WHI

We have previously identified several variants associated with blood cell traits using whole exome sequencing in 761 African Americans coupled with imputation in > 13,000 African

Americans with GWAS data from genome-wide Affymetrix 6.0 genotyping (Auer et al. 2012). The samples are drawn from several cohorts including WHI, ARIC, CARDIA and JHS. Association analyses were performed separately for WHI and CARe cohorts (ARIC, CARDIA and JHS) and subsequently meta-analyzed across the two. Due to the ascertainment of variants through whole exome sequencing, 56% of our analyzed variants had MAF < 5%.

Here, we use meta-analysis results from our previous study as a gold standard to define true positives and investigate the *p*-values in the WHI cohort using our EM-LRT-Dose and standard Dosage method, which had been adopted by the original study. We did not keep a copy of the posterior probabilities because of the large number of samples imputed and because standard analyses do not involve the posterior probabilities. Therefore, this is a real data example of scenario II. Table 5 presents all variants with MAF < 5% reported to reach genome-wide significant threshold in the original study, comparing *p*-values from our EM-LRT-Dose and the standard Dosage method. We notice that EM-LRT-Dose generated slightly more significant *p*-values at the associated variants in three out of the four tests performed. In one case (snp.177015 with white blood cell count [WBC]), the *p*-value from EM-LRT-Dose (*p*-value = 4.72×10^{-8}) reached the conventionally employed genome-wide significance threshold of 5×10^{-8} while that from Dosage was marginally genome-wide insignificant (*p*-value = 6.11×10^{-8}).

3.5 Discussion

It is crucial to take imputation uncertainty into consideration when performing association testing. Existing methods have focused on common variants, which have been the focus of the past wave of GWAS using HapMap-based imputation. With the deluge of next generation sequencing data being generated, increasingly denser reference panels are allowing imputation of a much larger number of variants, including an increasing number of relatively

rare or poorly imputed variants. It is thus highly warranted to re-visit potential strategies for post-imputation association analysis and to seek more powerful or efficient statistical methods.

In this chapter, we have proposed EM-LRT methods explicitly incorporating marker level imputation quality statistic into association tests. We considered two scenarios: when posterior probabilities of all potential genotypes are available and when only dosages are available. We evaluated the performance of the proposed methods along with existing alternatives using simulation studies and by application to real data sets.

In scenario I, our proposed EM-LRT-Prob demonstrated nearly identical performance as the Mixture model, which has been shown to be the best post-imputation association method particularly when imputation uncertainty is high (Zheng et al. 2011). While our EM-LRT-Prob effectively also fits a mixture model (therefore in terms of the underlying statistical model essentially the same as the Mixture method adopted in Zheng et al. 2011), we have proposed and implemented a much more computationally efficient algorithm to fit the model. Mixture method (Zheng et al. 2011) used *R* function *optim*() to find ML estimates. Technically, the *optim*() function uses numerical differentiation to obtain ML estimates based on the score function and Hessian matrix, which is considerably slower than our proposed EM algorithm. To quantify the computational efficiency, we conducted association testing on a CLHNS data set of 1,717 subjects and 13,801 SNPs, using EM-LRT-Prob and Mixture methods with the same starting values (the Mixture method tends to run even slower without using the suggested starting values). We observed that the association tests required 279 seconds computing time and 0.91 GB RAM for EM-LRT-Prob and 1,505 seconds computing time and 1.23 GB RAM for the Mixture method on a 2.93 GHz Intel® Xeon® Processor X5670. Computing time scales linearly with sample size for both EM-LRT-Prob and the Mixture method (Figure 5).

In scenario II, the Dosage method has been shown analytically as the optimal one dimensional summary statistic for association testing in a typical linear model (Liu et al. 2013b). In this chapter, we extended the utility of this optimal one-dimensional measure by employing it together with the imputation quality measure R^2 first to sample posterior probabilities (in an attempt to rescue as much full information as possible) and then to conduct association testing on the sampled probabilities using our proposed EM-LRT method.

Our simulations suggest an advantage of the proposed methods over the standard Dosage method when imputation quality is relatively low, where imputation quality is measured by R^2, the squared Pearson correlation between the imputed dosages and the unknown true genotypes. Since the calculation of R^2 requires true genotypes, it is not available in practice and imputation software provides an estimate based on the observed dispersion in imputed genotypes over its expected value. Such an estimate (Rsq in MaCH (Li et al. 2010), MaCH-Admix (Liu et al. 2013a), minimac (Howie et al. 2012), R2 for BEAGLE (Browning and Browning 2008) and INFO for IMPUTE/IMPUTE2 (Marchini et al. 2007; Howie et al. 2009)) has been widely used for the assessment of imputation quality and for post-imputation quality control. However, as shown in Figure 6 (based on the CLHNS data), MaCH Rsq is not a perfect measure of R^2. For example, it has been reported earlier to have the tendency of underestimating true quality for common variants (Gao et al. 2012; Liu et al. 2012). We also observed the tendency of over-estimation towards the lower end of the MaCH Rsq. Therefore, we still recommend post-imputation quality filtering before application of our methods. We suggest application to variants with estimated $R^2 > 0.1$, which is less stringent than what is typically recommended (Liu et al. 2012; Duan et al. 2013a), but above which imputation quality is typically under- rather than over- estimated. To further examine the effectiveness of the filtering threshold, we quantified

type I error rate via simulations for varying R^2 (four values examined: 0.05, 0.1, 0.3, 0.5) in combination with varying MAF (three values examined: 2.5%, 5% or 10%). Specifically, for each R^2 and MAF combination, we simulated A (A=2500) (exchangeable) groups of data sets under the null hypothesis. For each group, we simulated B (B=2000) data sets (again, under the null hypothesis) and calculated the *p*-values by applying all methods to each of the B=2000 simulated data sets. We then calculated the group-specific type I error as the proportion of B=2000 *p*-values (in that group) below the significance threshold of 0.05. We therefore obtained A=2500 type I error estimates. Finally, we conducted the following one-sample t-test on these 2500 type I error estimates H_0: type I error ≤ 0.05 vs. H_1: type I error > 0.05. Significant results from the t-test indicate inflated type I error. Results are shown in Table 6. As we can see the results suggest that the mixture model based methods (EM-LRT-Dose, EM-LRT-Prob, and Mixture) have inflated type I error when $R^2 \leq 0.1$, which is likely caused by the tendency of the mixture model over-fitting the data based on additional d.f. compared to the null model.

In summary, we have proposed likelihood-ratio tests based on expectation maximization algorithms for post-imputation association testing. Simulation and real data analyses show our methods have protected type I error. In addition, simulation and real data results suggest slightly enhanced statistical power of our EM-LRT methods over a standard Dosage method, which has been shown to be the optimal one dimensional statistic for post-imputation association testing; and computationally more efficient (average more than fivefold reduction in computing time) than the Mixture method, which has been shown to be the most powerful at increased computational costs for variants imputed with high level of uncertainty. We anticipate our methods will replace the Mixture method for the analysis of low frequency variants or those imputed with high uncertainty. We envision our methods being applied on a larger scale for

GWASs with imputation from sequencing based reference panels, including in the public domain, the 1000 Genomes Project (Altshuler et al. 2010a; Altshuler et al. 2012), the UK10K Project (Futema et al. 2012), and the reference haplotypes assembled by the International Haplotype Consortium (Marchini 2013) as well as study specific reference panels (Auer et al. 2012; Fuchsberger et al. 2012; Liu et al. 2012; Duan et al. 2013a; Kang et al. 2013; Bizon et al. 2014). Our methods are implemented in software package EM-LRT, freely available at http://www.unc.edu/~yunmli/emlrt.html.

Table 1 Rejection Sampling vs. Dosage Approximation for f_{i1} Estimation

		Rejection Sampling		Dosage Approximation	
R^2	MAF Cutoff	MSE	Runtime (Sec.)	MSE	Runtime (Sec.)
0.95	20%	1.01E-02	5.82	1.42E-02	7.60E-04
0.75	30%	4.49E-02	3.04	6.39E-02	6.80E-04
0.50	30%	5.98E-02	2.3	6.29E-02	7.14E-04
0.30	30%	5.71E-02	2.42	6.11E-02	8.10E-04
0.25	30%	5.32E-02	2.61	6.06E-02	8.36E-04
0.20	25%	3.82E-02	2.41	3.93E-02	7.56E-04
0.10	20%	1.69E-02	2.35	1.99E-02	7.04E-04

MAF: Minor allele frequency
MSE: Mean square error

Table 2 Type I Error Rate at Significance Level = 5E-02

R^2	MAF	Dosage	EM-LRT-Dose	EM-LRT-Prob	Mixture	Truth
	0.2	4.99E-02	5.01E-02	5.01E-02	5.01E-02	4.99E-02
	0.1	5.01E-02	5.02E-02	5.03E-02	5.03E-02	4.99E-02
0.3	0.05	4.99E-02	5.01E-02	5.01E-02	5.01E-02	4.99E-02
	0.025	5.01E-02	5.03E-02	5.03E-02	5.03E-02	5.01E-02
	0.01	4.98E-02	4.96E-02	4.96E-02	4.96E-02	4.99E-02
	0.2	5.00E-02	5.02E-02	5.03E-02	5.03E-02	5.00E-02
	0.1	4.99E-02	5.02E-02	5.03E-02	5.03E-02	5.00E-02
0.2	0.05	5.00E-02	5.03E-02	5.03E-02	5.03E-02	5.01E-02
	0.025	4.99E-02	5.03E-02	5.03E-02	5.03E-02	5.01E-02
	0.01	5.00E-02	5.02E-02	5.01E-02	5.01E-02	4.99E-02
	0.2	5.00E-02	5.08E-02	5.05E-02	5.05E-02	5.01E-02
	0.1	5.00E-02	5.06E-02	5.08E-02	5.08E-02	5.00E-02
0.1	0.05	5.01E-02	5.11E-02	5.13E-02	5.13E-02	5.01E-02
	0.025	5.01E-02	5.15E-02	5.14E-02	5.14E-02	5.01E-02
	0.01	5.00E-02	5.09E-02	5.07E-02	5.07E-02	4.98E-02

Table 3 Type I Error Rate at Significance Level = 5E-05

R^2	MAF	Dosage	EM-LRT-Dose	EM-LRT-Prob	Mixture	Truth
	0.2	4.85E-05	5.00E-05	4.94E-05	4.94E-05	5.24E-05
	0.1	5.09E-05	4.98E-05	5.23E-05	5.23E-05	5.43E-05
0.3	0.05	4.71E-05	5.03E-05	5.07E-05	5.07E-05	5.41E-05
	0.025	4.95E-05	4.92E-05	4.80E-05	4.80E-05	4.97E-05
	0.01	5.35E-05	4.79E-05	4.69E-05	4.69E-05	4.97E-05
	0.2	4.58E-05	4.57E-05	4.58E-05	4.58E-05	5.00E-05
	0.1	4.67E-05	4.71E-05	4.84E-05	4.84E-05	5.30E-05
0.2	0.05	5.08E-05	5.09E-05	5.05E-05	5.05E-05	5.14E-05
	0.025	5.23E-05	5.02E-05	5.09E-05	5.09E-05	5.06E-05
	0.01	4.78E-05	4.41E-05	4.26E-05	4.26E-05	4.92E-05
	0.2	4.93E-05	5.53E-05	5.19E-05	5.19E-05	5.02E-05
	0.1	5.08E-05	5.27E-05	5.24E-05	5.24E-05	5.35E-05
0.1	0.05	5.05E-05	5.09E-05	5.23E-05	5.22E-05	4.85E-05
	0.025	4.98E-05	5.15E-05	5.04E-05	5.04E-05	4.93E-05
	0.01	5.02E-05	5.05E-05	4.95E-05	4.95E-05	5.27E-05

Table 4 Associated Variants with $R^2 \leq 0.3$ in the CLHNS Study

		P-values				
Coordinate*	R^2	Dosage	EM-LRT-Dose	EM-LRT-Prob	Mixture	Truth[#]
chr16:82646152	0.251	**2.13E-11**	**1.45E-11**	4.68E-11	4.68E-11	6.77E-20
chr16:82650717	0.282	**2.88E-11**	**1.83E-11**	5.87E-11	5.87E-11	1.35E-21
chr16:82663288	0.268	**2.67E-10**	7.78E-10	**6.46E-10**	**6.46E-10**	2.16E-25
chr16:82670249	0.270	2.04E-08	**1.45E-08**	**1.59E-08**	1.61E-08	1.72E-12
chr16:82670539	0.249	9.26E-09	1.27E-08	**4.79E-09**	**4.83E-09**	1.25E-12
chr16:82670636	0.230	1.22E-08	2.01E-08	**7.32E-09**	**7.33E-09**	1.78E-12

*: Coordinates are in genome build 37.
Bold and underlined: The most significant *p*-value among the four methods.
Bold but not underlined: The second most significant *p*-values among the four methods.
[#]Truth: established by regressing phenotype on true genotypes.

Table 5 Associated Variants with MAF < 5% in the WHI Study

			P-value	
SNP	Trait	Meta *p*-value	Dosage	EM-LRT-Dose
snp.684276	hematocrit	5.70E-11	5.85E-11	8.94E-11
snp.177048	log(WBC)	3.00E-13	3.95E-08	2.73E-08
snp.177015	log(WBC)	4.30E-13	6.11E-08	4.72E-08
snp.41127	platelet	1.50E-11	2.52E-08	3.71E-09

Underlined: The most significant *p*-value among the two methods.

Table 6 One-sample T-test for Type I Error

		P-values			
MAF	Method	R^2=0.05	R^2=0.1	R^2=0.3	R^2=0.5
0.025	Dosage	6.75E-01	2.72E-01	4.97E-01	7.33E-01
0.05		2.03E-01	6.38E-01	9.21E-01	1.88E-01
0.1		9.62E-01	8.75E-01	2.01E-01	6.78E-01
0.025	EM-LRT-Dose	8.96E-183	3.00E-65	2.00E-01	9.05E-01
0.05		6.15E-216	5.11E-35	3.89E-01	7.33E-03
0.1		1.40E-69	9.73E-12	4.73E-03	4.17E-02
0.025	EM-LRT-Prob	2.34E-111	2.51E-55	2.49E-01	8.69E-01
0.05		1.54E-174	3.38E-40	2.50E-01	4.81E-03
0.1		4.24E-134	1.04E-24	5.01E-04	2.91E-02
0.025	Mixture	5.45E-111	3.72E-55	2.52E-01	8.70E-01
0.05		2.94E-174	4.79E-40	2.55E-01	4.98E-03
0.1		7.60E-134	1.22E-24	5.30E-04	2.95E-02

Underlined: *p*-value < 5E-4.

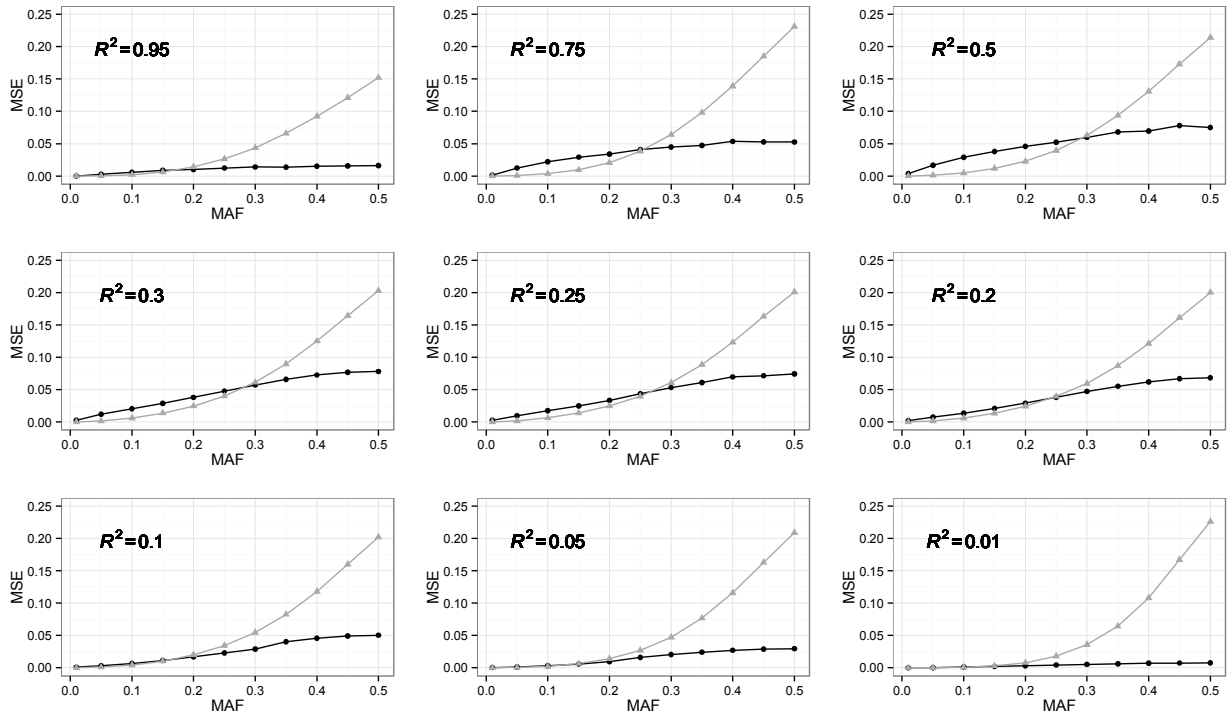

Figure 1 MAF Threshold: Rejection Sampling (Black) vs. Dosage Approximation (Grey)

MSE (Y-axis) between sampled genotype probability $\hat{f}_1$ and true f_1 using rejection sampling (black) and dosage approximation (grey) is compared across a spectrum of R^2.

