Global Analysis of Methylation Profiles from High Resolution CpG Data

Ni Zhao[1], Douglas A. Bell[2], Arnab Maity[3], Ana-Maria Staicu[3], Bonnie R. Joubert[4], Stephanie J. London[4], Michael C. Wu[1, *]

[1]Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109
[2] Environmental Genomics Group, Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709
[3]Department of Statistics, North Carolina State University, Raleigh, NC 27695
[4]Epidemiology Branch/Genetics, Environment & Respiratory Disease Group, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709

*Address for Correspondence:
Michael C. Wu
Biostatistics and Biomathematics Program
Public Health Sciences Division
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North, M3-C102
P.O. Box 19024
Email: mcwu@fhcrc.org

Abstract

New high throughput technologies are now enabling simultaneous epigenetic profiling of DNA methylation at hundreds of thousands of CpGs across the genome. A problem of considerable practical interest is identification of large scale, global changes in methylation that are associated with environmental variables, clinical outcomes, or other experimental conditions. However, there has been little statistical research on methods for global methylation analysis using technologies with individual CpG resolution. To address this critical gap in the literature, we develop a new strategy for global analysis of methylation profiles using a functional regression approach wherein we approximate either the density or the cumulative distribution function (CDF) of the methylation values for each individual using B-spline basis functions. The spline coefficients for each individual are allowed to summarize the individual's overall methylation profile. We then test for association between the overall distribution and a continuous or dichotomous outcome variable using a variance component score test which naturally accommodates the correlation between spline coefficients. Simulations indicate that our proposed approach has desirable power while protecting type I error. The method was applied to detect methylation differences, both genome wide and at LINE1 elements, between the blood samples from rheumatoid arthritis patients and healthy controls and to detect the epigenetic changes of human hepatocarcinogenesis in the context of alcohol abuse and hepatitis C virus infection. A free implementation of our methods in the R language is available in the Global Analysis of Methylation Profiles (GAMP) package at `http://research.fhcrc.org/wu/en.html`.

Key Words: Density approximation; Epigenome wide association study; Global testing; Spline smoothing; Variance component testing.

Conflict of Interest Statement: The authors declear no conflict of interest.

1 Introduction

Recent advances in high-throughput biotechnology have culminated in the development of large scale epigenome wide association studies (EWAS) in which the DNA methylation at hundreds of thousands of CpGs along the genome can be simultaneously measured across a large number of samples [Rakyan et al., 2011, Bibikova et al., 2011, Sandoval et al., 2011, Rakyan et al., 2011]. EWAS have resulted in the identification of differentially methylated CpGs associated with differences in disease states, clinical outcomes, environmental exposures, or other experimental conditions [Joubert et al., 2012, Shen et al., 2013, Heyn et al., 2013, 2012]. These discoveries can provide a breadth of information from fundamental insights into the mechanisms underlying complex disease and to potential biomarkers for diagnosis or prognosis [Laird et al., 2003, Attar, 2012]. Despite many successes, analysis of EWAS remains challenging. In addition to open questions concerning preprocessing and normalization [Dedeurwaerder et al., 2011, Teschendorff et al., 2013], association analysis with outcome variables is also difficult. Standard analysis proceeds via individual CpG analysis wherein the association between each CpG and an outcome variable (e.g. disease state, environmental exposure, etc.) is assessed one-by-one. After computing a p-value for each CpG, multiple testing criteria such as the false discovery rate (FDR) or Bonferroni corrections are applied. CpGs surviving this correction are called differentially methylated and followed for validation and interpretation. Recently, alternative approaches based on pathway analysis have also been applied and largely mimic the analyses conducted for gene expression data.

Although individual CpG analysis has been extremely successful in identifying individual CpG sites associated with a variety of outcomes, a question of considerable interest lies in whether there is global differential methylation across the entire epigenome [Eng et al., 2000]. For example, global hypomethylation is believed to occur in cancer [Brothman et al., 2005, Kim et al., 2006] and has been reported in age-related frailty [Bellizzi et al., 2012].

The traditional analytical chemistry-based approaches (e.g. high performance liquid chromatography and mass spectrometry) for evaluating global methylation profiles are highly quantitative and reproducible [Song et al., 2005, Beck and Rakyan, 2008] but require large amounts of DNA and provide no information on the distribution of methylation across the genome.

A popular surrogate approach to evaluating global methylation uses polymerase chain reaction (PCR)-based assays that measure methylation of repetitive Alu elements and long interspersed nucleotide elements (LINE), which sample CpG methylation at hundreds of thousands of repeat regions located across the genome. Such methods have been used for a wide range of diseases and experimental conditions [Chalitchagorn et al., 2004, Bollati et al., 2007, Sharma et al., 2008, Figueiredo et al., 2009], and correlate well with analytical chemistry methods [Lisanti et al., 2013]. Changes in methylation at repeat elements may have a significant biological meaning because repeat element hypomethylation, particularly in tumors, may be associated with retrotransposon reactivation and subsequent genome instability [Yang et al., 2004]. However, these technologies are limited only to repeat regions. New technologies, such as high-resolution methylation microarrays, can provide genome-wide methylation profile with more thorough coverage and higher resolution than repeat element based methods.

Three important methylation platforms used in more and more EWAS studies include: 1) Infinium HumanMethylation450 BeadChip from Illumina which has a whole genome coverage of 485,000 CpG, near 99% of RefSeq genes, 96% of CpG islands, and other expert selected content [Dedeurwaerder et al., 2011]; 2) Reduced representation bisulfite sequencing (RRBS) which provides methylation data on 5-10 million CpG dispersed around the genome [Meissner et al., 2005]; 3) Whole genome bisulfite sequencing (WGBS) which provides methylation data at all mappable CpGs in the genome. The increasing availability of such array or sequencing based approaches calls for statistical methods to detect methylation differences across the entire genome or a large subset of methylation markers, such as markers that are within repeat regions, or are restricted to specific genomic context such as Cpg islands (CGIs), CGI shore, non-CGI regions or within a biological pathway.

