Bioinformatics, 2016, 1–9
doi: 10.1093/bioinformatics/btw497
Advance Access Publication Date: 4 August 2016
Original Paper

OXFORD

Genome analysis

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

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

[1]Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Cancer Center, Dalian Medical University, Dalian 116044, China and [2]Department of Statistics, University of Auckland, Auckland 1142, New Zealand

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 7, 2016; revised on July 11, 2016; accepted on July 15, 2016

Abstract

Motivation: DNA methylation is an important epigenetic modification that has essential role 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 treatment conditions. Although numerous tools have been proposed to detect differentially methylated single CpG site (DMC) between samples, methods for direct DMR detection, especially for complex study designs, are largely limited.

Results: We present a new software, GetisDMR, for direct DMR detection. We use beta-binomial regression to model the whole-genome bisulfite sequencing data, where variations in methylation levels and confounding effects have been accounted for. We employ a region-wise test statistic, which is derived from local Getis-Ord statistics and considers the spatial correlation between nearby CpG sites, to detect DMRs. Unlike existing methods, that attempt to infer DMRs from DMCs based on empirical criteria, we provide statistical inference for direct DMR detection. Through extensive simulations and an application to two mouse datasets, we demonstrate that GetisDMR achieves better sensitivities, positive predictive values, more exact locations and better agreement of DMRs with current biological knowledge.

Availability and Implementation: It is available at https://github.com/DMU-lilab/GetisDMR.

Contacts: y.wen@auckland.ac.nz or zhiguangli@dlmedu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 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 (Deaton and Bird, 2011; Ehrlich, 2002; Li *et al.*, 1993; Santos *et al.*, 2002; Suzuki and Bird, 2008). The presence of large scale aberrant DNA methylation pattern, typically with site-specific hypermethylation in tumor suppressor genes and global hypo-methylation in oncogenes compared to normal tissue, is a hallmark feature of various types of cancers (Ehrlich, 2002; Sharma *et al.*, 2010).

The whole-genome bisulfite sequencing (WGBS), which combines the bisulfite treatment with next generation sequencing, becomes the state-of-the-art technology in investigating DNA methylation pattern at single base resolution with relatively high coverage across multiple samples. The bisulfite 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 (Clark *et al.*, 2006). Methylation proportion at each CpG site is usually defined as the

proportion of molecules with cytosine methylated ($\frac{C}{C+T}$) and is used to summarize the pattern of DNA methylation (Akalin *et al.*, 2012; Dolzhenko and Smith, 2014; Schultz *et al.*, 2012).

Over the past few years, a number of approaches have been proposed for assessing differentially methylated regions (DMRs) from WGBS data. One of the most straightforward method is to use the Fisher's Exact Test to compare the methylation proportions among different treatment groups at each CpG site (Lister *et al.*, 2009). Recently, Saito *et al.* developed the ComMet, which is built based on hidden Markov models and designed to detect DMRs between a pair of samples (Saito *et al.*, 2014). Though these methods are easy to implement and can compare a pair of samples obtained either directly from the experiment or by pooling together samples under the same experimental condition, these methods do not take the between sample variations into account and cannot adjust for confounding effects when replicates are available (Hansen *et al.*, 2012; Jaffe *et al.*, 2012).

Converging evidences suggest that close-by CpG sites tend to have similar methylation proportions (Hansen *et al.*, 2012; Hebestreit *et al.*, 2013). With the assumption that methylation proportions change smoothly along the genome and the adjacent CpG sites have similar methylation proportions, various smoothing based methods have been proposed to detect DMRs (Hansen *et al.*, 2012). 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 proportion estimates, especially for CpG sites with low coverage. For example, the BSmooth method first estimates the methylation proportions with a local-likelihood smoother, and then performs the statistical test using a signal-to-noise statistic (Hansen *et al.*, 2012). 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 first employs a local smoothing technique and then detects DMRs based on the smoothed methylation proportion estimates (Hebestreit *et al.*, 2013). 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 be differentially methylated, and thus it is only suitable for detecting DMRs in targeted bisulfite sequencing data. Though smoothing based methods make use of information from adjacent CpG sites, in most cases they require biological replicates and thus cannot be applied to the datasets without replicates. Currently, the WGBS is quite costly, which prohibits the obtainment of multiple replicates for different experimental conditions given limited budget (Hirst and Marra, 2010; Stevens *et al.*, 2013). It is quite common that some of the biological replicates are combined into one sample before library generation for sequencing experiments (Laurent *et al.*, 2010; Saito *et al.*, 2014). Moreover, there are situations where biological replicates are hard to obtain, especially in retrospective studies (Beyan *et al.*, 2012).

