RESEARCH

Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics

Yalu Wen[1*], Fushun Chen[2], Qingzheng Zhuang[2], Yan Zhuang[2] and Zhiguang Li[2†]

[*]Correspondence: y.wen@auckland.ac.nz
[1]Department of Statistics, University of Auckland, 38 Princes Street, 1010 Auckland, New Zealand
Full list of author information is available at the end of the article
[†]Corresponding authors

Abstract

DNA methylation is an important epigenetic modification that has essential roles in gene regulation, cell differentiation, and cancer development. Bisulfite sequencing is a widely used technique to obtain genome-wide DNA methylation profiles, and one of the key tasks of analyzing bisulfite sequencing data is to detect differentially methylated regions (DMRs) among samples under different conditions. Although numerous tools have been proposed to detect differentially methylated single CpG site(DMC) between paired samples, methods for direct DMR detection, especially for complex study designs, are largely limited. Here, we present a new software, GetisDMR, for direct DMR detection. We use the beta-binomial regression to model the whole-genome bisulfite sequencing data, where variations in methylation levels and confounding effects have been accounted for. We further employ a region-wise test statistic, which is derived based on the local Getis-Ord statistics and takes the spatial correlation between nearby CpG site into consideration, to detect DMRs. Unlike the existing methods, that attempt to infer DMRs from DMCs based on empirical criteria, we provide statistical inference for DMR detection. Through extensive simulations and an application to two mouse datasets, we demonstrate that GetisDMR consistently achieves better sensitivity, positive predictive values, more exact locations of DMRs, and better agreement of DMRs with our current biological knowledge.

Keywords: Differentially methylated region; Getis-Ord local statistics; DNA methylation

Introduction

DNA methylation is a stable epigenetic modification that plays a key role in numerous biological processes, such as genomic imprinting, regulation of gene expression, cell differentiation, development and carcinogenesis [1–5]. For example, it is believed that DNA methylation regulates the transcription of genes mainly through two ways. Firstly, cytosine methylation can physically impede the binding of transcriptional proteins to the gene [6], and secondly methylated DNA can be bound by methyl-binding domain proteins (MBDs)[7, 8]. MBD proteins recruit additional proteins, including histone deacetylases and other chromatin remodeling proteins, which enables the extensive cross-talk between DNA methylation and chromatin modifying histone marks[9, 10]. The presence of large scale aberrant DNA methylation pattern, typically with site-specific hyper-methylation in tumor suppressor genes and global hypo-methylation in oncogenes compared to normal tissue, is a hallmark feature of various types of cancers[2, 11].

The whole-genome bisulfite sequencing (WGBS), which combines the bisulphite treatment with next generation sequencing (BS-Seq), becomes the state-of-the-art technology in investigating DNA methylation pattern at single base resolution with relatively high coverage across multiple samples. The bisulphite treatment converts un-methylated cytosines to uracils, while leaving the methylated cytosines unchanged. Thus, it allows for the discrimination between methylated and unmethylated CpG sites [12]. A WGBS experiment typically involves many distinct cells with potentially different methylation state at a particular cytosine, and therefore methylation level at each CpG site defined as the proportion of molecules with cytosine methylated ($\frac{C}{C+T}$) is used to summarize the pattern of DNA methylation [13–15].

Over the past few years, a number of approaches have been proposed for assessing differentially methylated region (DMR) from WGBS data. One of the most straightforward method is to use the Fisher's Exact Test to compare the methylation levels among different groups at each CpG site [16]. Recently, Saito *et al.* developed the ComMet, which is built based on hidden Markov models (HMMs) and designed to detect DMR between a pair of samples [17]. Though these methods are easy to implement and can compare a pair of two samples obtained either directly from the experiment or by pooling samples under the same experimental condition, these methods do not take the between sample variations into account and can't adjust for confounding variables when biological replicates are available[18, 19].

Converging evidences suggest that the close-by CpG sites tend to have similar methylation levels [18, 20]. With the assumption that methylation levels change smoothly along the genome and the adjacent CpG sites have similar methylation levels, various smoothing based methods have been proposed to detect DMRs[18]. Although the smoothing procedures adopted by these methods may differ in details, all of them employ local averaging to improve the precision of the methylation level estimates, especially for CpG sites with low coverage. For example, the BSmooth method first estimates the methylation levels with a local-likelihood smoother, and then performs the statistical test using a signal-to-noise statistics[18]. A DMR is claimed when groups of consecutive CpGs with the signal-to-noise statistics larger than a cutoff selected based on its marginal distribution. BiSeq is another method, which first employs a local smoothing technique and then detects the DMR based on the smoothed methylation level estimates[20]. The key difference between BiSeq and BSmooth is that BiSeq adopts a hierarchical testing procedure to detect DMRs and takes the spatial correlations among p-values of adjacent CpG sites into account. The BiSeq requires the specification of a set of candidate regions that may differentially methylated, and thus it is only suitable for detecting DMRs in targeted bisulfite sequencing data. Though smoothing based methods make use of the information from adjacent CpG sites, they always requires biological replicates and thus can't be applied to the datasets without replicates. Currently, the WGBS is quite costly, which prohibits the obtainment of multiple replicates for both of the experimental conditions given limited budget[21, 22]. It is quite common that some of the biological replicates are combined into one sample before library generation for sequencing experiments[17, 23]. Moreover, there are situations where biological replicates are hard to obtain, especially in retrospective studies[24].

