Powerful Association Tests to Detect Disease-Related DNA Methylation Regions

Xiaoyu Cai[1], Lo-Bin Chang[1], and Chi Song*[2]

[1]Department of Statistics, The Ohio State University
[2]College of Public Health, Division of Biostatistics, The Ohio State University

Abstract

In epigenetics, researchers are often interested in detecting differential DNA methylation associated with phenotypes such as complex human diseases. We propose two powerful statistical tests $\mathrm{AF_b}$ and AF that can detect disease-related DNA methylation regions. Our methods are based on adaptive combination of marginal tests in generalized functional linear models, assuming continuity in the effects of the methylation sites in the pre-defined genomic regions (such as genes). Simulation studies based on real human genome properties show that our methods carry out high statistical power for various simulation models and signal proportions. Experiments on schizophrenia data show that the detected genes by our methods are remarkably consistent with previously reported genes related to schizophrenia. Many of the previous reports were based on evidence other than DNA methylation, which demonstrates the potential of our methods in detecting novel associated DNA methylation regions.

1 Introduction

Understanding the etiology of complex diseases is one of the major challenges for biomedical research. In the past decades, disease-associated single nucleotide variants (SNVs), the genetic determinants, have been extensively detected and studied by numerous genome-wide association studies (GWASs). Having realized that the genetic components may only explain a small proportion of complex human diseases, researchers are now increasingly interested in seeking disease-related epigenetic components, including DNA methylation and histone modification. DNA methylation is a biological process during which methyl groups are added to cytosines at the sites of CpG dinucleotides, which presumably leads to repression of gene transcription when located near a gene promoter. Evidently, proper DNA methylation plays a critical role in regulating gene expression for cell differentiation and embryonic development, which is to some degree heritable across generations [1]. As shown in a large amount of research, abnormal DNA methylation could give rise to severe adverse consequences, including human diseases such as cancer [2], muscular dystrophy, ICF syndrome, and immunological defects [3], as well as birth defects [4].

During recent years, technology advancements have promoted the development of epigenome-wide association studies (EWASs). Methylation profiles with better resolution reveal that the

*Correspondence to: Chi Song, College of Public Health, Division of Biostatistics, The Ohio State University, 1841 Neil Ave., 208E Cunz Hall, Columbus, OH 43210. E-mail: song.1188@osu.edu

relationship among methylation, gene expression, and human diseases is more complicated than that was anticipated in the beginning. The functions of methylation vary with context [5]. Different methylation regions of a gene (e.g. exonic and intronic portions of gene body) may have different effects on the gene expression and disease development, and to large extend this regulation mechanism is unknown [6, 7]. There is an urgent need for statistical methods that can identify disease-associated gene methylation profiles while treating the effects of different methylation sites differently within each gene.

Commonly used methylome profiling technologies can be coarsely categorized as capture-based and bisulfite (BS) conversion based. Capture-based technologies rely on the pulldown of methylated DNAs. Fragments with any methylated CpGs are pulled down by either methyl-binding proteins or immunoprecipitation. BS conversion technologies are based on bisulfite treatment of DNA, which converts unmethylated cytosines to uracils while leaving methylated cytosines unchanged. In the amplification stage, uracils are amplified as thymines, and thus can be distinguished from methylated cytosines. At each CpG site, the intensities of methylation ($\mathcal{M}$) and unmethylation ($\mathcal{U}$) can be estimated for bulk tissue samples. DNA methylation level of each sample can then be assessed as methylation proportion $\frac{\mathcal{M}}{\mathcal{M}+\mathcal{U}}$. BS conversion is commonly used in EWAS, because of its potential to produce CpG-resolution or base-resolution methylation profiles.

Increasingly dense BS methylation arrays have been developed. The Infinium HumanMethylation27 BeadChip array covers 27,578 CpG sites located in or near CpG islands within the promoter regions of 14,475 genes [8]. The Infinium HumanMethylation450 BeadChip array contains 485,577 sites (3,901 non-CpG loci) covering 99 % of RefSeq genes and 96 % of CpG islands [9]. The Infinium MethylationEPIC BeadChip array accesses >850,000 methylation sites, with 90 % of the HumanMethylation450 content covered [10]. However, even the MethylationEPIC array only covers 4 % of all CpG sites, and is less targeted at gene bodies, where methylation is hypothesized to be the most variable [11]. Sequencing, therefore, is undoubtedly preferred for future EWAS because of its better resolution and larger coverage over the genome [12].

Bisulfite sequencing (BS-seq) couples BS conversion and next-generation sequencing (NGS), making it possible to produce single-base resolution methylation profiles across the entire genome [13]. Currently, the most readily available BS-seq platform is reduced representation BS-seq (RRBS), which covers 5 - 10 million CpG sites [14]. Whole-genome bisulfite sequencing (WGBS) is able to generate methylation data at all CpGs in the genome, but for now, it is only feasible in studies with small sample sizes due to its relatively high cost.

Over the past decade, many statistical methods for BS-seq data have been proposed to detect differentially methylated cytosines (DMCs) or differentially methylated regions (DMRs). Like GWAS, EWAS also started from testing each CpG site serially using classical hypothesis testing methods, such as Fisher's exact test (FET) [15, 16, 17, 18, 19] and logistic regression [19, 20], and adjusting for multiple comparisons.

Neither FET nor logistic regression takes biological variability into account [21]. To overcome this limitation, BSmooth [22] models the methylation level at a CpG site by a smoothly varying function of its location and compares two groups by a signal-to-noise statistic similar to t-test. Besides accounting for biological variation, smoothing also reduces the sequencing coverage requirement for BS-seq and takes spatial correlation across nearby CpG sites into consideration. BiSeq [23] is another commonly-used smoothing-based method. Smoothing is carried out on predefined CpG clusters, and smoothed methylation level is modeled by a beta distribution. The mean parameter of beta distribution is further modeled by the generalized linear model

(GLM) with a “probit” link.

Beta-binomial methods belong to another category that accounts for biological variability. For a particular CpG site, given the total number of reads, the number of methylated reads follow a binomial distribution with probability of success being the underlying methylation level. This methylation level is further assumed beta-distributed, so the dispersion parameter of the beta distribution accounts for biological variation within groups. Based on this model, some methods compare the mean parameters of beta distributions between case and control groups to detect DMCs. MethylSig [24] conducts the log-likelihood ratio test. Local information can be incorporated in estimation to increase power when the sample size is small. DSS [25] estimates the dispersion parameter by an empirical Bayes approach, or a shrinkage approach when the number of replicates is small. Differential methylation is determined by the P-value of the Wald test. DSS-single [26] shares the same framework, but further model the mean as a function of location. This function is estimated by a smoothing procedure, so within-group biological variation can be accounted for by borrowing information from nearby CpG sites when there are no replicates. Empirical Bayes approach is also applied by MOABS [27], which identifies DMCs by credible methylation difference (CDIF), a new metric developed upon the credible interval of the mean difference between two groups. Other methods model the relationship between the mean parameter and experimental factors/covariates by GLM, in order to allow for more general experimental designs. RADMeth [28] uses the “logit” link function and tests whether the full model is significantly better than the reduced model (without any factors) by the log-likelihood ratio test. DSS-general [29] uses the “arcsin” link function, and a linear combinations of GLM coefficients is tested by the Wald test. MACAU [30] generalizes the beta-binomial model by adding a term accounting for population structure. GetisDMR [31] is similar to RADMeth in employing the “logit” link and log-likelihood ratio test, but differentially methylated regions (DMRs) are detected by Getis-Ord statistic, a widely-used quantity in spatial statistics.

Besides biological variability, the spatial correlation among CpG sites is another important factor to be considered. One solution is to use the hidden Markov model (HMM), where the Markov chain is used to model methylation levels of CpG sites as states (hypermethylation, hypomethylation, and no change) and emission probabilities are their chance of being DMCs among samples. HMM methods include ComMet [20], HMM-Fisher [32] and HMM-DM [33].

