Detection of Differentially Methylated Regions using Kernel Distance and Scan Statistics

Fengjiao Hu, Hongyan Xu, Varghese George
Department of Biostatistics and Epidemiology,
Medical College of Georgia, Georgia Regents University, Augusta, GA 30912

Abstract

Researchers in genomics are increasingly interested in DNA methylation that is altered in disease since epigenetic changes may be susceptible to modification by environmental factors. We propose two different approaches to test for differentially methylated regions (DMRs) that account for correlations among CpG sites within DMRs, one using a kernel distance statistic and the other using a binomial spatial scan statistic. In the first approach, the kernel distance statistic is calculated as a function of the difference in methylation rates between the treatment and control groups for each CpG site, incorporating the correlations among the sites using the kernel function. The binomial scan statistic approach compares the likelihood ratios of the two groups with moving windows along the genome, using a mixed-effect model to account for correlation among CpG sites within each window. Both methods allow for adjusting for covariates. Simulation results indicate that both methods are robust with good power and good control of Type I error. The binomial scan statistic approach appears to have higher power, while the method based on kernel distance statistic is computationally faster.

Key Words: Kernel distance, Binomial spatial scan statistic, differentially methylated regions, CpG sites, mixed-effect model

1. Introduction

Extensive investigations have been performed through genome-wide association studies (GWASs) on the genetic risk of diseases in recent years. However, genetic loci through GWASs can only explain a small proportion of phenotypic variation for most common diseases (Hindorff et al., 2009). On the other hand, people realized that diseases are not only affected by genetic factors, but also non-genetic sources, such as environmental factors. This causes increasingly interested in exploring non-genetic sources, including epigenetic changes, especially DNA methylation at CpG sites, which has important implications on diseases.

In order to detect difference in DNA methylation, methylation data from Next-Generation Sequencing (NGS) have been used for statistical analysis. NGS coupled with bisulphite treatment of DNA converts unmethylated cytosines to uracils and leave methylated cytosines intact. This results in counts of uracil (unmethylated) and cytosine (methylated) at each CpG site for every sample. The total count of uracils and cytosines is the sequencing coverage at each CpG site, which could be different for each individual. Individuals with large sequencing coverage could have undue influence in statistical analysis. In order to avoid that, methylation rates have been suggested for analysis, which

is estimated as the ratio of methylated alleles over the sum of methylated and unmethylated alleles at a given site.

Methylation rates are continuous when measured across a large number of cells (Eckhardt et al., 2006). Also methylation rates at CpGs sites could be affected by those at nearby CpG sites, and have complicated correlation structure (Leek et al., 2010). Considering these, recent research focus has expanded to patterns of methylation in clusters of CpG sites, in order to detect differentially methylated regions (DMRs) in the genome.

Statistical methods have been developed to detect DMRs, including some general approaches for bump detection, such as bump-hunting techniques (Jaffe et al., 2012). Yip et al. (2014) develops a statistical method based on Jaffe et al. (2012). The main advantage of this method is that it not only uses percent methylation values but also locations of CpG sites. After scanning the genome from the beginning to the end using a sliding window which contains a fixed number of CpG sites, Ansari-Bradley test is used to find the region that has significant differences of distance distribution between two groups. However, since Ansari-Bradley test has the assumption that observations are independent, Yip et al. (2014)'s approach does not account for the correlation between CpG sites. Besides that, both methods only use methylation rates for analysis, and ignore the binomial distribution of methylation counts.

Some methods are specific for detecting DMRs based on bisulfite sequencing, for example, the two widely used packages, BSmooth (Hansen et al., 2012) and Biseq (Hebestreit et al., 2013). Both methods use functional data analysis to capture the slowly changed methylation levels over a region observed in the data. After that, BSmooth tests the group differences via a test statistic that similar to a *t*-test for each CpG site, DMRs are defined as adjacent CpG sites with absolute *t*-statistics above a defined threshold with permutations for significance testing. However, this method depends on the pre-defined threshold for absolute *t*-statistic, which would hinder automated analysis and possible leading to biased conclusion. BiSeq is a package that uses a false discovery rate procedure to control the expected proportion of incorrectly rejected regions. The main advantage of BiSeq compare to BSmooth is increased power by this hierarchical procedure. Also it takes spatial dependence into account. Eventually, the significant target regions are trimmed to the actually DMRs.

However sometimes, the data do not show as smooth of a function, and as such the wavelets are suggested to use by Ryu et al. (2014). Their generalized integrated function test, estimates subject-specific functional profiles first by using wavelets, and the average profile within groups is calculated. An ANOVA-like test is used to compare groups for a region, by comparing the overall functional relationship to the average curve within each group. This method mainly focused on testing for differential methylation of a region, which needs other tools to identify the candidate region first. Besides that, this method has limitation that can only be used for regions without missing data.