Regression based methods have also been proposed to detect DMRs. For example, MethylKit assumes that the number of methylated reads follow a binomial distribution and models the methylated reads within the logistic regression framework[13]. The p-values were calculated and multiple comparisons were adjusted using a sliding linear model method. As methylation levels vary significantly across individuals, failing to consider the variability across individuals may result in inflated type-I error[18, 19]. Beta-binomial regression has been recently introduced to model methylation levels in WGBS data, as it can take both the sampling and epigenetic variations into account[15, 25, 26]. For example, DSS method uses a lognormal-beta-binomial Bayesian hierarchical model to describe the methylated counts, and the DMR is defined as the CpG site with p-value less than a pre-specified threshold[25]. The DSS method allows information sharing across different CpG sites to improve precision of the test, but the correlation of p-values for proximal sites is not explicitly considered in the DMR detection which may reduce both the sensitivity and specificity of the test. The methylSig method also models the methylated reads using a beta-binomial distribution and the likelihood ratio test is used to detect differentially methylated CpG (DMC) site[26]. Although the methylSig can be used to identify DMC, it does not have the mechanism to detect DMRs which is of more biological relevance. RADMeth adopts a beta-binomial regression to calculate the p-value of each CpG site and then combines the information from p-values within 200 base pairs(bp)[15]. Beta-binomial regression based methods can explicitly take both the epigenetic and sampling variations into account, but they mainly focus on detecting DMCs and have limited power of identifying DMRs. They usually pre-specify a certain length for the DMR and then combines the information within the window to infer the significance of the detected DMRs[13, 15]. However, compelling evidences suggest that DMR can ranges from a few base pairs to thousands of base pairs, and a fixed length of DMR certainly contradicts with our knowledge[27]. Compared with detecting DMCs, DMR detection has several advantages. Firstly, locating the regions with multiple DMCs are one of the most basic goals for methylation studies. Secondly, as pointed out by Bock, after adjusting for multiple comparisons for DMC detection, only the strongest differences tend to remain significant[28]. Targeting at detecting DMR rather than single CpG site can substantially reduce the number of hypothesis being tested and thus increases statistical power[20, 28].

To overcome the current limitations, we developed GetisDMR, a genome-wide methylation analysis tool for data obtained from both WGBS and reduced representation bisulfite sequencing (RRBS) experiments. GetisDMR utilizes a beta-binomial regression to account for both epigenetic and sampling variations, and when biological replicates are not available the method assumes the methylated reads following binomial distribution. Therefore, the method can be applied to the situations with and without biological replicates. GetisDMR further uses a local Getis-Ord statistic, which is widely used in identifying statistically significant spatial clusters of high/low values (hot spots), to detect DMRs [29–32]. Our proposed method allows the data to determine the length of identified DMRs. In the following sections, we first lay out the details of the method, and then we compare our method with ComMet, BSmooth and DSS through simulation studies [17, 18, 25]. We further apply our method to two public available mouse dataset[33, 34], and finally we briefly discuss our results.

Method

Beta-binomial Regression

We use a beta-binomial distribution to characterize the methylation data. We assume at each CpG site, the number of methylated reads follows a binomial distribution,$X_{ijk} \sim \text{Binomial}(N_{ijk}, p_{ijk})$, where p_{ijk}, X_{ijk} and N_{ijk} represent the methylation level, the number of methylated reads, and the total number of reads at the k^{th} CpG site of j^{th} sample in the i^{th} treatment group.

To consider the biological variability between different samples under the same treatment condition, we assume the methylation level follows a beta distribution (i.e. $p_{ijk} \sim Beta(\alpha_{ik}, \beta_{ik})$). Therefore, the number of methylated reads (X_{ijk}) at each site has a beta-binomial distribution with probability mass function:

$$P(X_{ijk} = x) = \binom{N_{ijk}}{x} \frac{\Gamma(\alpha_{ik} + x)\Gamma(\beta_{ik} + N_{ijk} - x)\Gamma(\alpha_{ik} + \beta ik)}{\Gamma(N_{ijk} + \alpha_{ik} + \beta_{ik})\Gamma(\alpha_{ik})\Gamma(\beta_{ik})} \tag{1}$$

,where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.

The mean and variance of X_{ijk} are $N_{ijk}\mu_{ik}$ and $N_{ijk}\mu_{ik}(1-\mu_{ik})(1+(N_{ijk}-1)\phi_{ik})$ respectively, where $\mu_{ik} = \dfrac{\alpha_{ik}}{\alpha_{ik} + \beta_{ik}}$ and $\phi_{ik} = \dfrac{1}{\alpha_{ik} + \beta_{ik} + 1}$.

We use beta-binomial regression to model the methylation levels of each CpG site given different treatment conditions and covariates. Specifically, given the number of methylated reads at each CpG sites follows a beta-binomial distribution specified in equation (1), we assume $\mu_{ik} = g(X\eta)$, where g is a logic link function, X is an $s \times t$ design matrix and η is a $t \times 1$ vector of regression parameters. The regression parameters (η) can be interpreted as the log odds ratio for each additional unit increase in the explanatory variable (X).

The beta-binomial regression model is fit for each CpG site, and the parameters (i.e. ϕ_{ik}, μ_{ik} and η) are estimated through maximizing the likelihood function. The significance of the treatment effect is tested using likelihood ratio test.

It is noteworthy that in the case where biological replicates are not available, we assume methylated reads follow a binomial distribution, which is a special case of beta-binomial distribution(i.e $\phi_{ik} = 0$). Under this situation, a χ^2 test of independence or a Fisher's Exact Test will be conducted to test the significance of the treatment effect at each CpG site.