In this paper, we develop two new, related methodologies for the global analysis of methylation profiles (GAMP), either across the epigenome or restricted to a large number of CpGs. The intuition behind our approach is that global methylation differences may be observable through differences in the overall distribution of CpG methylation levels, yet changes in a select, small subset of CpGs (which fails to reflect "global" methylation differences) will not dramatically change the entire distribution. Consequently, for our first strategy, we approximate the density of the methylation distribution for each individual using B-Spline basis functions [Ramsay and Silverman, 2005]. For our second strategy, we approximate the cumulative distribution function (CDF) of the methylation distribution for each individual using B-spline basis functions. Then for both approaches, we summarize the entire distribution of methylation values using the estimated B-spline coefficients. To test for differential global methylation, we employ a variance component test [Lin, 1997] previously used for regression based analysis of gene expression [Goeman et al., 2004, Liu et al., 2007, 2008] and genetic variants [Kwee et al., 2008, Wu et al., 2010, 2011]. In this paper, we mainly focus on methylation data obtained through Illumina HumanMethylation450 microarray, but the method is directly appliable to other high-throughput methylation platforms.

Our approach offers a number of attractive features. First, since we are using a more robust summary measure rather than the original CpGs, the approach is therefore targeted towards comprehensive, modest changes in methylation globally. Furthermore it is robust to very strong differential methylation in a few CpGs of interest – while interesting this scenario may not reflect true global differential methylation. Second, we will employ a computationally fast variance component test from the kernel machine framework which accommodates the high degree of correlation between spline coefficients while allowing for covariate adjustment. Finally, our variance component testing approach can be used for a range of outcome types including continuous, dichotomous, survival [Cai et al., 2011, Lin et al., 2011], and multivariate [Maity et al., 2012] outcomes while adjusting for covariates. The ability to adjust for covariates and confounders is an important feature given recent concerns regarding the need for controlling cell type effects [Houseman et al., 2012].

2 Methods

The idea behind our approach is that large scale, global differences in methylation will be reflected in differences between individuals in the distribution of their CpG methylation measurements. Thus, our general approach is to approximate the distribution of methylation values for each individual and then test for an association between the distributions and an outcome variable. In this method section, we focus first on approximating each individual's methylation profile using either the density or the cumulative density function (CDF). Subsequently we will describe the hypothesis testing procedure using the variance component test in Section 2.2.

2.1 Estimation of the Methylation Distribution

2.1.1 Estimation of the Density for Each Sample

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

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

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

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

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

Briefly, B-splines are a sequence of joined polynomial segments between a series of knots which are used to model functional data. Between each pair of knots the curves are modeled as a polynomial of some order greater than 1. For a pre-specified number of interior knots R and order L of the polynomials, the total number of B-spline basis functions is given by

$p = R + L$. We model the true methylation profiles $H_i(\cdot)$ by

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

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

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

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

Two important issues in this context are the number and placement of the knots, and the choice of the penalty parameter λ. Since methylation percentages are between 0 and 1 and approximately bimodal, we place more knots in the areas with strong curvature (closer to 0 and 1) and fewer knots in between. In general, we observed that 25-35 knots with polynomial order 4 seems to be a reasonable model for the data. Regarding the choice of λ, there are many available data based methods such as leave-one-out cross-validation, generalized cross-validation and the restricted maximum likelihood criteria, see e.g., [Ruppert and Carroll, 2003]. In this article we use generalized cross-validation (GCV) method to select λ.

Although $\widehat{H}_i(\cdot)$ can be thought of as an approximation of the density for the methylation values, strictly speaking, adjustments are needed to ensure that it has the properties of

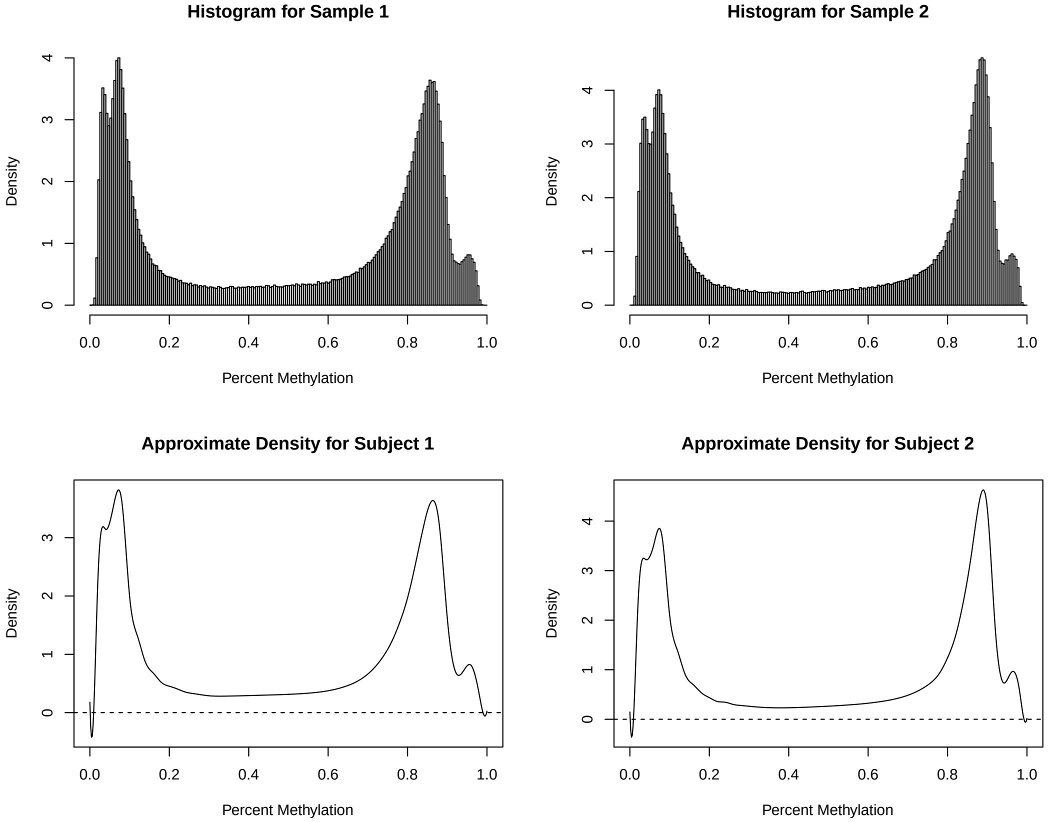

Figure 1: Example histograms for two samples and their corresponding B-spline approximated densities.

being a probability density function. However, since we are simply using the profile of the histogram as a tool for summarizing the entire profile of methylation values, this is not necessary from the perspective of testing.

