Xiaoqing Yu* and Shuying Sun

HMM-DM: identifying differentially methylated regions using a hidden Markov model

DOI 10.1515/sagmb-2015-0077

Abstract: DNA methylation is an epigenetic modification involved in organism development and cellular differentiation. Identifying differential methylations can help to study genomic regions associated with diseases. Differential methylation studies on single-CG resolution have become possible with the bisulfite sequencing (BS) technology. However, there is still a lack of efficient statistical methods for identifying differentially methylated (DM) regions in BS data. We have developed a new approach named HMM-DM to detect DM regions between two biological conditions using BS data. This new approach first uses a hidden Markov model (HMM) to identify DM CG sites accounting for spatial correlation across CG sites and variation across samples, and then summarizes identified sites into regions. We demonstrate through a simulation study that our approach has a superior performance compared to BSmooth. We also illustrate the application of HMM-DM using a real breast cancer dataset.

Keywords: bisulfite sequencing; cancer; differential methylation; hidden Markov model.

1 Introduction

DNA methylation is an epigenetic process that adds a methyl group to the 5′ position of cytosine at a CG dinucleotide (i.e. a cytosine is located next to a guanine nucleotide). Differentially methylated (DM) regions can serve as novel biomarkers for disease diagnosis and potential targets for demethylation drug development in cancer studies (Strathdee and Brown, 2002; Wei et al., 2003). Therefore, in recent studies, DM region identification has received more and more attention. To identify differential methylation patterns between any two groups of samples, it is essential to obtain methylation signals at the single CG site level. A commonly used technology that measures methylation at the single CG site level is high-throughput bisulfite sequencing (BS), which combines bisulfite treatment with next-generation sequencing to provide single base, quantitative methylation signals. First, sodium bisulfite treatment specifically converts unmethylated cytosine to uracil (later read as thymine), leaving the methylated cytosine unaffected (Krueger et al., 2012). This change is then measured by next-generation sequencing, such as whole-genome bisulfite sequencing (WGBS) and targeted BS (Meissner et al., 2008). DNA sequencing reads may be subsequently mapped via bisulfite-conversion-aware aligners, such as BRAT (Harris et al., 2010), Bismark (Krueger and Andrews, 2011), BS Seeker (Chen et al., 2010), BISMA (Rohde et al., 2010), SAAP-RRBS (Sun et al., 2012), BSMAP (Xi and Li, 2009), PASS-bis (Campagna et al., 2013), and RRBSMAP (Xi et al., 2012). For each sequenced CG site, these aligners generate the total number of cytosine (C) and the number of thymine (T) aligned to the position among genomic DNA sequences, and the methylation signal of a specific CG site is calculated as C/(C+T). With genome-wide methylation signals measured at the single CG level, the detection of DM regions with fine resolution becomes possible.

***Corresponding author: Xiaoqing Yu,** Department of Biostatistics, Yale University, New Haven, CT 06511, USA, e-mail: yuxq1120@gmail.com
Shuying Sun: Department of Mathematics, Texas State University, San Marcos, TX 78666, USA