Detection of Differentially Methylated Regions

It is well known that methylation levels are strongly spatially correlated. Hebestreit*et al.* have shown that local smoothing can reduce the variance of the methylation levels, especially for lowly covered CpG sites[20]. Although local smoothing has the potential to increase the power, it is not applicable when biological replicates are not available as the variance of methylation levels can't be estimated based on smoothed methylation levels. What we have observed is that the spatial correlation also preserved in the test statistics (Figure 1). To make use of this information, we proposed a Getis-Ord local statistics based method to detect DMRs. The rationale of using such a statistic is that most of the genetic regions are not differentially methylated, and from the perspective of spatial statistics identifying DMRs is similar to hot-spot detection, where local Getis-Ord statistics have been widely used.

In our method, we first transform the p-values into z-scores ($z_k = \Phi^{-1}(1-p_k)$, with Φ^{-1} being the inverse of a standard normal distribution function), which should approximately follow a normal distribution as most of the CpG sites are not differentially methylated. We further use the Getis-Ord local statistics to detect the DMRs based on z-scores.

The Getis-Ord local statistics is defined as

$$G_i^* = \frac{\sum_{k=1}^{n} \omega_{ik} z_k - \bar{z} \sum_{k=1}^{n} \omega_{ik}}{S\sqrt{\frac{[n\sum_{k=1}^{n} \omega_{ik}^2 - (\sum_{k=1}^{n} \omega_{ik})^2]}{n-1}}} \tag{2}$$

,where $\bar{z} = \sum_k z_k/n$, $S = \sqrt{\sum_k z_k^2/n - \bar{z}^2}$, and ω_{ik} is the weight parameters between i^{th} CpG site and the k^{th} CpG sites [29, 30].

The correlation of the local Getis-Ord local statistics between the i^{th} CpG site and the j^{th} CpG sites is

$$corr(G_i^*, G_j^*) = \frac{n\sum_k \omega_{i,k}\omega_{j,k} - W_i^* W_j^*}{\sqrt{[nS_{1i} - (W_i^*)^2][nS_{1j} - (W_j^*)^2]}} \tag{3}$$

,where $W_i^* = \sum_k \omega_{ik}$ and $S_{1i} = \sum_k \omega_{ik}^2$.

We further denote $\vec{G^*} = [G_1^*, G_2^*, G_3^*..., G_K^*]'$, and therefore asymptotically $\vec{G^*} \sim MVN(\vec{0}, \Xi)$ with $\Xi_{i,j} = corr(G_i^*, G_j^*)$.

The weight parameters (ω_{ik}) for the Getis-Ord local statistics G_i^* is determined based on the correlation between the z-scores. The correlations between the z-scores are estimated using the empirical variogram ($corr(z_i, z_j) = (var(z) - \gamma(d))/var(z)$,where z_i and z_j is d base pair apart and $\gamma(d)$ is the semivariogram). Let Σ represent the estimated variance covariance matrix, Σ_i be the covariance vector between i^{th} CpG site and the rest of the CpG sites, and $\Sigma_{-i,-i}$ be the variance-covariance matrix between all CpG sites except the i^{th} CpG site. The weight parameters (ω_{ik}) is determined by $\Sigma_i \Sigma_{-i,-i}^{-1}$.

For the whole genome sequencing data, the estimation of semivariogram at every possible value of distance between two CpG sites and the inverse of the covariance matrix are computationally demanding. However, as shown in Figure 1, the correlation reduces substantially as physical distance between the two CpG sites increases. A set of commonly used models in spatial statistics, such as exponential model, Spherical variogram, and Matern class of models, can be used to model this relationship. To reduce the computational burden, we only consider a limited range over which approximately a linear relationship between the physical distance and correlation exists. Therefore, instead of calculating empirical semivariogram for every possible distances, here we use a linear function to model the relationship between semivariogram and the physical distance. We further employ a data-adaptive method to determine the range over which the linear model predicts the correlation reasonably well. Specifically, we randomly select several chuck of genomes with 100000 bps in length, and we fit linear models with distances increasing continuously. The distance (d_{select}) that gives the maximum adjusted R^2 is chosen. The correlation between CpG sites with distance larger than d_{select} will be zero, which substantially reduces the computational burden. The correlation between z-scores of two CpG sites is calculated as below:

$$corr(z_k, z_j) = \begin{cases} x = & 0 \text{ if } d > d_{select} \\ y = & f(d) \text{ if } d \leq d_{select} \end{cases}$$

,where $d = |pos_k - pos_j|$, pos_k is the physical location of the k^{th} CpG Site, and $f(d)$ is a linear function which captures the relationship between correlation and physical distance.

To detect DMRs, we define a region-based spatial statistics, $G_{pos_i,pos_j} = \frac{\sum_{j \in S} G_j^*}{|S|}$, where S is the set of CpG site with physical location within (pos_i, pos_j), and $|S|$ is the total number of CpG site in S. To determine the boundary of methylated region to be evaluated, we also adopt a data-adaptive method. Instead of evaluating every region with a fixed length, we only evaluate the significance of the region given a pre-determined number of CpG site with G_i^* larger than a pre-specified cutoff value. Asymptotically, $G_{pos_i,pos_j} \sim N(0, \psi^2)$, where $\psi^2 = \sum_{(i,j) \in S} cov(G_i^*, G_j^*)$. The significance of the test statistics (G_{pos_i,pos_j}) can also be evaluated through permutation test.