Regression based methods have also been proposed to detect DMRs. For example, MethylKit assumes that the number of methylated reads follows a binomial distribution, and models the methylated reads within the logistic regression framework (Akalin *et al.*, 2012). The *P*-values are calculated and multiple comparisons are adjusted using a sliding linear model method. As methylation proportions vary significantly across individuals, failure to consider the variability across individuals may result in inflated type-I error (Hansen *et al.*, 2012; Jaffe *et al.*, 2012). Beta-binomial regression has been recently introduced to model methylation proportions in WGBS data, as it can take both the sampling and epigenetic variations into account (Dolzhenko and Smith, 2014; Feng et al., 2014; Park *et al.*, 2014). For example, DSS method uses a lognormal-beta-binomial Bayesian hierarchical model to describe the methylated reads, and the DMR is defined as the CpG site with *P*-value less than a pre-specified threshold (Feng *et al., 2014*). 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 modeled in the DMR detection. This 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 single CpG (DMC) site (Park *et al.*, 2014). Although the methylSig can be used to identify DMCs, 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*-values of each CpG site and then combines the information from *P*-values within 200 base pairs(bp) (Dolzhenko and Smith, 2014). 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 (Akalin *et al.*, 2012; Dolzhenko and Smith, 2014). However, compelling evidences suggest that the length of DMRs can range from a few base pairs to thousands of base pairs, and a fixed length of DMR certainly contradicts with the existing biological knowledge (Sun *et al.*, 2014). Compared with detecting DMCs, DMR detection has several advantages. First, locating the regions with multiple DMCs are one of the most basic goals for methylation studies. Second, as pointed out by Bock, after adjusting for multiple comparisons for DMC detection, only the strongest differences tend to remain significant (Bock, 2012). Targeting at detecting DMR rather than single DMC can substantially reduce the number of hypothesis being tested and thus increases the statistical power (Bock, 2012; Hebestreit *et al.*, 2013).

To overcome the current limitations, we develop GetisDMR, a genome-wide methylation analysis tool for direct DMR detection. GetisDMR utilizes a beta-binomial distribution (with biological replicates) or binomial distribution (without biological replicates) to model the methylated reads. When bioloigcal replicates are available, it also adopts a regression framework to account for the potential confounding effects. It further uses a local Getis-Ord statistic, which is widely used in identifying statistically significant spatial clusters of high/low values (hot spots) (Bhunia et al., 2013; Getis and Ord, 1992; Ord and Getis, 1995, 2001), to detect DMRs. The length of detected DMRs is determined by the data, and the statistical inference of the detected DMRs is also provided. 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 (Feng *et al.*, 2014; Hansen *et al.*, 2012; Saito *et al.*, 2014). We further apply our method to two public available mouse datasets (Hon *et al.*, 2013; Lister *et al.*, 2013), and finally we briefly discuss our results.

2 Methods

2.1 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 proportion, the number of methylated reads, and the total number of reads at the kth CpG site of jth sample in the ith treatment group.

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

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

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 $\mu_{ik} = \frac{\alpha_{ik}}{\alpha_{ik}+\beta_{ik}}$ and $\phi_{ik} = \frac{1}{\alpha_{ik}+\beta_{ik}+1}$.

We use beta-binomial regression to model the methylation proportions at each CpG site given different treatment conditions and covariates. Specifically, given the number of methylated reads at each CpG site follows a beta-binomial distribution specified in Equation (1), we assume $\mu_{ik} = g(Z\eta)$, where g is a logistic link function, Z is a design matrix and η is a 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 (i.e. Z).

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 the likelihood ratio test.

It is noteworthy that in the case where biological replicates are not available, we assume the number of methylated reads follows 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.

2.2 Detection of differentially methylated regions

It is well known that methylation proportions are strongly spatially correlated. Hebestreit *et al.* have shown that local smoothing can reduce the variance of methylation proportions, especially for lowly covered CpG sites (Hebestreit *et al.*, 2013). 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 proportions cannot be estimated based on smoothed methylation proportions. What we have observed is that the spatial correlations among near-by CpG sites are also preserved in the test statistics (Supplementary Fig. S1). To make use of this information, we propose a local Getis-Ord statistic based method to detect DMRs. The rationale of using such a statistic is that most of the genetic regions are not differentially methylated, and identifying DMRs is similar to hot-spot detection from the perspective of spatial statistics, 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 local Getis-Ord statistics to detect DMRs based on z-scores.

The local Getis-Ord statistic is defined as

$$G_k^* = \frac{\sum_{l=1}^{n} \omega_{kl}z_l - \bar{z}\sum_{l=1}^{n}\omega_{kl}}{S\sqrt{\frac{[n\sum_{l=1}^{n}\omega_{kl}^2 - (\sum_{l=1}^{n}\omega_{kl})^2]}{n-1}}} \quad (2)$$

where n is the total number of CpG sites, $\bar{z} = \sum_k z_k/n$, $S = \sqrt{\sum_k z_k^2/n - \bar{z}^2}$, and ω_{kl} is the weight parameters between kth CpG site and the lth CpG site (Getis and Ord, 1992; Ord and Getis, 1995).