2.1.2 Estimation of the CDF for Each Sample

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

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

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

where $\{t_k, k = 1, \ldots, B\}$ form an equally spaced grid of B points in $[0, 1]$. In constructing a basis for the CDF, we again use a grid of 35 knots between 0 and 1 due to the nature of methylation data, but in contrast to modeling the density function, we space the knots evenly since the difference in curvature is no longer as apparent. Because the CDF is smoother than the histogram, we also considered a less dense knot placement scheme in which 15 or 25 knots were used to construct a basis for the CDF.

We again assume a B-Spline basis representation for the true CDF with order 4 basis functions and write

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

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

2.2 Variance Component Testing for Differences in Approximated Distributions

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

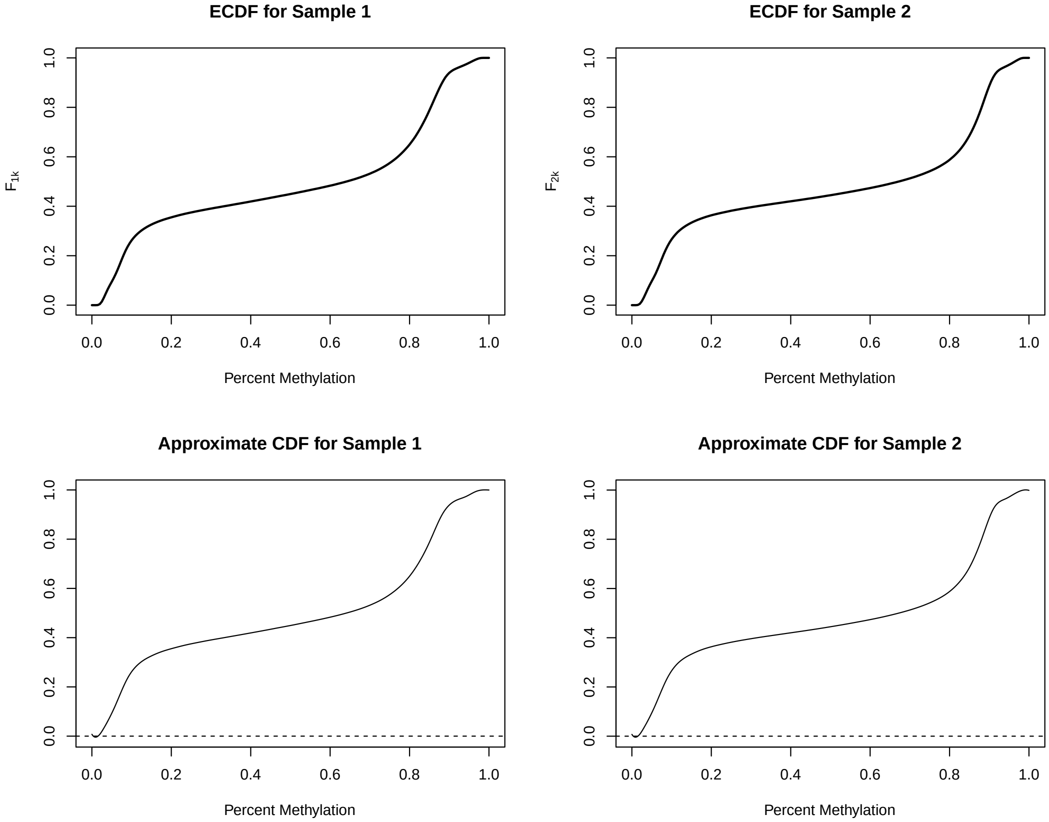

Figure 2: Example ECDFs for two samples and the approximated B-spline approximations.

et al., 2008, Wu et al., 2010, 2011].

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

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

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

for continuous outcomes and the logistic model

$$\mathrm{logit}[P(y_i = 1)] = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \sum_{j=1}^{p} \beta_j c_{ij} \tag{2}$$

for dichotomous outcomes, where we define α_0 to be an intercept, $\boldsymbol{\alpha}$ and β_j to be the regression coefficients corresponding to additional covariates X and each B-spline coefficient, and ε_i to be a random error with mean 0 and variance σ^2. To test for an association between $\mathbf{C}$ and $\mathbf{y}$ corresponds to testing:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0. \tag{3}$$

In principle, this can be done using a p degree of freedom test, but the $\mathbf{C}$ tend to be highly correlated and p can be large such that power is low. An alternative approach is to assume that the β_j follow some arbitrary distribution $G(\cdot)$ with mean 0 and variance τ. Then τ indexes the significance of the entire group of B-spline coefficients and then testing (3) is equivalent to testing

$$H_0 : \tau = 0, \tag{4}$$

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

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

where $\widehat{\alpha}_0$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\sigma}$ are estimated under (4). Similarly, for dichotomous outcomes we can construct the score statistic

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

where $\widehat{\mathbf{y}} = \text{logit}^{-1}(\widehat{\alpha}_0 + \mathbf{Z}\widehat{\boldsymbol{\alpha}})$ and both $\widehat{\alpha}_0$ and $\widehat{\boldsymbol{\alpha}}$ are again estimated under the null, (4).

Under the null hypothesis, Q asymptotically follows a mixture of chi-squares distributions. In particular, $Q \sim \sum \lambda_\ell \chi_1^2$ where λ_ℓ are the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{C}\mathbf{C}'\mathbf{P}_0^{1/2}$ and $\mathbf{P}_0 = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for continuous outcomes and $\mathbf{P}_0 = \mathbf{D} - \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$ for dichotomous outcomes with $\mathbf{D} = \text{diag}\{\widehat{y}_i(1 - \widehat{y}_i)\}$. This distribution can be approximated use moment matching methods[Liu et al., 2007, 2009] or exact approaches[Davies, 1973, 1980, Duchesne and Lafaye De Micheaux, 2010] allowing for easy p-value computation.

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

2.3 Simulations

Simulations were conducted to evaluate the type I error and power of the proposed global methylation profile analysis approach, in comparison to other traditional approaches for detecting distributional differences or mean shift between two distributions. Two simulation scenarios were considered: 1) global methylation profile is associated with the outcome, without any additional covariates, 2) additional covariates are associated with both the methylation profile and the outcome. Results are presented for the simulations with a dichotomous outcome, such as case-control status. Notice that in the second simulation scenario, the additional covariates qualify as confounders.