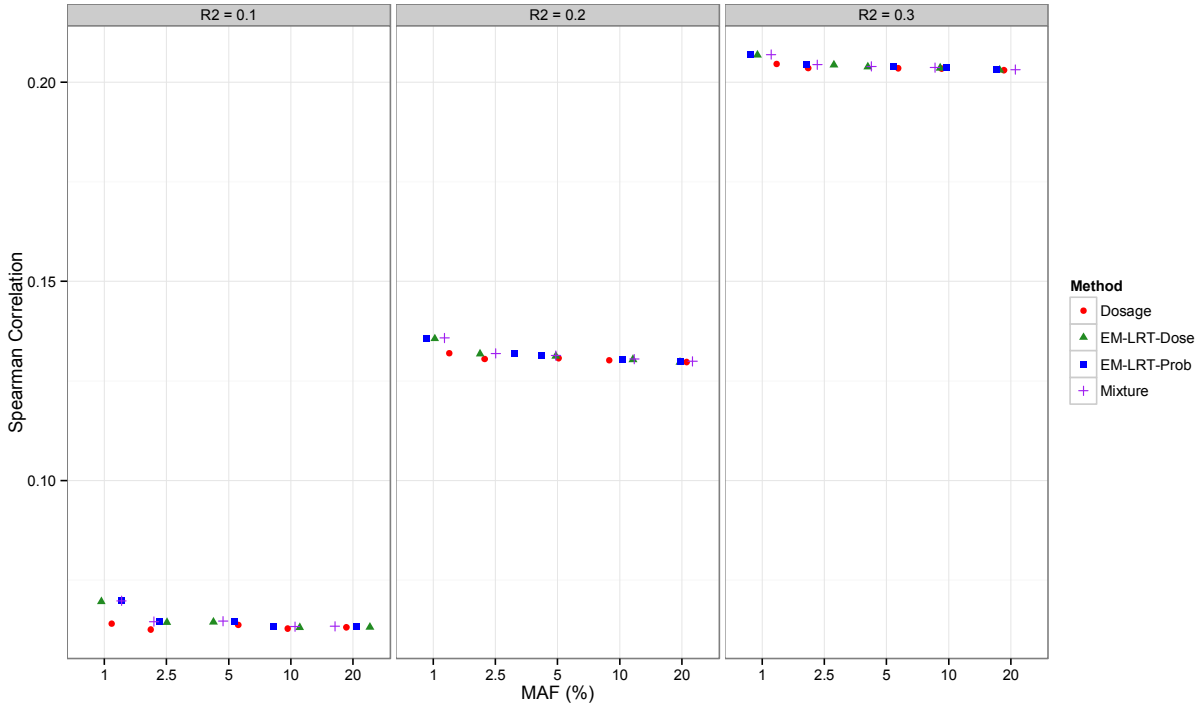

Figure 2 Spearman Correlation with Gold Standard *P*-values

Spearman correlation (Y-axis) between gold standard *p*-values and *p*-values from different methods is displayed across a spectrum of MAF and R^2.

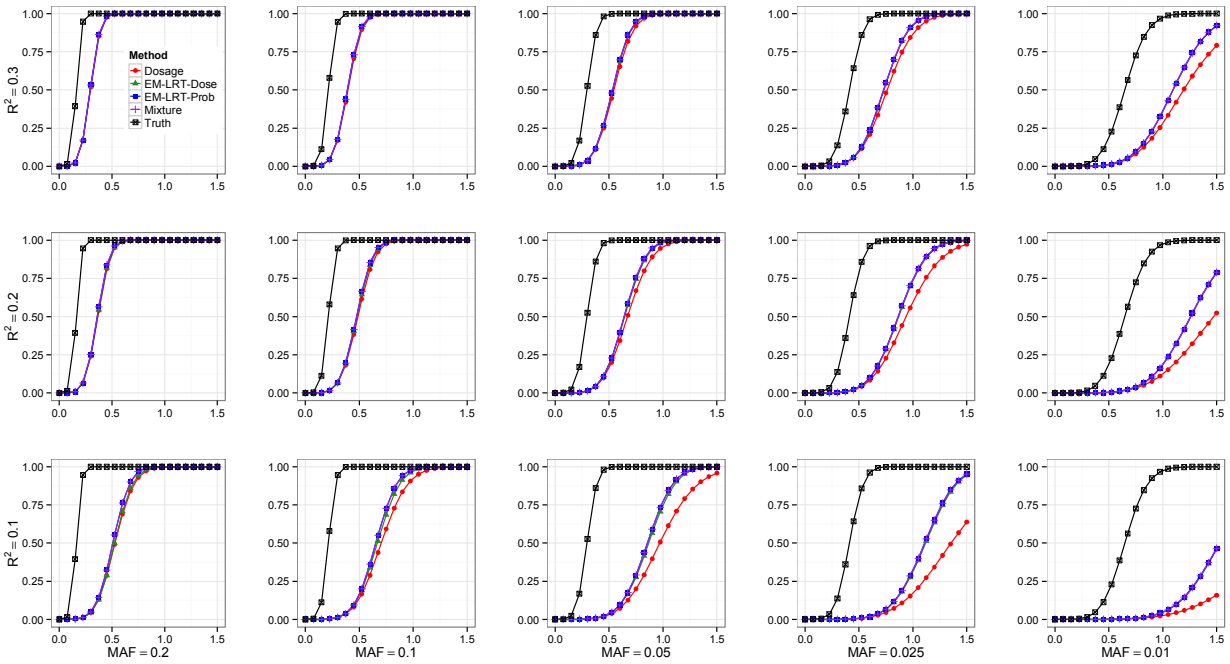

Figure 3 Power Comparison

Statistical power (Y-axis) of different methods is shown across a spectrum of R^2 and MAF.

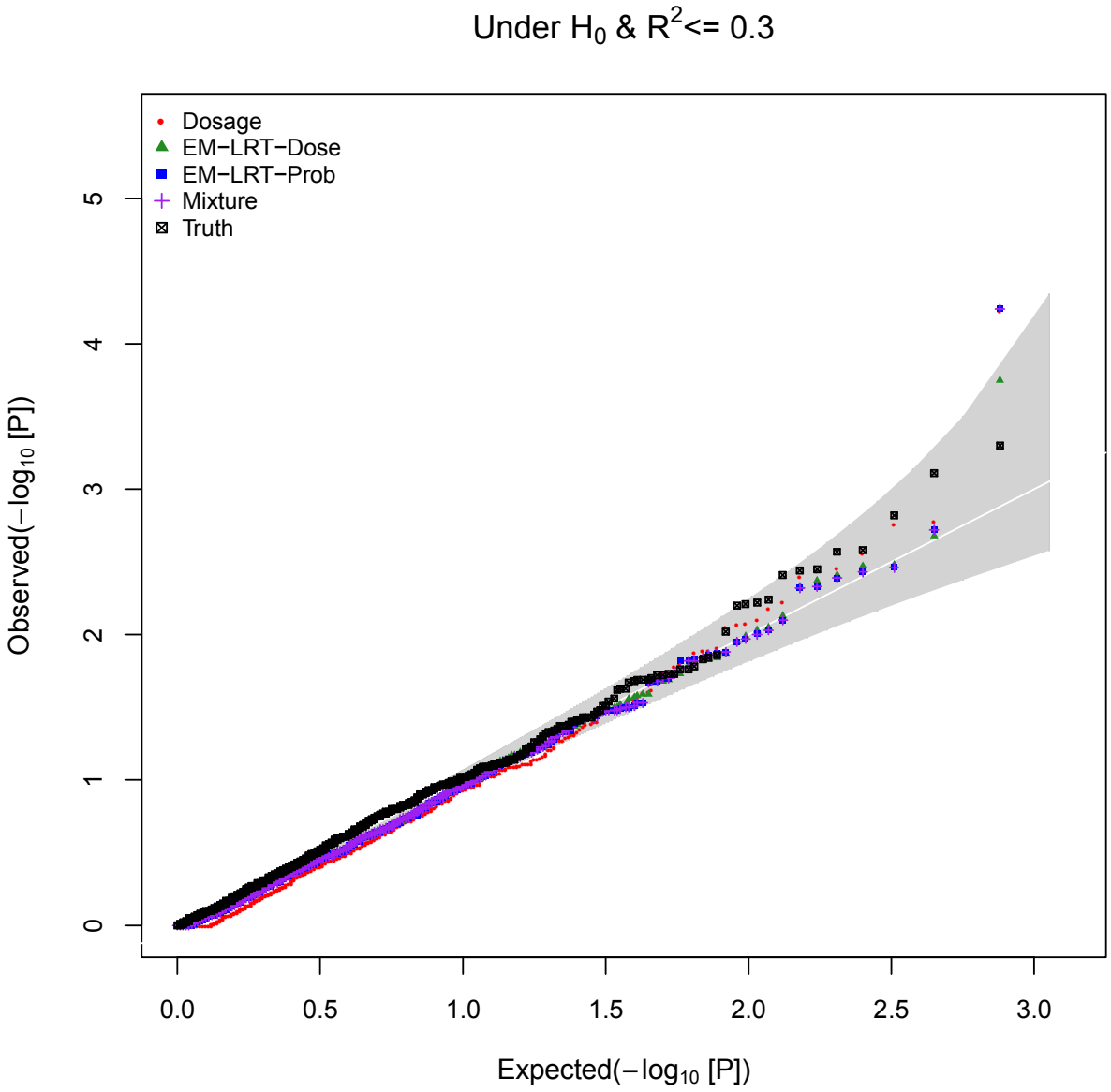

Figure 4 Q-Q Plot for Null Variants with Low Imputation Quality in the CLHNS Study

The observed (Y-axis) vs. expected (X-axis) $-\log_{10}[p\text{-values}]$ are shown for 1,135 SNPs in the CLHNS data set. These SNPs are considered to be under the null hypothesis (true p-value > 5×10^{-6}), and all have low imputation quality ($R^2 < 0.3$).

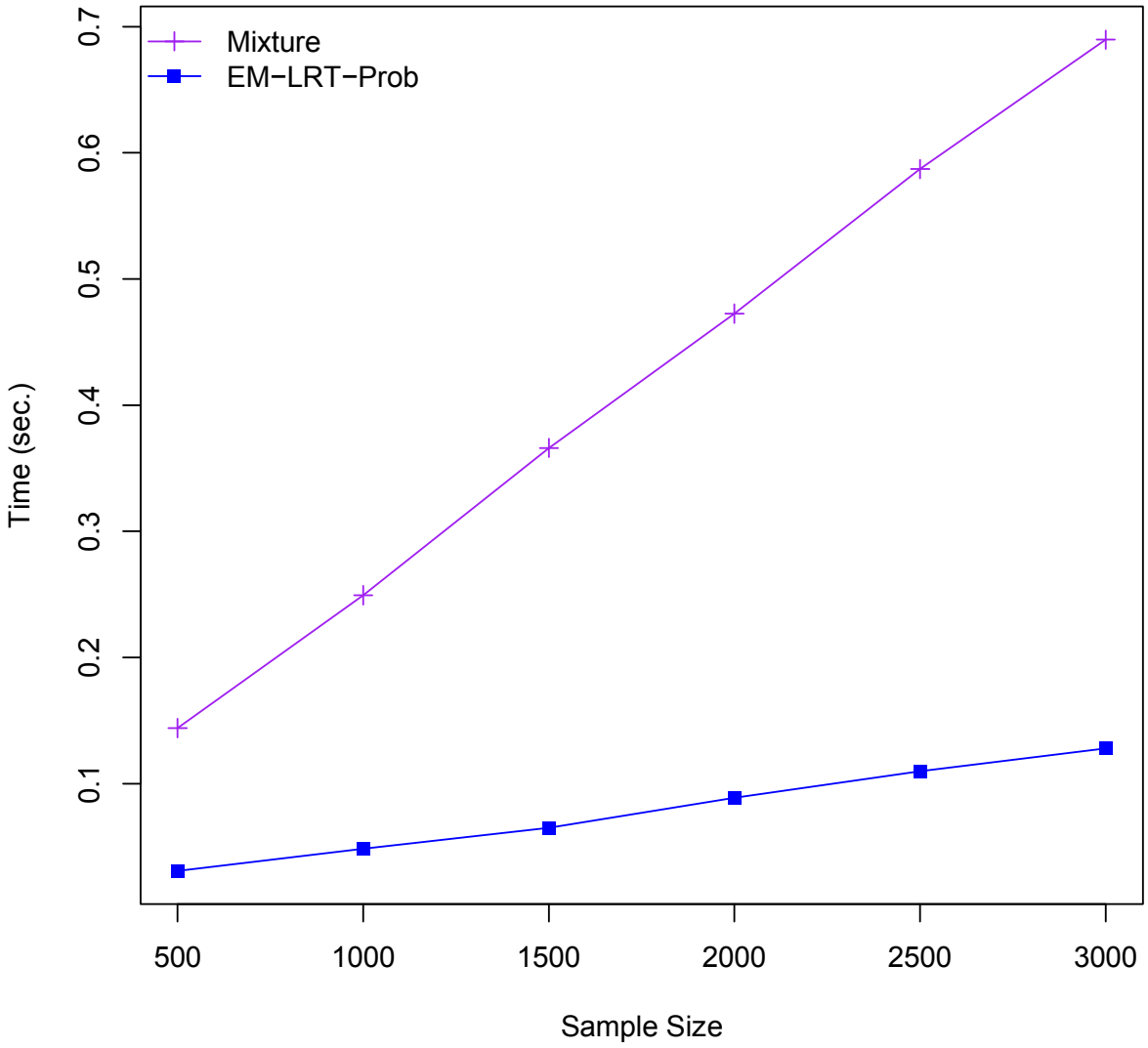

Figure 5 Computing Time: Mixture Method vs. EM-LRT-Prob

The computing time of the Mixture method and our proposed EM-LRT-Prob method is displayed across a range of sample sizes. For each sample size, computing time is averaged across 2,000 simulated data sets.