In this paper, we propose two methods, one based on kernel distance and the other based on scan statistic. The purpose is using both methods to detect DMRs along the whole genome based on methylation data from NGS. The main advantages of our methods are not only that both can detect DMRs without pre-defined regions, but also that our methods can account for correlation among CpG sites, and can adjust for covariates and other confounding factors. This is very important for methylation data, since the methylated rates have been shown to be strongly associated with covariates, such as age

(Bell et al., 2012; Teschendorff et al., 2010) and gender (Kibriya et al., 2011; Liu et al., 2010).

2. Methods

Test of whether methylation rates are different between two groups (case and control) can be done by using kernel distance statistic and binomial spatial scan statistic. Kernel distance statistic is a nonparametric method, which can be expressed as a quadratic function with differences of methylation rates between case and control groups for each CpG site, with the tri-weight kernel function (Schaid et al., 2013) to adjust between sites correlation.

Binomial spatial scan statistic is a likelihood-based method, which compares the likelihood ratio between cases and controls with moving windows. In order to adjust the between sites correlation, the mixed-effect model will be used.

Before calculating kernel distance and scan statistics, logistic regression of methylation rates is considered to adjust for covariates. The main advantage of logistic regression is that it allows for the inclusion of sample specific covariates, thus has the ability to adjust for confounding variables and batch effect.

Besides that, the methylation counts and sequencing coverage need to be adjusted based on Xu et al. (2013) for both methods. Considering the natural groups in the specimen and among the methylation loci, Xu et al. (2013)'s approach is used to adjust the clustering structure. The design effect is calculated for both groups, by treating the NGS reads at a specific CpG site as a cluster within each individual. It is calculated based on Rao and Scott (1992), which is the ratio of estimated variance of methylation rate with clustering and without clustering, reflecting the variance inflation due to clustering. Here the estimated variance of proportion without considering clusters is based on a binomial distribution. This method has advantage of no specific model assumption for the intra-cluster correlation.

2.1 Kernel Distance Statistic

Kernel distance statistic is a quadratic function calculated with differences of methylated rates for two groups at each CpG site. It can be expressed as a quadratic kernel statistic $Q = \boldsymbol{\delta}'\boldsymbol{A}\boldsymbol{\delta}$, with kernel matrix $\boldsymbol{A}$ to adjust correlation between CpG sites and multiple scaling factors for kernel function to find potential DMRs.

2.1.1 Adjusting Methylation Rates

In order to calculate kernel distance statistic, the difference of methylation rates $\boldsymbol{\delta}$ need to be calculated for each CpG site. Considering the unequal sequencing coverage for all individuals in a group at each CpG site, NGS reads at each CpG site within an individual is treated as a cluster, and clustered data analysis method used in Xu et al. (2013) is adopted here. The design effect is calculated, and then used for adjusting coverage and methylation counts at each CpG site for every group.

Suppose m_{kij} is the count of the methylation molecular at CpG site j of individual i in group k, here $k = A$ for cases and $k = U$ for controls. Suppose m_{kij} follows Binomial distribution $m_{kij} \sim B(C_{kij}, p_{kij})$, where c_{kij} is the coverage, and p_{kij} is the true

methylation rate at CpG site j for individual i in group k, with $n_k = \sum_{i=1}^{k} n_{ki}$ is the total number of CpG sites in the genome for all individuals in group k.

To calculate the design effect, first, calculate the overall methylated counts at CpG site j in case and control group, respectively, ignoring the clustering within individuals. That is, $m_{Aj} = \sum_{i=1}^{n_A} m_{Aij}$ and $m_{Uj} = \sum_{i=1}^{n_U} m_{Uij}$, where A is the set of all cases and U is the set of all controls. Then the sample methylation proportions in case and control group are given by $\hat{\beta}_{Aj} = \frac{m_{Aj}}{C_{Aj}}$ and $\hat{\beta}_{Uj} = \frac{m_{Uj}}{C_{Uj}}$, with $C_{Aj} = \sum_{i=1}^{n_A} c_{Aij}$ and $C_{Uj} = \sum_{i=1}^{n_U} c_{Uij}$. The variances of the sample methylation proportions are given by $\hat{V}(\hat{\beta}_{Aj}) = \frac{n_A \sum_{i=1}^{n_A} (m_{Aij} - c_{Aij}\hat{\beta}_{Aj})^2}{(n_A - 1) C_{Aj}^2}$ and $\hat{V}(\hat{\beta}_{Uj}) = \frac{n_U \sum_{i=1}^{n_U} (m_{Uij} - c_{Uij}\hat{\beta}_{Uj})^2}{(n_U - 1) C_{Uj}^2}$. However, without clustering, the variances of the sample methylation proportion from a binomial distribution would be $\hat{V}_B(\hat{\beta}_{Aj}) = \frac{\hat{\beta}_{Aj}(1-\hat{\beta}_{Aj})}{C_{Aj}}$ and $\hat{V}_B(\hat{\beta}_{Uj}) = \frac{\hat{\beta}_{Uj}(1-\hat{\beta}_{Uj})}{C_{Uj}}$, therefore, the design effect because of clustering are $d_{Aj} = \frac{\hat{V}(\hat{\beta}_{Aj})}{\hat{V}_B(\hat{\beta}_{Aj})}$ and $d_{Uj} = \frac{\hat{V}(\hat{\beta}_{Uj})}{\hat{V}_B(\hat{\beta}_{Uj})}$.