Our test statistic explicitly makes use of the correlation between nearby CpG sites, which can improve the power of the test statistics. This is especially true when some CpG sites have relatively low coverage and the z-scores are not quite accurate for such sites. Instead of testing the significance of each CpG site, we focus on detecting DMRs, which can substantially reduce the number of hypothesis being tested and boost the power of DMR detection. Moreover, as the spatial correlations between z-scores are used to increase the precision of the estimates, our method can also be directly applied to the situation where biological replicates are not available (i.e. 1 treatment vs 1 control).

Simulation

We evaluated the performance of the proposed GetisDMR method with three commonly adopted methods (i.e BSmooth[18], ComMet[17] and DSS [25]) under various conditions. In simulation one, we compared the performance of GetisDMR with ComMet when biological replicates are not available. In simulation two, we compared the performance of GetisDMR with BSmooth, ComMet and DSS when biological replicates were available. In the third simulation, we compared the GetisDMR method with the other three methods when confounders/other covariates that also influenced the methylation levels were present. The performance of these methods were evaluated based on both sensitivity(P(dectected DMR|true DMR)) and positive predictive value(PPV,P(true DMR|detected DMR)). Similar to the definition used in ComMet[17], a true positive DMR was defined as a true DMR that overlapped with a detected DMR in a certain proportion of their lengths. We further defined the sensitivity as the proportion of true DMRs being detected. The PPV was defined as the proportion of detected DMRs that overlapped with the true DMRs larger than a certain proportion of their lengths. For all the below simulations, we evaluated the sensitivity and PPV with 50% overlaps.

Scenario I

In the first set of simulations, we evaluated the performance of GetisDMR under a variety of conditions where the biological replicates were not available. Most of the current available methods are designed for the situation where there are biological replicates for each experimental condition. However, in practice, due to budget and other issues[21, 22], biological replicates are not always available, which makes the methods such as BSmooth not applicable. To the best of our knowledge, ComMet, which employs a hidden Markov chain model to detect DMRs, is the only method that has been claimed to be able to detect DMRs without biological replicates. Therefore, in this set of simulations we compared the sensitivity and PPV of our method with the ComMet. To mimic the methylation levels and the spatial correlations between adjacent CpG sites, we used a real dataset from a whole genome bis-seq experiment[33] and placed methylation level differences of various intensities and lengths. Specifically, we chose one sample from the experiment to serve as control(the cortex sample), and kept the methylation level of each CpG site of the control sample the same as the original data. Another sample from the experiment(the brain sample) was served as case, where the methylation levels of each CpG site were simulated. To simplify the simulation, without loss of generality we only focused on chromosome 19. In total, we put 400 DMRs on chromosome 19 with half of the DMRs were up-regulated and the other half were down-regulated. Because of the fact that the length of DMRs reported in the literature usually ranges from a few hundreds to a few thousands bps[27], the lengths of the DMRs were sampled from a truncated Gaussian distribution ($L \sim N(150, 100^2)$, with $50 < L < 4000$). Within each DMR region, the fraction of DMCs were varied from 0.7 to 1, and for each DMC site the differences in methylation level between the case and control samples were varied from 0.1 to 0.4. For the other CpG sites, the methylation levels in the case sample were set the same as those in the control sample. The number of methylated reads for each CpG site in both case and control samples were simulated from binomial distribution with the total number of reads at each CpG site equal to the total number of reads from each sample at the same CpG site. For each of the simulated model, we generated 50 replicates, and we analyzed each replicate by using the proposed GetisDMR method and the ComMet[17].

Scenario II

In the second set of simulations, we evaluated the performance of GetisDMR under a variety of conditions where biological replicates were available, and we further compared the performance of GetisDMR with ComMet, BSmooth and DSS[17, 18, 25]. Similar to Scenario I, we used a real dataset from a whole genome bis-seq experiment to mimic the methylation levels and the spatial correlations between adjacent CpG sites[33]. We randomly chose 6 samples to serve as control samples, and the remaining samples to serve as case samples. The methylation levels in control samples were set the same as the methylation level in one of the randomly selected control sample. Similar to Scenario I, we put 400 DMRs on chromosome 19 with half of the DMRs up-regulated and the remaining down-regulated. The length of the DMRs were sampled from a truncated Gaussian distribution ($L \sim N(150, 100^2)$, with $50 < L < 4000$). Within each DMR region, we varied the fraction (ranging from 0.7 to 1) of DMCs and the differences (ranging from 0.1 to 0.4) in methylation levels between cases and controls. For non-differentially methylated regions, the methylation levels in the case samples were set the same as those in the control samples. The number of methylated reads for each CpG site in both case and control samples were simulated from binomial distribution with the total number of reads at each CpG site equal to the total number of reads from each sample at the same CpG site. For each scenario considered we generated 50 replicates, and analyzed each replicate by using the proposed method, the ComMet[17], the BSmooth[18], and the DSS[25].

Scenario III

In this set of simulations, we evaluated the performance of the proposed method when confounders/covariates were present. Similar to Scenario II, 6 samples were served as controls, and the remaining samples were served as cases. The methylation levels at each CpG site were simulated using $log(\frac{p_{tjk}}{1-p_{tjk}}) = \mu_{tk} + X_{tj}\beta$, where p_{tjk} is the methylation level of sample j ($j = 1, 2, 3, \ldots, 6$) in group t ($t = 0$, control and $t = 1$, case) at CpG site k, and X_{tj} is the covariate vector for sample j in group t. For simplicity, we only simulated one binary covariate, and we varied its effect size(log odds ratio) ranging from 0.1 to 0.5. The μ_{0k} for control samples was set at $log(\frac{p_{0k}}{1-p_{0k}})$ where p_{0k} is the observed methylation level in one of the randomly selected control sample from the experiment. The μ_{1k} for DMCs in the case samples was set at $log(\frac{p_{0k}}{1-p_{0k}}) + d$, where d was set at 0.2 and 0.3. The μ_{1k} for the other sites in the case samples was set at $log(\frac{p_{0k}}{1-p_{0k}})$. Within each DMR region, the fraction of DMCs was set at 0.8 and 1. Similar to Scenario II, we put 400 DMRs on the data, and the length of each DMR was sampled from $L \sim N(150, 100^2)$, with $50 < L < 4000$. For each scenario considered we generated 50 replicates, and analyzed each replicate by using the proposed GetisDMR method, the ComMet[17], the BSmooth[18], and the DSS[25].