Most aforementioned methods also proposed approaches of various types for defining DMRs from DMCs. For example, Bsmooth combines DMCs whose test statistic is larger than a threshold; DSS also put thresholds on region length and CpG numbers; eDMR and RADMeth employ the Stouffer-Liptak test. Robinson et al. [21] and Shafi et al. [34] summarized DMR defining approaches in their reviews.

In addition to finding DMRs by combining DMCs, another strategy is to detect DMRs in predefined regions (e.g., genes, CpG islands). QDMR [35], CpG_MPs [36], and SMART [37] use the entropy to quantify methylation level variation in a region among samples, and DMRs are determined by a threshold for entropy. A few DMR analysis pipelines, such as COHCAP [38], DMAP [39], and swDMR [40], provide a flexible selection of statistical tests (including ANOVA, FET, t-test, Wilcoxon test, etc.) for different experimental designs. Regions satisfying certain criteria are claimed as DMRs. Park and Lin [41] proposed BCurve, which detects DMRs by Bayesian credible bands. For each sample, the methylation level over a region is modeled as a smoothing function of CpG locations by B-spline basis functions. Credible bands of methylation levels are estimated by a Bayesian shrinkage approach for cases and controls, respectively. Regions are identified as DMRs if credible bands do not overlap. Wu et al. [42] proposed to use aSPUw [43] test, a weighted version of aSPU [44], to detect associated CpG sites within a gene region. This method adaptively combines score statistics for CpG sites to minimize the power loss due to nonassociated CpG sites. Zhao et al. [45] proposed the global analysis of methylation profiles (GAMP), which employs the well-known SKAT [46]. Two methods, GAMP_{cdf} and GAMP_{pdf}, were developed to detect methylation differences over a large number of CpG sites or across the epigenome. They approximated the cumulative distribution function (CDF) or the density of methylation distribution for each sample by B-spline basis functions, respectively. Methylation profiles represented by functional basis coefficients are included in GLMs and tested by SKAT.

Statistical challenges are presented despite the existence of a great number of DMC or DMR detection approaches. First, most methods lack the flexibility of adjusting for covariates. It is known that age and ancestral population are two major confounding factors for DNA methylation. Yet, many methods mentioned earlier are designed for case-control studies and unable to incorporate age and population structure into comparison. Second, a great number of the above methods start from detecting DMCs site by site. On one hand, this strategy ignores the potential spatial correlation among CpG sites. On the other hand, multiple test adjustment is required afterwards, leading to potential power loss especially when the number of CpG sites is large.

We propose to model the association between a trait and methylation levels of multiple CpG sites in a region using generalized functional linear models. To detect trait-associated regions, we propose to use the adaptive Fisher (AF) method [47] to combine multiple tests of functional model coefficients according to their significance. Its weighted version, weighted adaptive Fisher (wAF) [48], has high power in detecting disease-associated genes. In this article, we conduct simulation studies and a real data analysis to show our newly proposed method, AF_{b}, has good power for detecting differentially methylated regions. Specifically, AF_{b} adapts to different proportions of nonzero effects in the functional space (i.e. proportions of basis functions associated with the trait). For the ease of discussion, we refer to the scenario of large proportion as a dense scenario, and the scenario of small proportion as a sparse scenario.

2 Methods

Consider n independent subjects. Assume that K CpG sites are located in a region (e.g., a known gene or a CpG island) with locations $0 < t_1 \leq \cdots \leq t_K \leq T$ ordered in terms of the cumulative count of base pairs. For subject i, let Y_i denote a trait, $\boldsymbol{M}_i = (M_i(t_1), ..., M_i(t_K))^T$ denote methylation levels of the K CpG sites, and $\boldsymbol{C}_i = (C_{i1}, ..., C_{iJ})^T$ denote J covariates. We use the following generalized functional linear model to describe the association between the trait and this region,

$$h\Big(E(Y_i)\Big) = \beta_0 + \int_0^T \beta(t) M_i(t) dt + \sum_{j=1}^{J} \alpha_j C_{ij}, \tag{1}$$

where $M_i(t), t \in [0, T]$ is the methylation function for subject i, $\beta(t)$ is the methylation effect function over the region $[0, T]$, and $h(\cdot)$ is taken as the logit link function for binary traits or the identity link function for continuous traits.

We assume methylation effects at $t_1, \cdots, t_K$ are a discrete realization of an underlying continuous function $\beta(t)$ over $[0, T]$. Here, we model $\beta(t)$ as a linear combination of basis functions

$$\beta(t) = \sum_{m=1}^{K_b} \gamma_m b_m(t), \tag{2}$$

where $b_m(t), m = 1, 2, \cdots, K_b$ is a series of B-spline basis functions [49]. We choose $K_b \ll K$ to reduce dimension and improve computation efficiency. Using numerical integration and equation (2), the model (1) can be approximated as follows:

$$\begin{aligned} h\Big(E(Y_i)\Big) &= \beta_0 + \int_0^T \beta(t) M_i(t) dt + \sum_{j=1}^{J} \alpha_j C_{ij} \\ &\approx \beta_0 + \sum_{k=1}^{K} \beta(t_k) M_i(t_k) \Delta t_k + \sum_{j=1}^{J} \alpha_j C_{ij} \qquad (3) \\ &= \beta_0 + \sum_{k=1}^{K} \Big[\sum_{m=1}^{K_b} \gamma_m b_m(t_k) \Big] M_i(t_k) \Delta t_k + \sum_{j=1}^{J} \alpha_j C_{ij} \\ &= \beta_0 + \sum_{m=1}^{K_b} \gamma_m \Big[\sum_{k=1}^{K} b_m(t_k) M_i(t_k) \Delta t_k \Big] + \sum_{j=1}^{J} \alpha_j C_{ij} \\ &= \beta_0 + \sum_{m=1}^{K_b} \gamma_m X_{im} + \sum_{j=1}^{J} \alpha_j C_{ij}, \qquad (4) \end{aligned}$$

where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK_b})^T = \big(\sum_{k=1}^{K} b_1(t_k) M_i(t_k) \Delta t_k, \ldots, \sum_{k=1}^{K} b_{K_b}(t_k) M_i(t_k) \Delta t_k \big)^T$, and

$$\Delta_{t_k} = \begin{cases} t_2 - t_1 & k = 1, \\ \frac{1}{2}(t_{k+1} - t_{k-1}) & k = 2, \ldots, K-1, \\ t_K - t_{K-1} & k = K. \end{cases}$$

We are interested in detecting any possible association between the trait and a CpG site in the region. Specifically, we test

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\gamma} \neq \mathbf{0}, \qquad (5)$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K_b})^T$. In the following subsections, the proposed adaptive fisher methods involve the score statistics $\boldsymbol{U} = (U_1, \ldots, U_{K_b})^T$ and the estimated covariance matrix $\boldsymbol{V} = \widehat{Cov}(U|H_0) = \{V_{ij}\}_{i,j=1}^{K_b}$, which are given by

$$\boldsymbol{U} = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{Y_i})(\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i), \qquad (6)$$

and

$$\boldsymbol{V} = \hat{\sigma}^2 \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i)(\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i)^T. \qquad (7)$$

where $\hat{\mu}_{Y_i} = h^{-1}(\hat{\beta}_0 + \sum_{j=1}^{J} \hat{\alpha}_j C_{ij})$ with $\hat{\beta}_0$ and $\hat{\alpha}_j,\ j = 1, 2, ..., J$ being the maximum likelihood estimators, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{Y_i}(1 - \hat{\mu}_{Y_i})$ for binary traits and $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\mu}_{Y_i})^2$ for continuous traits, and $\hat{\boldsymbol{X}}_i = (\hat{X}_{i1}, ..., \hat{X}_{iK_b})^T$ with $\hat{X}_{ik}$ being the predictive value of X_{ik} from a linear regression model with the covariates C_{ij}'s as predictors.