2.3.1 Simulations without additional covariates

We first conducted simulations under simple situation when the case-control status depends only on the methylation profile but no additional covariates. We simulated methylation profiles for $N = 40, 60, 100$ and 500 individuals with half as cases and half as controls.

Although the score test used in the proposed global methylation profile analysis allows for rapid analytical p-value calculation, approximating the methylation density or CDF using B-splines is still relatively time-consuming when the number of CpGs is large. This makes simulations difficult under scenarios in which we require large numbers of simualtions, e.g. in assessing type I error. Specifically, it takes approximately two days to simulate and conduct the analysis for 5000 data sets with $N = 500$ and 485,000 CpGs (any single data set would take less than a minute to run). Therefore, we conducted simulations using only 10,000 CpG markers for each individual.

To mimic real methylation data, we employed the inverse logit transformation of a multivariate normal distribution. Specifically, for the control group, the methylation values for all CpG markers were simulated as $M_0 = \text{logit}^{-1}(t_0)$ where logit^{-1} is the inverse logit function $\text{logit}^{-1}(t_0) = \exp(t_0)/(1 + \exp(t_0))$. $t_0 \sim N(0, 3)$ where the standard deviation 3 is to ensure that the methylation profiles have enriched "0"(unmethylated) and "1"(methylated) values as in real data. Methylation for cases were simulated as a mixture of two distributions: a proportion of $1 - p$ of the CpG markers come from the same distribution M_0 as in the control group and the remaining p proportion of CpG markers have methylation values from a different distribution. We considered two different case scenarios:

- Case scenario 1: $M_1 = \text{logit}^{-1}(t_1), t_1 \sim (1 - p) * N(0, 3) + p * N(0.3, 3)$
- Case scenario 2: $M_2 = \text{logit}^{-1}(t_2), t_2 \sim (1 - p) * N(0, 3) + p * N(0, 3.15)$

We focus on these two case scenarios because in the first scenario the methylation profiles from the cases and the controls have different means while in the second scenario the global methylation in the two groups have similar means but different variance. When $p = 0$, the methylation profiles from the cases are the same as from the controls and thus type I error can be evaluated. By changing the values of p, we can evaluate the power under different association strength.

For each of the two simulation scenarios, we applied both the density and CDF based global analysis approaches to evaluate the association between case-control status and methylation profiles. Specifically, we used B-splines to approximate the density or the CDF of each individual's methylation distribution. For the density estimation, we constructed histograms using 200 evenly spaced bins between 0 and 1. The knots for the B-spline were spaced at intervals of 0.02 between 0 and 0.3 and between 0.7 and 1. Since the region in the center is less variable, knots were placed at intervals of 0.1 in length. This generated totally 35 knots between 0 and 1 for density estimation. For the CDF estimation, we estimated the ECDF at 1000 evenly spaced points between 0 and 1. 35 knots for B-spline estimation were spaced between 0 and 1 evenly. Because the CDF is smoother than the density function, we also considered using fewer knots (15 and 25) for the B-spline estimation. After estimating the spline coefficients for approximating the density and the CDF, we applied the variance component score test to evaluate the association between the spline coefficients and the outcome variable. For each sample size, 2000 simulations were conducted for type I error that the mixture proportion $p = 0$ and 500 simulations were conducted for power evaluation with $p = 0.125, 0.25, 0.375$ and 0.5.

We compared the power of the two proposed global analyses to traditional approaches for overall methylation differences: t-test which captures the mean shift in two distributions and Wilcoxon rank sum test which tests whether two distributions have the same mean ranks. The Kolmogorov-Smirnov (K-S) test is another commonly used test that quantifies distributional differences. However, the Kolmogorov-Smirnov test considers each CpG marker as a sampling unit and its naive application is not valid [Goeman and Bühlmann, 2007]. In particular, the sample size for the K-S test would be 485,000 rather than the number of samples. This implies that an inference would generalize to the population of CpGs rather than the population of samples or individuals.

2.3.2 Simulation with additional covariates

Additional simulations were conducted to evaluate the performance of our proposed tests in situations when there are additional covariates, especially confounders. Specifically, we simulated two groups of methylation values and the dichotomous outcome was constructed based on the methylation group label and additional covariates. Again,10,000 CpG markers were considered for each individual. In the first group of methylation profile, each CpG methylation value was simulated the same way as in the control group in section 2.3.1: e.g., $M_0 = \text{logit}^{-1}(t_0), t_0 \sim N(0, 3)$. The methylation values from the second group was simulated as a mixture of two distributions with half of the markers simulated the same way as M_0 and half the markers simulated as logit^{-1} transformation of another distribution $N(0.3, 3)$. Additional covariate X was simulated as $N(0, 0.5)$ for the first methylation group and $N(0.5, 0.5)$ for the second methylation group. Notice that the distribution of X is related to the methylation group label, and thus X would be a potential confounder if it is also associated with the simulated dichotomous outcome.

- $M_0 = \text{logit}^{-1}(t_0), t_0 \sim N(0, 3); X \sim N(0, 0.5)$
- $M_1 = \text{logit}^{-1}(t_1), t_1 \sim 0.5N(0, 3) + 0.5N(0.3, 3); X \sim N(0.5, 0.5)$

The dichotomous outcome, i.e, the case/control status, was simulated according to success probability π that is based on both the methylation group label and additional covariate X: $\pi = \text{logit}^{-1}[b * I(M \in M_1) + X]$. Type I error can be evaluated when $b = 0$ that the methylation profile is unrelated to the outcome. Power of the proposed tests can be evaluated by changing the values of $b = 1, 2, 3, 4$ and 5 with different sample sizes $N = 40, 60, 100$, and 500.

2.4 Data Applications

We illustrate our proposed methods for global analysis of methylation profiles via application to two real data sets. We applied the proposed density and CDF based approaches to evaluate the global methylation differences across the entire epigenome. To facilitate interpretation, we also considered restricting our analysis to analyze the CpGs falling within relevant genomic features, such as the CpG islands (CGI), CGI shores, CGI shelfs, non-CGI and gene bodies. Additionally, we also carried out a restricted analysis of the CpG sites that were located within repeated LINE1 elements, which have been examined in many population studies and are presumed to reflect global methylation levels. With an estimated 500,000 copies per genome, the LINE1 elements make up about 17% of the human genome [Rodić and Burns, 2013]. Although the percentage of active LINE1 elements is unclear, LINE1 elements are believed to be responsible for most reverse transcription in the genome, including retrotransposition of Alu elements [Okada et al., 1997] and the creation of processed pseudogenes [Wei et al., 2001]. Thus, LINE1 elements are very important for genomic stability.