The correlation of the local Getis-Ord statistics between the kth CpG site and the mth CpG sites is

$$\text{corr}(G_k^*, G_m^*) = \frac{n\sum_l \omega_{kl}\omega_{ml} - W_k^* W_m^*}{\sqrt{[nS_{1k} - (W_k^*)^2][nS_{1m} - (W_m^*)^2]}} \quad (3)$$

where $W_k^* = \sum_l \omega_{kl}$ and $S_{1k} = \sum_l \omega_{kl}^2$.

We further denote $\vec{G^*} = [G_1^*, G_2^*, G_3^* \ldots, G_K^*]'$, and therefore asymptotically $\vec{G^*} \sim MVN(\vec{0}, \Xi)$ with $\Xi_{k,l} = \text{corr}(G_k^*, G_l^*)$. The Getis-Ord statistics detect those CpG sites with values higher/lower in magnitude than we would expect on a random basis. A high/low value indicates that its neighborhood CpG sites also have high/low values, and vice versa. A value near zero indicates there is no apparent concentration (i.e. the neighborhood CpG sites have a range of values, and there is no apparent hot-spots and thus no DMRs).

The weight parameters (ω_{kl}) for the local Getis-Ord statistics G_k^* is determined based on the correlation between z-scores. The correlations between z-scores are estimated using the empirical variogram ($\text{corr}(z_k, z_l) = (\text{var}(z) - \gamma(d))/\text{var}(z)$, where z_k and z_l is d bp apart and $\gamma(d)$ is the semivariogram). Let Σ represent the estimated variance covariance matrix, Σ_k be the covariance vector between kth CpG site and the rest of the CpG sites, and $\Sigma_{-k,-k}$ be the variance-covariance matrix between all CpG sites except the kth CpG site. The weight parameters (ω_{kl}) is determined by $\Sigma_k\Sigma_{-k,-k}^{-1}$.

For the whole genome sequencing data, the estimation of semivariogram at every possible distances between two CpG sites and the inverse of the covariance matrix are computationally demanding. However, as shown in Supplementary Figure S1, the correlation reduces substantially as physical distance between 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 spline was used to capture the relationship between the physical distance and correlation. The correlations between CpG sites with distance larger than d_{select} are set at zero, which substantially reduces the computational cost. The correlation between z-scores of two CpG sites is calculated as below:

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

where $d = |\text{pos}_k - \text{pos}_l|$, pos_k is the physical location of the kth CpG site, and $f(d)$ is a spline function which captures the relationship between correlation and physical distance.

To detect DMRs, we define a region-based spatial statistics, $G_{\text{pos}_i,\text{pos}_j} = \frac{\sum_{l\in S} G_l^*}{|S|}$, where S is the set of CpG sites with physical location within (pos_i, pos_j), and $|S|$ is the total number of CpG sites in S. To determine the boundary of methylated region to be evaluated, we 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 sites with G_k^* larger than a cut-off value. Asymptotically, $G_{\text{pos}_i,\text{pos}_j} \sim N(0, \psi^2)$, where $\psi^2 = \sum_{(k,l)\in S}\text{cov}(G_k^*, G_l^*)$. The significance of the test statistics ($G_{\text{pos}_i,\text{pos}_j}$) can also be evaluated through permutation test.

Our test statistic explicitly makes use of correlations between nearby CpG sites, which can improve the power of the test. 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 single 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 case versus 1 control).

3 Simulation

We evaluated the performance of the proposed GetisDMR method with three commonly adopted methods [i.e ComMet (Saito *et al.*, 2014), BSmooth (Hansen *et al.*, 2012) and DSS (Feng *et al., 2014*)] under various conditions. In simulation one, we compared the performance of GetisDMR with ComMet in the absence of biological replicates. 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 proportions 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 (Saito *et al.*, 2014), 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 sensitivity as the proportion of true DMRs being detected. Specificity is usually used to evaluate the false positive rate. However, for the whole genome dataset, a method with a high specificity may still have a large amount of false positive findings. For example, with 99% of specificity, we may still have over thousands of false positive findings for the WGBS analyses. Therefore, we decided to report PPV instead of specificity to reflect the false positive rate. 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.