The design effect is then used to adjust the methylation counts, and total coverage in cases and controls, eventually have $\tilde{m}_{Aj} = \frac{m_{Aj}}{d_{Aj}}$, $\tilde{m}_{Uj} = \frac{m_{Uj}}{d_{Uj}}$ and $\tilde{C}_{Aj} = \frac{C_{Aj}}{d_{Aj}}$, $\tilde{C}_{Uj} = \frac{C_{Uj}}{d_{Uj}}$. The estimated methylation rate at CpG site j is $\tilde{\beta}_{Aj} = \frac{\tilde{m}_{Aj}}{\tilde{C}_{Aj}}$ for cases and $\tilde{\beta}_{Uj} = \frac{\tilde{m}_{Uj}}{\tilde{C}_{Uj}}$ for controls. Define the average methylation rate at CpG site j for cases is $p_{dj} = \frac{\tilde{\beta}_{Aj}}{\sum_j \tilde{\beta}_{Aj}}$ and $p_{cj} = \frac{\tilde{\beta}_{Uj}}{\sum_j \tilde{\beta}_{Uj}}$ for controls, then $\delta_j = p_{dj} - p_{cj}$ is the difference of the methylation rate between two groups at CpG site j, and would be used in kernel distance statistic.

2.1.2 Choice of Kernels

Kernel distance statistic is calculated with kernel matrix $\boldsymbol{A}$, which is used to represent the correlation between the two CpG sites. Generally, the correlation of methylation decreases as the distance of the two CpG sites increases. Therefore the kernel matrix should be based on a function that determines how rapid the correlation decreases to 0 as the distance increases. Here we define the tri-weight function $A_{jl} = \left(1 - \left(d'_{jl}\right)^2\right)^3$, if $d'_{jl} \geq 1$ and 0 otherwise (Schaid et al., 2013), where $d'_{jl} = d_{jl}/\tau$ is a scaled distance based on unknown scaling factor τ, and d_{jl} measures the distance between CpG site j and site l.

Here unknown scaling factor τ represents the cluster size, however it is difficult to predict the size of DMRs and the number of DMRs along the genome. Therefore the tri-weight function is used over a range of scaled distances as suggested in Schaid et al. (2013).

The use of multiple scaling factors makes it inappropriate to calculate p-values based on the approximated scaled chi-square distribution, and permutation of case and control status is required instead. Also in order to avoid multiple testing problems caused by multiple scaling factors, the minimum p-value is used.

After find minimum p-value, the scaling factor that corresponding to the minimum p-value is recorded as the length of DMR, τ^*, and the corresponding kernel distance can be calculated as,

$$Q(\tau^*) = \sum_{j=1}^{m}\sum_{l=1}^{m}(A_{jl}(\tau^*)(p_{Aj} - p_{Uj})(p_{Al} - p_{Ul})),$$

with percent contribution to $Q(\tau^*)$ at each CpG site calculated as $U_j(\tau^*)/Q(\tau^*)$, where $U_j(\tau^*) = \sum_{l=1}^{m}(A_{jl}(\tau^*)(p_{Aj} - p_{Uj})(p_{Al} - p_{Ul}))$. Then the distribution of methylation rates can be plotted based on the percent contribution $U_j(\tau^*)/Q(\tau^*)$ versus CpG site j, which can give us a graphical view of potential DMRs.

2.1.3 Adjusting for covariates

Here the distance statistic considering covariates is proposed for methylation data, based on Schaid et al. (2013). Let x_{ki} represents covariate of individual i in group k, the logistic regression

$$\log\left(\frac{m_{kij}}{c_{kij} - m_{kij}}\right) = \beta_{0k} + \beta_{1k}x_{ki}$$

is used to fit all the data for both groups, and calculate the fitted odds for methylation at CpG site j for individual i in group k, then can get the corresponded adjusted expected methylation rate is

$$\hat{p}_{kij} = \frac{\hat{m}_{kij}}{\hat{c}_{kij}} = \frac{\exp(\hat{\beta}_{0k} + \hat{\beta}_{1k}x_{ki})}{1 + \exp(\hat{\beta}_{0k} + \hat{\beta}_{1k}x_{ki})},$$

and then the difference of observed and expected methylated counts at CpG j for individual i in group k is calculated as residual $r_{kij} = m_{kij} - \hat{p}_{kij}c_{kij}$.