2.4.1 Methylation Study on Rheumatoid Arthritis and Healthy Controls

Methylation changes are believed to play a key role in rheumatoid arthritis and contribute to underlying inflammation and joint damage [Nakano et al., 2013]. Recently, a study [Liu et al., 2013] examined methylation differences between arthritis patients and healthy controls. The Illumina HumanMethylation450 array was used to measure methylation levels at approximately 485,000 CpGs genome wide in blood from 354 arthritis patients and 335 healthy controls. We obtained the data from the Gene Expression Omnibus (GEO) [Edgar et al., 2002] (`http://www.ncbi.nlm.nih.gov/geo/`) under accession number GSE42861. Both the density and CDF based approaches were applied to test for associations between methylation profiles and disease status. For our analysis, we restricted attention to the approximately 470,000 autosomal CpGs, of which approximately 75,000 CpGs are mapped to LINE1 elements.

We first applied both the density and CDF based analysis procedures to the data set to test for global differential methylation between rheumatoid arthritis and controls over all autosomal CpGs with adjustment for sample specific cell mixture proportions, which were estimated via Houseman's algorithm [Houseman et al., 2012] using reference information on cell-specific methylation signatures for the major cell types in blood. For the density based approach, we computed the relative histogram using 200 evenly spaced breaks between 0 and 1, and then we approximated the density using B-splines with knots placed at intervals of 0.02 between 0 and 0.3 and between 0.7 and 1.0. Between 0.3 and 0.7, knots were evenly spaced at intervals of 0.10. For the CDF based approach, we computed the ECDF for each sample at a grid of 1000 values between 0 and 1 and then approximated the CDF using B-splines with knots placed at 35 evenly spaced intervals between 0 and 1. For both approaches, GCV was used to estimate the B-spline smoothing parameter λ. The approximate densities and CDFs are shown in Figure 5. We then used the variance component test under a logistic model to regress a binary indicator for whether each subject was an arthritis patient on the B-spline coefficients from the approximate density or from the approximate CDF. The same approaches were used to test for differential methylation profiles between the rheumatoid arthritis and healthy controls for CpG markers that are mapped to repeat LINE1 elements and different genomic features, such as CGI, CGI shore, CGI shelf, non-CGI and gene body. Additionally, we also restricted our analysis to the 1498 CpG markers that are in the rheumatoid arthritis pathway (KEGG 05323).

2.4.2 Epigenetic Changes of Alcohol Abuse and Hepatitis Infection during Human Hepatocarcinogenesis

Hepatocellular carcinoma (HCC) constitutes $\sim 80\%$ of liver cancers and is the second most common cause of cancer deaths worldwide. A recent study [Hlady et al., 2014] used the Illumina HumanMethylation450 array to measure the DNA methylation level in HCC samples and healthy controls under exposure to environmental (broadly defined) agents and and lifestyle variables, including alcohol abuse and human hepatitis C virus (HCV) infection. To illustrate our approaches, we downloaded the data from GEO (accession number GSE60753) and restricted subsequent analyses to surgically resected normal and HCC human samples, excluding samples from metastatic tumors, cultured cells, cell lines or non-cancer cirrhotic samples. We then applied our proposed methods to explore possible global methylation differences between the 34 healthy liver samples and 32 surgically resected HCC samples. We examined global methylation differences among the HCC tissue samples arising from HCV infection (12 of 32 samples) and alcohol abuse (15 of 32 samples).

Both the density and CDF based approaches were applied to detect global methylation differences between HCC patients and healthy controls and between the HCC patients in the setting of alcohol abuse and the HCC patients with HCV infections over all autosomal CpGs. There was no significant difference between the different HCC groups with respect to gender and TNM stage (table 1 in Hlady et al. [2014]) and therefore, these features were not controlled for in our association test. The density and CDF estimations were conducted the same way as in the analysis of the rheumatoid arthritis data (Section 2.4.1). GCV was used to estimate the B-spline coefficients. As in the previous application, we also applied our tests to CpG markers that are within the LINE1 elements, and different genomic context such as CGI, CGI shore, CGI shelf or gene body.

Table i: Empirical type I error at $\alpha = 0.05$ level under different simulation scenarios.

No Covariate						
n	Density	CDF	CDF$_{25}$	CDF$_{15}$	t-test	Rank sum
40	0.054	0.057	0.055	0.054	0.050	0.050
60	0.054	0.051	0.052	0.051	0.046	0.044
100	0.050	0.056	0.056	0.054	0.051	0.051
500	0.050	0.053	0.053	0.051	0.056	0.053
With Covariate						
n	Density	CDF	CDF$_{25}$	CDF$_{15}$	t-test	Rank sum
40	0.048	0.056	0.057	0.055	0.265	0.223
60	0.051	0.057	0.057	0.055	0.360	0.318
100	0.044	0.052	0.052	0.054	0.556	0.501
500	0.043	0.043	0.043	0.044	0.997	0.995

3 Results

3.1 Type I Error Simulation

The type I error for all simulations are presented in Table i. When there is no additional covariates (section 2.3.1), all methods, including our proposed density and CDF based analysis approaches, the t-test and the Wilcoxon rank sum test can correctly control type I error at $\alpha = 0.05$, even when the sample size is modest. However, when there is an additional covariate (section 2.3.2), the t-test and rank sum test had seriously inflated type I error because they failed to adjust for the potential confounding effect from X. The proposed global methylation profile analyses still have valid type I error.

3.2 Power Simulation

The power result for the simulations when there is no additional covariate is presented in Figure 3. We considered two simulation scenarios where in the first scenario the cases and controls have different mean methylation values and in the second simulation scenario, despite the distributional differences between the cases and controls, there is very limited change in the average methylation levels.