3.1 Scenario I

In the first set of simulations, we evaluated the performance of GetisDMR under a variety of conditions where 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 (Hirst and Marra, 2010; Stevens *et al.*, 2013), biological replicates are not always available, which makes many smoothing based methods not applicable. Therefore, in this set of simulations we compared the sensitivity and PPV of our method with ComMet, which employs a hidden Markov chain model to detect DMRs. At the time when the manuscript was written, ComMet is the only available software that has been claimed to be able to detect DMRs without biological replicates. To mimic the methylation proportions and the spatial correlations between adjacent CpG sites, we used a real dataset from a WGBS experiment (Hon *et al.*, 2013) and placed methylation proportion differences of various intensities. Specifically, we chose one sample from the experiment to serve as the control (the cortex sample), and kept the methylation proportion 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 the case, where the methylation proportions 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 up-regulated and the other half 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 (Sun *et al.*, 2014), 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, the fraction of DMCs were varied from 0.7 to 1, and for each DMC the differences in methylation proportions between the case and control samples were varied from 0.1 to 0.4. For the other CpG sites, the methylation proportions 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 method (Saito *et al.*, 2014).

3.2 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 (Feng *et al., 2014*; Hansen *et al.*, 2012; Saito *et al.*, 2014). Similar to Scenario I, we used a real dataset from a WGBS experiment to mimic the methylation proportions and the spatial correlations between adjacent CpG sites (Hon *et al.*, 2013). We randomly chose 6 samples to serve as control samples, and the remaining samples to serve as case samples. The methylation proportions in control samples were set the same as the methylation proportion 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 lengths of the DMRs were sampled from a truncated Gaussian distribution ($L \sim N(150, 100^2)$, with $50 < L < 4000$). Within each DMR, we varied the fraction (ranging from 0.7 to 1) of DMCs and the differences (ranging from 0.2 to 0.3) in methylation proportions between cases and controls. For non-differentially methylated regions, the methylation proportions 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 (Saito *et al.*, 2014), the BSmooth (Hansen *et al.*, 2012) and the DSS (Feng *et al.*, 2014).

3.3 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 proportions at each CpG site were simulated using $\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \mu_{ik} + X_{ij}\beta$, where p_{ijk} is the methylation proportion of sample j ($j = 1, 2, 3, , 6$) in group i ($i = 0$, control and $i = 1$, case) at CpG site k, and X_{ij} is the covariate vector for sample j in group i. 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 $\log\left(\frac{p_{0k}}{1-p_{0k}}\right)$ where p_{0k} is the observed methylation proportion in one of the randomly selected control sample from the experiment. The μ_{1k} for DMCs in the case samples was set at $\log\left(\frac{p_{0k}}{1-p_{0k}}\right) + d$, where d

was set at 0.2 or 0.3. The μ_{1k} for the other sites in the case samples was set at $\log\left(\frac{p_{0k}}{1-p_{0k}}\right)$ (i.e. $d=0$). Within each DMR, the fraction of DMCs was set at 0.8. Similar to Scenario II, we put 400 DMRs on the data, and the lengths of DMRs were 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 (Saito *et al.*, 2014), the BSmooth (Hansen *et al.*, 2012) and the DSS (Feng *et al.*, 2014).

4 Results

4.1 Scenario I

The sensitivity and PPV of Scenario I are summarized in Figure 1 (the differences in methylation proportions are 0.2 and 0.3) and Supplementary Figure S2 (the differences in methylation proportions are 0.1 and 0.4). As expected, with the increase in the differences of methylation proportions between the case and control groups, the sensitivity and PPV increased for both of the methods. For example, with 90% DMCs in a DMR, 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 proportions 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 proportions increased. Consistent with what we had expected, as the fraction of DMCs in a DMR increased, the sensitivity and PPV increased as well. It is worth noting that when the differences in methylation proportions are small, the increase in the fraction of DMCs within a DMR increases both the sensitivity and PPV for GetisDMR, while it has little effect on ComMet (Supplementary Fig. S2). This could be largely explained by the fact that when biological replicates are 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_k^*) and thus boosts the power for DMR detection. As shown in Figure 1, both the sensitivity and PPV of the proposed GetisDMR method were significantly higher than those of ComMet under all the situations considered in this set of simulations. We also varied the definition of a true positive DMR to assess the performance of the proposed model. Specifically, we evaluated and compared the sensitivity and PPV between the two methods with 10%, 25%, 75% and 90% overlaps. The trends are similar to Figure 1 and Supplementary Figure S2 (supplementary Table S1).

4.2 Scenario II

The sensitivity and PPV of Scenario II are summarized in Figure 2. Similar to the results from Scenario I, both sensitivity and PPV increased with the increase in the fraction of DMCs and the differences of methylation proportions. As expected, given the same level of differences in methylation proportions 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 proportion were relatively high. This is partly due to the fact that when biological replicates are available, the GetisDMR takes both the biological and sampling variations into account. However, under the setting of Scenario II, the methylation proportions were set the same for all the samples in each group (i.e. no biological variability), indicating a binomial distribution should be sufficient to model the data. However, GetisDMR assumes there