Results

Scenario I

The sensitivity and PPV of Scenario I are summarized in Figure 2. As expected, with the increase in the differences of methylation levels between the treatment and control groups, the sensitivity and PPV increased for both of the methods. For example, with 90% DMCs in a DMR region, the sensitivity of GetisDMR increased from 0.0063 to 0.51 and the PPV increased from 0.29 to 0.92 as the differences in methylation levels increased from 0.1 to 0.4. Similarly, the sensitivity of ComMet increased from 0.0022 to 0.22 and the PPV increased from 0.017 to 0.78 as the differences in methylation levels increased. Consistent with what we had expected, as the fraction of DMCs in a DMR region increased, the sensitivity and PPV increased as well. It is worth noting that when the differences in methylation levels were small, the increase in the fraction of DMCs within a DMR region will increase both the sensitivity and PPV for GetisDMR, while it has little effect for ComMet. This could be largely explained by the fact that when biological replicates were not available, GetisDMR utilizes spatial correlations between z-scores(calculated from the Fisher's Exact Test) from nearby CpG sites to stabilize the estimators(G_i^*) and thus boosts the power for DMR detection. As shown in Figure 2, both the sensitivity and PPV of the proposed method were higher than that of ComMet under all the situations considered in this set of simulations.

Scenario II

The sensitivity and PPV of Scenario II are summarized in Figure 3. Similar to the results from Scenario I, both sensitivity and PPV increased with the increase in the fraction of DMCs within a DMR region and the increase in the differences of methylation levels. As expected, given the same level of differences in methylation levels and DMC fractions, both the sensitivity and PPV were higher when biological replicates were available. We noticed that while the PPV of GetisDMR was always significantly higher than that of ComMet, the sensitivity of GetisDMR was slightly lower than that of ComMet when the differences of methylation level were relatively high. This is partly due to the fact that when biological replicates were available, the GetisDMR takes both the biological and sampling variations into account. However, under the setting of Scenario II, the methylation levels were set the same for all the samples in both the treatment and control groups(i.e. no biological variability), indicating a binomial distribution should be sufficient to model the data. However, GetisDMR assumes there is biological variability and estimates an additional parameter to account for it, which could result in the loss of efficiency and this is especially true when the differences in methylation level are high. On contrary, ComMet assumes no biological variability and pools samples under the same experimental condition, which boosts the power of DMR detection under the current setting. Indeed, when there is no biological variability, a logistic regression has higher power than that of the beta-binomial regression. However, in most of cases we don't know if the biological variability is present, and failing to consider such a variation can result in inflated type I error. Nevertheless, as the differences in sensitivity between the two methods are not substantial (Figure 3), we recommend to use the beta-binomial regression to account for the potential biological variability. The BSmooth attained lower sensitivity and PPV than both of the proposed method and the ComMet among all the situations except the case when the difference in methylation levels was small(i.e. $d = 0.1$). While the sensitivity of the DSS method is lower than both GetisDMR and the ComMet, the PPV of the DSS method is higher than that of the ComMet adn BSmooth methods but lower than that of the GetisDMR method.

Scenario III

The sensitivity and PPV of Scenario III are summarized in Figure 4. As expected, with the increase in the effects of covariates, both the sensitivity and PPV for ComMet decreased. For example, when the odds ratio for the binary confounding variable changed from 1.11 to 1.64, with 80% CpG sites being differentially methylated within a DMR region and the differences in methylation level setting at 0.3, the sensitivity for ComMet changed from 0.83 to 0.75 and the PPV changed from 0.32 to 0.19, whereas under the same setting the sensitivity for GetisDMR changed from 0.87 to 0.86 and the PPV changed from 0.81 to 0.80. The performance of

GetisDMR is largely robust against the presence of other confounding variables, and this could be explained by the fact that GetisDMR adopts a beta-binomial regression in which the covariates can be explicitly modeled. When other factors(e.g. age) affect the methylation levels, the treatment effect estimates can be substantially biased when such factors are not taken into account. The ComMet method ignores the potential effects of confounding factors, and pools the data under the same experiment condition into one sample. Therefore, it is subject to low power as the effects of confounding factors increase. Similar to Scenario II, the BSmooth method performed worse than ComMet and GetisDMR, but its performance was relative robust against the effects of confounding variables. DSS method tended to have higher PPV than the ComMet and BSmooth, but its sensitivity was lower than sensitivities from both the ComMet and BSmooth methods. Among all the situations considered in this set of simulations, the GetisDMR had higher sensitivity and PPV than all the other methods regardless of the effects of confounding variables, the differences in methylation levels and the fraction of DMCs within a DMR.