In the first simulation scenario (Figure 3: left panel), the proposed CDF based analysis approach had similar power as the t-test and the Wilcoxon rank sum test and reported higher power than the density based approach. The t-test and Wilcoxon rank sum test had high power because the major difference between cases and controls in this simulation scenario lies in the mean shift. The CDF based approach tended to yield higher power than the density based approach in this scenario, partly because the CDF counts the proportion of markers with methylation values below each threshold and can capture the mean shift better than the density. The CDF approach with 15 or 25 knots had almost identical power as using 35 knots, supporting the observation that CDF is smoother than the density and can be summarized using fewer knots. In the second simulation scenario (Figure 3: right panel) where the major difference between cases and controls lies in the variance , the density based test was the most powerful; the CDF based approach had lower but still adequate power. The t-test and rank sum test, which are designed to capture the central tendency of two distributions, reported power only at the type I error level. Density based approach is better in capturing the distributional differences without mean/median shifts.

Figure 4 summarized the power result for simulations when there are additional covariates. Only the results from the proposed tests were included in this figure as the t-test and the Wilcoxon rank sum test failed to control type I error. Consistent with the previous simulations with no additional covariates, the CDF based approach reported higher power than the density based approach because of its greater ability to detect mean shifts in methylation profiles.

3.3 Data Analysis Results

In illustrating our methods on the rheumatoid arthritis data set, we applied the proposed density and CDF based global methylation profile tests to evaluate whether the methylation profile is associated with arthritis after adjusting for estimated cell proportions. Across

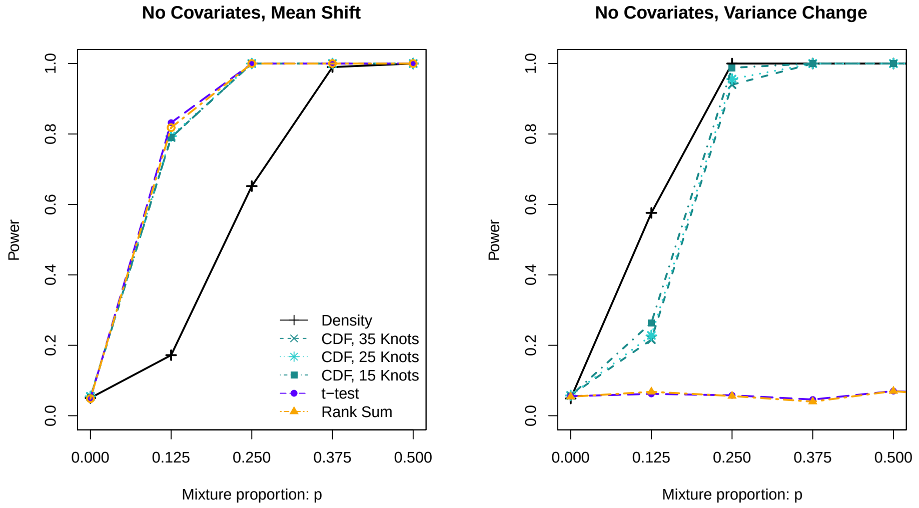

Figure 3: Simulated type I error and power for the proposed methylation profile test in situations without additional covariate in comparison with t-test and Wilcoxon rank sum test. Left panel: simulation scenario that the overall average mehtylation levels differ in cases and controls. Right panel: simulation scenario that the overall methylation in the cases and controls have different variances but similar means. Sample size $N = 100$.

the whole autosome, both the density and CDF based approaches yielded highly significant results with p-values 3.495×10^{-14} and 1.208×10^{-11} respectively. The significance remained when we restricted our analysis to the CpGs in the KEGG pathway KO05323 (Rheumatoid arthritis) with p-values from the density based method computed as 7.44×10^{-14} and from CDF based approach as 9.03×10^{-11}. Similarly significant results were obtained when the analysis was restricted to markers that are within the LINE1 elements or within different genomic features, such as CGIs, CGI shore, CGI shelf, non-CGI regions, and gene bodies (all p-values $< 10^{-7}$). This result suggests consistent methylation profile differences between rheumatoid arthritis and healthy controls across the whole epigenome.

We further illustrated our approach by assessing the global methylation difference between HCC patients and healthy controls in liver tissues. The methylation profile of the autosomal CpGs is significantly associated with the HCC disease status using both the density based approach (p-value = 0.040) and CDF based approach (p-value = 0.0088). The

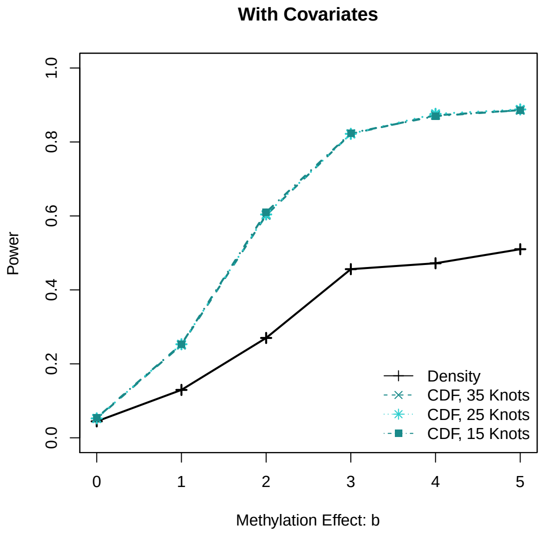

Figure 4: Simulated type I error and power for the proposed methylation profile test in situations with additional covariate in comparison with t-test and Wilcoxson rank sum test. Sample size $N = 100$.

significance was mainly due to the significant difference between the methylation profiles in the non-CGI regions (p-values = 0.0001 and 0.0002 for the density and CDF based approaches) and the CGI shelf regions (p-values = 0.0026 and 0.0030 for the density and CDF based approaches) of the HCC patients and healthy controls. The methylation profiles at the CGI and CGI shore regions were not significantly different. When we limited the analysis to the markers that are in the repeated LINE1 elements, the results remained significant (p-values $= 5.04 \times 10^{-5}$ and 0.0001 for density and CDF based approaches respectively). Additionally, we evaluated methylation profile differences between the alcohol abusive HCC patients and HCV infected HCC patients: there was no significant difference between the methylation profiles, either genome wide or limited to any of the tested genomic regions. Our results suggest that although large scale differences in the overall methylation distribution exist between HCC patients and healthy controls, the global methylation profiles were similar among the HCC patients with different disease etiologies.