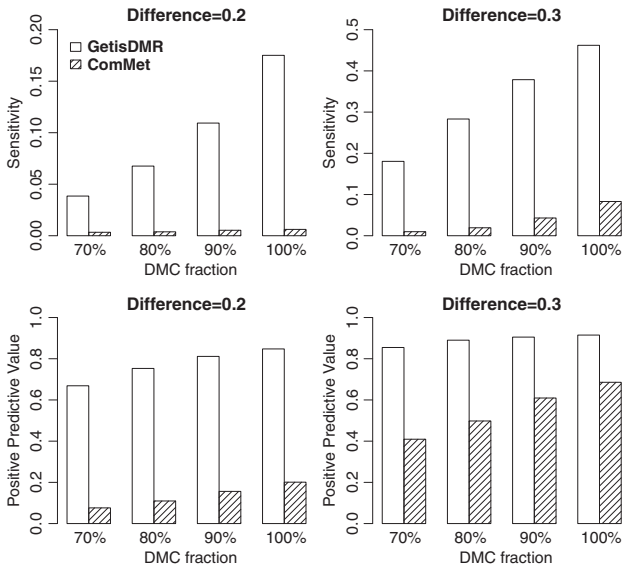

Fig. 1. Sensitivity and Positive Predictive Value of Scenario I

is biological variability and estimates an additional parameter to account for it, which could result in the loss of efficiency. On contrary, ComMet assumes no biological variability and pools together 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 do not know if the biological variability is present, and failing to consider such a variation can result in false positive findings (Scenario III). Nevertheless, as the differences in sensitivity between the two methods are not substantial (Fig. 2), we recommend to use the beta-binomial regression to account for the potential biological variability. The BSmooth method attained lower sensitivity and PPV than both of the proposed GetisDMR method and the ComMet method among all the situations except the case when the difference in methylation proportions was small (i.e. $d=0.1$, results are shown in Supplementary Fig. S3). While the sensitivity of the DSS method was lower than both GetisDMR and the ComMet, the PPV of the DSS method was higher than that of the ComMet and BSmooth methods but lower than that of the GetisDMR method. We also varied the definition of a true positive DMR and evaluated the sensitivity and PPV with 10%, 25%, 75% and 90% overlaps. The trends are similar to Figure 2 and Supplementary Figure S3 (Supplementary Table S2). In summary, GetisDMR has significantly lower rate of false positive findings, and it has higher or comparable sensitivity among all the situations considered in this set of simulations.

4.3 Scenario III

The sensitivity and PPV of Scenario III are summarized in Figure 3. As expected, with the increase in the effects of covariates, both the sensitivity and PPV for ComMet decreased. For example, when the odds ratio (OR) for the binary confounding variable changed from 1.11 (log OR = 0.1) to 1.65 (log OR = 0.5), with 80% CpG sites being differentially methylated within a DMR and the differences in methylation proportion 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