Real Data Application

The mouse brain and kidney dataset

We applied the proposed method to a public available mouse dataset[33] to investigate the methylation patterns in the bone marrow and kidney tissues of adult mice. The dataset, which included WGBS data from 17 C57Bl/6 mouse tissues, was downloaded from the Gene Expression Omnibus (GEO ID: GSE42836). In data preprocessing, the segemhl[35] software was used to map the 36 bp single-end reads to mm9 genome, and only reads with unique mapping positions were kept for further analyses. The hits on the positive stand cytosine and negative strand cytosine at one CpG site were summed together to get the total number of reads and methylated reads. We applied our GetisDMR method to compare the methylation levels between the bone marrow and kidney tissues. In total, our method has detected 116912 DMRs. The median length of the identified DMR is 458, and the median number of CpG sites within a DMR region is 8 (the details of the identified regions were summarized in supplementary Table S1). To investigate whether the identified DMRs are located in the genomic regions that are relevant to the normal functioning or development of kidney and/or bone marrow, the Genomic Regions Enrichment of Annotations Tool (GREAT) was used to infer the biological functions based on identified DMRs with more than 15 CpG sites harbored ($n = 15172$). Two types of annotations, gene expressions at different tissues of various mouse development stages and the mouse phenotypes that are affected upon gene malfunctioning, were used to explore the biological functions. We further ranked significance of enrichment according to the binomial distribution-based p-values obtained from the GREAT analyses. The details of the analyses are summarized in supplementary Table S2. Using gene expression annotation, we found that the genes presumably controlled by the detected DMRs tended to be over-expressed in kidney. Among the top terms, TS23_visceral organ, TS23_metanephros, TS23_renal-urinary system, and TS23_renal cortex, are related to the forming of normal kidney structures during mouse development at Theiler stage (TS) 23. Among the top 29 combinations of tissues and development stages that exhibit the most significant enrichment, 59% of them are related to kidney or bone marrow systems (Figure 5A, Table S2). More than 50% terms are directly related to kidney or bone marrow functioning even tracing down to the 60^{th} term(Figure 5B, Table S2). Using phenotype annotation, we found that genes located on or near the detected DMRs were even more enriched in either bone marrow or kidney systems. Among the top 29 terms, 83% of them are directly related to the two tissues(Figure 5C). More than 77% of the terms are related to the two tissues when we trace down to the 60^{th} term (Figure 5D).

The mouse frontal cortex dataset

We also applied our GetisDMR method to compare methylation patterns between neuron and non-neuron samples from mouse frontal cortex in a study of methylation profiles of mammalian brain[34]. The dataset was downloaded from GEO(GEO ID: GSE47966), and we used the same strategy as Lister *et al.* to obtain the number of reads and methylated reads[34]. To reduce the effects of confounding variables, we adjusted for age and gender (6 week and 12 month old females, and 7 week old male) in our analyses. Totally, 371092 DMRs were detected with the median width of 630 bp and median number of CpG sites of 7 (the details of identified DMRs are summarized in Table S3). The DMRs, comprised of more than 25 CpG sites($n = 12422$), were used to explore the biological functions via the GREAT database(Table S4). We also ranked the significance of enrichment according to the binomial distribution-based p-values from GREAT analyses using both gene expression and mouse phenotype annotations. As shown in Figure 6, among the top 29 terms, 76% and 59% of them are directly related to neural system function according to gene expression and mouse phenotype annotation, respectively. When we trace down to the 60^{th} highest enrichment term, the percentage of terms directly related to neural systems still reaches around 72% for gene expression and 61% for mouse phenotype annotation (Figure 6).

Discussion

In this work, we present a novel statistical method(i.e. GetisDMR) to detect DMRs from both the RRBS and WGBS datasets. The proposed method utilizes the beta-binomial regression model to account for confounding effects, biological and sampling variations. It further uses a local Getis-Ord statistic to combine information from nearby CpG sites. The region-wise overall test statistic allows for the detection of DMRs directly, which reduces the number of hypothesis being tested and increases the power of the proposed method. Through extensive simulations, we have demonstrated the proposed method had comparable or higher sensitivity and positive predictive values in detecting DMRs than ComMet, BSmooth and DSS[17, 18, 25]

One strength of the proposed method is that it utilizes the beta-binomial regression framework and models the methylation levels corresponding to experimental as well as other independent and potential confounding factors. As found by Boks *et al.* that DNA methylation levels could be influenced by confounder[36], such as age and gender, and thus it is important to take these confounding variables into consideration to avoid bias effect estimates. The beta-binomial regression model can also account for the biological variability and therefore reduces the rate of false positive compared with a standard binomial regression model. Compared with the beta regression model, which is a widely used technique to model outcomes within the range between 0 and 1, the beta-binomial regression increases power even at moderate coverage as it takes the coverage depth into consideration.

Another strength of our method is that we adopts region-wise test statistics based on local Getis-Ord statistic to directly detect DMRs. Currently, most of the existing methods focus on detecting DMCs and then identify DMRs based on some pre-specified criteria. Although the GetisDMR can be used to detect DMCs as we performed the beta-binomial regression for each CpG site independently, our method focuses on detecting DMRs directly and we provide statistical inference for the detected DMRs. It has been reported by various researchers that DNA-methylation levels are spatially correlated along the genome[37, 38], and our own data also shows that the p-values from the beta-binomial regression are spatially correlated. In the GetisDMR method, we first derived a z-score and then we employed the local Getis-Ord statistics to account for the spatial correlation in DMR detection. Detecting DMRs from bis-seq data is similar to that of hot spot detection, where local Getis-Ord statistics have been widely used. Indeed, most of the local Getis-Ord statistics calculated based on the z-scores are approximately normally distributed as most of the CpG sites are not DMCs. Any regions with abnormal large or small local Getis-Ord statistics may indicate a spatial hot spot and in our case it indicates a DMR. One of the challenges of defining the local Getis-Ord statistics is to calibrate the spatial correlation among z-scores. Although spatial correlations decrease with the increase in the distance between CpG sites, the magnitude of the correlation is quite data dependent. In our method, instead of pre-specifying a weight function to account for the correlation, we use a kernel function to capture this relationship and let the data to determine the magnitude of the parameters. This makes our method robust to various datasets with different underlying spatial correlation structures among nearby CpG sites. We further showed the asymptotic distribution of the region-wise test statistic (i.e. $\vec{G^*} \sim MVN(\vec{0}, \Xi)$), and controlled the region-wise false discovery rate. Therefore, our method can well control the false positive rate and it provides statistical inference not only for DMCs but also DMRs.