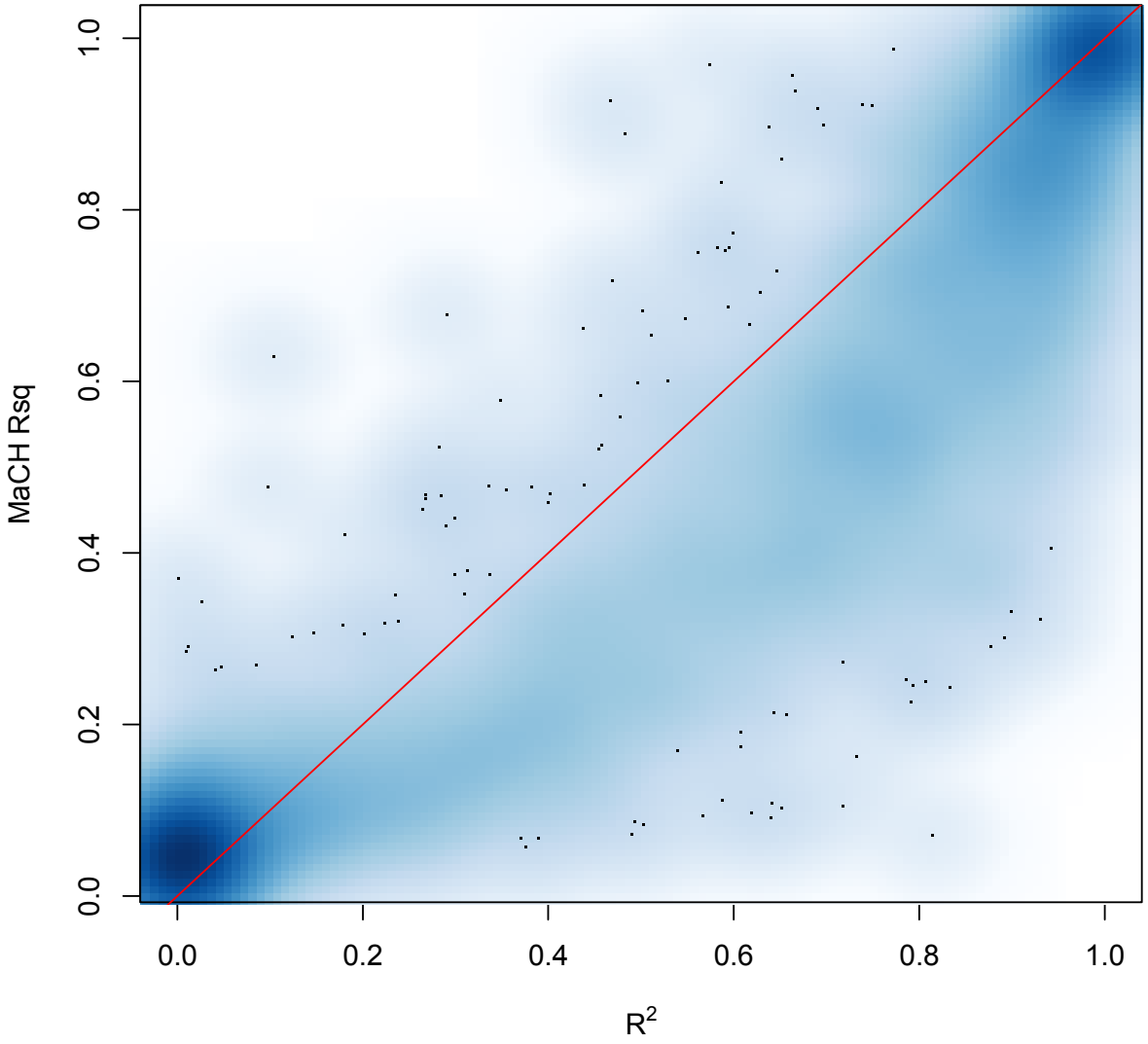

Figure 6 Estimated vs. True Imputation (Rsq vs. R^2)

The MaCH estimated imputation quality Rsq (Y-axis) is plotted against the true imputation quality R^2 (X-axis), which were calculated between genotype data from exome chip array and imputed genotype data (dosages). The red 45-degree line represents perfect estimation. A smooth density scatter plot is employed such that darker color corresponds to larger density and individual dots represent outliers.

CHAPTER 4:FUNMETHYL

4.1 Introduction

Most EWASs are conducted by testing the association between DNA methylation level (response variable) and quantitative traits (explanatory variables) CpG site by CpG site across the genome. However, because of the correlation structure among the sites and because many of them fall in naturally defined regions (e.g., belonging to the same gene; belonging to the same regulatory region such as an enhancer or DNAse hypersensitivity site), it is conceptually straightforward to imagine achieving enhanced statistical power by performing region-based test (e.g., simultaneously testing multiple sites together) especially when there are multiple small or moderate signals in that region. In this chapter, we propose to perform association testing between DNA methylation variants in a region (explanatory variable) and quantitative trait (response variable). Instead of collapsing DNA methylation variants or building a kernel matrix, the DNA methylation variants of an individual are treated as a realization of a stochastic process in the functional data analysis (Fan et al. 2013). Specifically, we consider every individual's DNA methylation levels in a region as a stochastic process and we further use the functional data analysis techniques to estimate the DNA methylation function in that region. Next, the DNA methylation effect in the model is expanded as a combination of basis functions and coefficients. Finally, to test the association between the DNA methylation variants and quantitative traits, we test if the coefficients of DNA methylation effect are all 0.

4.2 Methods

Assume *p* CpG sites with ordered physical locations $0 \le t_1 \le t_2 \le \ldots \le t_p = T$. For the ease of notation, we normalize the location to [0,1]. Next, for *i*-th individual, let Y_i denote quantitative trait, $M_i = \left(m_i(t_1), \ldots, m_i(t_p)\right)$ (can be either β-value or *M*-value) denote DNA methylation at *p* CpG sites and $Z_i = \left(z_{i1}, \ldots, z_{ic}\right)'$ denote a $c \times 1$ covariate vector. For continuous trait, the functional linear model is

$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)\,dt + Z_i'\gamma + \varepsilon_i$$

where α is the overall mean, $\beta(t)$ is the DNA methylation effect of DNA methylation function $X_i(t)$ at the location *t*, γ is a $c \times 1$ regression coefficient vector of covariates, and $\varepsilon_i \sim N\left(0, \sigma^2\right)$. Similarly, for dichotomous trait, the functional linear model can be easily modified as follows:

$$\text{logit}\, P(Y_i = 1) = \alpha + \int_0^1 X_i(t)\beta(t)dt + Z_i'\gamma$$

4.2.1 Estimation of DNA Methylation Function $X_i(t)$

In this section, we introduce three different techniques used for estimating $X_i(t)$. First, DNA methylation function can be represented as $X_i(t) = \sum_{k=1}^{K_x} c_{ik}\phi_k(t)$, where c_{ik} is a series of coefficients, $\phi_k(t)$ is a series of basis functions, and $k = 1, 2, \ldots, K_x$ with K_x being the number of basis function. The choice of basis function is flexible and it can be either B-spline or Fourier basis function. Further, DNA methylation function can be written in matrix format $X_i(t) = c_i \cdot \phi(t)$ where $c_i = \left(c_{i1}, \ldots, c_{iK_x}\right)$ and $\phi(t) = \left(\phi_1(t), \ldots, \phi_{K_x}(t)\right)'$. To estimate the coefficients, we employ the ordinary least squares (OLS) method. That is, the sum of squared error is

minimized across the region

$$SSE(c_i) = \sum_{j=1}^{p} \left[m_i(t_j) - X_i(t_j) \right]^2$$
$$= \sum_{j=1}^{p} \left[m_i(t_j) - c_i \cdot \phi(t_j) \right]^2$$

Solving for *c*, the DNA methylation function can be estimated as

$$\hat{X}_i(t) = M_i \Phi \left(\Phi' \Phi \right)^{-1} \cdot \phi(t)$$

where Φ is a $p \times K_x$ matrix containing the values $\phi_k(t_j)$. This technique is simple but only ensures the estimated function gives a good fit to data. Moreover, it may result in the estimated function excessively wiggly or locally variable when data are loosely sampled. Therefore, the technique provides a good compromise between goodness of fit and smoothness is absolutely desired. To illustrate the concept, one can think of minimizing mean squared error (MSE), the sum of the squared bias of the estimator and the variance. A technique that minimizes only the squared bias or the variance is not optimal. As a matter of fact, MSE can be dramatically minimized by sacrificing some bias in order to minimize sampling variance. Hence, the smoothness is imposed on the estimated function in the second technique.

As mentioned earlier, instead of purely minimizing the sum of squared error across the region, the roughness also needs to be minimized together. Consequently, the penalized sum of squared error is used

$$PENSSE(c_i) = \sum_{j=1}^{p} \left[m_i(t_j) - X_i(t_j) \right]^2 + \lambda \int \left[D^2 X_i(t) \right]^2 dt$$
$$= \sum_{j=1}^{p} \left[m_i(t_j) - c_i \cdot \phi(t_j) \right]^2 + \lambda c_i' R c_i$$

where λ is the smoothing parameter, $\int [D^2 X_i(t)]^2 dt$ is a measure of total curvature, and $R = \int D^2\phi(t) \cdot D^2\phi'(t) dt$ is the roughness penalty matrix. Again, solving for c, the estimated DNA methylation function is

$$\hat{X}_i(t) = M_i \Phi (\Phi'\Phi + \lambda R)^{-1} \phi(t)$$

Besides, the smoothing parameter λ needs to be determined and can be selected using generalized cross-validation (GCV) (Craven and Wahba 1978).

Last but not least, DNA methylation function can be estimated by directly using the discrete realization $(m_i(t_1),\ldots,m_i(t_p))$. This is perhaps the simplest way of estimating the DNA methylation function and hence, $\int_0^1 X_i(t)\beta(t)dt \approx \sum_{j=1}^{p} m_i(t_j)\beta(t_j)$.

4.2.2 Estimation of DNA Methylation Effect $\beta(t)$

DNA methylation effect can be estimated as $\beta(t) = (\varphi_1(t),\ldots,\varphi_{K_b}(t))(b_1,\ldots,b_{K_b})' = \varphi'(t)b$

where $\varphi(t) = (\varphi_1(t),\ldots,\varphi_{K_b}(t))'$ is spline basis vector, $b = (b_1,\ldots,b_{K_b})'$ is parameter vector, and K_b is the number of basis function. If K_b is a large number, the model may overfit the data considerably. To overcome this problem, the parameters are penalized and this leads to a more robust model. We adopt the truncated power series basis $\varphi(t) = (\varphi_1(t),\ldots,\varphi_{K_b}(t))' = (1, t, (t-\kappa_3)_+,\ldots,(t-\kappa_{K_b})_+)'$ and assume $b \sim N(0, D)$, where κ_k are knots in the interval [0,1], $(t-\kappa)_+$ is an indicator function, taking value of 1 if $t > \kappa$ and 0 if $t \le \kappa$, and D is a penalty matrix

$$D = \begin{bmatrix} 0_{2\times 2} & 0_{2\times(K_b-2)} \\ 0_{(K_b-2)\times 2} & I_{(K_b-2)\times(K_b-2)} \end{bmatrix}$$

4.2.3 Penalized Functional Linear Model

Thus far, we have introduced three techniques for estimating DNA methylation function and also one for expanding DNA methylation effect. By putting them together, the proposed penalized functional linear model can be written as follows:

$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)dt + Z_i'\gamma + \varepsilon_i$$
$$= \alpha + V_i'b + Z_i'\gamma + \varepsilon_i$$

where $V_i' = \begin{cases} \sum_{j=1}^{p} m_i(t_j)\varphi'(t) & , \text{ No-smoothing} \\ M_i\Phi(\Phi'\Phi)^{-1}\int_0^1 \phi(t)\varphi'(t)\,dt & , \text{ Least-square} \\ M_i\Phi(\Phi'\Phi+\lambda R)^{-1}\int_0^1 \phi(t)\varphi'(t)\,dt & , \text{ Roughness-penalty} \end{cases}$ and $b \sim N(0, D)$.

4.2.4 Hypothesis Testing

Because testing the association between DNA methylation at *p* CpG sites and the quantitative traits is of our interest, we perform the association testing with the hypothesis testing $H_0 : b = (b_1,\ldots,b_{K_b})' = 0$ vs. $H_1 : b = (b_1,\ldots,b_{K_b})' \neq 0$. We propose to use F-test for testing $H_0 : b = 0$ vs. $H_1 : b \neq 0$. The F-test compares two models, reduced model and full model, where reduced model is under the null hypothesis and nested in the full model. Moreover, under the null hypothesis, the test statistic follows an *F* distribution with $(K_b, n - K_b - c - 1)$ degrees of freedom.

4.3 Application to ARIC

The Atherosclerosis Risk in Communities Study (ARIC) is a prospective cohort study of cardiovascular disease risk in four U.S. communities. Between 1987 and 1989, 7,082 men and 8,710 women aged 45–64 years were recruited from Forsyth County, North Carolina; Jackson, Mississippi (African Americans only); suburban Minneapolis, Minnesota; and Washington

County, Maryland. The Illumina Infinium HumanMethylation450 Beadchip array (HM450) was used to measure DNA methylation (Illumina Inc.; San Diego, CA, USA). The platform detected methylation status of 485,577 CpG sites by sequencing-based genotyping of bisulphite-treated DNA. In this ARIC, 2,918 samples had DNA methylation data. Among these samples, 57 did not have body mass index (BMI) data, 51 did not have waist circumference (WC) data, and 472 did not have any or complete covariate data needed for confounder adjustment, leaving a final sample size $N = 2,105$. Further, there were total of 20,330 genes analyzed in our analysis and the number of CpG sites in genes varies from 1 to 1,940 (Min = 1, Median = 15, Mean = 20.88, Max = 1,940).

Before performing association study, we first removed the batch effect by directly regressing the β-values on the chip, where the batch effect is accounted for, and saved the residuals for the subsequent analysis. Here, we considered every single gene as one region and for each gene, if the number of CpG site > 5 we performed the proposed penalized functional region-based analysis and existing region-based analysis, directly testing the association on each gene using F-test, SKAT, and SKAT-O (Wu et al. 2011; Lee et al. 2012). In addition, the single-probe analysis was also performed, which identified a gene as causal when at least one p-value is significant at the Bonferroni-corrected level. When performing association testing across different genes, we directly used the residuals in place of β-values. Specifically, we used BMI and WC as response variable in the model, whereas the residual was the independent variable adjusting for covariates, including smoking, sex, age, center, leisure time physical activity, drinking, white blood cell count, plate and chip (array) row, and the 10 principal component (PC) scores from the Illumina Infinium HumanExome Beadchip genotype array, to account for potential confounding by genetic ancestry.

In addition, we removed the probes identified in single-probe analysis from the gene, and again performed the region-based analysis. Ideally, a powerful region-based method is capable of identifying the region even if the strong signal(s) is removed.

4.4 Real Data Simulation

4.4.1 Empirical Type I Error

To evaluate the empirical type I error, for *i*-th individual the following model was first fit for generating quantitative trait:

$$y_i = \sum_{k=1}^{9} z_{ik}\gamma_k + \varepsilon_i$$

where z_{i1} to z_{i9} are smoking, sex, age, center, leisure time physical activity, drinking, white blood cell count, plate and chip (array) row, first principal component (PC) from the Illumina Infinium HumanExome Beadchip genotype array (Grove et al. 2013), and $\varepsilon_i \sim N(0, 0.75)$. Further, we set $(\gamma_1,\gamma_2,\ldots,\gamma_9)$ = (-0.46,0.28,-0.03,-0.07,-0.10,0.09,0.16,-6.22,-0.03) and considered sample size N = 250, 500, 1000, and 2000. Next, 10^5 sets of quantitative trait were generated and therefore not associated with any DNA methylation variants. Next, we randomly selected one gene *BCL6* (p = 53) and performed penalized functional region-based analysis as well as existing region-based analysis using the generated quantitative trait and the DNA methylation data of gene *BCL6*. Besides, single-probe analysis was performed. Consequently, 10^5 sets of test statistic as well as *p*-value were produced and stored for each region-based method whereas $10^5 \times 53$ were produced and stored for single-probe method.

Finally, the empirical type I error rate was calculated as the average probability of *p*-value less than a given α level for region-based methods, and it was calculated as the average probability of *at least* one *p*-value less than a bonferroni-corrected given α level (α = 0.05, 0.01,

0.005, 0.001) for single-probe method.

4.4.2 Empirical Average Power

For *i*-th individual and *g*-th gene, the following model was first fit for identifying the top causal CpG probes and estimated covariate effects:

$$Y_i = \sum_{j=1}^{P} m_{ij}\beta_j + \sum_{k=1}^{18} z_{ik}\gamma_k$$

where Y_i is BMI, m_{ij} and β_j are the (DNA methylation) residual (see **Application to ARIC Study**) and DNA methylation effect at *j*-th CpG site, z_{i1} to z_{i18} are smoking, sex, age, center, white blood cell count, drinking, leisure time physical activity, plate and chip (array) row, 10 principal components (PCs), γ_k are the corresponding covariate effects, and *P* is the number of CpG sites in *g*-th gene. Second, after the model is fit, the sign of DNA methylation effects $\text{sign}\left(\hat{\beta}_j\right)$ and the estimated covariate effects $\hat{\gamma}_k$ were recorded. Next, the following model was used for generating quantitative trait:

$$y_i = \sum_{j\in C} m_{ij}b_j + \sum_{k=1}^{18} z_{ik}\hat{\gamma}_k + \varepsilon_i$$

where $b_j \sim U\left(\max\left(0, d-1\right), d\right) \times \text{sign}\left(\hat{\beta}_j\right)$ is the DNA methylation effect controlling by *d*, *C* is the indices of causal CpG sites selected, $\varepsilon_i \sim N\left(0, 2\sigma\right)$, σ is the average standard deviation of DNA methylation across the causal CpG sites. Therefore, the generated quantitative trait was associated with the causal CpG sites selected. *C* is determined by the top CpG sites and causal rate specified.