2.1 Smoothed Adaptive Fisher Method

Let the standardized score statistics be

$$\tilde{U}_k = U_k / \sqrt{V_{kk}},\ k = 1, \cdots, K_b, \qquad (8)$$

where V_{kk} is the k^{th} diagonal element of $\boldsymbol{V}$. Note that the P-value for testing a γ_k marginally based on the test statistic $\tilde{U}_k$ is approximately equal to $p_k = 2\big(1 - \Phi(|\tilde{U}_k|)\big)$. Consider the order statistics of p_k's, $p_{(1)} \leq \cdots \leq p_{(K_b)}$, in ascending order (i.e., the marginal tests are ordered from most to least significant). Let

$$R_{(k)} = -\log\ p_{(k)}, \tag{9}$$

for all $k = 1, \cdots, K_b$ so we have $R_{(1)} \geq \cdots \geq R_{(K_b)}$. Let $\boldsymbol{S} = (S_1, ..., S_{K_b})^T$ be the partial sums of $R_{(1)}, ..., R_{(K_b)}$, i.e.

$$S_k = \sum_{l=1}^{k} R_{(l)},\ k = 1, \cdots, K_b. \tag{10}$$

For each S_k, we calculated its P-value by

$$P_{S_k} = \Pr(S_k \geq s_k), \tag{11}$$

where s_k is be observed value of S_k. The smoothed adaptive Fisher statistic is defined by

$$T_{\text{AF}_\text{b}} = \min_{1 \leq k \leq K_b} P_{S_k}, \tag{12}$$

and the critical region is given by $T_{\text{AF}_\text{b}} < T_\alpha$, where the threshold T_α is determined by the distribution of the test statistic T_{AF_b} and the significant level α.

2.2 Adaptive Fisher Method

Alternatively, we could also put no assumption of smoothness on methylation effects, and thus model the association between disease status and methylation levels by the following generalized linear model,

$$h\Big(E(Y_i)\Big) = \beta_0 + \sum_{k=1}^{K} \beta_k M_i(t_k) + \sum_{j=1}^{J} \alpha_j C_{ij}, \tag{13}$$

which has a similar form as in model (4). Thus, we can likewise test

$$H_0 : \boldsymbol{\beta} = \boldsymbol{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}. \tag{14}$$

using the adaptive fisher statistic T_{AF} calculated by using equations (6) - (12) with β_k, $M_i(t_k)$, and K in place of γ_k, X_{ik}, and K_b, respectively for all k. This is a special case of the weighted adaptive Fisher test [48] with constant weights.

2.3 Computation

We use the following procedure to assess P_{S_k} in (11) and find the null distributions of T_{AF_b} in (12).

1. Calculate $\boldsymbol{U}$, $\boldsymbol{V}$ and $\tilde{\boldsymbol{U}} = (\tilde{U}_1, \cdots, \tilde{U}_{K_b})^T$ by equation (6) - (8).

2. Denote $\boldsymbol{E} = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n)^T$, where $\boldsymbol{e}_i = \boldsymbol{X}_i - \hat{\boldsymbol{X}}_i$, $i = 1, \cdots, n$. Permute the rows of $\boldsymbol{E}$ for a large number B times, obtaining $\boldsymbol{E}^{(b)} = (\boldsymbol{e}_1^{(b)}, \cdots, \boldsymbol{e}_n^{(b)})^T$. Compute $\boldsymbol{U}^{(b)}$, $\boldsymbol{V}^{(b)}$ and $\tilde{\boldsymbol{U}}^{(b)} = (\tilde{U}_1^{(b)}, \cdots, \tilde{U}_{K_b}^{(b)})^T$ based on $\boldsymbol{E}^{(b)}$, $b = 1, \cdots, B$.

3. Follow equation (9) and (10) to calculate $\boldsymbol{S}^{(b)} = (S_1^{(b)}, ..., S_{K_b}^{(b)})^T$, $b = 0, 1, 2, .., B$.

4. For a fixed $b^* \in \{0, 1, 2, ...B\}$,

$$P_{S_k}^{(b^*)} = \frac{1}{B+1} \sum_{b=0}^{B} \mathbb{I}\{S_k^{(b)} \geq S_k^{(b^*)}\}.$$

5. For each $\boldsymbol{S}^{(b)}$, $T_{\mathrm{AF_b}}^{(b)} = \min_{1 \leq k \leq K_b} P_{S_k}^{(b)}$, $b = 0, 1, 2, ..., B$.

6. The P-value of $\mathrm{AF_b}$ test can be approximated by

$$\widehat{\Pr}\{T_{\mathrm{AF_b}} \leq T_{\mathrm{AF_b}}^{(0)} | H_0\} = \frac{1}{B+1} \sum_{b=0}^{B} \mathbb{1}\{T_{\mathrm{AF_b}} \leq T_{\mathrm{AF_b}}^{(0)}\},$$

   where $T_{\mathrm{AF_b}}^{(0)} = \min_{1 \leq k \leq K_b} P_{S_k}^{(0)}$ is the observed value of the $\mathrm{AF_b}$ statistic and $\mathbb{1}(\cdot)$ is the indicator function.

3 Results

To evaluate the performance of AF methods, we conduct both simulation studies and real-data application. In the simulation studies, we compare AF, $\mathrm{AF_b}$ with GAMP and aSPUw, two popular methods to detect DMRs for pre-defined genomic regions. In the real-data application, we apply AF methods on the Whole Genome Profiling to Detect Schizophrenia Methylation Markers data, which is publicly available in the database of Genotypes and Phenotypes (dbGaP) with study accession: phs000608.v1.p1. We refer to it as Swedish SCZ data because the study collects samples from national population registered in Sweden. The software package of our proposed method is available at `https://github.com/cxystat/AFb`.

3.1 Simulation Studies

We simulate BS-seq methylation data with methylation levels quantified as estimated methylation proportions. To mimic real methylaton data, we collect location information of CpG sites and genes from the UCSC Human Genome build hg38. Let L be the length of a CpG island and T be the total length from the start of a CpG island to the end of its nearest downstream gene. In this simulation study, we take $(T, L) = (39139, 329), (99557, 576)$, or $(233783, 970)$, which are the first quartiles, the medians, the third quartiles of L and T, calculated using the data set.

Based on CpG intensities, we categorize human genome into regions of three types: CpG island, CpG shore and CpG desert. The regions of CpG islands are extracted from UCSC Genome Brower; CpG shores are defined as the 2,000 bp flanking regions upstream and downstream of CpG islands; the rest regions are CpG deserts. For each sequence from the data set, we can calculate the frequency of CpG sites on the region of each category (the number of CpG's within the region of the category divided by the length of the region), and then we take the median of the frequencies for each category. Let p_I, p_S and p_D be the medians for the CpG island, CpG shore and CpG desert categories, respectively. We have $p_I = 0.0944$, $p_S = 0.0190$ and $p_D = 0.0125$.

Next, we generate 1,000 replications for each of the three pair values of (T, L). In each replication, we simulate the locations of CpG sites in a sequence of total length T with a CpG island of length L, a CpG shore of length 2,000, and a CpG desert of length $T - L - 2,000$. Going over from the first location to the last, the simulation algorithm generates a Bernoulli(p) random variable to determines whether a locus is a CpG site, where p is equal to p_I (p_S or p_D) if the locus is in the region of the CpG island (CpG shore or CpG desert). Denote the locations of the CpG sites as $t_1 < \cdots < t_K$. We compute the expected total numbers of CpG sites, denoted by μ_N, for $(T, L) = (39139, 329)$, $(T, L) = (99557, 576)$, and $(T, L) = (233783, 970)$, and get $\mu_N = 528$, $\mu_N = 1,300$, and $\mu_N = 3,003$, respectively.