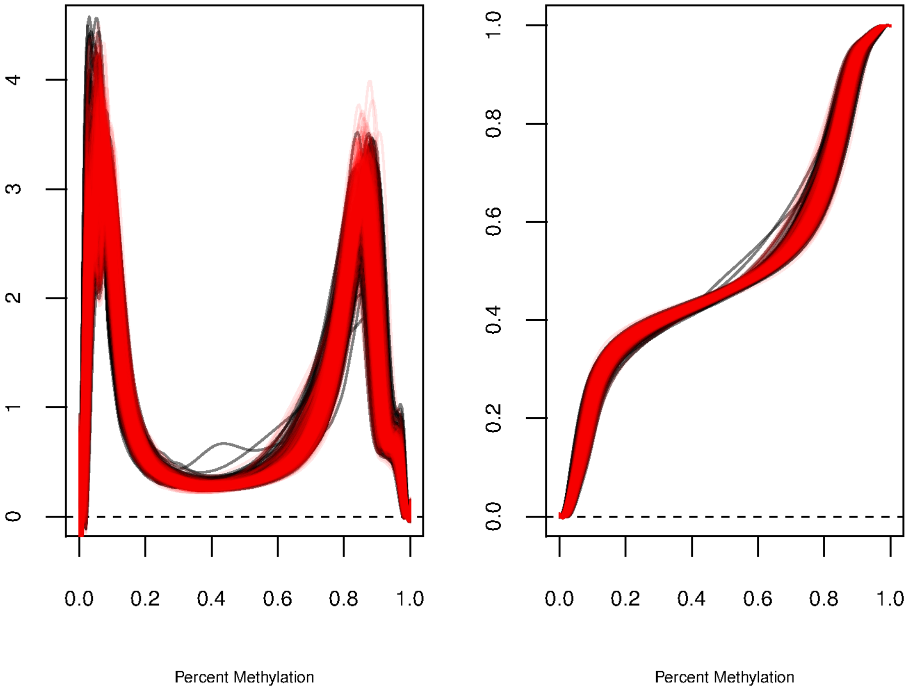

Figure 5: Approximate densities and CDFs from the rheumatoid arthritis study. Red curves are the methylation profiles from the rheumatoid arthritis patients and black curves are the methylation profiles from healthy controls .

4 Discussion

In this article, we propose two new strategies for global analysis of methylation profiles which are based on approximation of either the density or the CDF of the methylation values for each individual. Specifically, by indexing each individual's methylation distribution using B-spline basis coefficients, we summarize the methylation profile for each individual so that we can test for association between the overall methylation distribution and an outcome variable, while adjusting for additional covariates, by simply testing the spline coefficients. This functional approximation can comprehensively capture the distributional differences which are difficult to represent using a single or few statistics, such as mean or variance. For example, in contrast to t-test or Wilcoxon rank sum test, by using the B-spline coefficients, we can detect functional differences in methylation distributions, with or without mean changes. Although the proposed method tests the global null hypothesis, a key advantage of the proposed method is that we are essentially applying smoothing when we approximate the density or the CDF using B-splines. Therefore, this reduces the influence of single (or a few) probes strongly associated with the outcome.

Overall, of the two proposed methods, the CDF based approach tends to have higher power when the methylation levels have different global means while the density based approach tends to have higher power when the methylation profiles have functional differences other than a mean shift, such as in situations when we have different variances or when the methylation distribution comes from a mixture of several distributions. In the real data analysis that compares the global epigenetic changes between newborn babies and nonagenarians, the result that the CDF based approach obtained significant result while the density based approach failed to be significant is consistent with the previous observation that the cord blood from newborns tend to have higher global methylation level than the peripheral blood mononuclear from nonagenarians.

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

Our proposed methodology opens doors to new areas of research. First, we proposed ways to evaluate the global methylation profiling using data obtained through high throughput array or sequencing based technology. Compared with analytical chemistry-based or repeat element based approaches, the new technology provides data with individual CpG resolution and more thorough coverage. In this paper we consider data obtained from Illumina HumanMethylation450 platform; however, the same strategy can be used in sequencing based methylation profiling studies. Second, although we focus on testing global methylation across all CpGs, the approach can be restricted to specific subsets of CpGs such as CpGs falling within specific epigenetically relevant features (e.g. CpG islands, promoters, repeats, etc.) or the CpGs within a particular gene pathway thereby enabling a set or pathway based analysis that tests the global null hypothesis but is more geared towards a true pathway effect. However, caution needs to be taken as our approach is designed to detect global distributional differences and the density or CDF approximation may not be adequate when the number of CpGs is not large enough, e.g., when there are only 5 or 10 CpGs in the set of your interest. The method requries at least 50 CpGs.

Further, while we have explored the relationship between global methylation and a single dichotomous or continuous outcome, alternative outcome types are possible and warrant further exploration. Finally, while our work focuses on testing the overall methylation distributions, the idea of using a functional regression approach to summarize the overall distribution can also allow for understanding the relationship between outcome variables and other covariates while in the presence of global methylation differences, i.e. adjusting for the effect of methylation. This is important since methylation can serve as a potential confounder in biological models and failure to adjust for its effect may lead to biased or false conclusion. Such explorations remain for future research.

References

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

S. Beck and V. K. Rakyan. The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, 24(5):231–237, May 2008.

D. Bellizzi, P. D'Aquila, A. Montesanto, A. Corsonello, V. Mari, B. Mazzei, F. Lattanzio, and G. Passarino. Global DNA methylation in old subjects is correlated with frailty. *Age (Dordr)*, 34(1):169–179, Feb 2012.

Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

Valentina Bollati, Andrea Baccarelli, Lifang Hou, Matteo Bonzini, Silvia Fustinoni, Domenico Cavallo, Hyang-Min Byun, Jiayi Jiang, Barbara Marinelli, Angela C Pesatori, et al. Changes in dna methylation patterns in subjects exposed to low-dose benzene. *Cancer research*, 67(3):876–880, 2007.

Arthur R Brothman, Gregory Swanson, Teresa M Maxwell, Jiang Cui, Kelley J Murphy, Jennifer Herrick, VO Speights, Jorge Isaac, and L Ralph Rohr. Global hypomethylation is common in prostate cancer cells: a quantitative predictor for clinical outcome? *Cancer genetics and cytogenetics*, 156(1):31–36, 2005.

Tianxi Cai, Giulia Tonini, and Xihong Lin. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*, 67(3):975–986, 2011.

Krisanee Chalitchagorn, Shanop Shuangshoti, Nusara Hourpai, Narisorn Kongruttanachok, Pisit Tangkijvanich, Duangporn Thong-ngam, Narin Voravud, Virote Sriuranpong, and Apiwat Mutirangura. Distinctive pattern of line-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene*, 23(54):8841–8846, 2004.