In order to calculate kernel distance statistic, the design effect in Xu et al. (2013) are calculated first as in Section 2.1.1 to adjust the residuals and sequencing coverage, we have $\tilde{r}_{Aj} = \frac{r_{Aj}}{d_{Aj}}$, $\tilde{r}_{Uj} = \frac{r_{Uj}}{d_{Uj}}$ and $\tilde{C}_{Aj} = \frac{C_{Aj}}{d_{Aj}}$, $\tilde{C}_{Uj} = \frac{C_{Uj}}{d_{Uj}}$. The estimated adjusted methylation rate at CpG site j is $\tilde{\beta}_{Aj} = \frac{\tilde{r}_{Aj}}{\tilde{C}_{Aj}}$ for cases and $\tilde{\beta}_{Uj} = \frac{\tilde{r}_{Uj}}{\tilde{C}_{Uj}}$ for controls. Define the average adjusted methylation rate at CpG site j for cases is $p_{dj} = \frac{\tilde{\beta}_{Aj}}{\sum_j \tilde{\beta}_{Aj}}$ and $p_{cj} = \frac{\tilde{\beta}_{Uj}}{\sum_j \tilde{\beta}_{Uj}}$ for controls, then $\delta_j = p_{dj} - p_{cj}$ is the difference of the adjusted methylation rate between two groups at CpG site j, and would be used in kernel distance statistic with kernel matrix defined in Section 2.1.2.

2.1.4 Conclusions for Kernel Distance Statistic

Kernel distance statistic has advantage of being fast in computing (Schaid et al., 2013), however, the power of kernel distance statistic might be reduced since it is strongly depends on the pre-defined scale parameter τ to reflect the unknown value of cluster size. If the values of τ is not close to the actual size, it might have difficulty to detect the real DMR.

2.2 Binomial Spatial Scan Statistic

Besides kernel distance statistic, scan statistic is another method that can be used to detect DMRs, which is based on comparing the likelihood ratio of methylation rates between the case and control groups. And this method uses moving windows along the

genome, with multiple window sizes, which will help to reflect location of the DMRs, and eventually increase the power compare to kernel distance statistic.

Scan statistic was first studied by Naus (1965) to detect clusters in a point process in a one-dimensional setting. He applied the idea of maximum frequency to the case of ungrouped data and proposed a 'scan' test with the null hypothesis of a purely random Poisson process.

However, it is well known that methylation rates does not follow a uniform distribution. Therefore, reasonable method is needed to take into account the underlying distribution of methylation rates. Kulldorff (1997) described a likelihood-based scan statistic, and was extended to detect genetic variants by Ionita-Laza et al. (2012), considering the Bernoulli distribution of variants at each position for each individual. The scan statistic was calculated based on the likelihood ratio of the frequencies of variants carried among cases and controls within a window versus outside the window. And the scan statistic was calculated for each window with moving windows across the whole genome. Then the maximum scan statistic over the windows of all possible sizes, is defined as the global statistic. However, the approach by Ionita-Laza et al. (2012) cannot be adapted for methylation data, since methylated counts at each CpG site for every individual follows a binomial distribution instead, after considering sequencing coverage.

Here binomial scan statistic is proposed and calculated based on the adjusted methylation counts and sequencing coverage. Since methylation rate for each CpG site is affected by those of closed-by CpG sites in the region, the correlation of methylation counts are adjusted by mixed-effect model first. And then the methylation counts and sequencing coverage are adjusted by "design", after considering unequal coverage for every individual at each CpG site.

2.2.1 Binomial Scan Statistic

To account for the correlation of CpG sites in each moving window, a random intercept and slope mixed-effect logistic model is considered to model methylation counts at each CpG site for every individual,

$$\log\left(\frac{m_{kij}}{c_{kij}-m_{kij}}\right)=\beta_{0k}+\beta_{1k}s_j+\nu_{0ki}+\nu_{1ki}s_j+e_{kij},$$

where s_j represents the distance of CpG site j from the start point of the specific window.

In the mixed-effect logistic model setting, the random effect $\boldsymbol{\nu}_k=\begin{pmatrix}\nu_{0ki}\\ \nu_{1ki}\end{pmatrix}$ is assumed to vary independently across individuals, with $\boldsymbol{\nu_k}\sim N\left(0,\begin{pmatrix}\sigma^2_{\nu_{0ki}} & \sigma_{\nu_{0ki}}\sigma_{1ki}\\ \sigma_{\nu_{0ki}}\sigma_{1ki} & \sigma^2_{\nu_{1ki}}\end{pmatrix}\right)$, and is independent with the error e_{kij}, which is assumed to vary independently across CpG sites within an individual, with $e_{kij}\sim N(0,\sigma_e^2)$.

Next the fitted odds of methylation counts can be calculated for CpG j of individual i in group k. Similar to Section 2.1.3, we can get the corresponding adjusted expected methylation rate $\hat{p}_{kij}$, and the difference of observed and expected methylated counts at CpG j for individual i in group k is calculated as residual $r_{kij}=m_{kij}-\hat{p}_{kij}c_{kij}$.