Next, we generate correlated methylation proportions by the following steps:

1. Generate a realization $Z_0(-b), Z_0(-b+1), \cdots, Z_0(T)$ of the first-order autoregressive AR(1) process
$$Z_0(t) = \psi Z_0(t-1) + \epsilon(t),\ t = -b+1, \cdots, T \tag{15}$$
where $\psi = 0.99^{1/10}$, $b = 1000$, and $Z_0(-b)$ and $\epsilon(t)$'s are i.i.d. $N(0, 0.15^2)$. Note that $t = -b, -b+1, \cdots, 0$ is a burn-in period, and only $Z_0(t_k)$'s will be used in step 3.

2. For each subject i, we generate $\boldsymbol{Z}_i = (Z_i(t_1), \ldots, Z_i(t_K))^T$ from $N(\mathbf{0}, \Sigma)$, where Σ has a AR(1) structure with $\Sigma_{kk'} = Cov(Z_i(t_k), Z_i(t_{k'})) = \rho^{|k-k'|}$, $\rho = 0.9$.

3. Transform linear combinations of $Z_0(t)$ and $Z_i(t)$ at locations $t_1, \ldots, t_K$ by the inverse tangent function. For subject i,
$$\xi_i(t_k) = \arctan(wZ_0(t_k) + (1-w)Z_i(t_k) + c + \epsilon_{ik}),\ k \in \{1, ..., K\}, \tag{16}$$
where $\omega = 0.9$, $c = -1.8$, and ϵ_{ik}'s are i.i.d. $N(0, 0.05^2)$ perturbations (e.g. measurement errors). The weighted sums of $Z_0(t)$ and $Z_i(t)$ allow methylation profiles for the n subjects to have individual variations upon a baseline level. The constant c and weight ω are chosen to match the empirical distribution of real methylation proportions.

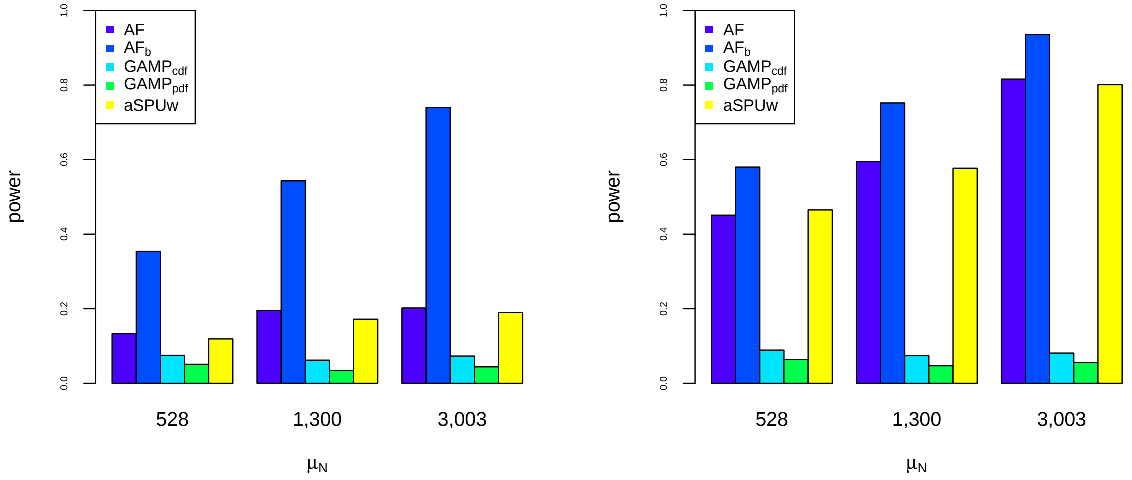
4. Rescale ξ to get the methylation proportions
$$M_i(t_k) = \frac{\xi_i(t_k) - \min_k \xi_i(t_k)}{\max_k \xi_i(t_k) - \min_k \xi_i(t_k)},\ k \in \{1, \ldots, K\}. \tag{17}$$

To construct methylation effect function $\beta(t)$, we discuss three methods in the following three subsections, which are based on B-spline basis functions, Fourier basis functions, and a autoregressive (AR) model. Finally, we generate binary trials, Y_i, $i = 1, \cdots, 1000$, using equation (1) and numerical integration [equation (3)], where the link function h is the logit function.

We compare the power of five methods: AF, AF$_{\mathrm{b}}$, GAMP$_{\mathrm{cdf}}$, GAMP$_{\mathrm{pdf}}$ and aSPUw, where the power is estimated empirically based on the $1,000$ replications. Note that the AF$_{\mathrm{b}}$ test is conducted with B-spline basis functions, regardless of the construction methods of $\beta(t)$ used to simulate data. For the second and third construction methods, the effect models used to generate data are different from that used for the AF$_{\mathrm{b}}$ tests. By choosing different construction methods, we can see how robust (or how sensitive) the performance of AF methods is to the effect models.

3.1.1 B-spline Basis Effects

In this simulation, the effect function is constructed using equation (2) with B-spline basis functions defined in Chambers and Hastie [49]. The number of the basis functions $K_b = \lceil 0.2K \rceil + 2$, where $\lceil x \rceil$ is the smallest integer no less than x. Among these K_b γ_m's, we randomly select



(a) Dense Scenario

(b) Sparse Scenario

Figure 1: Power comparison of the five methods. Data is generated using B-spline basis functions. (a) Power against varying expected number of CpG sites μ_N in dense scenario, with effect proportion $\pi = 20\%$ and effect size $\delta = 0.05$. (b) Power against varying expected number of CpG sites μ_N in sparse scenario, with effect proportion $\pi = 1\%$ and effect size $\delta = 0.3$. In both scenarios, $T \in \{39139,\ 99557,\ 233783\}$ and thus $\mu_N \in \{528,\ 1300,\ 3003\}$.

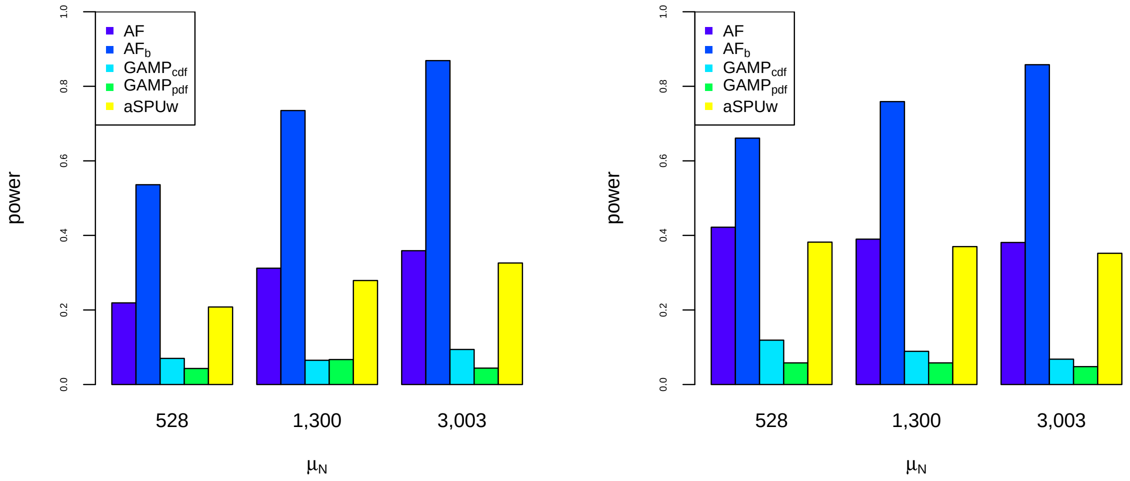
$\lceil \pi K_b \rceil$ of them to be nonzero and generated from $U[-\delta, \delta]$. For dense scenario, $\pi = 20\%$ and $\delta = 0.05$; for sparse scenario, $\pi = 1\%$ and $\delta = 0.3$.