In order to evaluate the statistical power objectively, 500 genes were randomly selected to capture different possible scenarios. Because investigating the empirical power when there are multiple low or moderate signals in the region is of our ultimate interest, a set of DNA

methylation effect was specified $d = 0.5, 1, 1.5$ as well as causal rate $r = 20\%, 30\%, 50\%$. In addition, different sample size $N = 250, 500, 1000$ was considered. For each sample size, 2000 sets of quantitative trait were generated across 500 genes with specified DNA methylation effect and causal rate. Next, we fit the proposed penalized region-based models, existing region-based models, and single-probe model with the generated quantitative trait, residual, and covariates. Finally, for each region-based method, 2000 sets of test statistics as well as *p*-values were produced and stored across 500 genes. Similarly, 2000 sets were also produced and stored over CpG sites across 500 genes produced by single-probe method.

For region-based methods, we defined the average power as the average probability of *p*-values less than a bonferroni-corrected given α level. For single-probe method, we defined the average power as the average probability of *at least* one *p*-value in a gene less than a bonferroni-corrected given α level (i.e., 0.05, 0.01, 0.005).

4.5 Results

4.5.1 Application to ARIC

Throughout the rest of the chapter, "F_pNS" denotes the F-test of penalized no-smoothing model, "F_B&pLS" and "F_F&pLS" denote the F-test of penalized least-square model with B-spline and Fourier basis function, respectively, and similarly "F_B&pRP" and "F_F&pRP" denote the F-test of penalized roughness-penalty model with B-spline and Fourier basis function, respectively.

Five penalized functional region-based tests (F_pNS, F_B&pLS, F_F&pLS, F_B&pRP and F_F&pRP) as well as three existing region-based tests (F-test, SKAT, and SKAT-O) were applied on 17,728 genes interrogated by at least 5 probes. When applying the proposed tests, we set $K_x = K_b = min(35, p\text{-}1)$ because 35 was considered large enough to prevent undersmoothing

(Goldsmith et al. 2010), and *p*-1 was used for preventing the number of Fourier basis functions exceeding *p* – the number of Fourier basis function specified will increase by 1 when it is even in our program. Besides, single-probe test was applied on every probe within these genes. However, we have not yet got the permission for publishing the results of the application to ARIC so only Q-Q plot and real data based simulation results are shown in this section.

4.5.2 Q-Q Plot

To evaluate the distribution of *p*-value under the null hypothesis, we first shuffled the quantitative traits BMI and WC and then conducted the association testing using all the tests considered. As a result, the *p*-values generated by the proposed region-based tests in Q-Q plots adhere to an expected uniform distribution, indicating the test statistics are well behaved (Figure 6). On the other hand, SKAT and SKAT-O constantly generated *p*-value larger than expected, which shows the conservativeness of these two tests and the results match to what we observed in the real data analysis. In addition, the *p*-values generated by single-probe test in Q-Q plots almost adhere to the expected uniform distribution; however, the results also show the conservativeness of this test especially with trait WC. This is not surprising because some CpG sites were potentially correlated as they were affected by the same environmental or biological factors.

4.5.3 Empirical Type I Error

The empirical type I error rates of the proposed penalized functional region-based, existing region-based tests, and single-probe test based on gene *BCL6* are reported in Table 7. For each sample size considered, 10^5 data sets were generated. Results of four different significance levels $\alpha = 0.05, 0.01, 0.005, 0.001$ are reported.

Except the single-probe test, all the tests considered have empirical type I error rates around

the nominal α levels. Therefore, the proposed functional region-based tests as well as existing region-based tests have proper control of type I error rate across a spectrum of sample sizes and significance levels. In addition, single-probe test has smaller type I error rates across all sample sizes and significance levels. In general, all the tests considered are very robust and can be useful in the large-scale or whole epigenome-wide association studies.

4.5.4 Empirical Average Power

The power performance of the proposed penalized functional region-based, existing region-based tests, and single-probe tests are compared based on the simulated quantitative traits and the DNA methylation residuals of ARIC study. For each sample size, 2000 data sets were generated with DNA methylation effect $d = 0.5, 1, 1.5$ and causal rate $r = 20\%, 30\%, 50\%$. The results of the proposed penalized region-based tests are compared with those of existing region-based and single-probe tests in Figure 7-9.

In Figure 7, all causal CpG sites have small DNA methylation effect ($d = 0.5$); when all causal CpG sites have moderate DNA methylation effect ($d = 1$), the results are shown in Figure 8; when all causal CpG sites have large DNA methylation effect ($d = 1.5$), the results are shown in Figure 9. When DNA methylation effect is small ($d = 0.5$), the penalized functional region-based tests and F-test have higher power than that of SKAT, SKAT-O, and single-probe test, except that single-probe test has slightly higher power when both sample size and causal rate are small ($N = 250$; $r = 20\%$). When DNA methylation effect is moderate or large ($d = 1$ or 1.5), the penalized functional region-based tests and F-test always have higher power than that of SKAT, SKAT-O, and single-probe test.

F-test has higher power than all the tests considered, except when the DNA methylation effect size, sample size, and causal rate are small ($d = 0.5$; $N = 250$; $r = 20\%$). On the contrary,

single-probe test generally has smaller power but it has the highest power in such scenario. This is mainly because there are likely one or two causal CpG sites in that scenario, which favors the model of single-probe test. SKAT and SKAT-O have minimal power.

In total, we compared five F-test statistics of penalized functional region-based models: four are based on the combination of least-square and roughness-penalty models and B-spline and Fourier basis functions, and one is based on the no-smoothing model. In general, the five F-test statistics of the penalized functional region-based models have similar power, although the three tests (F_B&pLS, F_F&pLS, F_F&pRP) have slightly higher power and the test of no-smoothing model (F_pNS) has lower power, and the rest test (F_B&pRP) has power right in the middle. In addition, as shown in the real data analysis, these five F-test statistics have very similar power level; therefore, the proposed penalized functional region-based tests do not strongly depend on whether the DNA methylation data is smoothed or not, and which basis functions are used.

4.6 Discussion

In this chapter, we have developed penalized functional region-based models for testing the association between a quantitative trait and multiple DNA methylation variants in a region. Because of the correlation structure among the DNA methylation variants, we considered the observed DNA methylation levels as realization of continuous DNA methylation functions $X_i(t)$ at location *t*. We applied two popular smoothing techniques (least-square and roughness-penalty) for estimating the DNA methylation functions based on B-spline or Fourier basis function. Moreover, we considered using the discrete DNA methylation levels directly for estimating the functions. Then, the estimated DNA methylation functions were used as explanatory variable in the penalized functional regression model adjusting for covariates.

Because the ultimate goal is to test whether a quantitative trait is associated with the DNA methylation variants in a region, it is straightforward to test the association between a quantitative trait and DNA methylation functions. We proposed to use F-test for testing this association. Next, because the model may overfit the data due to the large number of basis function specified, we used penalized spline to estimate the functional parameter by assuming the parameter follows a normal distribution with covariance matrix being penalty matrix that corresponds to the particular spline basis. Specifically, the level of smoothing is estimated using Restricted Maximum Likelihood (REML) in an associated mixed effect model. The methods proposed are implemented in the statistical package R.

From both application to ARIC and simulation study, the proposed penalized functional region-based tests have higher power than that of SKAT and SKAT-O in every scenario considered. The proposed penalized functional region-based tests use the correlation of methylation between adjacent CpG sites in the region. Although SKAT and SKAT-O take pair-wise correlation into account via the kernel matrix, the higher order correlation is not modeled and this could be the reason the proposed tests have higher power. The performance of the proposed penalized functional region-based tests is not necessarily related to whether the DNA methylation data is smoothed and which basis functions are used. Moreover, roughness-penalty smoothing technique takes extra time for finding the smoothing parameter compared to least-square and no-smoothing technique. Therefore, no-smoothing is considered as a good compromise between performance and computational cost.

Among the two basis functions, B-spline basis function leads to more consistent results. For example, from the real data application with trait BMI, 37.7% of the genes have the exact same *p*-values from F_B&pLS and F_B&pRP. Among these, the smoothing parameter λ in the

roughness-penalty model with B-spline basis function are very small (Mean = 0.04), leading the F-test statistics F_B&pLS and F_B&pRP are nearly identical. On the other hand, only 3.8% of the genes have exact same *p*-values from F_F&pLS and F_F&pRP. Among these, the λ in roughness-penalty model with Fourier basis function are very large (Mean = 1.54×10^{8}).

Finally, we envision the proposed penalized functional tests can be applied to not only candidate genes analysis but also epigenome-wide association studies because they have proper control of type I error rates at all levels (α = 0.05, 0.01, 0.005, 0.001) and also attractive performance in power. We hope this method can help finding more DNA methylation variants associated with the quantitative trait and further resolving the missing heritability problem.

Table 7 Empirical Type I Error (Gene *BCL6*)

Significance Level	Sample Size	F_pNS	F_B&pLS	F_B&pRP	F_F&pLS	F_F&pRP	F-test	SKAT	SKAT-O	Single-Probe
0.05	250	0.0506	0.0503	0.0507	0.0503	0.0508	0.0507	0.0497	0.0484	0.0271
	500	0.0502	0.0504	0.0504	0.0509	0.0501	0.0499	0.0490	0.0494	0.0266
	1000	0.0503	0.0499	0.0499	0.0509	0.0504	0.0505	0.0496	0.0503	0.0291
	2000	0.0511	0.0511	0.0511	0.0507	0.0506	0.0496	0.0496	0.0495	0.0336
0.01	250	0.0099	0.0101	0.0101	0.0101	0.0100	0.0104	0.0097	0.0094	0.0062
	500	0.0100	0.0097	0.0097	0.0102	0.0103	0.0105	0.0093	0.0093	0.0060
	1000	0.0103	0.0096	0.0096	0.0099	0.0098	0.0101	0.0102	0.0097	0.0071
	2000	0.0105	0.0103	0.0103	0.0105	0.0101	0.0102	0.0094	0.0098	0.0078
0.005	250	0.0052	0.0052	0.0051	0.0052	0.0050	0.0048	0.0050	0.0047	0.0032
	500	0.0051	0.0049	0.0049	0.0051	0.0051	0.0051	0.0047	0.0047	0.0033
	1000	0.0051	0.0047	0.0047	0.0047	0.0048	0.0051	0.0048	0.0048	0.0038
	2000	0.0052	0.0050	0.0050	0.0050	0.0050	0.0049	0.0046	0.0047	0.0041
0.001	250	0.0011	0.0010	0.0012	0.0010	0.0011	0.0009	0.0009	0.0008	0.0007
	500	0.0010	0.0011	0.0011	0.0010	0.0010	0.0012	0.0008	0.0010	0.0007
	1000	0.0010	0.0009	0.0009	0.0009	0.0010	0.0011	0.0010	0.0011	0.0009
	2000	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009	0.0009	0.0011	0.0010

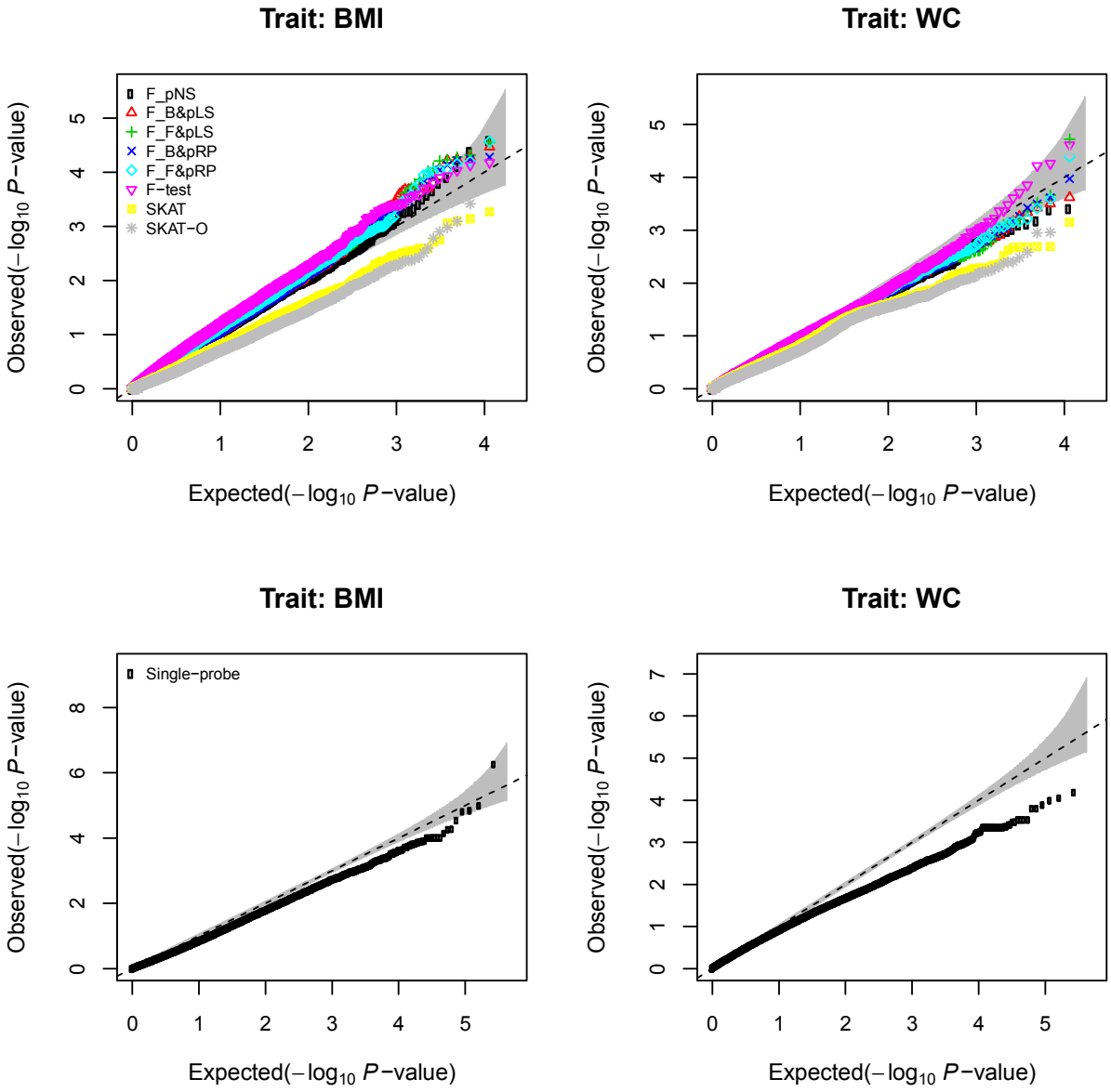

Figure 7 Q-Q Plot of *P*-values Generated by Region-based Tests and Single-probe Test

The observed (Y-axis) vs. expected (X-axis) $-\log_{10}(p\text{-values})$ generated by region-based tests (top) and single-probe test (bottom) using quantitative traits BMI (left) and WC (right) are shown.