It is worth mentioning that the proposed method is quite flexible to the study design as the DMR detection only requires p-values. For example, for studies where replicates are not available, the Fisher's Exact Test could be used to calculate p-values for each CpG site. We could then use the same procedure to detect DMRs. For studies with more than 2 treatment groups, the likelihood ratio tests could be performed to assess the treatment effects and the same local Getis-Ord statistic based procedure could be used to detect DMRs. For longitudinal studies or studies with clustered effects, robust estimation of the beta-binomial model parameters can be used and the same local Getis-Ord statistic based procedure could be employed to infer DMRs[39].

In real data application, we applied the proposed method to two public available datasets[33, 34]. The first dataset is designed to investigate methylation levels at different tissues of mouse, and it only has one sample per tissue(i.e. one sample from the bone marrow tissue and one sample from kidney tissue)[33]. In total, our method identified 116912 DMRs. The further analyses using the GREAT revealed that most of genes located either on or close to the identified DMRs are related to kidney or bone marrow systems. Using gene expression annotation, more than 50% terms selected by the GREAT is related to the two tissues. Similarly, using the phenotype annotation, 77% of terms are directly associated with the two tissues. We also applied our method to compare methylation levels between neuron and non-neuron samples from mouse frontal cortex[34], where the effects of age and gender have been controlled for. In total, we have identified 371092 DMRs and most of them are biologically relevant. Using the GREAT, the percentages of terms directly related to neural systems reach 72% and 61% for gene expression and mouse phenotype annotations, respectively. In both scenarios, the DMRs detected by our GetisDMR method showed significant enrichment of genes in the desired tissues, as well as the direct association with the expected mouse phenotypes. Although further studies are needed to confirm the biological functions of these identified DMRs, our findings shed light on the mehtylation patterns in different mouse tissues.

In conclusion, we have developed a powerful method to detect DMRs for the analysis of both RRBS and WGBS datasets. The GetisDMR method detects DMRs based on region-wise statistics, which utilize the spatial correlation between nearby CpG site. Our method achieves relatively high sensitivity and PPV, and it has the potential to be applicable for more sophisticated study designs, and studies without biological replicates.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

YW designed and implemented the algorithm. FC carried out the simulation studies. FC, QZ, ZY and ZL analyzed the two real datasets. All authors read and approved the final manuscript for publication.

Acknowledgments

The Project was supported by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry and the National Natural Science Foundation of China (Award No. 81502887). and FRDF.

Author details

[1]Department of Statistics, University of Auckland, 38 Princes Street, 1010 Auckland, New Zealand. [2]Stem Cell XXX, Dalian Medical University, No.8 Lv Shun Nan Road, XXX, 116000 Dalian, China.

References

1. Li, E., Beard, C., Jaenisch, R.: Role for dna methylation in genomic imprinting. Nature **366**(6453), 362–5 (1993)
2. Ehrlich, M.: Dna methylation in cancer: too much, but also too little. Oncogene **21**(35), 5400–13 (2002)
3. Santos, F., Hendrich, B., Reik, W., Dean, W.: Dynamic reprogramming of dna methylation in the early mouse embryo. Dev Biol **241**(1), 172–82 (2002)
4. Suzuki, M.M., Bird, A.: Dna methylation landscapes: provocative insights from epigenomics. Nat Rev Genet **9**(6), 465–76 (2008)
5. Deaton, A.M., Bird, A.: Cpg islands and the regulation of transcription. Genes Dev **25**(10), 1010–22 (2011)
6. Choy, M.K., Movassagh, M., Goh, H.G., Bennett, M.R., Down, T.A., Foo, R.S.: Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics **11**, 519 (2010)
7. Hendrich, B., Bird, A.: Identification and characterization of a family of mammalian methyl-cpg binding proteins. Mol Cell Biol **18**(11), 6538–47 (1998)
8. Bird, A.P., Wolffe, A.P.: Methylation-induced repression–belts, braces, and chromatin. Cell **99**(5), 451–4 (1999)
9. Kassner, I., Barandun, M., Fey, M., Rosenthal, F., Hottiger, M.O.: Crosstalk between set7/9-dependent methylation and artd1-mediated adp-ribosylation of histone h1.4. Epigenetics Chromatin **6**(1), 1 (2013)
10. Shen, H., Laird, P.W.: Interplay between the cancer genome and epigenome. Cell **153**(1), 38–55 (2013)
11. Sharma, S., Kelly, T.K., Jones, P.A.: Epigenetics in cancer. Carcinogenesis **31**(1), 27–36 (2010)
12. Clark, S.J., Statham, A., Stirzaker, C., Molloy, P.L., Frommer, M.: Dna methylation: bisulphite modification and analysis. Nat Protoc **1**(5), 2353–64 (2006)
13. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., Mason, C.E.: methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. Genome Biol **13**(10), 87 (2012)