As BS technologies have been widely used, a number of computational tools for DM region (DMR) identification have been developed (Robinson et al., 2014). Many early approaches for DNA methylation studies perform Fisher's exact test (Gu et al., 2010; Li et al., 2010; Lister et al., 2011; Bock et al., 2012; Stockwell et al., 2014.) or set arbitrary cutoffs for differential methylation in large sliding windows (Lister et al., 2009; Laurent et al., 2010). These methods are fairly straightforward, but they fail to take into account two important features of DNA methylation. First, the methylation levels of neighbor CG sites may be spatially correlated (Eckhardt et al., 2006), but the signals can also change within hundreds of base pairs (bps). Second, there is a large amount of variation for methylation signals across samples within the same biological group, especially in cancer samples. Recently, more complex statistical algorithms accounting for some of these features have been published. Based on the statistical elements they use, these algorithms can be classified into five categories. First, regression based methods include methylKit (Akalin et al., 2012), BiSeq (Hebestreit et al., 2013), and RADmeth (Dolzhenko and Smith, 2014). For example, methylKit first models the methylation differences between two groups for each CG site using a binomial distribution within a logistic regression framework, and then corrects the multiple hypothesis testing with a sliding linear model (Wang et al., 2011). Second, bump-hunting based methods include Bumphunter (Jaffe et al., 2012) and BSmooth (Hansen et al., 2012). BSmooth accounts for the spatial correlation by smoothing the methylation signals via a local-likelihood estimation for each sample, and tests for group difference using a modified *t*-test for each CG. Then, bump hunting is performed on the t-statistics to search for DM regions. Third, β-binomial distribution based methods include BiSeq, MOABS (Sun et al., 2014), DSS (Feng et al., 2014), methylSig (Park et al., 2014), and RADmeth. As one of the earliest methods based on β-binomial, BiSeq constrains the analysis on target regions enriched of frequently covered CG sites, instead of processing all sequenced CG sites. Within these target regions, smoothed methylation signals are modeled based on β-binomial assumptions. Group differences are then tested via Wald-test and later the significant target regions are trimmed to obtain differentially methylated regions. Fourth, HMM-based methods include Bisulfighter (Saito et al., 2014) and MethPipe (Song et al., 2013). These two methods take into account the spatial correlation between CG sites via a hidden Markov model (HMM); however, they are only designed for comparing two samples. Fifth, methods based on other statistical models/tests include DAMP (Jayanth and Puranik, 2011), COHCAP (Warden et al., 2013), eDMR (Li et al., 2013), and a method by Xu et al. (2013). DAMP and COHCAP incorporate Fisher's exact test, χ^2-test, and ANOVA; eDMR is based on a bimodal normal distribution and uses a Stouffer-Liptak test to adjust for correlation among adjacent CG sites.

Although the above algorithms can handle certain common issues in DNA methylation to some extent, they have limitations. For example, some are only designed for either whole-genome BS or targeted BS data. In addition, the default parameter settings may not be suitable for a specific dataset, and the parameters defined by users can largely influence the accuracy of analysis results. In particular, the length of identified DM regions is sensitive to the smoothing-window-size selection, and thus a wider window usually lowers the sensitivity for small DMRs. Moreover, most of these algorithms are designed for testing differences between normal and cancer specimens, and the variation across samples at a single CG site is not well accounted for. It is known that the cross-sample variation of methylation level is usually large in cancer samples. Therefore, it is difficult for these methods to handle the DM CG sites with large within-group variation when comparing different cancer samples or tissues.

To address the above challenges in DM region identification, we propose a hidden Markov model-based method (HMM-DM) that can detect DM regions from BS data obtained from different protocols. The HMM-DM approach first identifies DM CG sites accounting for spatial correlation across CG sites and variation across samples. It then summarizes adjacent DM CG sites into DM regions. The main methodological contributions of HMM-DM are: (1) it can robustly identify DM regions with various lengths and DM singletons; (2) methylation variation across samples in the same group is well accounted for; and (3) it is suitable for both whole genome and targeted bisulfite methylation sequencing data. We demonstrate the advantages of HMM-DM by applying it to simulated data and comparing it with the most commonly cited method named BSmooth. We also apply HMM-DM to a published breast cancer dataset and report our findings. In addition, our group has developed another hidden Markov model-based method named HMM-Fisher (Sun and Yu, 2016a,b). The manuscript that introduces HMM-Fisher has been submitted and can be found at the HMM-Fisher web page. In another research article (Yu and Sun, 2016), we have thoroughly compared our two HMM-based methods with methylKit, BSmooth, and BiSeq using both simulation datasets and real datasets.

2 Methods

2.1 Real methylation data

To train and test our model, we use publicly available DNA methylation sequencing data (*GSE27003*) (Sun et al., 2011) generated using the reduced representation bisulfite sequencing (RRBS) protocol (Gu et al., 2010, 2011) from eight breast cancer cell lines, including four estrogen receptor positive (ER^+) and four negative (ER^-) samples. We use the software package BRAT (Harris et al., 2010) to trim off bases with low quality from both ends of the reads and then align trimmed reads. Methylation levels are obtained for all CG sites in eight samples using the BRAT *acgt-count* function. CG sites that are covered in less than three samples in either group are removed as low coverage sites, leaving 77,822 CG sites on chromosome 1 for further analysis.