In Figure 1, $\mathrm{AF_b}$ has the best power among the five methods in both dense and sparse scenarios. In dense scenario, $\mathrm{AF_b}$ has the best performance. Its power is more than 0.2 larger compared to the second best method for all three lengths. AF has the second best performance, followed tightly by aSPUw. $\mathrm{GAMP_{cdf}}$ and $\mathrm{GAMP_{pdf}}$ have the least favorable power. In sparse scenario, $\mathrm{AF_b}$ still has the most competitive power, though not as superior as in dense scenario. AF and aSPUw have similar power, which is about 0.1 smaller than $\mathrm{AF_b}$. Two GAMP methods have the smallest power. Different from the other three methods, the power of two GAMP methods does not increase as the expected number of CpG sites increases. Type I errors for all methods are well-controlled empirically.

3.1.2 Fourier Basis Effects

In order to see how sensitive the $\mathrm{AF_b}$ performance is to the basis functions used to generate the data, we use the Fourier basis to simulate data instead. In other words, the effect function is generated using equation (2) with $b_0(t_k), \cdots, b_{K_b}(t_k)$, $k = 1, \cdots, K$ being Fourier basis functions [50]: $b_1(t) = a/\sqrt{2}$, $b_{2r}(t) = a \sin r\theta t$, $b_{2r+1}(t) = a \cos r\theta t$, where $a = \sqrt{2/(t_K - t_1)}$, $\theta = \pi a^2$, $r = 1, \cdots, R$ and K_b is the smallest odd number no less than $0.2K$. $\lceil \pi K_b \rceil$ of γ_m's are randomly selected to be nonzero and generated from $U[-\delta, \delta]$. For dense scenario, $\pi = 20\%$ and $\delta = 1$; for sparse scenario, $\pi = 1\%$ and $\delta = 5$. Notice that when carrying out the $\mathrm{AF_b}$ test, we still use the B-spline basis functions to calculate the test statistic given that the underlying basis functions are unknown in practice.

Figure 2 shows that $\mathrm{AF_b}$ still has the best performance among the five methods for all three



(a) Dense Scenario

(b) Sparse Scenario

Figure 2: Power comparison of the five methods. Data is generated using Fourier basis functions. (a) Power against varying expected number of CpG sites μ_N in dense scenario, with effect proportion $\pi = 20\%$ and effect size $\delta = 1$. (b) Power against varying expected number of CpG sites μ_N in sparse scenario, with effect proportion $\pi = 1\%$ and effect size $\delta = 5$. In both scenarios, $T \in \{39139,\ 99557,\ 233783\}$ and thus $\mu_N \in \{528,\ 1300,\ 3003\}$.

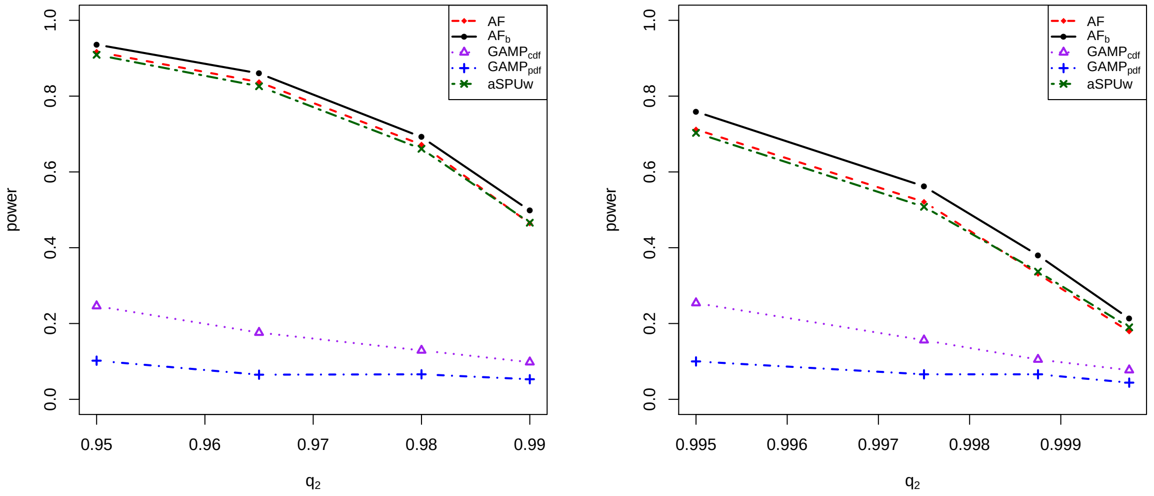
lengths in both scenarios. In dense scenario, the power of $\mathrm{AF_b}$ is about 0.3 to 0.4 larger than the second best method AF. AF, and aSPUw have similar power, with AF slightly better than aSPUw. $\mathrm{GAMP_{cdf}}$ and $\mathrm{GAMP_{pdf}}$ have the least favorable power. In sparse scenario, the trend is very similar with the dense scenario. $\mathrm{AF_b}$ has the highest power. AF and aSPUw rank second and third, with very close power. Different from the dense scenario, the power of AF and aSPU does not improve with the increasing length of target region. Two GAMP methods have the smallest power. Since GAMP is designed for the global profile over the entire methylome, difference in the chosen region might not be large enough to be detected by GAMP. In other words, a gene-length region is not ideal for GAMP to be effective and powerful. Type I errors for all methods are well-controlled empirically.

3.1.3 Autoregressive Effects

In reality, methylation effects usually may not be perfectly smooth as a linear combination of basis functions. In the last subsection, we evaluate the performance of five methods when the effect function is not generated based on basis functions. Here, we construct the effect function using the following AR(1) model (instead of basis functions)

$$\tilde{\beta}(t) = \phi\tilde{\beta}(t-1) + \epsilon(t),\ t \in \{1-b, \ldots, T\}, \tag{18}$$

where $t \in \{1-b, \cdots, 0\}$ is the burn-in period, $b = 1000$, $\phi = 0.95^{1/20}$, and the initialization $\tilde{\beta}(-b)$ and white noise $\epsilon(t)$'s are i.i.d. from $N(0, 0.005^2)$. Among the K CpG sites $\{t_1, \cdots, t_K\}$, we use a discrete Markov chain of length K with two states to select sites that have nonzero effects. Denote the two states as 0 (no effect) and 1 (having effect). The transition probabilities are $P(1,1) = q_1,\ P(1,0) = 1 - q_1,\ P(0,1) = 1 - q_2,\ P(0,0) = q_2$. Let $\mathcal{S}$ be the set of the site indexes with state value 1. We let $\beta(t) = \tilde{\beta}(t)$ if $t \in \mathcal{S}$ and $\beta(t) = 0$ if $t \notin \mathcal{S}$.



(a) $q_1 = 0.8$

(b) $q_1 = 0.975$

Figure 3: Power comparison of the five methods. Data is generated using AR(1) model. (a) Power against varying transition probability q_2 when $q_1 = 0.8$. $q_2 \in \{0.95, 0.965, 0.98, 0.99\}$. (b) Power against varying transition probability q_2 when $q_1 = 0.975$. $q_2 \in \{0.995, 0.9975, 0.99875, 0.99975\}$. In both cases, $T = 39139$.

Figure 3 shows the power of the five methods for different transition probabilities. In Figure 3(a), $q_1 = 0.8$, which is equivalent to 5 consecutive CpG sites on average carrying effects. In Figure 3(b), $q_1 = 0.975$, which is equivalent to averagely 40 consecutive effect carriers. Comparing the two cases, effects are more scattered in (a) while are more regional in (b). Within each case, we evaluate the power for different q_2 values. Note that as q_2 increases, on average there are more CpG sites having no effects in the region. Power curves in Figure 3 indicate that $\mathrm{AF_b}$ performs best with the highest power in both cases. AF and aSPUw still have the second and third best performance, whose power are very similar. The gap between $\mathrm{AF_b}$ and AF are much less than in Figure 1 and Figure 2. The performances of $\mathrm{GAMP_{cdf}}$ and $\mathrm{GAMP_{pdf}}$ are inferior, but the power of $\mathrm{GAMP_{cdf}}$ improves faster when there are more sites having effects.