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

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

Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélene Denis, Christos Sotiriou, and François Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, 2011.

Pierre Duchesne and Pierre Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

Charis Eng, James G Herman, and Stephen B Baylin. A bird's eye view of global methylation. *Nature Genetics*, 24(2):101–102, 2000.

Jane C Figueiredo, Maria V Grau, Kristin Wallace, A Joan Levine, Lanlan Shen, Randala Hamdan, Xinli Chen, Robert S Bresalier, Gail McKeown-Eyssen, Robert W Haile, et al. Global dna hypomethylation (line-1) in the normal colon and lifestyle characteristics and dietary and genetic factors. *Cancer Epidemiology Biomarkers & Prevention*, 18(4):1041–1049, 2009.

Jelle J. Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007. doi: 10.1093/bioinformatics/btm051.

Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

Holger Heyn, Ning Li, Humberto J Ferreira, Sebastian Moran, David G Pisano, Antonio Gomez, Javier Diez, Jose V Sanchez-Mut, Fernando Setien, F Javier Carmona, et al. Distinct dna methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*, 109(26):10522–10527, 2012.

Holger Heyn, F Javier Carmona, Antonio Gomez, Humberto J Ferreira, Jordana T Bell, Sergi Sayols, Kirsten Ward, Olafur A Stefansson, Sebastian Moran, Juan Sandoval, et al. Dna methylation profiling in breast cancer discordant identical twins identifies dok7 as novel epigenetic biomarker. *Carcinogenesis*, 34(1):102–108, 2013.

Ryan Hlady, Rochelle Tiedemann, William Puszyk, Ivan Zendejas, Lewis Roberts, Jeong-Hyeon Choi, Chen Liu, and Keith Robertson. Epigenetic signatures of alcohol abuse and hepatitis infection during human hepatocarcinogenesis. *Oncotarget*, 5(0), 2014.

Eugene A Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

Bonnie R Joubert, Siri E Håberg, Roy M Nilsen, Xuting Wang, Stein E Vollset, Susan K Murphy, Zhiqing Huang, Cathrine Hoyo, Øivind Midttun, Lea A Cupul-Uicab, et al. 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*, 120:1425–31, 2012.

Young-In Kim, Anna Giuliano, Kenneth D Hatch, Achim Schneider, Magdy A Nour, Gerard E Dallal, Jacob Selhub, and Joel B Mason. Global dna hypomethylation increases progressively in cervical dysplasia and carcinoma. *Cancer*, 74(3):893–899, 2006.

Lydia Coulter Kwee, Dawei Liu, Xihong Lin, Debashis Ghosh, and Michael P Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, 2008.

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

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

Xinyi Lin, Tianxi Cai, Michael C Wu, Qian Zhou, Geoffrey Liu, David C Christiani, and Xihong Lin. Kernel machine snp-set analysis for censored survival outcomes in genome-wide association studies. *Genetic epidemiology*, 35(7):620–631, 2011.

S. Lisanti, W. A. Omar, B. Tomaszewski, S. De Prins, G. Jacobs, G. Koppen, J. C. Mathers, and S. A. Langie. Comparison of methods for quantification of global DNA methylation in human cells and tissues. *PLoS ONE*, 8(11):e79044, 2013.

Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

Dawei Liu, Debashis Ghosh, and Xihong Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics*, 9(1):292, 2008.

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

Y. Liu, M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, K. Shchetynsky, A. Scheynius, J. Kere, L. Alfredsson, L. Klareskog, T. J. Ekstrom, and A. P. Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.*, 31(2):142–147, Feb 2013.

Arnab Maity, Patrick F Sullivan, and Jun-ing Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36:686–95, 2012.

A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*, 33(18):5868–5877, 2005.

K. Nakano, J. W. Whitaker, D. L. Boyle, W. Wang, and G. S. Firestein. DNA methylome signature in rheumatoid arthritis. *Ann. Rheum. Dis.*, 72(1):110–117, Jan 2013.

N Okada, M Hamada, I Ogiwara, and K Ohshima. Sines and lines share common 39 sequences: a review. *Gene*, 205:229–243, 1997.

Vardhman K Rakyan, Thomas A Down, David J Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8): 529–541, 2011.

James Ramsay and BW Silverman. *Functional data analysis*. Wiley Online Library, 2005.

N. Rodić and K. H. Burns. Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? *PLoS Genet.*, 9(3):e1003402, Mar 2013.

Wand Ruppert and RJ Carroll. *Semiparametric regression*. Cambridge University Press, 2003.

Juan Sandoval, Holger Heyn, Sebastian Moran, Jordi Serra-Musach, Miguel A Pujana, Marina Bibikova, and Manel Esteller. Validation of a dna methylation microarray for 450,000 cpg sites in the human genome. *Epigenetics*, 6(6):692–702, 2011.

Priyanka Sharma, Jitender Kumar, Gaurav Garg, Arun Kumar, Ashok Patowary, Ganesan Karthikeyan, Lakshmy Ramakrishnan, Vani Brahmachari, and Shantanu Sengupta. Detection of altered global dna methylation in coronary artery disease patients. *DNA and cell biology*, 27(7):357–365, 2008.

Jing Shen, Shuang Wang, Yu-Jing Zhang, Hui-Chen Wu, Muhammad G Kibriya, Farzana Jasmine, Habibul Ahsan, David PH Wu, Abby B Siegel, Helen Remotti, et al. Exploring genome-wide dna methylation profiles altered in hepatocellular carcinoma using infinium humanmethylation 450 beadchips. *Epigenetics*, 8(1):0–1, 2013.

L. Song, S. R. James, L. Kazim, and A. R. Karpf. Specific method for the determination of genomic DNA methylation by liquid chromatography-electrospray ionization tandem mass spectrometry. *Anal. Chem.*, 77(2):504–510, Jan 2005.

Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–196, 2013.

W. Wei, N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, J. D. Boeke, and J. V. Moran. Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.*, 21(4):1429–1439, Feb 2001.

Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

Allen S Yang, Marcos RH Estécio, Ketan Doshi, Yutaka Kondo, Eloiza H Tajara, and Jean-Pierre J Issa. A simple method for estimating global dna methylation using bisulfite pcr of repetitive dna elements. *Nucleic acids research*, 32(3):e38–e38, 2004.