2.2 Simulation data

Because methylation patterns in real samples are complex and difficult to simulate, all DM regions are simulated based on methylation levels and variation statuses of the "control group" of a real dataset. In particular, we take the first 10,000 CG sites of the four ER^+ samples from the data described earlier as a control group. In order to mimic the true methylation signals and variation in real sequencing data, the methylation levels of the test group are simulated using the control group as a background. Specifically, DM regions in the test group are obtained by adding differential methylation signals with various lengths and intensities to the background. Simulated DM regions are generated this way to preserve the natural changes in methylation patterns across CG sites and the variation patterns among samples. First, CG sites are categorized into four types of regions based on their methylation levels and variation statuses in the control group (see Supplementary Figures 1 and 2 in the supplemental file posted on the HMM-DM website at https://github.com/xxy39/HMM-DM): H, High-methylation regions with small between sample variation; L, low-methylation regions with small between sample variation; M-H, High-methylation regions with large between sample variation; and M-L, low-methylation regions with large between sample variation. This step generates 2,459 regions. Second, from the regions generated above, we randomly choose 80 DM regions with various methylation statuses and sizes (1–76 CG sites) to create methylation differences (see Supplementary Table 1). These DM regions cover a total of 929 DM CG sites. Third, methylation levels for the test group in these DM regions are sampled from uniform distributions (see the Supplementary Table 2). Because the region types are defined based on the control group, to create a contrast in the simulation data we generate low methylation levels for the test group if the control group has high methylation levels (i.e. H and M-H regions); we generate high methylation levels for the test group if the control group has low methylation levels (i.e. L and M-L regions). In addition, to ensure a true difference in DM regions with larger variation and/or smaller sizes, we use more stringent uniform distributions for H-M and H-L DM regions and DM regions with ≤ 3 CG sites (see the Supplementary file for more information: https://github.com/xxy39/HMM-DM).

2.3 Hidden Markov model

A HMM consists of hidden states and observed data. The sequence of hidden states is modeled by a Markov process, where the probability of the present state only depends on the previous state and all other preceding states are irrelevant. To identify DM regions, we design a first-order hidden Markov model assuming that the hidden differential methylation state at each CG site depends on the differential methylation state of the immediately preceding CG site. To build a hidden Markov chain that integrates information from all samples in two groups, we first define observations and hidden states.

2.3.1 Observations

Observations are a $P\times L$ matrix of observed methylation signals/ratios for P samples at L CG sites, $O=\{o_1, o_2, \ldots, o_p, \ldots, o_P\}$, with $o_p=\{o_{p,1}, o_{p,2}, \ldots, o_{p,L}\}$. Each observation ranges from 0 to 1, as the methylation signal at each CG site is calculated as the ratio of the number of reads with a methylated C to the total number of reads covering this CG site.

2.3.2 Hidden states

We use h_i to denote the hidden differential methylation state at the ith CG site (Figure 1). There are three possible hidden states for each CG, Hyper (hypermethylated), EM (equally methylated), and Hypo (hypomethylated). EM corresponds to the situation where all samples from two different groups have similar methylation levels. Hyper represents the situation where samples in group 1 generally have higher methylation levels than samples in group 2, while Hypo represents the reverse pattern. The initial probabilities of three hidden states are $P(h_1=\text{Hyper})=P(h_1=\text{EM})=P(h_1=\text{Hypo})=1/3$. With observations and hidden states established, the probability of the observed data is

$$\begin{aligned} P(o_{1,1}, \ldots, o_{L,P}) &= P(o_{1,1}, \ldots, o_{L,P} | h_1, \ldots, h_L)\cdot P(h_1, \ldots, h_L) \\ &= \left[\prod_{p=1}^{P}\prod_{i=1}^{L} P(o_{pi}|h_i)\right]\cdot\left[P(h_1)\prod_{i=2}^{L} P(h_i|h_{i-1})\right]. \end{aligned}$$

The other key features of HMM are described below.