Figure 1-3 show that $\mathrm{AF_b}$ has the highest empirical power for all three types of effects: B-spline basis, Fourier basis and autoregressive effects. Good performance of $\mathrm{AF_b}$ in both dense and sparse scenarios shows that it exhibits adaptability to different effect proportions. Thus, $\mathrm{AF_b}$ is a competitive method for detecting disease-associated methylation regions.

3.2 Real Data Application

We apply AF methods on the Swedish SCZ data. MBD-seq is employed to collect methylation profiles. Raw data is processed by the pipeline of Aberg et al. [51]. 1,459 subjects remain after quality control, including 741 schizophrenia (SCZ) cases and 718 controls.

The pipeline uses a nonnegative coverage at each CpG site to measure its methylation level. The coverage is estimated based on the number of sequence reads and the sample-specific fragment size distribution. Intercorrelated CpG sites are combined into blocks. For blocks containing multiple CpGs, the methylation level is the average estimated coverage. After this data reduction, 28,217,444 CpG sites are combined into 5,074,538 blocks.

The distribution of estimated coverage is highly right-skewed, so we log-transform the data to stabilize the variance. Moreover, we regress out seven principal components (PCs) from the log-coverages to eliminate potentially unmeasured confounders. In other words, $M_i(t_k)$ in model (1) is the residual of subject i's log-coverage at block k after adjusting for the PCs.

We apply AF methods on each of the 19,429 autosomal genes and their flanking regions (upstream 10,000 bp and downstream 5,000 bp) in the reference genome of UCSC Human Genome build hg19. We take disease status as the outcome and take age, gender, batch number, amount of starting material for MethylMiner, and the quantity of methylation-enriched DNA captured as covariates. We estimate P-values using a step-up procedure [44]. All genes start with $B = 100$ permutations for rough estimates of P-values. For genes with P-values smaller than or equal to $5/B$, we increase B by 10 times and redo the permutation. We repeat this procedure until all remaining P-values are greater than $5/B$, or reaches the precision level of 1×10^{-7}.

$\mathrm{AF_b}$ detects six significant genes (P-values $\leq 2.5 \times 10^{-6}$), which are listed in Table 1. Among them, FOXP1 and HOXB4 are previously reported associated with SCZ. DNA methylation of HOXB4 is correlated with the changes in hippocampal volume, suggesting it may explain SCZ-related neurodegeneration [56]. Ingason et al. [55] identifies FOXP1 by an animal experiment and confirms its human ortholog's association with SCZ by a GWAS. Results of AF tests provide additional evidence of this association between FOXP1 and SCZ, and different from previous studies, our result also suggests that the association might come from methylation in addition to SNV. FOXP1 is also identified as associated with attention deficit disorder with hyperactivity (ADHD) [54], cognitive disorders [61], and autism [62].

Some of the listed genes are related to other neurological diseases. ATP8B4 and GANC are associated with panic disorder (PD), which causes panic attacks when there is no real danger [52]. The authors identified the two genes by whole-exome sequencing of a Japanese family with several PD patients. A subsequent association test on a Japanese PD case-control study reveals that an SNV (chr15: 42631993, T > C) in GANC is a potential pathogenic variant. Because PD is well-known comorbidity to SCZ that is often overlooked [63, 64], our finding may add to the knowledge of how DNA methylation of these genes contributes to the development of PD and SCZ beyond SNVs.

Cecil et al. [53] considers a CpG site in PRDM13 (chr6: 100061307) as a differentially methylated probe to childhood maltreatment, which is a key risk factor to psychiatric vulnerability. Since DNA methylation can be acquired beyond heritage, our finding might suggest that DNA methylation of PRDM13 mediates the effect of childhood experience on the risk of SCZ, which needs to be further investigated.

Table 2 shows the nine significant genes identified by AF, among which PRDM13, HOXB4, and FOXP1 are also detected by $\mathrm{AF_b}$. Among the other six genes, SATB1 and ETS2 are found to be associated with SCZ in the existing literature. SATB1 is identified by two EWASs [57, 58], whereas the gene expression level of ETS2 (which could be regulated by DNA methylation) is detected to be associated with SCZ in a microarray meta-analysis study [60].

CDKN2A, a gene containing a type II diabetes (T2D) risk marker identified by previous GWAS, is detected to be associated with SCZ by our AF method. Similarly, Hansen et al. [59] finds another T2D at-risk variant in TCF7L2 (rs7903146 [T]) also increases the risk of SCZ. The comorbidity of T2D and SCZ has been studied [65], suggesting that CDKN2A is also a candidate risk gene for SCZ.

Table 1: Significant Genes Identified by AF_b in Swedish SCZ Study

Gene	**P-value**	**Related Disease**	**Function**
ATP8B4	3.0×10^{-7}	PD [52]	Phospholipid transport in the cell membrane.
PRDM13	3.0×10^{-7}	Psychiatric vulnerability [53]	
FOXP1	7.0×10^{-7}	ADHD [54] SCZ [55]	Regulation of gene transcription during development and adulthood.
GANC	9.0×10^{-7}	PD [52]	Encodes a key enzyme in glycogen metabolism.
HOXB4	1.0×10^{-6}	SCZ [56]	Transcription factor involved in development.
HOXA7	1.3×10^{-6}		DNA-binding transcription factor, may regulate gene expression, morphogenesis and differentiation.

Table 2: Significant Genes Identified by AF in Swedish SCZ Study

Gene	P-value	Related Disease	Function
PRDM13	4.0×10^{-7}	Psychiatric vulnerability [53]	
HOXB4	4.0×10^{-7}	SCZ [56]	Transcription factor involved in development.
HOXB3	4.0×10^{-7}		
SATB1	6.0×10^{-7}	SCZ [57, 58]	Recruits chromatin-remodeling factors to regulate chromatin structure and gene expression.
CDKN2A	6.0×10^{-7}	T2D [59]	Important tumor suppressor gene.
ETS2	6.0×10^{-7}	SCZ [60]	Transcription factor which regulates genes involved in development and apoptosis.
FOXP1	7.0×10^{-7}	ADHD [54] SCZ [55]	Regulation of gene transcription during development and adulthood.
B4GALT5	9.0×10^{-7}		
DHRS9	9.0×10^{-7}		May function as a transcriptional repressor in the nucleus.

Table 3: Computation Time

Method	AF	$\mathrm{AF_b}$	aSPUw
Time	219s	66s	2986s

4 Conclusion and Discussion

In this paper, we propose two methods $\mathrm{AF_b}$ and AF for detecting disease-associated DNA methylation regions. We demonstrate its competitive statistical power by simulation studies. By applying our methods to an SCZ dataset, we successfully identify differentially methylation genes. Most of our detected genes have been reported to be associated with SCZ or other related events (such as PD and childhood maltreatment). More interestingly, the previous report of our detected genes was based on other evidence, such as SNV association or gene expression studies. This highlights the potential of our proposed methods in detecting novel disease-related DNA methylation. While considering the fact that DNA methylation can be altered by environmental factors beyond inheritance, our method could help explain gene-environment interaction in disease etiology and could help identify potential drug targets since many DNA methylation sites are more actionable than other genomic features (e.g. SNV).

Besides statistical power, compared with AF and aSPUw, $\mathrm{AF_b}$ also improves computation efficiency. In Table 3, we compare the running time to scan 242 genes on chromosome 21 for the 1459 subjects in Swedish SCZ study using $\mathrm{AF_b}$, AF and aSPUw with 1,000 permutations. The computation is paralleled on a compute node with two Intel Xeon E5-2680 v4 processors (28 cores in total) and 128G memory. AF methods are much faster than aSPUw. $\mathrm{AF_b}$ is three times faster than AF because of the dimension reduction using basis functions. In our data example, we reduced the dimension to one fifth of the original number of CpG sites. In practice, we can also set a maximum for the number of basis functions, so the efficiency of $\mathrm{AF_b}$ can be further improved for genes with a large number of CpG sites.