Considering the different sequencing coverage for every individual at each CpG site, the residuals and sequencing coverage are adjusted by the design effect based on Xu et al. (2013), as in Section 2.1.1. Eventually we have adjusted residuals $\tilde{r}_{Aj}$ ($\tilde{r}_{Uj}$) and sequencing coverage $\tilde{C}_{Aj}$ ($\tilde{C}_{Uj}$) for cases (controls) at each CpG site. Here we assume $\tilde{r}_{Aj}\sim B(\tilde{C}_{Aj}, p_A)$ and $\tilde{r}_{Uj}\sim B(\tilde{C}_{Uj}, p_U)$, with p_A and p_U are true methylation rates for cases and controls. Considering $\tilde{r}_{kj}\sim B\left(\tilde{C}_{kj}, p_k\right)$, then the likelihood of $\tilde{r}_{kj}$ is

$$f\left(\tilde{r}_{kj}\right) = \binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} p_k^{\tilde{r}_{kj}} (1-p_k)^{\tilde{C}_{kj}-\tilde{r}_{kj}}$$
$$= \binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} exp\left\{\tilde{C}_{kj}\left(\frac{\tilde{r}_{kj}}{\tilde{C}_{kj}}\log\left(\frac{p_k}{1-p_k}\right) + \log(1-p_k)\right)\right\}.$$

For a specific window, after adjusting for between CpG sites correlation by using mixed-effect model, the residuals $(\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks})$ for the s consecutive CpG sites are assumed to be independent. Then the joint likelihood of residuals over continuous s CpG sites in the defined region for group k is the product of the likelihoods of the s CpG site, which can be expressed as,

$$f(\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks}) = \prod_{j=1}^{s}\binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} exp\left\{\tilde{C}_{kj}\left(\frac{\tilde{r}_{kj}}{\tilde{C}_{kj}}\log\left(\frac{p_k}{1-p_k}\right) + \log(1-p_k)\right)\right\}$$
$$= \prod_{j=1}^{s}\binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} exp\left\{\sum_{j=1}^{s}\tilde{C}_{kj}\left(\frac{\sum_{j=1}^{s}\tilde{r}_{kj}}{\sum_{j=1}^{s}\tilde{C}_{kj}}\log\left(\frac{p_k}{1-p_k}\right) + \log(1-p_k)\right)\right\}.$$

From this likelihood, we can see the distribution of adjusted residuals is from a one-parameter exponential family $y\sim 1EXP(\eta, \phi, T, B_e, a)$ with

$$T(\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks}) = \frac{\sum_{j=1}^{s}\tilde{r}_{kj}}{\sum_{j=1}^{s}\tilde{C}_{kj}}$$
$$\eta = log\left(\frac{p_k}{1-p_k}\right) \rightarrow p_k = \frac{\exp(\eta)}{1+\exp(\eta)}$$
$$B_e(\eta) = -\log(1-p_k) = \log(1+e^{\eta})$$
$$\phi = \frac{1}{\sum_{j=1}^{s}\tilde{C}_{kj}} \text{ with } a(\phi) = 1$$
$$g_e(x) = (B_e')^{-1} = \log(x) - \log(1-x)$$

and the log-likelihood $l(\eta; y) = (\eta T(y) - B_e(\eta))/\phi$ after ignoring additive constant that do not depend on η.

Based on this likelihood function, we can find maximum likelihood estimator (MLE) of parameter η in one-parameter exponential family $y_i\sim 1EXP(\eta, \phi_i, T, B_e, a)$ as $\hat{\eta} = g_e(T^*(y))$, where $g_e = (B_e')^{-1}$ (Agarwal et al., 2006).

Let η_A and η_U be the MLE parameters for the data with two groups in the same specified region. In order to test the hypothesis $H_1: \eta_A \neq \eta_U$ versus $H_0: \eta_A = \eta_U$, the ratio of the likelihood under H_1 versus the likelihood under H_0 can be used as a test statistic, with the log of the test statistic given by

$$\Delta = \kappa(T_A, \Phi_A) + \kappa(T_U, \Phi_U) - \kappa(T, \Phi),$$

where $\kappa(x,y) = (xg_e(x) - B_e(g_e(x)))/y$ and $\frac{1}{\Phi} = \frac{1}{\Phi_A} + \frac{1}{\Phi_U}$, $T = b_A T_A + (1-b_U)T_U$ with $b_A = \frac{1}{\Phi_A}/(\frac{1}{\Phi_A} + \frac{1}{\Phi_U})$.