14. Schultz, M.D., Schmitz, R.J., Ecker, J.R.: 'leveling' the playing field for analyses of single-base resolution dna methylomes. Trends Genet **28**(12), 583–5 (2012)
15. Dolzhenko, E., Smith, A.D.: Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics **15**, 215 (2014)
16. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., Ecker, J.R.: Human dna methylomes at base resolution show widespread epigenomic differences. Nature **462**(7271), 315–22 (2009)
17. Saito, Y., Tsuji, J., Mituyama, T.: Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. Nucleic Acids Res **42**(6), 45 (2014)
18. Hansen, K.D., Langmead, B., Irizarry, R.A.: Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol **13**(10), 83 (2012)
19. Jaffe, A.E., Feinberg, A.P., Irizarry, R.A., Leek, J.T.: Significance analysis and statistical dissection of variably methylated regions. Biostatistics **13**(1), 166–78 (2012)
20. Hebestreit, K., Dugas, M., Klein, H.U.: Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics **29**(13), 1647–53 (2013)
21. Hirst, M., Marra, M.A.: Next generation sequencing based approaches to epigenomics. Brief Funct Genomics **9**(5-6), 455–65 (2010)
22. Stevens, M., Cheng, J.B., Li, D., Xie, M., Hong, C., Maire, C.L., Ligon, K.L., Hirst, M., Marra, M.A., Costello, J.F., Wang, T.: Estimating absolute methylation levels at single-cpg resolution from methylation enrichment and restriction enzyme sequencing methods. Genome Res **23**(9), 1541–53 (2013)
23. Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., Wei, C.L.: Dynamic changes in the human methylome during differentiation. Genome Res **20**(3), 320–31 (2010)
24. Beyan, H., Down, T.A., Ramagopalan, S.V., Uvebrant, K., Nilsson, A., Holland, M.L., Gemma, C., Giovannoni, G., Boehm, B.O., Ebers, G.C., Lernmark, A., Cilio, C.M., Leslie, R.D., Rakyan, V.K.: Guthrie card methylomics identifies temporally stable epialleles that are present at birth in humans. Genome Res **22**(11), 2138–45 (2012)
25. Feng, H., Conneely, K.N., Wu, H.: A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res **42**(8), 69 (2014)
26. Park, Y., Figueroa, M.E., Rozek, L.S., Sartor, M.A.: Methylsig: a whole genome dna methylation analysis pipeline. Bioinformatics **30**(17), 2414–22 (2014)
27. Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A., Li, W.: Moabs: model based analysis of bisulfite sequencing data. Genome Biol **15**(2), 38 (2014)
28. Bock, C.: Analysing and interpreting dna methylation data. Nat Rev Genet **13**(10), 705–19 (2012)
29. Getis, A., Ord, J.K.: The analysis of spatial association by use of distance statistics. Geographical Analysis **24**(3), 189–206 (1992)
30. Ord, J.K., Getis, A.: Local spatial autocorrelation statistics - distributional issues and an application. Geographical Analysis **27**(4), 286–306 (1995)
31. Ord, J.K., Getis, A.: Testing for local spatial autocorrelation in the presence of global autocorrelation. Journal of Regional Science **41**(3), 411–432 (2001)
32. Bhunia, G.S., Kesari, S., Chatterjee, N., Kumar, V., Das, P.: Spatial and temporal variation and hotspot detection of kala-azar disease in vaishali district (bihar), india. BMC Infect Dis **13**, 64 (2013)
33. Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., Ren, B.: Epigenetic memory at embryonic enhancers identified in dna methylation maps from adult mouse tissues. Nat Genet **45**(10), 1198–206 (2013)
34. Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghighi, F.G., Sejnowski, T.J., Behrens, M.M., Ecker, J.R.: Global epigenomic reconfiguration during mammalian brain development. Science **341**(6146), 1237905 (2013)
35. Otto, C., Stadler, P.F., Hoffmann, S.: Fast and sensitive mapping of bisulfite-treated sequencing data. Bioinformatics **28**(13), 1698–704 (2012)
36. Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S., Ophoff, R.A.: The relationship of dna methylation with age, gender and genotype in twins and healthy controls. PLoS One **4**(8), 6767 (2009)
37. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., Beck, S.: Dna methylation profiling of human chromosomes 6, 20 and 22. Nat Genet **38**(12), 1378–85 (2006)
38. Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B., Feinberg, A.P.: Comprehensive high-throughput arrays for relative methylation (charm). Genome Res **18**(5), 780–90 (2008)
39. Pashkevich, M.A., Kharin, Y.S.: Robust estimation and forecasting for beta-mixed hierarchical models of grouped binary data. SORT **28**(2), 125–160 (2004)

Figures

Figure 1: Spatial correlation among nearby CpG sites

Figure 2: Sensitivity and Positive Predictive Value of Simulation 1.

Figure 3: Sensitivity and Positive Predictive Value of Simulation 2.

Figure 4: Sensitivity and Positive Predictive Value of Simulation 3.

Figure 5: Biological annotations of GetisDMR identified DMRs from bone marrow and kidney methylome data.

Figure 6: Biological annotations of GetisDMR identified DMRs from neuron and non-neuron methylome data data.

Additional Files

DMRs from bone marrow and kidney methylome data.
Biological annotations of DMRs from bone marrow and kidney methylome data.
DMRs from neuron and non-neuron methylome data.
Biological annotations of DMRs from neuron and non-neuron methylome data.