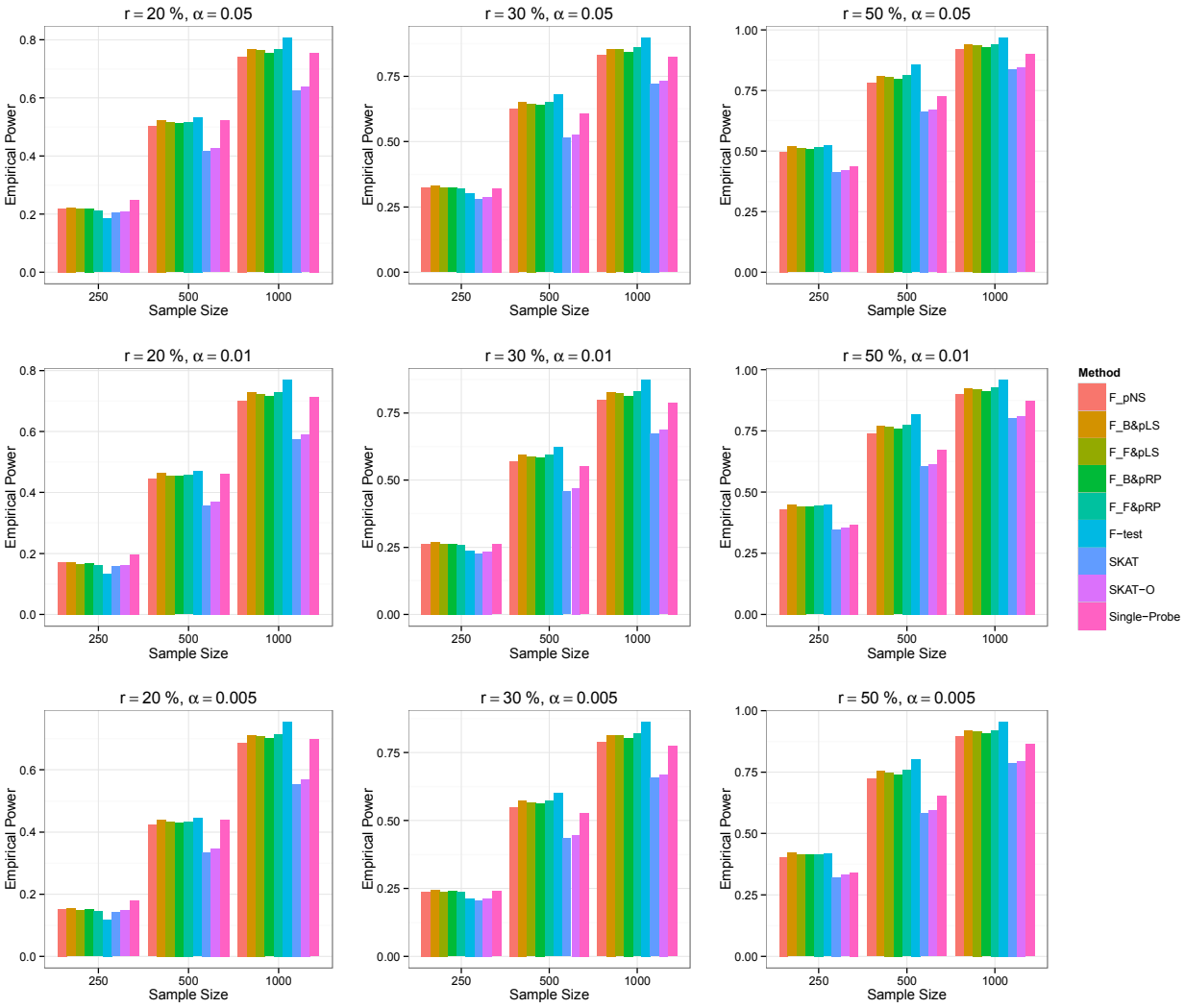

Figure 8 Empirical Power with Small DNA Methylation Effect (*d* = 0.5)

Empirical power (Y-axis) of different methods is shown across a spectrum of sample size (X-axis), causal rate (r), and significance level (α) with small DNA methylation effect.

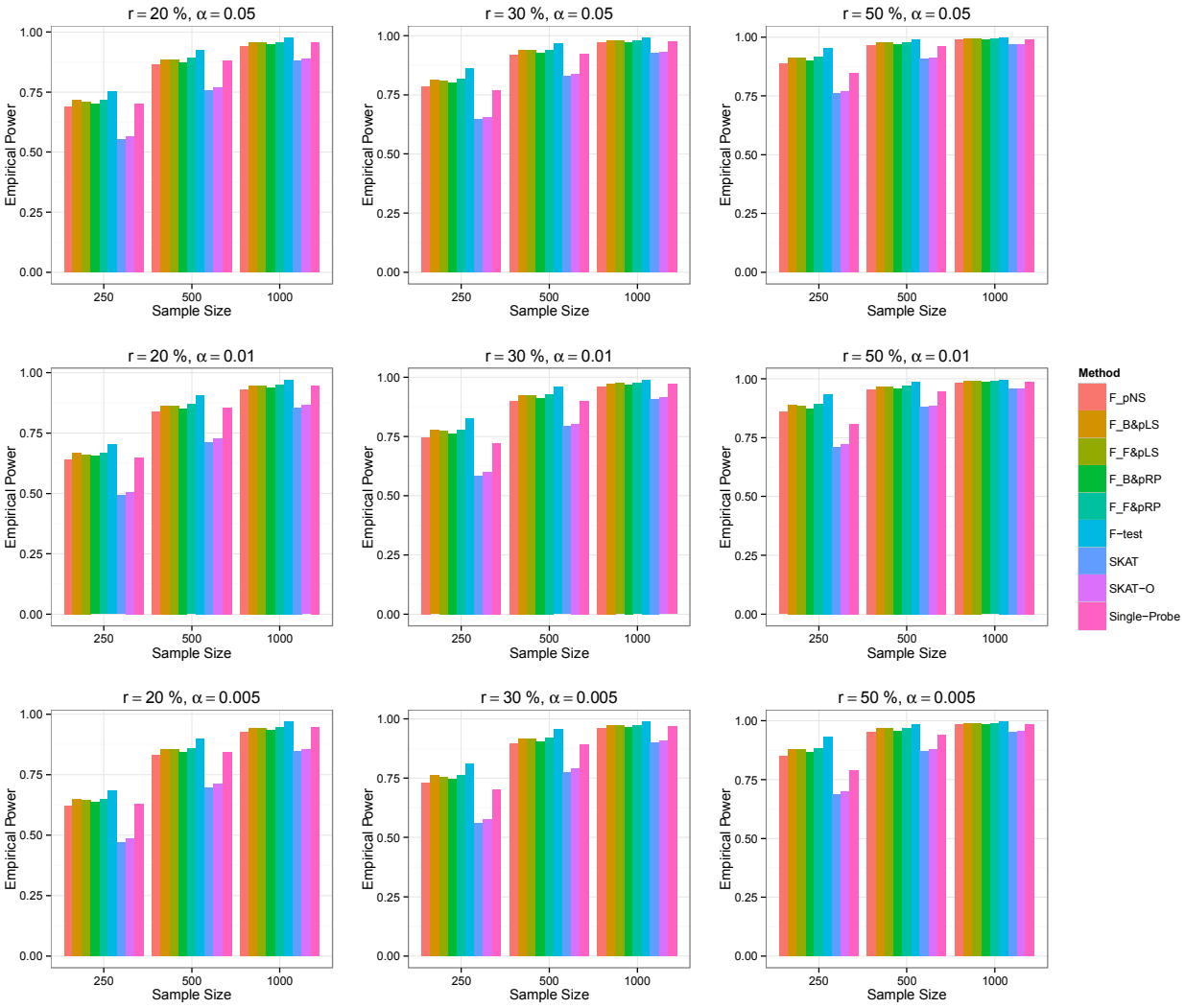

Figure 9 Empirical Power with Moderate DNA Methylation Effect ($d = 1$)

Empirical power (Y-axis) of different methods is shown across a spectrum of sample size (X-axis), causal rate (r), and significance level (α) with moderate DNA methylation effect.

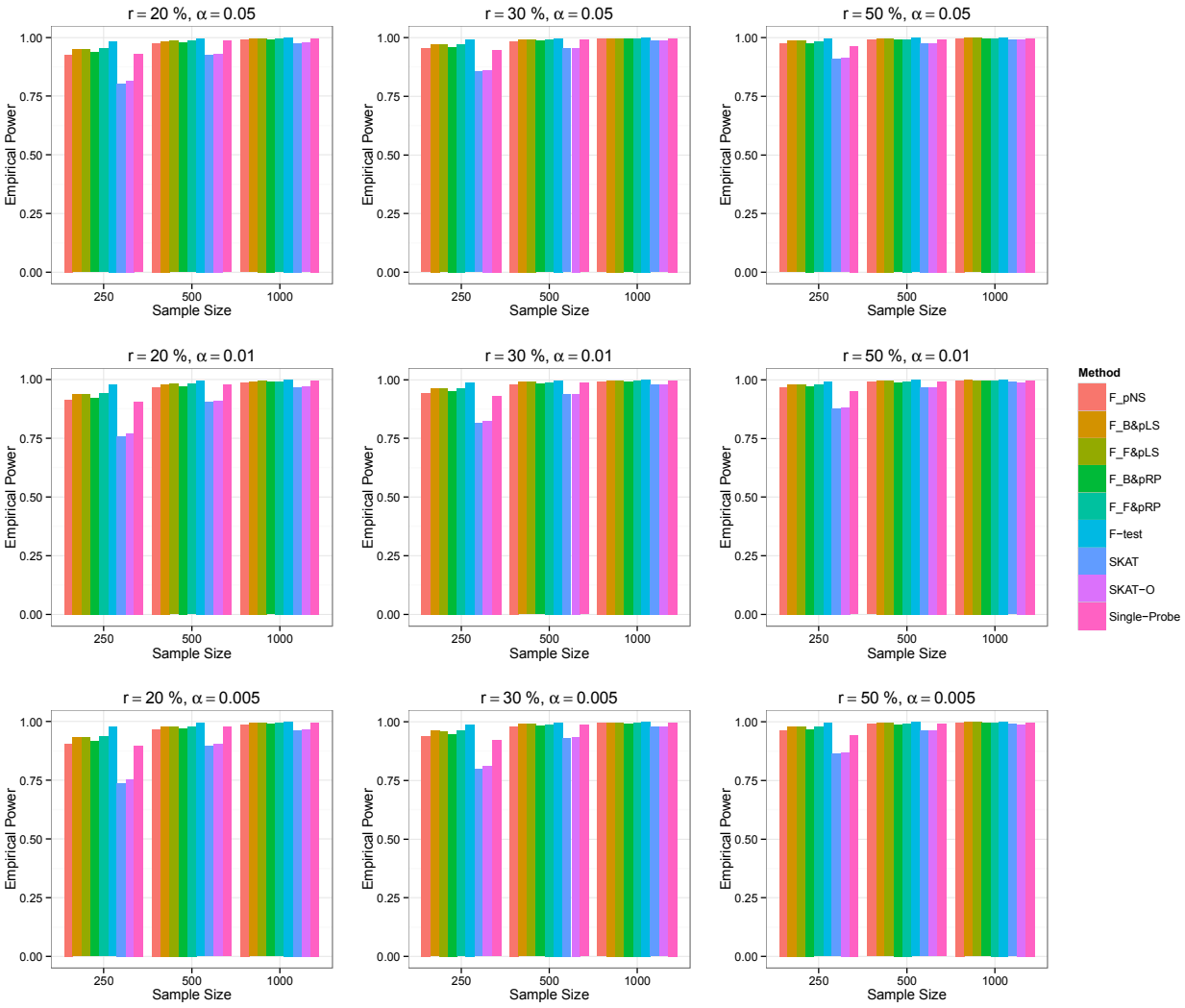

Figure 10 Empirical Power with Large DNA Methylation Effect (*d* = 1.5)

Empirical power (Y-axis) of different methods is shown across a spectrum of sample size (X-axis), causal rate (r), and significance level (α) with large DNA methylation effect.

CHAPTER 5:ACROSS-PLATFORM IMPUTATION OF DNA METHYLATION LEVELS USING PENALIZED FUNCTIONAL REGRESSION

5.1 Introduction

SNP imputation is a standard procedure used to resolve inconsistencies between genotyping arrays and to increase the resolution of data collected in GWASs (Li et al. 2009). Therefore, we propose the application of a similar concept to impute data in DNA methylation profiles from a subset of probes. Although DNA methylation does not exhibit as clear correlation structure as LD blocks among SNPs (Eckhardt et al. 2006), we observe both local and nonlocal correlations between probes. In this chapter, we develop a penalized functional regression model (Goldsmith et al. 2010) which uses functional predictors to capture these non-local correlations. Our study demonstrates that this model can impute a HM27 data set into a HM450 data set effectively and accurately. We describe the details of our methodological framework in 5.2 Methods. In 5.4 Application to AML Data Set, we apply this approach to a large-scale methylation data set from acute myeloid leukemia patients.

5.2 Methods

We employed the penalized functional regression model (Goldsmith et al. 2010), with minor modification detailed below to quantify the relationship between DNA methylation from HM450 probes and DNA methylation density function estimated from HM27 probes together with other covariates [See **Selection of Local Covariates**]. Specifically, assume for each target HM450 probe (target probe), we observed data $[Y_i, X_i(t), Z_i]$ across all individuals i = 1, 2, ..., N,

where N is the number of samples, Y_i is the DNA methylation level for the i-th sample at the target HM450 probe measured as $\log_2[\beta/(1-\beta)]$, $X_i(t)$ is the i-th sample specific density function of T_i, the DNA methylation level for the i-th sample at HM27 probes, and Z_i is the vector of covariates for the i-th sample. We used the following generalized functional linear model

$$Y_i \sim EF(\mu_i, \eta)$$

$$g(\mu_i) = \alpha + \int_0^1 X_i(t)\beta(t)dt + Z_i\gamma$$

Here, $EF(\mu_i, \eta)$ denotes an exponential family distribution with mean μ_i and dispersion parameter η, $g(\cdot)$ is a link function, α is the overall mean, $\beta(t)$ is the effect of density function $X_i(t)$ when $T_i = t$, and γ is a regression coefficient vector of covariates.

5.2.1 Estimation of $X_i(t)$

To improve imputation, we incorporated functional predictors into our model to capture information such as non-linear relationship from non-local probes. We estimated the DNA methylation function $X_i(t)$ for a particular target probe with the DNA methylation data from HM27 probes in the same group as the target probe. Assume the target probe is in group g and there are q HM27 probes in the same group. The observed DNA methylation data is denoted as $t_i = (t_1^g, \ldots, t_q^g)$, where t_j^g is the DNA methylation value at j-th HM27 probe in group g and j = 1, …, q. Instead of estimating $X_i(t)$ by expanding into the PC basis obtained from its covariance matrix (Goldsmith et al. 2010), we estimated the density function by using *R* function *density*() with the observed data t_i so that $X_i(t) = f_{T_i}(t)$.

5.2.2 Estimation of $\beta(t)$

$\beta(t)$ was expanded by a linear spline basis $\beta(t) = b_1 + b_2 t + \sum_{k=3}^{K_b} b_k (t - \gamma_k)_+$, where γ_k are knots in the interval [0,1] and $(t - \gamma_k)_+$ is an indicator function, taking value of 1 if $t > \gamma_k$ and 0 if $t \le \gamma_k$. We further defined a spline basis vector $\phi(t) = \{\phi_1(t), \phi_2(t), \ldots, \phi_{K_b}(t)\} = \{1, t, (t-\gamma_3)_+, \ldots, (t-\gamma_{K_b})_+\}$ and a coefficient vector $b = (b_1, \ldots, b_{K_b})'$. Further, smoothing was induced by assuming $b \sim N(0, D)$ where D is a penalty matrix corresponding to the particular spline basis $\phi(t)$.

Finally, we had $\int_0^1 X_i(t)\beta(t)dt = \int_0^1 f_{T_i}(t)\varphi(t)bdt = \int_0^1 f_{T_i}(t)\varphi(t)dt \cdot b$. For ease of notation, we denoted $J_{X\phi}$ as the $n \times K_b$ matrix with the (i,k)-th entry equal to $\int_0^1 f_{T_i}(t)\phi_k(t)dt$ and Z as the $n \times p$ matrix with the i-th row equal to Z_i where p is the number of covariates. The model can be written in matrix format as

$$Y \mid X(t) \sim EF(\mu, \eta)$$

$$g(\mu) = [1, J_{X\varphi}, Z][\alpha, b, \gamma]$$

$$b \sim N(0, D)$$

which is a mixed effect model with K_b random effects b and penalty matrix

$$D = \begin{bmatrix} 0_{2\times 2} & 0_{2\times(K_b-2)} \\ 0_{(K_b-2)\times 2} & I_{(K_b-2)\times(K_b-2)} \end{bmatrix}$$

Typically, $K_b = 35$ is sufficient to avoid undersmoothing in most applications.

5.2.3 Selection of Local Covariate

We exploited linear correlation with neighboring probes by including DNA methylation values of HM27 probes near the target HM450 probe as local covariates Z in our imputation model. For simplicity, we selected the five nearest upstream and the five nearest downstream probes to each target probe as these local covariates.

5.2.4 Grouping Probes

Based on the assumption that probes with similar properties tend to show similar methylation profiles, we divided the probes into several property groups. Here we divided the probes among five groups according to their relative location to a CpG island, labeled "CpG Island," "North Shore," "South Shore," "North Shelf," and "South Shelf" (Bibikova et al. 2011). Probes may also be categorized according to other properties, such as their relative location to a gene (Bibikova et al. 2011).