Here we have $\Phi_A = \frac{1}{\sum_{j=1}^{s} \tilde{c}_{Aj}}$, $\Phi_U = \frac{1}{\sum_{j=1}^{s} \tilde{c}_{Uj}}$ and $T_A = \frac{\sum_{j=1}^{s} \tilde{r}_{Aj}}{\sum_{j=1}^{s} \tilde{c}_{Aj}}$, $T_U = \frac{\sum_{j=1}^{s} \tilde{r}_{Uj}}{\sum_{j=1}^{s} \tilde{c}_{Uj}}$ for cases and controls, with

$$\begin{aligned}\Delta &= \kappa(T_A, \Phi_A) + \kappa(T_U, \Phi_U) - \kappa(T, \Phi) \\ &= \frac{T}{\Phi}\left(r_A \log\left(\frac{r_A}{b_A}\right) + \left(\frac{b_A}{T} - r_A\right)\log\left(1 - T\frac{r_A}{b_A}\right) + (1-r_A)\log\left(\frac{1-r_A}{1-b_A}\right)\right. \\ &\qquad \left. + \left(\frac{1-b_A}{T} - 1 + r_A\right)\log\left(1 - T\frac{1-r_A}{1-b_A}\right)\right) - \frac{1-T}{\Phi}\log(1-T)\end{aligned}$$

where $b_A = \frac{\sum_{j=1}^{s} \tilde{c}_{Aj}}{\sum_{j=1}^{s} \tilde{c}_{Aj} + \sum_{j=1}^{s} \tilde{c}_{Uj}}$, $r_A = \frac{\sum_{j=1}^{s} \tilde{r}_{Aj}}{\sum_{j=1}^{s} \tilde{r}_{Aj} + \sum_{j=1}^{s} \tilde{r}_{Uj}}$ and $T = \frac{\sum_{j=1}^{s} \tilde{r}_{Aj} + \sum_{j=1}^{s} \tilde{r}_{Uj}}{\sum_{j=1}^{s} \tilde{c}_{Aj} + \sum_{j=1}^{s} \tilde{c}_{Uj}}$, $\Phi = \frac{1}{\sum_{j=1}^{s} \tilde{c}_{Aj} + \sum_{j=1}^{s} \tilde{c}_{Uj}}$.

2.2.2 Adjusting for covariates

In order to adjust for covariates for each individual, such as age and gender, the following mixed-effect logistic regression with random intercept and slope is considered:

$$\log\left(\frac{m_{kij}}{c_{kij} - m_{kij}}\right) = \beta_{0k} + \beta_{1k} x_{1ki} + \beta_{2k} s_j + \nu_{0ki} + \nu_{1ki} s_j + e_{kij},$$

where x_{1ki} denotes covariate for individual i in group k; and s_j represents the distance of CpG site j from the start point of the specific window.

The residuals can be calculated based on this mixed-effect model, and also residuals and sequencing coverage are adjusted, and then used for calculating scan statistic as in Section 2.2.1.

2.2.3 Conclusions for Scan Statistic

The scan statistic is calculated for each window using moving windows with variable window (VW) size approach across the whole genome. And DMR is the window with the highest binomial scan statistic. For each window *W* of size *w*, the binomial scan statistic can be calculated. The scan statistic for window size *w* (LR_w) is the highest value for the scan statistic for windows of size *w*. And then the maximum of LR_w over all values of *w* is used as global statistic.

The window size *w*=1,2,…,*m*/2 is recommended in Ionita-Laza et al. (2012). However, the LR_w calculation is unstable if the frequency of methylated counts within a given window is 0, for either cases or controls. This can be avoided by adding a pseudo-count of 1 to the adjusted methylated and unmethylated counts for each CpG site, equivalent to assuming a uniform prior distribution for methylation across the different sites.

Since the distribution of binomial scan statistic is unknown, an approximate *p*-value for the window with the largest LR_w is calculated by permutation of the case-control status of the subjects.

Binomial scan statistic has potential advantage of improved power, since the moving window with multiple window size can solve the difficulty of determining the value of τ

in the kernel distance method. However, since the mixed-effect model needs to be applied for each moving window, also the scan statistic is a likelihood based method, it could be time-consuming in computation.

3. Simulation

The main focus for simulation is to compare the two methods based on kernel distance and binomial scan statistic, with respect to statistical validity, power and computational efficiency.

For simplicity, (i) we will not include any covariates, and (ii) we will use equal sample size for cases and controls in the simulation model. For the power comparisons at various alternative hypotheses and various significant levels, we assume that there is only one DMR in the simulated genomic region, and all CpG sites within the region are equally spaced.

3.1 Simulation Parameters

Methylation counts at each CpG site for every individual are generated from $Bin(c_{kij}, p_{kij})$, $i = 1,2,\dots,2N$, $j = 1,2,..m$, $k = A, U$. Here the sequencing coverage c_{kij} is allowed to vary by sampling from a normal distribution $N(30,13)$, with a minimum of 5 based on the real data analysis from Xu et al. (2013). And the methylation rate p_{kij} is simulated based on the two-step procedure in Lacey et al. (2013) in order to model the spatial dependence for the methylation rates of close by CpG sites.