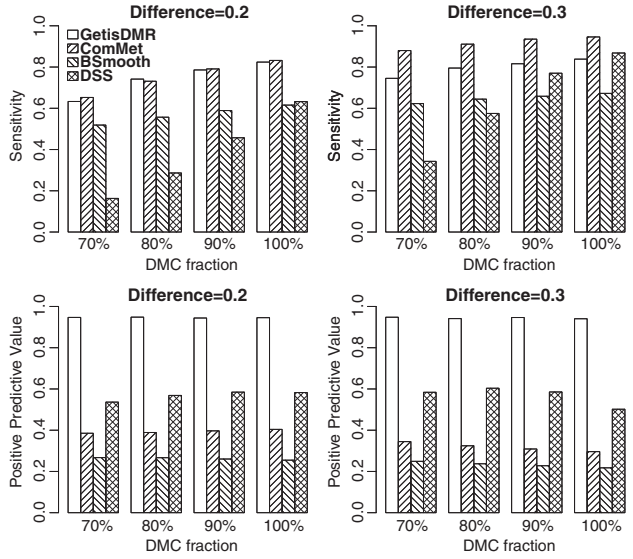

Fig. 2. Sensitivity and Positive Predictive Value of Scenario II

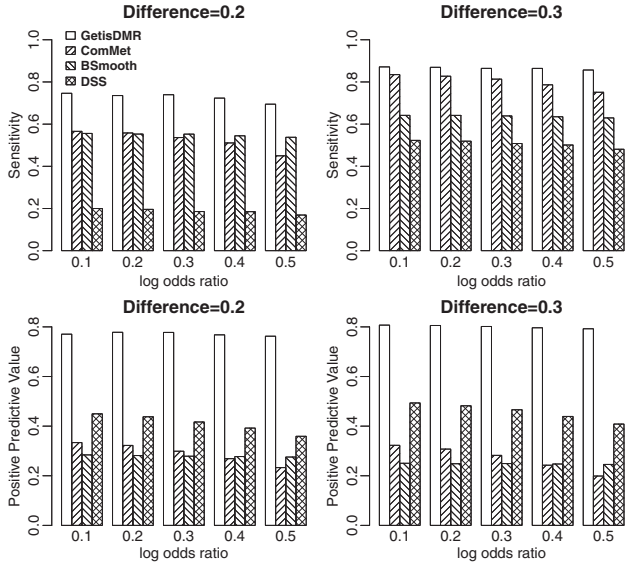

Fig. 3. Sensitivity and Positive Predictive Value of Scenario III

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 proportions, the treatment effect estimates can be substantially biased when such factors are not taken into account. The ComMet method ignores the potential confounding effects, and pools the data under the same experiment condition together into one sample. Therefore, it is subject to low power as the confounding effects increase. Similar to Scenario II, the BSmooth method performed worse than ComMet and GetisDMR, but its performance was relative robust against confounding effects. DSS method tended to have higher PPV than ComMet and BSmooth, but its sensitivity was lower than those of the ComMet and the BSmooth methods in most of the cases. Similar results hold when 100% CpG sites are differentially methylated within a DMR region (Supplementary Fig. S4). We also varied the definition of a true positive DMR (i.e. overlap proportions are set at 10%, 25%, 75% and 90%), and the trends are similar to Figure 3 and Supplementary Figure S4 (Supplementary Table S3). 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 proportions and the fraction of DMCs within a DMR.

5 Real data application

5.1 The mouse bone marrow and kidney dataset

We applied the GetisDMR method to a public available mouse dataset (Hon *et al.*, 2013) to investigate the methylation patterns in bone marrow and kidney tissues of adult mice. The dataset (GEO ID: GSE42836) includes the base-resolution methylomes of 17 mouse adult tissues spanning all three germ layers. Hon *et al.* found that most tissue-specific DMRs occur at distal cis-regulatory elements, and some tissue-specific DMRs mark vestigial enhancers that are dormant in adult tissues but active in embryonic development. They estimated that more than 6.7% of the mouse genome is variably methylated (Hon *et al.*, 2013). In our study, we only focus on comparing the methylation patterns between bone marrow and kidney tissues of adult mice. In data preprocessing, the segemhl (Otto *et al.*, 2012) software was used to map the single-end reads to mm9 genome, and only reads with unique mapping positions were kept for further analyses. The hits on the positive and negative strand cytosine at one CpG site were summed together to get the total number of reads and methylated reads. In total, our GetisDMR method has detected 116,912 DMRs. The median length of the identified DMR is 458 bp, and the median number of CpG sites within a DMR is 8 (the details of the identified regions were summarized in supplementary file 1). 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=15 172). 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 the 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 file 2. Using gene expression annotation, we found that the genes presumably controlled by the detected DMRs tended to be over-expressed in kidney or bone marrow tissues. 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 25 combinations of tissues and development stages that exhibit the most significant enrichment, 60% of them are related to kidney or bone marrow systems (Fig. 4A). About 50% terms are directly related to kidney or bone marrow functioning even tracing down to the 60th term (Supplementary Fig. S5A). 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 25 terms, 80% of them are directly related to the two tissues (Fig. 4B). About 80% of the terms are related to the two tissues when we trace down to the 60th term (Supplementary Fig. S5B).

For comparison purposes, we also analyzed this dataset using the ComMet method. GREAT database was used to explore the biological functions of the identified DMRs. The percentages of

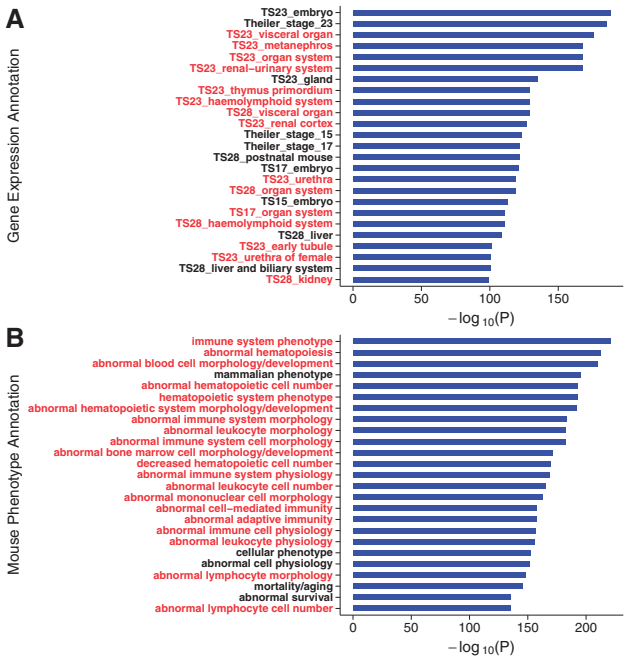

Fig. 4. Biological annotations of DMRs from bone marrow and kidney methylome data. The top 25 terms (bone marrow or kidney tissue related terms are highlighted in red) using the MGI gene expression annotation (**A**) and the mouse phenotype annotation (**B**)