2.3.3 Transition probabilities

The transition probability describes the probability of transferring from one differential methylation state to another between any two consecutive CG sites, $P(h_i|h_{i-1})=t_{h_{i-1},h_i}$, where $i=1, \ldots, L$. We use a vector $\theta^t=\{t_{j,k}\}=\{t_{1,1}, t_{1,2}, t_{1,3}, t_{2,1}, t_{2,2}, t_{2,3}, t_{3,1}, t_{3,2}, t_{3,3}\}$ to denote the transition probabilities between two states, where j and k are hidden states of two consecutive CG sites, respectively (Table 1).

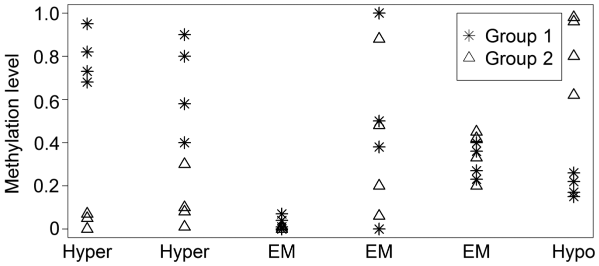

Figure 1: A hidden Markov model with examples of observations and corresponding hidden states.
The methylation levels of six CG sites in group 1 and 2 are represented in stars and triangles, respectively. The hidden states of six CG sites are denoted as “Hyper”, “EM”, and “Hypo”.

Table 1: Transition probabilities between two adjacent states h_{i-1} and h_i.

$h_i=k$, $h_{i-1}=i$	1 (Hyper)	2 (EM)	3 (Hypo)
1 (Hyper)	$t_{1,1}$	$t_{1,2}$	$t_{1,3}$
2 (EM)	$t_{2,1}$	$t_{2,2}$	$t_{2,3}$
3 (Hypo)	$t_{3,1}$	$t_{3,2}$	$t_{3,3}$

For any j=1, 2, 3, $t_{j,1}+t_{j,2}+t_{j,3}=1$.

2.3.4 Emission probabilities

Emission probabilities model the probability of observing methylation level at a CG site given a differential methylation state. For a given CG site, if there are similar methylation levels between two groups (EM), we consider the two groups as P independent samples and assume that their methylation levels follow the same β distribution. Alternatively, if there is differential methylation between two groups, then (o_{1i}, …, o_{Mi}) from group 1 are M independent samples whose methylation levels follow a β distribution with a specific shape, while (o_{M+1i}, …, o_{pi}) from group 2 are P-M independent samples whose methylation levels follow a β distribution with a different shape. Therefore, the distribution of o_{pi} conditional on h_i is

$$o_{p,i}\,|\,h_i,\theta_i^e \sim \begin{cases} \begin{cases} \beta(a_{i1},\ 1) \text{ for } h_i=\text{Hyper}, & p=1,\ 2,\ \dots,\ M \\ \beta(1,\ a_{i2}) \text{ for } h_i=\text{Hyper}, & p=M+1,\ M+2,\ \dots,\ P \end{cases} \\ \beta(a_{i3},\ a_{i4}) \text{ for } h_i=\text{EM}, \quad p=1,\ 2,\ \dots,\ M,\ M+1,\ \dots,\ P \\ \begin{cases} \beta(1,\ a_{i5}) \text{ for } h_i=\text{Hypo}, & p=1,2,\dots,M \\ \beta\left(a_{i6},1\right) \text{ for } h_i=\text{Hypo}, & p=M+1,\ M+2,\ \dots,\ P \end{cases} \end{cases}$$

where $\theta_i^e=(a_{i1},\ a_{i2},\ a_{i3},\ a_{i4},\ a_{i5},\ a_{i6})$ at CG site i. To reduce the number of parameters, we use "1" for the hyper and hypo states, and other parameters will be estimated using hyperpriors (see Section 2.4).