Briefly, first a sample of independent values is drawn from $Beta(\alpha, \beta)$ distribution. And then the vector of independent random variables X will be transformed into a vector of correlated random variables $X^* = 1 - \Phi\{C\Phi^{-1}(1 - X)\}$, where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution function with Cholesky decomposition C of the correlation matrix $\Sigma = CC'$. For correlation structure, all diagonal correlation equal 1, and all off-diagonal correlation equal the ratio of common value of ρ and the distance of the two CpG sites, in order to represent the fact that the correlation of methylation rates for two CpG sites decreases as the distance between them increases.

By using the above two-step procedure, the methylation rates p_{kij} for CpG sites will be generated as $p_{kij} \sim Beta(\alpha_U, \beta_U)$, $k = A, U$ for CpG site j of individual i under null hypothesis. Under alternative hypothesis, the methylation rates p_{kij} for CpG sites will be generated as $p_{kij} \sim Beta(\alpha_U, \beta_U)$, $k = A, U$ for CpG site j outsides of DMR. Within DMR, the methylation rates p_{kij} will be simulated as $p_{Uij} \sim Beta(\alpha_U, \beta_U)$, and $p_{Aij} \sim Beta(\alpha_A, \beta_A)$, Here $\alpha_A \neq \alpha_U$ or $\beta_A \neq \beta_U$, representing that methylation rates are different between cases and controls within DMR.

3.2 Simulation Results

Simulation was performed using a total sample size of 48, with 24 in each group; and also sample size of 60. The methylation rate was simulated for one region with 24 CpG sites, and 6 of which are in a DMR, with correlation of $\rho = 0.7$ or $\rho = 0.5$ for adjacent CpG sites. The correlation among non-adjacent sites were scaled down by dividing ρ by the distance between sites.

We set $\alpha_U = 0.1$, $\beta_A = \beta_U = 0.9$, and use different values of α_A to represent effect sizes. Based on the property of beta distribution, the mean of methylation rates in DMR increases as the value of α_A increases.

In order to present the effect of the parameters on the power, plots of power versus different values of α_A are presented in Figure 1. Here $\alpha_A = 0.1$ means the effect size is zero, and the powers for kernel distance and scan statistics are very close to the type I error of 0.05. The straight horizontal line is $y = 0.05$. This indicates that both methods have well-controlled type I error rate.

From Figure 1, we can also see that the powers for kernel distance and scan statistics increase as the effect sizes increase. And the scan statistic has better power than kernel distance statistic.