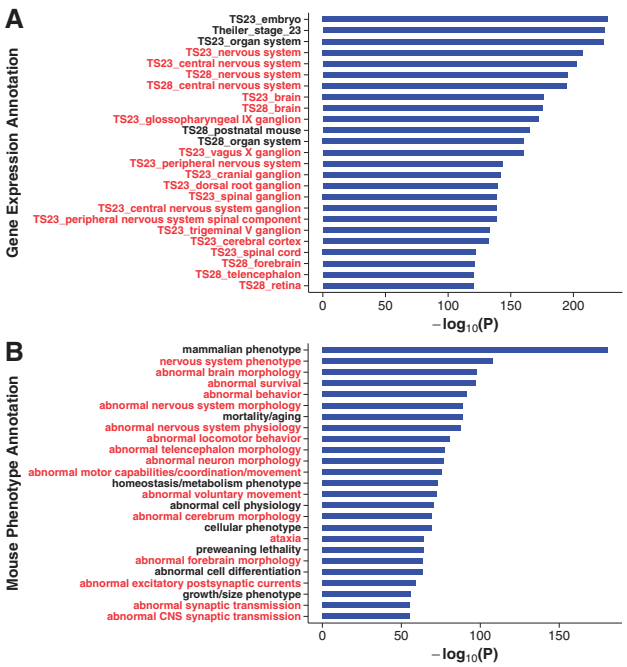

Fig. 5. Biological annotations of DMRs from neuron and non-neuron brain methylome data. The top 25 terms (neuron or non-neuron brain tissue related terms are highlighted in red) using the MGI gene expression annotation (**A**) and the mouse phenotype annotation (**B**)

enrichment terms that are related to the two tissues are summarized in Supplementary Figure S5. About 43% and 80% of the terms are related to the two tissues when we trace down to the 60th highest enrichment term using gene expression and mouse phenotype annotations, respectively. The results from ComMet method are similar to those of GetisDMR (Supplementary Fig. S5).

5.2 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 (Lister *et al.*, 2013). 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 (Lister *et al.*, 2013). To reduce the effects of confounding variables, we adjusted for age and gender (6 week and 12 month old females, and 7 week old males) in our analyses. Totally, 371 092 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 Supplementary file 3). The DMRs, comprised of more than 25 CpG sites($n = 12, 422$), were used to explore the biological functions via the GREAT database (Supplementary file 4). We 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 5, among the top 25 terms, 80% and 68% of them are directly related to neural system function according to gene expression and mouse phenotype annotations, respectively. When we trace down to the 60th highest enrichment term, the percentage of terms directly related to neural systems still reaches around 73% for gene expression and 57% for mouse phenotype annotations (Supplementary Fig. S6).

For comparison purposes, we also analyzed this dataset using ComMet, DSS and BSmooth. We explored the biological functions of the identified DMRs via the GREAT database. The percentages of enrichment terms that are related to neural systems are summarized in Supplementary Figure S6. When we trace down to the 60th highest enrichment term according to the gene expression annotation, the percentages of terms directly related to the neural system for DMRs identified by ComMet, DSS and BSmooth are 42%, 58% and 20%, respectively. When we trace down to the 60th highest enrichment term according to the mouse phenotype annotation, the percentages of terms directly related to the neural system for DMRs identified by ComMet, DSS and BSmooth are 5%, 25% and 22%, respectively. As shown in Supplementary Figure S6, the DMRs identified by the proposed method tend to have a better agreement with biological knowledge than the other three methods.

6 Discussion

In this work, we present a novel statistical method (GetisDMR) to detect DMRs from WGBS datasets. The proposed method utilizes the beta-binomial regression model to account for confounding effects, as well as biological and sampling variations. It further uses a local Getis-Ord statistic to combine information from nearby CpG sites to detect DMRs. 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 sensitivities and positive predictive values in detecting DMRs than ComMet, BSmooth and DSS(Feng *et al.*, 2014; Hansen *et al.*, 2012; Saito *et al.*, 2014)

One strength of the proposed method is that it utilizes the beta-binomial regression framework and models the methylation proportions corresponding to experimental as well as other independent and potential confounding factors. As found by Boks *et al.* (2009)that DNA methylation proportions could be influenced by

confounder, 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 account for biological variability and therefore reduces the rate of false positives compared with a standard logistic 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 the GetisDMR adopts region-wise test statistics based on local Getis-Ord statistics to directly detect DMRs. Currently, most of the existing methods focus on detecting DMCs and then identify DMRs based on some pre-specified empirical criteria. Although 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 provides statistical inference for direct DMR detection. It has been reported by various researchers that DNA-methylation proportions are spatially correlated along the genome (Eckhardt *et al.*, 2006; Irizarry *et al.*, 2008), and our own data shows that the *P*-values from the beta-binomial regression are also 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 WGBS 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. 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 for DMRs.