5.2.5 Quality Filter

When an imputation model is formed without sufficient information, it tends to be underfitted and yield inaccurate imputation results. It is therefore desirable to have quality metrics for gauging imputation quality. As such a quality metric, we proposed an under-dispersion measure defined as the ratio of the variance of fitted methylation values to its expected value (the variance of the true methylation values in the training set). If this ratio is below a certain threshold for a probe, it indicates an underfitted model for that probe, and we discard imputed values for the probe before subsequent analysis. A more exacting threshold ratio can provide more accurate results, although at the cost of fewer probes imputed.

5.2.6 Imputation Quality Assessment

We assessed imputation quality using fivefold cross-validation. Within each split, the full data set was divided into a training set accounting for 80% of the data and a testing set comprised of the remaining 20%. For each testing set, we only retained HM27 data, which contains a subset of HM450 probes, and masked methylation values of other HM450-specific probes. For the training set, we only used HM450 data. We treated methylation data on probes shared between HM27 and HM450 as predictors to impute methylation values at HM450-specific probes. Specifically, we fitted a generalized functional regression model based on the training set, learned the relationship between methylation values of shared and HM450-specific probes, and used the fitted model to impute the masked values of HM450 probes from the HM27 data in the testing set. Finally, we evaluated the imputation performance over splits by averaging quality measures.

As quality measures, we selected the mean squared error (MSE) and the squared Pearson correlation (R^2) between the imputed and the true methylation values in the testing sets. Although R^2 is a more intuitive measure of quality directly related to power and sample size in downstream analysis, we would like to note that this metric could easily be affected by a few outliers. Additionally, if the variance of methylation values for a specific probe is small, R^2 can be dramatically affected even by small imputation errors.

5.2.7 Simulation of Association Study

To assess the potential improvement of statistical power when using well-imputed methylation values for EWAS, we performed several simulated association studies. Specifically, we randomly selected 100 HM450 probes with imputation R between 0.4 and 0.5, and simulated a data set for each probe. For each probe, a trait value Y_i^* was simulated from the methylation

level of this probe according to the linear model $Y_i^* = c\beta_i^* + \varepsilon_i$ for sample *i*, where β_i^* is the true methylation β value, the effect size $c \in \{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$, and $\varepsilon_i \sim N\left(0, 2s_{\beta_i^*}\right)$, where $s_{\beta_i^*}$ is the sample standard deviation of β_i^*.

We repeated the simulation 2000 times. For each simulated data set, we performed association tests based on the true methylation values, as well as imputed values from the simple linear model and our proposed penalized functional model. The empirical power of each method was calculated as the proportion of observed *p*-values that fall below the significance threshold $\alpha = 0.05$. Finally, we evaluated the empirical power for each effect size *c* by averaging results from 100 probes.

5.3 Simulation Results

It is not surprising to find relatively little difference in the performance of the two models at the two ends of the distribution (Figure 10-11) because of probes that are either trivial or impossible to impute. Therefore, we focus on probes with imputation R between 0.4 and 0.5. As shown in Figure 12, using imputed values from penalized functional model for association tests is consistently more powerful than using values from simple linear model, while type I error rate was still properly controlled when *c* = 0. The results suggest that even using probes with moderate imputation quality can improve the statistical power of association test dramatically.

5.4 Application to AML Data Set

We evaluated our imputation model using DNA methylation data from TCGA acute myeloid leukemia (AML) samples (Cancer Genome Atlas Research Network 2013). The data set contains DNA methylation data of tumor tissues from 194 patients with AML and is one of the largest methylation data sets in TCGA project. All samples were evaluated using both HM27 and

HM450. We transformed the initial β values into M values, defined as the log 2 ratio of β and 1 – β, as the M values better follow the Gaussian distribution (Network 2013). Our goal is to impute the HM27 data set into a HM450 data set to get an expanded view of the epigenomic landscape. The data set is publicly available at the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/).

Since imputation of sporadic missing data is not the focus of this work, we removed all probes with at least one missing value for the sake of convenience. However, these missing values can be imputed with existing methods to generate data without missing values. Additionally, we removed 743 probes designed in HM27 but not in HM450. In total, our HM27 data set consisted of 20,794 probes and our HM450 data set consisted of 393,152 probes. The latter set contained all 20,794 probes in HM27, leaving the remaining 373,358 as our potential imputation targets.

We noted that as HM27 and HM450 employ different biochemical methods to measure methylation levels, platform-specific effects might negatively impact imputation performance. To alleviate this systematic effect, we fitted a LOESS regression model between two platforms, stratified by the number of CpGs in the probe (#CpG = 0,1,2,3,4,5,6,7+), using 14 randomly chosen samples and normalized the HM27 data against the HM450 data (Network 2013).

Most probes showed nearly constant methylation levels in populations, making imputation trivial for them. We therefore focused on probes showing large variation and chose the top 20,000 such probes to evaluate imputation quality. In the fivefold cross-validation experiment, 14 samples used for normalization were removed at first. Among the remaining 180, 144 individuals were chosen at random as the training set and 36 as the testing set within each split. The empirical cumulative distribution of imputation MSE and R^2 are shown in Figure 10-11. We compared the two models with and without functional predictors and found that

incorporating functional predictors leads to significantly improved imputation MSE and R^2 ($P<2.2 \times 10^{-16}$ for both metrics, paired Wilcoxon test). Table 8 summaries some basic statistics.

We used as an example the target probe cg00288598 to illustrate how the functional predictors improve imputation quality. As shown in Figure 13, the selected local probes showed much lower variation than the target probe, leading to an underfitted linear regression model and thus low imputation quality. In contrast, the methylation profile of the target probe is strongly associated with the distribution of methylation levels from all HM27 probes in its assigned North Shelf group, as indicated in Figure 14. Therefore after the functional predictors are added, the model can utilize the information from these non-local probes, including probes on different chromosomes, to alleviate the underfitting problem.

Because not all target probes can be imputed with the same level of accuracy, we tried to use the under-dispersion measure described in the 5.2 Methods section to filter out inaccurate imputation results. We examined the relationship between imputation MSE/R^2 and the under-dispersion measure. We observed a negative correlation between imputation MSE and this quality measure (Figure 15, Pearson correlation coefficient R = –0.65), and a positive correlation between imputation R^2 and the measure (Figure 16, Pearson correlation coefficient R = 0.93). Therefore when performing imputation, we can calculate the under-dispersion measure and use it to filter out low-quality imputation results. Figure 15-16 indicates that by choosing an appropriate threshold, we can remove most low-quality results while simultaneously retaining nearly all high-quality results.

5.5 Discussion

In summary, we propose a penalized functional regression framework for the across-platform imputation of methylation probes. Our real data analysis demonstrates that by

incorporating functional predictors, our model can produce accurate imputation results when the reference panel (training set) and target panel (testing set) characterize the same tissue under similar conditions. However, since DNA methylation profiles are highly tissue and condition-specific (Lister et al. 2009; Laurent et al. 2010; Varley et al. 2013), our method will not work well if the two data sets are from different tissues or very different conditions. Recent studies suggest some statistical models to predict methylation profile in target tissue from a surrogate tissue (Ma et al. 2014), which might be helpful in this case. Moreover, other systematic errors such as batch effect may also harm imputation quality. Therefore, we suggest using techniques such as principal component analysis to check for obvious discrepancies between reference and target panels before applying our method.

Since most CpG sites display stable DNA methylation levels, imputation error is low on average. However, researchers may consider dynamic CpG sites to be of more interest, as these sites often co-localize with key regulators such as enhancers and transcription factor binding sites (Ziller et al. 2013). Therefore, we calculate quality metrics for individual probes, facilitating the evaluation of imputation quality for each probe and removing probes with low imputation quality for downstream analysis. For probes showing large variation of methylation levels, we notice that even when incorporating functional predictor, the imputation quality is still low for a significant portion of these probes. Possible reasons for this are as follows: First, the DNA methylation profile alone may not provide sufficient information for accurate imputation. We may need to incorporate other information to improve the imputation quality, such as local DNA context and binding profile of regulatory proteins (Bhasin et al. 2005; Bock et al. 2006; Zheng et al. 2013), although this requires additional data source in the same or similar tissue type that are rarely available. Second, HM27 has a much lower resolution than HM450. As such, many

HM450 probes may not be highly correlated with HM27 probes, making them difficult to impute. We expect to observe better performance if we impute from a denser microarray. Third, our normalization procedure does not fully eliminate the inconsistency of measurements between HM27 and HM450, which also affects the performance of this model.

After accurate imputation, we can easily combine data from multiple platforms to obtain methylation levels of more CpG sites for downstream analysis, such as detecting methylation quantitative trait loci or EWASs (Rakyan et al. 2011; Heyn and Esteller 2012). We expect this higher-resolution exploration of the epigenome will lead to rapid advances in understanding the functional role of normal DNA methylation and the impact of its aberration.

Table 8 Quantiles of Imputation MSE and R^2

	Imputation MSE			Imputation R^2		
	Q1	Median	Q3	Q1	Median	Q3
Covariates only	0.0328	0.0582	0.0772	0.0485	0.1574	0.3760
Covariates + Functional Predictor	0.0292	0.0516	0.0708	0.1178	0.27	0.4536

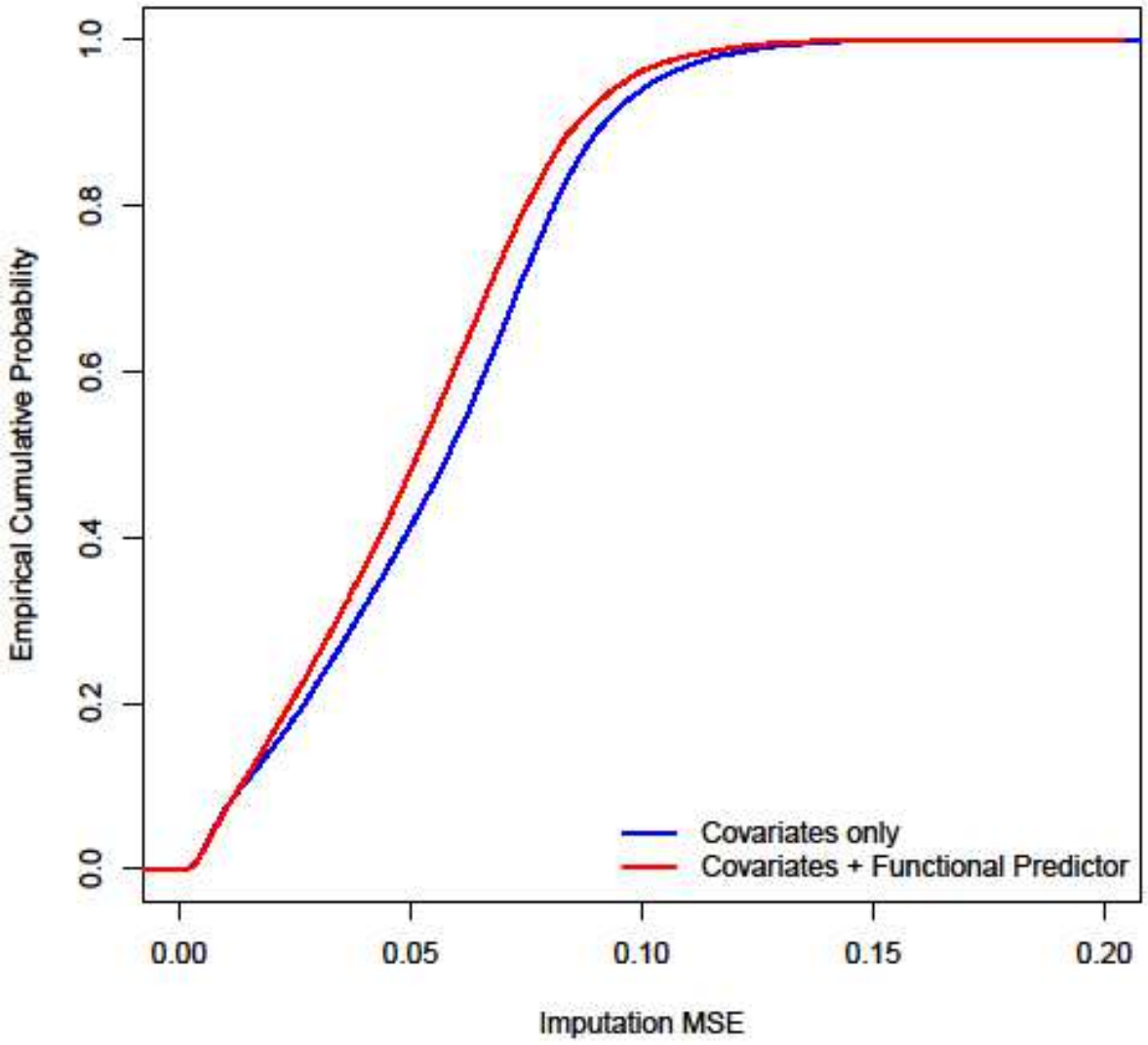

Figure 11 Empirical Cumulative Distribution of Imputation MSE for Probes Showing Large Variation in AML Data Set

The empirical cumulative distribution (Y-axis) of imputation MSE (X-axis) generated by Covariates only (blue) is compared with the one generated by Covariates + Functional Predictor (red).

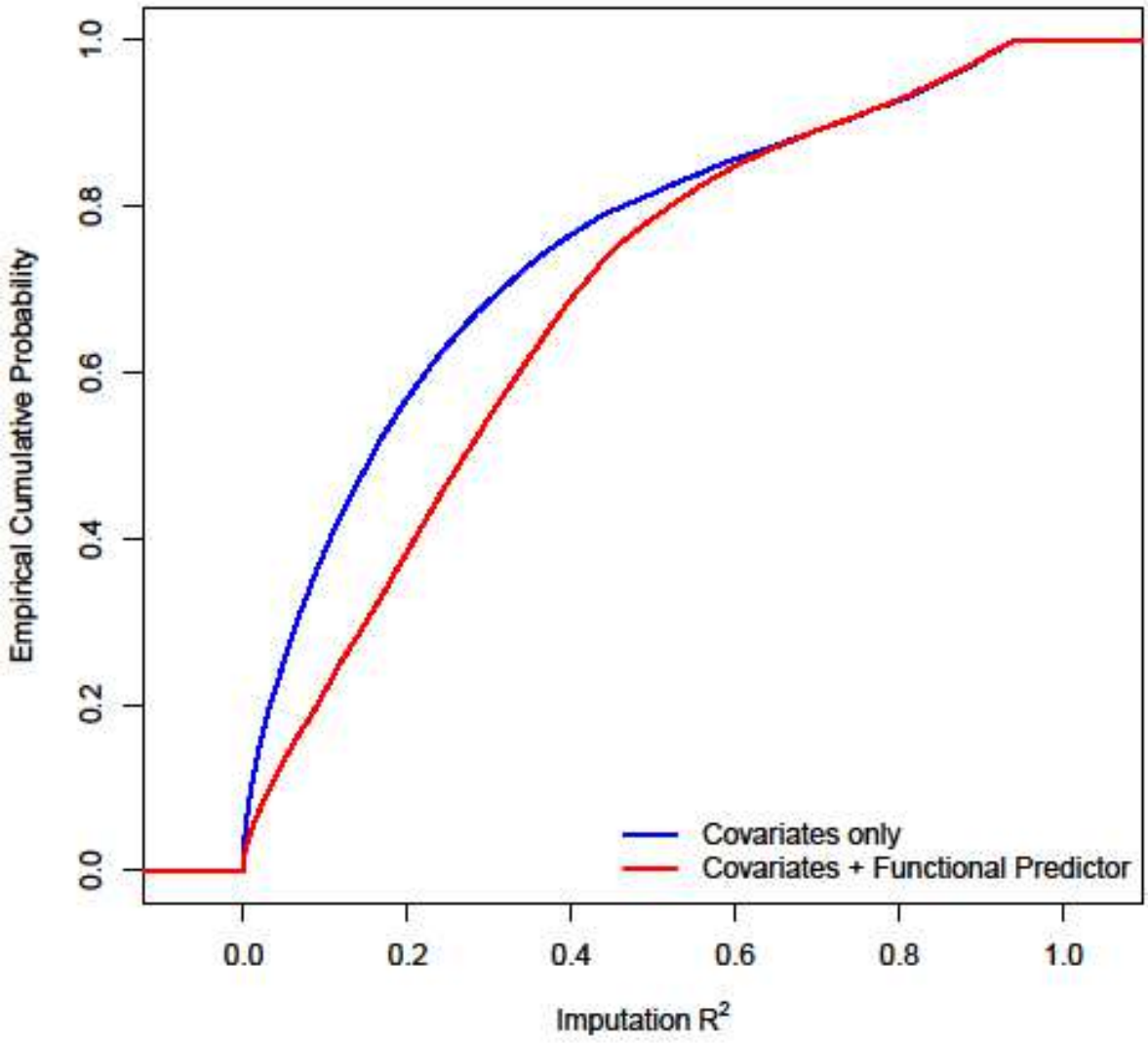

Figure 12 Empirical Cumulative Distribution of Imputation R^2 for Probes Showing Large Variation in AML Data Set

The empirical cumulative distribution (Y-axis) of imputation R^2 (X-axis) generated by Covariates only (blue) is compared with the one generated by Covariates + Functional Predictor (red).