Combining biomarkers into groups is a widely used strategy to improve the power of association tests in GWAS. Since BS-seq data is also single-nucleotide resolution, some methods for detecting SNV sets can also successfully detect DMRs with some appropriate adjustments. For instance, $\mathrm{AF_b}$, GAMP, and aSPUw are based on wAF, SKAT, and aSPU respectively. In contrast, burden tests, which simply pool methylation levels at all CpG sites in a region into an "epigenomic burden", may not work well in EWAS, if different subregions affect the trait differently, even in opposite directions. In contrast, $\mathrm{AF_b}$ has shown its robustness for different directions and proportions of methylation effects in simulation. Furthermore, B-spline basis functions $b_m(t)$'s can be carefully selected to cover different subregions of the genes, such as promoters, exons, and introns, as they may have distinct biological functions.

We mentioned several regression-based methods in the introduction section. Most of them regress methylation counts or proportions on experimental factors or covariates, while GAMP and aSPUw use methylation profiles and covariates as explanatory variables and trait as the response variable. AF methods also employ the latter modeling strategy. Besides being able to include covariates, this strategy is more flexible for different data types. For example, beta-binomial methods can only be applied to BS-seq but not MBD-seq data. AF methods, GAMP methods, and aSPUw, on the other hand, are not restricted by techniques used for methylation levels because they are used as regressors. We have applied AF methods on MBD-seq data in our data application and found potential SCZ-associated genes with differentially methylated profiles. AF methods could potentially be applied on data produced by third-generation sequencing (TGS) technologies, such as PacBio sequencing [66] and Nanopore sequencing [67], which are superior for simpler library preparation and of fewer artifacts and biases compared to BS-seq [68].

References

[1] Peter A Jones and Daiya Takai. The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070, 2001.

[2] Richard L Momparler and Veronica Bovenzi. Dna methylation and cancer. *Journal of cellular physiology*, 183(2):145–154, 2000.

[3] Guo-Liang Xu, Timothy H Bestor, Déborah Bourc'his, Chih-Lin Hsieh, Niels Tommerup, Merete Bugge, Maj Hulten, Xiaoyan Qu, James J Russo, and Evani Viegas-Péquignot. Chromosome instability and immunodeficiency syndrome caused by mutations in a dna methyltransferase gene. *Nature*, 402(6758):187, 1999.

[4] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[5] Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484, 2012.

[6] Fabienne Brenet, Michelle Moh, Patricia Funk, Erika Feierstein, Agnes J Viale, Nicholas D Socci, and Joseph M Scandura. Dna methylation of the first exon is tightly linked to transcriptional silencing. *PloS one*, 6(1):e14524, 2011.

[7] Xiaojing Yang, Han Han, Daniel D De Carvalho, Fides D Lay, Peter A Jones, and Gangning Liang. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell*, 26(4):577–590, 2014.

[8] Marina Bibikova, Jennie Le, Bret Barnes, Shadi Saedinia-Melnyk, Lixin Zhou, Richard Shen, and Kevin L Gunderson. Genome-wide dna methylation profiling using infinium® assay. *Epigenomics*, 1(1):177–200, 2009.

[9] 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.

[10] Inc Illumina. `https://www.illumina.com/techniques/microarrays/methylation-arrays.html`. Accessed December 2, 2019.

[11] Kevin Brennan and James M Flanagan. Epigenetic epidemiology for cancer risk: harnessing germline epigenetic variation. In *Cancer Epigenetics*, pages 439–465. Springer, 2012.

[12] James M Flanagan. Epigenome-wide association studies (ewas): past, present, and future. In *Cancer Epigenetics*, pages 51–63. Springer, 2015.

[13] Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215, 2008.

[14] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005.

[15] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315, 2009.

[16] Yingrui Li, Jingde Zhu, Geng Tian, Ning Li, Qibin Li, Mingzhi Ye, Hancheng Zheng, Jian Yu, Honglong Wu, Jihua Sun, et al. The dna methylome of human peripheral blood mononuclear cells. *PLoS biology*, 8(11), 2010.

[17] Claude Becker, Jörg Hagmann, Jonas Müller, Daniel Koenig, Oliver Stegle, Karsten Borgwardt, and Detlef Weigel. Spontaneous epigenetic variation in the arabidopsis thaliana methylome. *Nature*, 480(7376):245–249, 2011.

[18] Grant A Challen, Deqiang Sun, Mira Jeong, Min Luo, Jaroslav Jelinek, Jonathan S Berg, Christoph Bock, Aparna Vasanthakumar, Hongcang Gu, Yuanxin Xi, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nature genetics*, 44(1):23, 2012.

[19] Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E Garrett-Bakelman, Maria E Figueroa, Ari Melnick, and Christopher E Mason. methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology*, 13(10):R87, 2012.

[20] Sheng Li, Francine E Garrett-Bakelman, Altuna Akalin, Paul Zumbo, Ross Levine, Bik L To, Ian D Lewis, Anna L Brown, Richard J D'Andrea, Ari Melnick, et al. An optimized algorithm for detecting and annotating regional differential methylation. In *BMC bioinformatics*, volume 14, page S10. BioMed Central, 2013.

[21] Mark D Robinson, Abdullah Kahraman, Charity W Law, Helen Lindsay, Malgorzata Nowicka, Lukas M Weber, and Xiaobei Zhou. Statistical methods for detecting differentially methylated loci and regions. *Frontiers in genetics*, 5:324, 2014.

[22] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83, 2012.

[23] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653, 2013.

[24] Yongseok Park, Maria E Figueroa, Laura S Rozek, and Maureen A Sartor. Methylsig: a whole genome dna methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422, 2014.

[25] Hao Feng, Karen N Conneely, and Hao Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69, 2014.

[26] Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin, and Karen N Conneely. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic acids research*, 43(21):e141–e141, 2015.

[27] Deqiang Sun, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret A Goodell, and Wei Li. Moabs: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38, 2014.

[28] Egor Dolzhenko and Andrew D Smith. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, 15(1):215, 2014.

[29] Yongseok Park and Hao Wu. Differential methylation analysis for bs-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453, 2016.

[30] Amanda J Lea, Jenny Tung, and Xiang Zhou. A flexible, efficient binomial mixed model for identifying differential dna methylation in bisulfite sequencing data. *PLoS genetics*, 11(11), 2015.

[31] Yalu Wen, Fushun Chen, Qingzheng Zhang, Yan Zhuang, and Zhiguang Li. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local getis-ord statistics. *Bioinformatics*, 32(22):3396–3404, 2016.

[32] Shuying Sun and Xiaoqing Yu. Hmm-fisher: identifying differential methylation using a hidden markov model and fisher's exact test. *Statistical applications in genetics and molecular biology*, 15(1):55–67, 2016.

[33] Xiaoqing Yu and Shuying Sun. Hmm-dm: identifying differentially methylated regions using a hidden markov model. *Statistical applications in genetics and molecular biology*, 15(1):69–81, 2016.

[34] Adib Shafi, Cristina Mitrea, Tin Nguyen, and Sorin Draghici. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in bioinformatics*, 19(5):737–753, 2018.

[35] Yan Zhang, Hongbo Liu, Jie Lv, Xue Xiao, Jiang Zhu, Xiaojuan Liu, Jianzhong Su, Xia Li, Qiong Wu, Fang Wang, et al. Qdmr: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic acids research*, 39(9):e58–e58, 2011.

[36] Jianzhong Su, Haidan Yan, Yanjun Wei, Hongbo Liu, Hui Liu, Fang Wang, Jie Lv, Qiong Wu, and Yan Zhang. Cpg_mps: identification of cpg methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic acids research*, 41(1):e4–e4, 2013.

[37] Hongbo Liu, Xiaojuan Liu, Shumei Zhang, Jie Lv, Song Li, Shipeng Shang, Shanshan Jia, Yanjun Wei, Fang Wang, Jianzhong Su, et al. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic acids research*, 44(1):75–94, 2016.