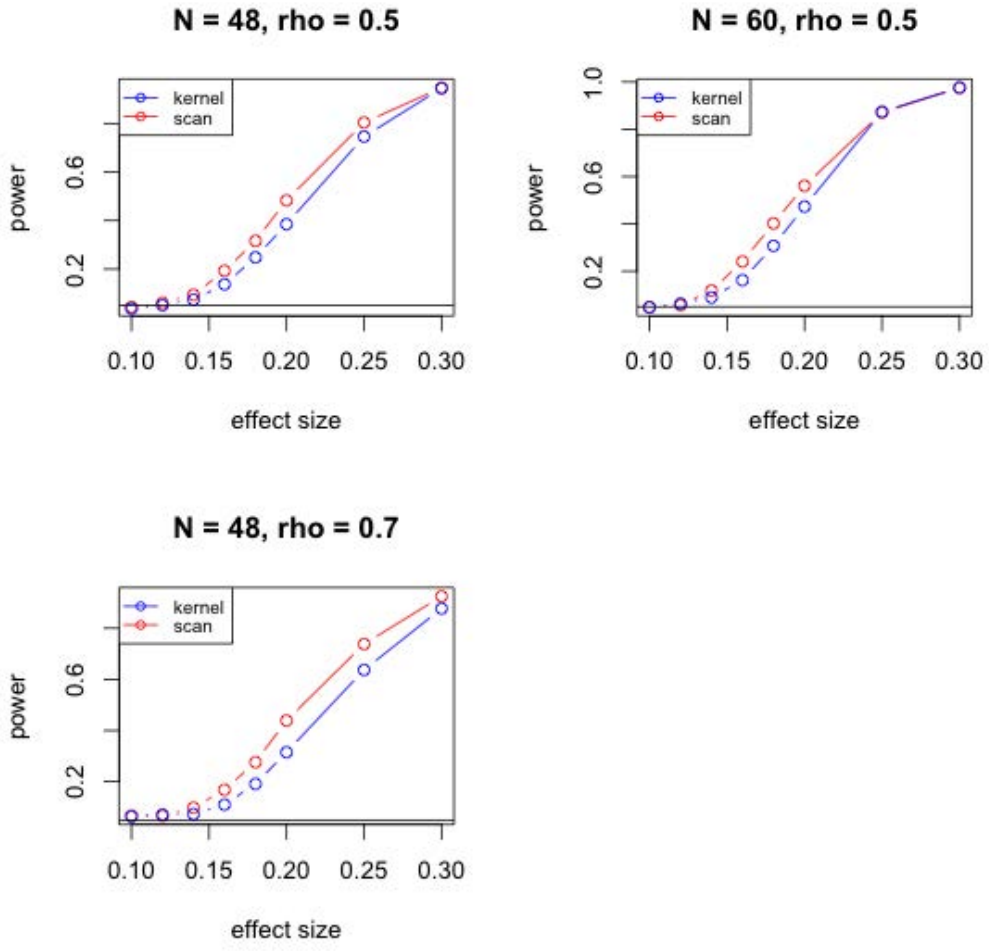

Figure 1: Power curves for simulation results

4. Discussion

Simulation results indicate that both methods are valid approach for detecting DMRs with reasonable power and good control of Type I error. The binomial scan statistic approach appears to have higher power than the kernel distance statistic approach. However, it has

limitation that it is computationally slow compared to the kernel distance statistic approach. More extensive simulations are being conducted to further compare these two approaches.

References

Agarwal, D., Phillips, J. M., & Venkatasubramanian, S. (2006). *The hunting of the bump: on maximizing statistical discrepancy*. Paper presented at the Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, Miami, Florida.

Bell, J. T., Tsai, P. C., Yang, T. P., Pidsley, R., Nisbet, J., Glass, D., . . . Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet, 8*(4), e1002629. doi: 10.1371/journal.pgen.1002629

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., . . . Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet, 38*(12), 1378-1385. doi: 10.1038/ng1909

Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol, 13*(10), R83. doi: 10.1186/gb-2012-13-10-r83

Hebestreit, K., Dugas, M., & Klein, H. U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics, 29*(13), 1647-1653. doi: 10.1093/bioinformatics/btt263

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A, 106*(23), 9362-9367. doi: 10.1073/pnas.0903103106

Ionita-Laza, I., Makarov, V., Consortium, A. A. S., & Buxbaum, J. D. (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am J Hum Genet, 90*(6), 1002-1013. doi: 10.1016/j.ajhg.2012.04.010

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol, 41*(1), 200-209. doi: 10.1093/ije/dyr238

Kibriya, M. G., Raza, M., Jasmine, F., Roy, S., Paul-Brutus, R., Rahaman, R., . . . Ahsan, H. (2011). A genome-wide DNA methylation study in colorectal carcinoma. *BMC Med Genomics, 4*, 50. doi: 10.1186/1755-8794-4-50

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods, 26*(6), 1481-1496. doi: Doi 10.1080/03610929708831995

Lacey, M. R., Baribault, C., & Ehrlich, M. (2013). Modeling, simulation and analysis of methylation profiles from reduced representation bisulfite sequencing experiments. *Stat Appl Genet Mol Biol, 12*(6), 723-742. doi: 10.1515/sagmb-2013-0027

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., . . . Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet, 11*(10), 733-739. doi: 10.1038/nrg2825

Liu, J., Morgan, M., Hutchison, K., & Calhoun, V. D. (2010). A study of the influence of sex on genome wide methylation. *PLoS One, 5*(4), e10028. doi: 10.1371/journal.pone.0010028

Naus, J. I. (1965). The Distribution of the Size of the Maximum Cluster of Points on a Line. *Journal of the American Statistical Association, 60*(310), 532-538. doi: 10.2307/2282688

Rao, J. N., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics, 48*(2), 577-585.

Ryu, D., Xu, H., George, V., Su, S., Wang, X., Shi, H., & Podolsky, R. H. (2014). Generalized Integrated Functional Test for Regional Methylation Rates.

Sato, F., Jin, Z., Schulmann, K., Wang, J., Greenwald, B. D., Ito, T., . . . Meltzer, S. J. (2008). Three-tiered risk stratification model to predict progression in Barrett's esophagus using

epigenetic and clinical features. *PLoS One, 3*(4), e1890. doi: 10.1371/journal.pone.0001890

Schaid, D. J., Sinnwell, J. P., McDonnell, S. K., & Thibodeau, S. N. (2013). Detecting genomic clustering of risk variants from sequence data: cases versus controls. *Hum Genet, 132*(11), 1301-1309. doi: 10.1007/s00439-013-1335-y

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., . . . Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res, 20*(4), 440-446. doi: 10.1101/gr.103606.109

Xu, H., Podolsky, R. H., Ryu, D., Wang, X., Su, S., Shi, H., & George, V. (2013). A method to detect differentially methylated loci with next-generation sequencing. *Genet Epidemiol, 37*(4), 377-382. doi: 10.1002/gepi.21726

Yip, W. K., Fier, H., DeMeo, D. L., Aryee, M., Laird, N., & Lange, C. (2014). A novel method for detecting association between DNA methylation and diseases using spatial information. *Genet Epidemiol, 38*(8), 714-721. doi: 10.1002/gepi.21851