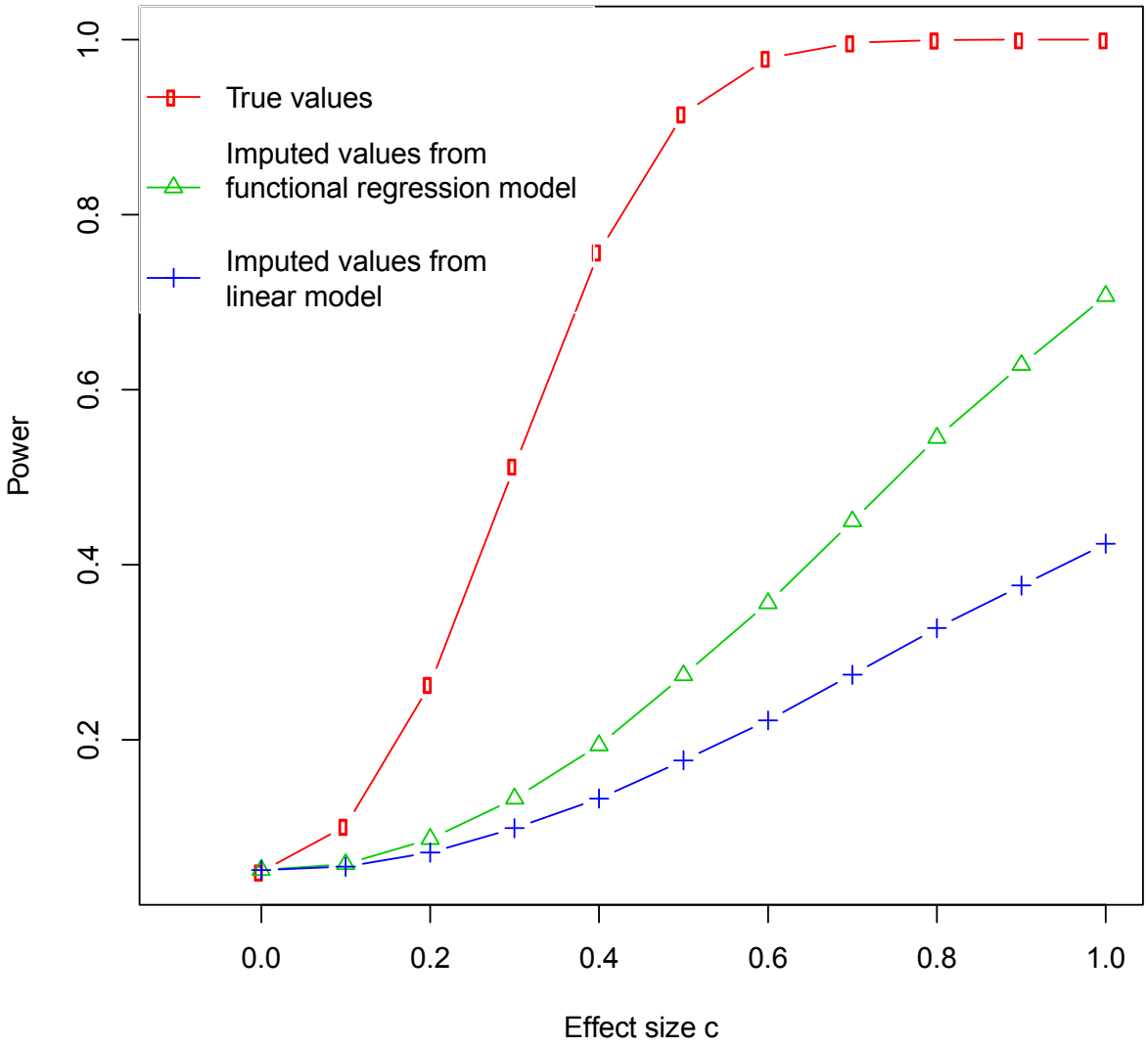

Figure 13 Empirical Power of Simulated Association Tests Across A Spectrum of Effect Size *c*

The empirical power (Y-axis) of association test using three different values, true values (red), imputed values from functional regression model (green), and imputed values from linear model (blue), across a spectrum of effect size b (X-axis) are compared.

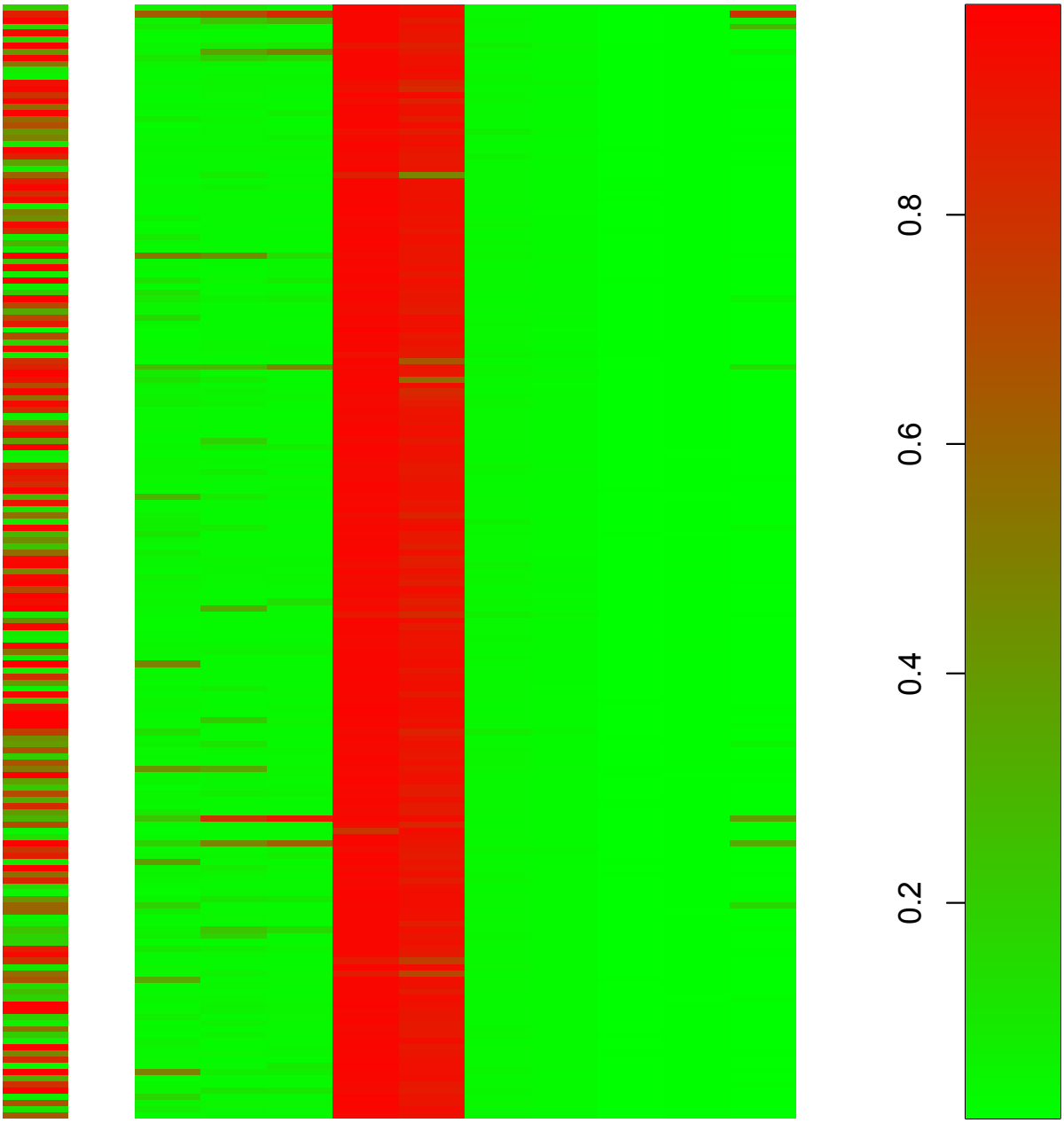

Figure 14 The DNA Methylation Profile of the Target Probe vs. 10 Selected Local Probes

The DNA methylation profile of the target probe cg00288598 (left) is compared with the profile of the 10 selected local probes (middle).

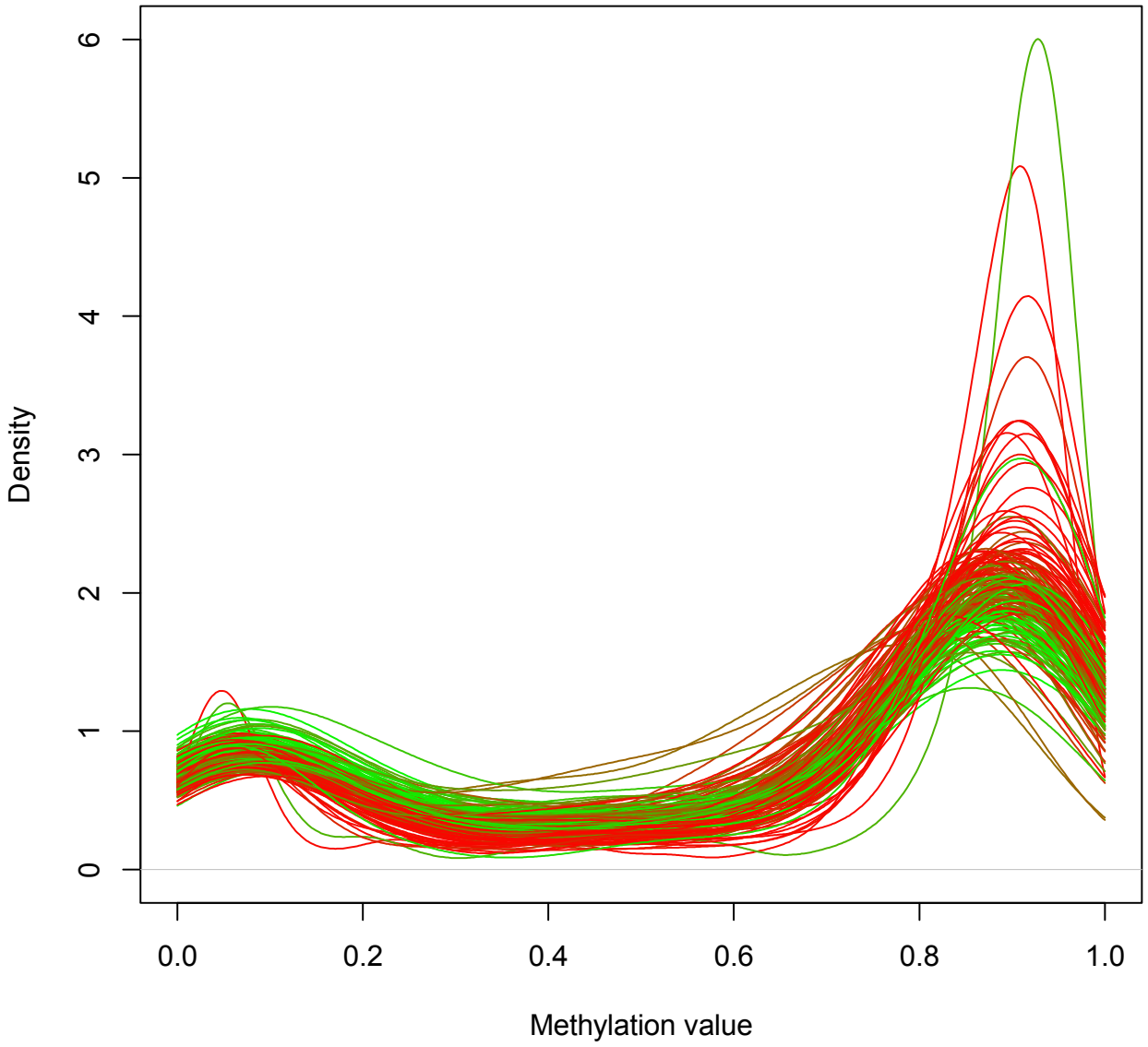

Figure 15 The Individual-specific Density Plot of DNA Methylation Level

The density curve of DNA methylation levels generated from HM27 probes in North Shelf regions. Each line represents one individual's DNA methylation density curve and is colored according to the DNA methylation level of the cg00288598 probe.

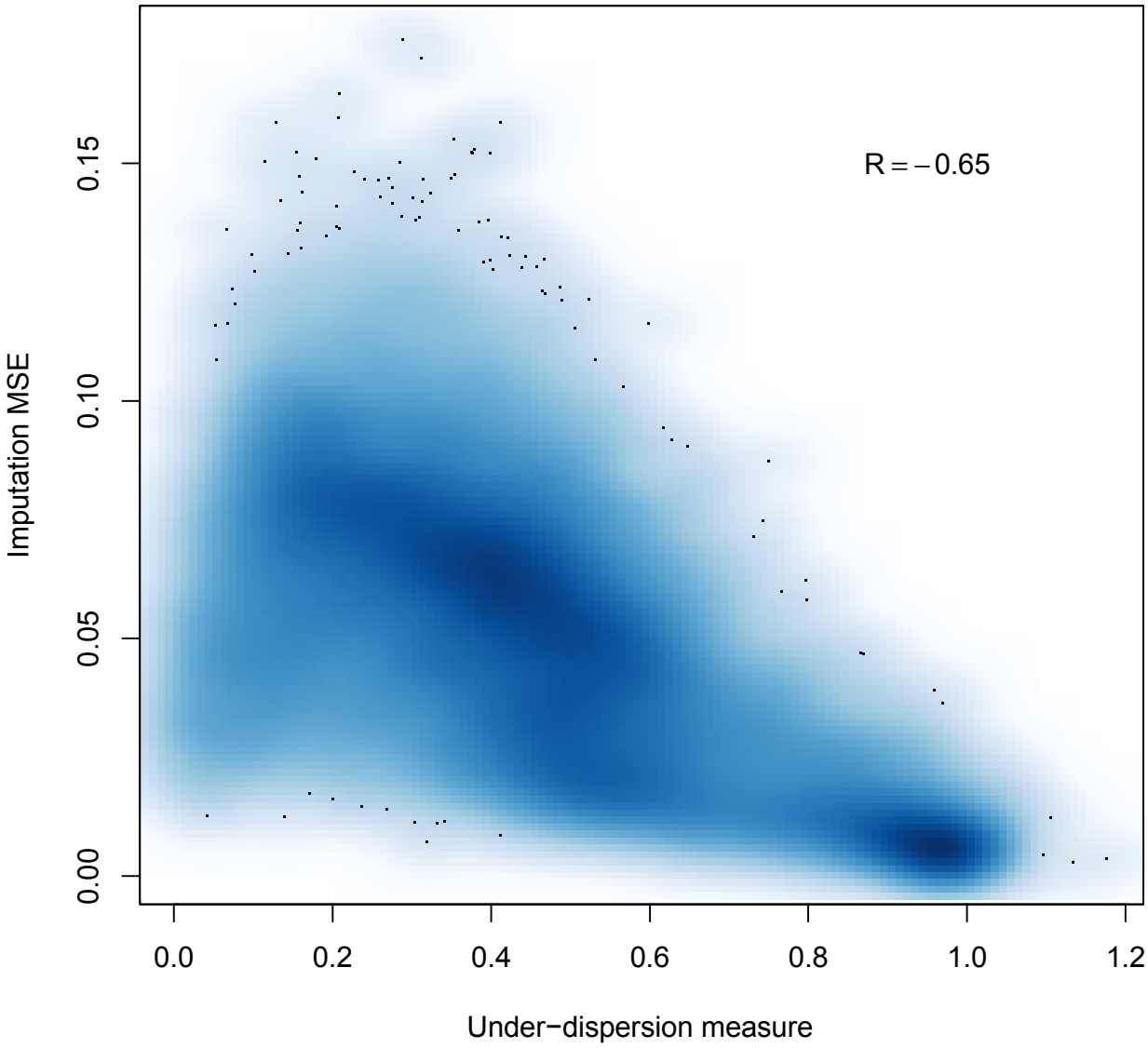

Figure 16 Scatter Plot of Imputation MSE vs. Under-Dispersion Measure

The imputation MSE (Y-axis) calculated between true and predicted DNA methylation level is plot against the under-dispersion measure (X-axis). The correlation between imputation MSE and under-dispersion measure is -0.65.