[38] Charles D Warden, Heehyoung Lee, Joshua D Tompkins, Xiaojin Li, Charles Wang, Arthur D Riggs, Hua Yu, Richard Jove, and Yate-Ching Yuan. Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis. *Nucleic acids research*, 41(11):e117–e117, 2013.

[39] Peter A Stockwell, Aniruddha Chatterjee, Euan J Rodger, and Ian M Morison. Dmap: differential methylation analysis package for rrbs and wgbs data. *Bioinformatics*, 30(13):1814–1822, 2014.

[40] Zhen Wang, Xianfeng Li, Yi Jiang, Qianzhi Shao, Qi Liu, BingYu Chen, and Dongsheng Huang. swdmr: a sliding window approach to identify differentially methylated regions based on whole genome bisulfite sequencing. *PloS one*, 10(7), 2015.

[41] Jincheol Park and Shili Lin. Detection of differentially methylated regions using bayesian curve credible bands. *Statistics in Biosciences*, 10(1):20–40, 2018.

[42] Chong Wu, Jun Young Park, Weihua Guan, and Wei Pan. An adaptive gene-based test for methylation data. In *BMC Proceedings*, volume 12, page 60. BioMed Central, 2018.

[43] Yiwei Zhang, Zhiyuan Xu, Xiaotong Shen, Wei Pan, Alzheimer's Disease Neuroimaging Initiative, et al. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96:309–325, 2014.

[44] Wei Pan, Junghi Kim, Yiwei Zhang, Xiaotong Shen, and Peng Wei. A powerful and adaptive association test for rare variants. *Genetics*, 197(4):1081–1095, 2014.

[45] Ni Zhao, Douglas A Bell, Arnab Maity, Ana-Maria Staicu, Bonnie R Joubert, Stephanie J London, and Michael C Wu. Global analysis of methylation profiles from high resolution cpg data. *Genetic epidemiology*, 39(2):53–64, 2015.

[46] 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.

[47] Chi Song, Xiaoyi Min, and Heping Zhang. The screening and ranking algorithm for change-points detection in multiple samples. *The annals of applied statistics*, 10(4):2102, 2016.

[48] Xiaoyu Cai, Lo-Bin Chang, and Chi Song. Association analysis of common and rare snvs using adaptive fisher method to detect dense and sparse signals. *arXiv preprint arXiv:1812.05188*, 2018.

[49] John M Chambers, Trevor J Hastie, et al. *Statistical models in S*, chapter 7. Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA, 1992.

[50] James O Ramsay and Bernard W Silverman. *Functional Data Analysis*, chapter 3, pages 45–46. Springer, New York, NY, 2005.

[51] Karolina A Aberg, Joseph L McClay, Srilaxmi Nerella, Lin Y Xie, Shaunna L Clark, Alexandra D Hudson, Jozsef Bukszár, Daniel Adkins, Swedish Schizophrenia Consortium, Christina M Hultman, et al. Mbd-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case–control samples. *Epigenomics*, 4(6):605–621, 2012.

[52] Yoshiro Morimoto, Mihoko Shimada-Sugimoto, Takeshi Otowa, Shintaro Yoshida, Akira Kinoshita, Hiroyuki Mishima, Naohiro Yamaguchi, Takatoshi Mori, Akira Imamura, Hiroki Ozawa, et al. Whole-exome sequencing and gene-based rare variant association tests suggest that pla2g4e might be a risk gene for panic disorder. *Translational psychiatry*, 8(1):41, 2018.

[53] Charlotte AM Cecil, Rebecca G Smith, Esther Walton, Jonathan Mill, Eamon J McCrory, and Essi Viding. Epigenetic signatures of childhood abuse and neglect: Implications for psychiatric vulnerability. *Journal of psychiatric research*, 83:184–194, 2016.

[54] Jessica Lasky-Su, Benjamin M Neale, Barbara Franke, Richard JL Anney, Kaixin Zhou, Julian B Maller, Alejandro Arias Vasquez, Wai Chen, Philip Asherson, Jan Buitelaar, et al. Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(8):1345–1354, 2008.

[55] Andrés Ingason, Ina Giegling, AM Hartmann, J Genius, B Konte, M Friedl, Stephan Ripke, PF Sullivan, D St Clair, DA Collier, et al. Expression analysis in a rat psychosis model identifies novel candidate genes validated in a large case–control sample of schizophrenia. *Translational psychiatry*, 5(10):e656, 2015.

[56] Md Ashad Alam, Hui-Yi Lin, Hong-Wen Deng, Vince D Calhoun, and Yu-Ping Wang. A kernel machine method for detecting higher order interactions in multimodal datasets: Application to schizophrenia. *Journal of neuroscience methods*, 309:161–174, 2018.

[57] Karolina A Aberg, Joseph L McClay, Srilaxmi Nerella, Shaunna Clark, Gaurav Kumar, Wenan Chen, Amit N Khachane, Linying Xie, Alexandra Hudson, Guimin Gao, et al. Methylome-wide association study of schizophrenia: identifying blood biomarker signatures of environmental insults. *JAMA psychiatry*, 71(3):255–264, 2014.

[58] Carolina Montano, Margaret A Taub, Andrew Jaffe, Eirikur Briem, Jason I Feinberg, Rakel Trygvadottir, Adrian Idrizi, Arni Runarsson, Birna Berndsen, Ruben C Gur, et al. Association of dna methylation differences with schizophrenia in an epigenome-wide association study. *JAMA psychiatry*, 73(5):506–514, 2016.

[59] Thomas Hansen, Andrés Ingason, Srdjan Djurovic, Ingrid Melle, Mogens Fenger, Omar Gustafsson, Klaus D Jakobsen, Henrik B Rasmussen, Sarah Tosato, Marcella Rietschel, et al. At-risk variant in tcf7l2 for type ii diabetes increases risk of schizophrenia. *Biological psychiatry*, 70(1):59–63, 2011.

[60] Mirko Manchia, Ignazio S Piras, Matthew J Huentelman, Federica Pinna, Clement C Zai, James L Kennedy, and Bernardo Carpiniello. Pattern of gene expression in different stages of schizophrenia: Down-regulation of nptx2 gene revealed by a meta-analysis of microarray datasets. *European Neuropsychopharmacology*, 27(10):1054–1063, 2017.

[61] Claire Bacon and Gudrun A Rappold. The distinct and overlapping phenotypic spectra of foxp1 and foxp2 in cognitive disorders. *Human genetics*, 131(11):1687–1698, 2012.

[62] Wei-Hsien Chien, SusanShur-Fen Gau, Chun-Houh Chen, Wen-Che Tsai, Yu-Yu Wu, Po-Hsu Chen, Chi-Yung Shang, and Chia-Hsiang Chen. Increased gene expression of foxp1 in patients with autism spectrum disorders. *Molecular autism*, 4(1):23, 2013.

[63] Lawrence A Labbate, P Christopherson Young, and George W Arana. Panic disorder in schizophrenia. *The Canadian Journal of Psychiatry*, 44(5):488–490, 1999.

[64] Nicholas Argyle. Panic attacks in chronic schizophrenia. *The British Journal of Psychiatry*, 157(3):430–433, 1990.

[65] Lisa Dixon, Peter Weiden, Janine Delahanty, Richard Goldberg, Leticia Postrado, Alicia Lucksted, and Anthony Lehman. Prevalence and correlates of diabetes in national schizophrenia samples. *Schizophrenia bulletin*, 26(4):903–912, 2000.

[66] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of dna methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461, 2010.

[67] Jared T Simpson, Rachael E Workman, PC Zuzarte, Matei David, LJ Dursi, and Winston Timp. Detecting dna cytosine methylation using nanopore sequencing. *Nature methods*, 14(4):407, 2017.

[68] Michael C Schatz. Nanopore sequencing meets epigenetics. *Nature methods*, 14(4):347, 2017.