It is worth mentioning that the proposed method is quite flexible to the study designs 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 (Pashkevich and Kharin, 2004).

In real data application, we applied the proposed method to two public available mouse datasets (Hon *et al.*, 2013; Lister *et al.*, 2013). The first dataset is designed to investigate methylation proportions 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 the kidney tissue) (Hon *et al.*, 2013). In total, our method detected 116 912 DMRs. Further analyses using the GREAT tool 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, about 50% terms selected by the GREAT is related to the two tissues. Similarly, using the phenotype annotation, 80% of terms are directly associated with the two tissues. We also applied our method to compare methylation proportions between neuron and non-neuron samples from mouse frontal cortex (Lister *et al.*, 2013), where the effects of age and gender have been controlled for. In total, we have detected 371 092 DMRs and most of them are biologically relevant. Using the GREAT, the percentages of terms directly related to neural systems reach 73% and 57% for gene expression and mouse phenotype annotations, respectively. In both scenarios, the DMRs detected by the 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 detected DMRs, our findings shed light on the methylation patterns in different mouse tissues.

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

Acknowledgements

The project was supported by the Faculty Research Development Funds from the University of Auckland, the Scientific Research Foundation for the Returned Overseas Chinese Scholars from State Education Ministry, the National Natural Science Foundation of China (Award No. 81502887 and 81472637), and the Pandeng Scholar Program from the Department of Education of Liaoning Province. We wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. We also want to thank the two anonymous reviewers whose comments helped improve and clarify this manuscript.

Conflict of Interest: none declared.

References

Akalin,A. *et al.* (2012) methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome Biol.*, **13**, R87.

Beyan,H. *et al.* (2012) Guthrie card methylomics identifies temporally stable epialleles that are present at birth in humans. *Genome Res.*, **22**, 2138–2145.

Bhunia,G.S. *et al.* (2013) Spatial and temporal variation and hotspot detection of Kala-azar disease in Vaishali district (bihar), india. *BMC Infect. Dis.*, **13**, 64.

Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.

Boks,M.P. *et al.* (2009) The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One*, **4**, e6767.

Clark,S.J. *et al.* (2006) DNA methylation: bisulphite modification and analysis. *Nat. Protoc.*, **1**, 2353–2364.

Deaton,A.M. and Bird,A. (2011) CPG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.

Dolzhenko,E. and Smith,A.D. (2014) Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinf.*, **15**, 215.

Eckhardt,F. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.

Ehrlich,M. (2002) DNA methylation in cancer: too much, but also too little. *Oncogene*, **21**, 5400–5413.

Feng,H. *et al.* (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.*, **42**, e69.

Getis,A. and Ord,J.K. (1992) The analysis of spatial association by use of distance statistics. *Geograph. Anal.*, **24**, 189–206.

Hansen,K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.

Hebestreit,K. *et al.* (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, **29**, 1647–1653.

Hirst,M. and Marra,M.A. (2010) Next generation sequencing based approaches to epigenomics. *Brief. Funct. Genomics*, **9**, 455–465.

Hon,G.C. *et al.* (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.*, **45**, 1198–1206.

Irizarry,R.A. *et al.* (2008) Comprehensive high-throughput arrays for relative methylation (charm). *Genome Res.*, **18**, 780–790.

Jaffe,A.E. *et al.* (2012) Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, **13**, 166–178.

Laurent,L. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.

Li,E. *et al.* (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Lister,R. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.

Ord,J.K. and Getis,A. (1995) Local spatial autocorrelation statistics – distributional issues and an application. *Geograph. Anal.*, **27**, 286–306.

Ord,J.K. and Getis,A. (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. *J. Regional Sci.*, **41**, 411–432.

Otto,C. *et al.* (2012) Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, **28**, 1698–1704.

Park,Y. *et al.* (2014) Methylsig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, **30**, 2414–2422.

Pashkevich,M.A. and Kharin,Y.S. (2004) Robust estimation and forecasting for beta-mixed hierarchical models of grouped binary data. *SORT*, **28**, 125–160.

Saito,Y. *et al.* (2014) Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Res.*, **42**, e45.

Santos,F. *et al.* (2002) Dynamic reprogramming of Dna methylation in the early mouse embryo. *Dev. Biol.*, **241**, 172–182.

Schultz,M.D. *et al.* (2012) 'leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.*, **28**, 583–585.

Sharma,S. *et al.* (2010) Epigenetics in cancer. *Carcinogenesis*, **31**, 27–36.

Stevens,M. *et al.* (2013) Estimating absolute methylation levels at single-CPG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res.*, **23**, 1541–1553.

Sun,D. *et al.* (2014) Moabs: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.

Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.