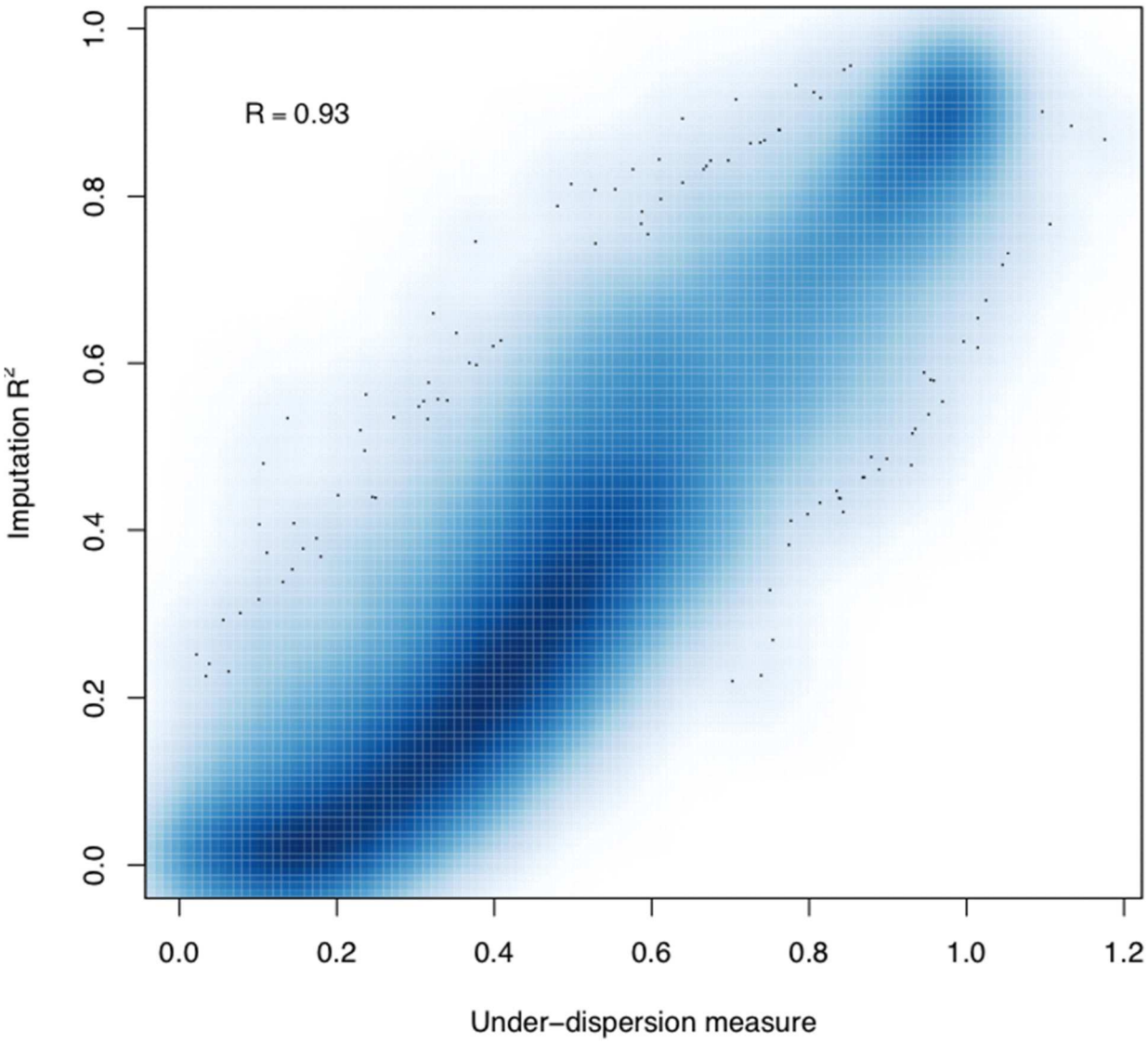

Figure 17 Scatter Plot of Imputation R^2 vs. Under-dispersion Measure

The imputation R^2 (Y-axis) calculated between true and predicted DNA methylation level is plot against the under-dispersion measure (X-axis). The correlation between imputation R^2 and under-dispersion measure is 0.93.

CHAPTER 6:CONCLUDING REMARKS

This document presents several novel methods for identifying truly associated rare variants and causal genes in genome-wide and epigenome-wide association studies, respectively. While each method is intended for a specific study design and involves a variety of statistical and computational tools, the central goal remains the same: to build an optimal model and maximize the statistical power. We have demonstrated that various statistical methodologies can be used to improve power for association studies while maintaining acceptable type I error rates. For GWASs, we proposed EM-LRT and found that when posterior probabilities of all potential genotypes are available, the proposed EM-LRT-Prob has nearly identical performance as Mixture method; however, it is much more computationally efficient. On the other hand, when only dosages are available, the proposed EM-LRT-Dose, which incorporates the information of imputation quality measure R^2, has enhanced power over the Dosage method for association analysis of variants with low frequency or imputation quality. For EWASs, we proposed penalized functional region-based tests and showed that they have higher power of identifying causal genes than the single-probe test, SKAT, and SKAT-O from both real data analysis and real data based simulation when there are multiple small or moderate signals in the region. That being said, single-probe test and the proposed penalized functional region-based tests have similar power performance when there are strong signals in the region.

In addition, geneticists are embraced nowadays by technological advances. For the study of DNA methylation, for example, technological advances constantly provide us with more

choices to measure DNA methylation patterns across the genome, including multiple commercial arrays, multiple sequencing-based technologies or protocols. As a result, large studies may have a mixture of old and new arrays, or a mixture of old and new technologies, on the large number of samples they investigate, which makes data analysis challenging. The proposed penalized functional regression model can predict methylation level at sites on a new array, the Illumina HumanMethylation450K Beadchip, reasonably well from those on an old array, the Illumina HumanMethylation27K Beadchip. However, since DNA methylation profile is highly tissue-specific and condition-specific, our method will not work well if these two data sets are from different tissues or quite different conditions.

REFERENCES

Acar, E. F., and L. Sun. 2013. A generalized kruskal-wallis test incorporating group uncertainty with application to genetic association studies. Biometrics 69:427–435.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour. 2006. Microarray data analysis: from disarray to consolidation and consensus. Nature reviews. Genetics 7:55–65.

Altshuler, D., E. Lander, and L. Ambrogio. 2010a. A map of human genome variation from population scale sequencing. Nature 476:1061–1073.

Altshuler, D. M., R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Altshuler, D. M., R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, P. E. Bonnen, et al. 2010b. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.

Auer, P. L., J. M. Johnsen, A. D. Johnson, B. a Logsdon, L. a Lange, M. a Nalls, G. Zhang, et al. 2012. Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. American journal of human genetics 91:794–808.

Aulchenko, Y. S., M. V Struchalin, and C. M. Van Duijn. 2010. ProbABEL package for genome-wide association analysis of imputed data.

Bergman, Y., and H. Cedar. 2013. DNA methylation dynamics in health and disease. Nature structural & molecular biology 20:274–81.

Berman, B. P., D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, et al. 2011. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. Nature Genetics.

Berndt, S. I., S. Gustafsson, R. Mägi, A. Ganna, E. Wheeler, M. F. Feitosa, A. E. Justice, et al. 2013. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nature genetics 45:501–12.

Bhasin, M., H. Zhang, E. L. Reinherz, and P. A. Reche. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Letters 579:4302–4308.

Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, et al. 2011. High density DNA methylation array with single CpG site resolution. Genomics 98:288–295.

Bird, A. 2002. DNA methylation patterns and epigenetic memory. Genes and Development.

Bizon, C., M. Spiegel, S. A. Chasse, I. R. Gizer, Y. Li, E. P. Malc, P. A. Mieczkowski, et al. 2014. Variant calling in low-coverage whole genome sequencing of a Native American population sample. BMC genomics 15:85.

Bock, C., M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genetics 2:0243–0252.

Browning, B. L., and S. R. Browning. 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics 84:210–223.

Cancer Genome Atlas Research Network. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England journal of medicine 368:2059–74.

Chambers, J. C., W. Zhang, J. Sehmi, X. Li, M. N. Wass, P. Van der Harst, H. Holm, et al. 2011. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nature genetics 43:1131–8.

Chen, W.-M., and G. R. Abecasis. 2007. Family-based association tests for genomewide association scans. American journal of human genetics 81:913–926.

Craven, P., and G. Wahba. 1978. Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31:377–403.

Croteau-Chonka, D. C., Y. Wu, Y. Li, M. P. Fogarty, L. a Lange, C. W. Kuzawa, T. W. McDade, et al. 2012. Population-specific coding variant underlies genome-wide association with adiponectin level. Human molecular genetics 21:463–71.

Dastani, Z., M. F. Hivert, N. Timpson, J. R. B. Perry, X. Yuan, R. A. Scott, P. Henneman, et al. 2012. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. PLoS Genetics 8.

Duan, Q., E. Y. Liu, P. L. Auer, G. Zhang, E. M. Lange, G. Jun, C. Bizon, et al. 2013a. Imputation of coding variants in African Americans: Better performance using data from the exome sequencing project. Bioinformatics 29:2744–2749.

Duan, Q., E. Y. Liu, D. C. Croteau-Chonka, K. L. Mohlke, and Y. Li. 2013b. A comprehensive SNP and indel imputability database. Bioinformatics 29:528–531.

Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics 38:1378–1385.

Fan, R., Y. Wang, J. L. Mills, A. F. Wilson, J. E. Bailey-Wilson, and M. Xiong. 2013. Functional linear models for association analysis of quantitative traits. Genetic epidemiology 37:726–42.

Ferrari, S., and F. Cribari-Neto. 2004. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics.

Fu, W., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Fuchsberger, C., B. Howie, M. Laakso, M. Boehnke, and A. GR. 2012. The value of population-specific reference panels for genotype imputation in the age of whole-genome sequencing. Presented at the 62nd Annual Meeting of The American Society of Human Genetics.

Futema, M., V. Plagnol, R. A. Whittall, H. A. W. Neil, and S. E. Humphries. 2012. Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. Journal of Medical Genetics.

Gao, X., P. Marjoram, R. Mckean-Cowdin, M. Torres, W. J. Gauderman, and R. Varma. 2012. Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels. Frontiers in Genetics.

Goldsmith, J., J. Feder, B. Caffo, D. Reich, and C. M. Crainiceanu. 2010. PENALIZED FUNCTIONAL REGRESSION Penalized Functional Regression.

Gonzalo, S. 2010. Epigenetic alterations in aging. Journal of applied physiology (Bethesda, Md. : 1985) 109:586–597.

Grove, M. L., B. Yu, B. J. Cochran, T. Haritunians, J. C. Bis, K. D. Taylor, M. Hansen, et al. 2013. Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. PLoS ONE 8.

Harris, R. A., T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, et al. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature biotechnology 28:1097–1105.

Heyn, H., and M. Esteller. 2012. DNA methylation profiling in the clinic: applications and challenges. Nature Reviews Genetics.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5.

Irizarry, R. A., C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, et al. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature genetics 41:178–186.

Jiao, S., L. Hsu, C. M. Hutter, and U. Peters. 2011. The use of imputed values in the meta-analysis of genome-wide association studies. Genetic Epidemiology 35:597–605.

Kang, J., K.-C. Huang, Z. Xu, Y. Wang, G. R. Abecasis, and Y. Li. 2013. AbCD: arbitrary coverage design for sequencing-based genetic studies. Bioinformatics (Oxford, England) 29:799–801.

Kutalik, Z., T. Johnson, M. Bochud, V. Mooser, P. Vollenweider, G. Waeber, D. Waterworth, et al. 2011. Methods for testing association between uncertain genotypes and quantitative traits. Biostatistics 12:1–17.

Laird, P. W. 2010. Principles and challenges of genomewide DNA methylation analysis. Nature reviews. Genetics 11:191–203.

Lander, E. S., and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–99.

Lange, L. a, D. C. Croteau-Chonka, A. F. Marvelle, L. Qin, K. J. Gaulton, C. W. Kuzawa, T. W. McDade, et al. 2010. Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. Human molecular genetics 19:2050–2058.

Laurent, L., E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, et al. 2010. Dynamic changes in the human methylome during differentiation. Genome Research 20:320–331.

Lee, S., M. C. Wu, and X. Lin. 2012. Optimal tests for rare variant effects in sequencing association studies. Biostatistics (Oxford, England) 13:762–75.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology 34:816–834.

Li, Y., C. Willer, S. Sanna, and G. Abecasis. 2009. Genotype imputation. Annual review of genomics and human genetics 10:387–406.

Lin, D. Y., Y. Hu, and B. E. Huang. 2008. Simple and Efficient Analysis of Disease Association with Missing Genotype Data. American Journal of Human Genetics 82:444–452.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322.

Liu, E. Y., S. Buyske, A. K. Aragaki, U. Peters, E. Boerwinkle, C. Carlson, C. Carty, et al. 2012. Genotype Imputation of Metabochip SNPs Using a Study-Specific Reference Panel of ~4,000 haplotypes in African Americans From the women's health initiative. Genetic Epidemiology 36:107–117.

Liu, E. Y., M. Li, W. Wang, and Y. Li. 2013a. MaCH-Admix: Genotype Imputation for Admixed Populations. Genetic Epidemiology 37:25–37.

Liu, K., A. Luedtke, and N. Tintle. 2013b. Optimal methods for using posterior probabilities in association testing. Human heredity 75:2–11.

Ma, B., E. H. Wilker, S. a. G. Willis-Owen, H.-M. Byun, K. C. C. Wong, V. Motta, a. a. Baccarelli, et al. 2014. Predicting DNA methylation level across human tissues. Nucleic Acids Research 1–14.

Marchini, J. 2013. A haplotype map derived from whole genome low-coverage sequencing of over 25,000 individuals. Presented at the 63rd Annual Meeting of The American Society of Human Genetics.

Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nature reviews. Genetics 11:499–511.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 39:906–913.

Morris, A., and E. Zeggini. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology 34:188–193.

Nelder, J. A., and R. Mead. 1965. A Simplex Method for Function Minimization. The Computer Journal 7:308–313.

Network, C. genome A. R. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43–9.

Pei, Y. F., L. Zhang, J. Li, and H. W. Deng. 2010. Analyses and comparison of imputation-based association methods. PLoS ONE 5.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81:559–575.

Rakyan, V. K., T. a Down, D. J. Balding, and S. Beck. 2011. Epigenome-wide association studies for common human diseases. Nature reviews. Genetics 12:529–41.

Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447:425–432.

Smith, Z. D., and A. Meissner. 2013. DNA methylation: roles in mammalian development. Nature reviews. Genetics 14:204–20.

Varley, K. E., J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, et al. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Research 23:555–567.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89:82–93.

Wu, Y., Y. Li, E. M. Lange, D. C. Croteau-Chonka, C. W. Kuzawa, T. W. McDade, L. Qin, et al. 2010. Genome-wide association study for adiponectin levels in Filipino women identifies CDH13 and a novel uncommon haplotype at KNG1-ADIPOQ. Human molecular genetics 19:4955–4964.

Zhang, P., X. Zhan, N. a Rosenberg, and S. Zöllner. 2013. Genotype imputation reference panel selection using maximal phylogenetic diversity. Genetics 195:319–30.

Zheng, H., H. Wu, J. Li, and S.-W. Jiang. 2013. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. BMC medical genomics 6 Suppl 1:S13.

Zheng, J., Y. Li, G. R. Abecasis, and P. Scheet. 2011. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genetic epidemiology 35:102–10.

Zhuang, J., M. Widschwendter, and A. E. Teschendorff. 2012. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC bioinformatics 13:59.

Ziller, M. J., H. Gu, F. Müller, J. Donaghey, L. T.-Y. Tsai, O. Kohlbacher, P. L. De Jager, et al. 2013. Charting a dynamic DNA methylation landscape of the human genome. Nature 500:477–81.