2.4 Estimating parameters

To estimate the parameters of HMM-DM, we use Bayesian methods by giving priors to unknown parameters and derive their conditional probabilities (posterior probabilities). All parameters are then sampled from their conditional probabilities.

2.4.1 Transition probability parameters $\boldsymbol{\theta}^t$

We count the numbers of transitions $y_{j,k}$ in the inferred states that fall into each category in Table 1, where $\sum_{j=1}^{3}\sum_{k=1}^{3} y_{j,k}=L-1$. For example, $y_{j,2}$=100 means that 100 CG sites change their differential methylation state from "Hyper" to "EM". Let each ($y_{j,1}$, $y_{j,2}$, $y_{i,3}$) follow a multinomial distribution,

$$\begin{cases} (y_{1,1},\ y_{1,2},\ y_{1,3})\sim\text{Multinomial}(y_{1,1}+y_{1,2}+y_{1,3},\ t_{1,1},\ t_{1,2},\ t_{1,3}) \\ (y_{2,1},\ y_{2,2},\ y_{2,3})\sim\text{Multinomial}(y_{2,1}+y_{2,2}+y_{2,3},\ t_{2,1},\ t_{2,2},\ t_{2,3}) \\ (y_{3,1},\ y_{3,2},\ y_{3,3})\sim\text{Multinomial}(y_{3,1}+y_{3,2}+y_{3,3},\ t_{3,1},\ t_{3,2},\ t_{3,3}) \end{cases}. \tag{1}$$

Let the prior of each ($t_{j,1}$, $t_{j,2}$, $t_{i,3}$) follow a Dirichlet distribution,

$$\begin{cases} (t_{1,1},\ t_{1,2},\ t_{1,3})\sim\text{Dirichlet}(\alpha_1,\ \alpha_2,\ \alpha_3) \\ (t_{2,1},\ t_{2,2},\ t_{2,3})\sim\text{Dirichlet}(\alpha_4,\ \alpha_5,\ \alpha_6) \\ (t_{3,1},\ t_{3,2},\ t_{3,3})\sim\text{Dirichlet}(\alpha_7,\ \alpha_8,\ \alpha_9) \end{cases}. \tag{2}$$

Thus, from (1) and (2), the posterior probabilities of θ^t become,

$$\begin{cases}(t_{1,1}, t_{1,2}, t_{1,3}|y_{1,1}, y_{1,2}, y_{1,3})\sim\text{Dirichlet}(\alpha_1+y_{1,1}, \alpha_2+y_{1,2}, \alpha_3+y_{1,3})\\(t_{2,1}, t_{2,2}, t_{2,3}|y_{2,1}, y_{2,2}, y_{2,3})\sim\text{Dirichlet}(\alpha_5+y_{2,1}, \alpha_6+y_{2,2}, \alpha_6+y_{2,3})\\(t_{3,1}, t_{3,2}, t_{3,3}|y_{3,1}, y_{3,2}, y_{3,3})\sim\text{Dirichlet}(\alpha_7+y_{3,1}, \alpha_8+y_{3,2}, \alpha_9+y_{3,3})\end{cases}, \tag{3}$$

which allows us to estimate the transition probabilities easily by sampling directly from Dirichlet distributions.

2.4.2 Emission probability parameters θ^e

The prior distribution of θ_i^e at the ith CG site is modeled as shown below,

$$\begin{array}{ll}\alpha_{i1}\sim\text{Uniform}[u_1, v_1] \text{ for } h_i=\text{Hyper}, & p=1, 2, \ldots, M\\ \alpha_{i2}\sim\text{Uniform}[u_2, v_2] \text{ for } h_i=\text{Hyper}, & p=M+1, M+2, \ldots, P\\ \varphi_i=\alpha_{i3}/(\alpha_{i3}+\alpha_{i4})\sim\beta(a_0+b_0) & \\ \gamma_i=\alpha_{i3}+\alpha_{i4}\sim\text{Uniform}(m, n) \text{ for } h_i=\text{EM}, & p=1, 2, \ldots, P.\\ \alpha_{i5}\sim\text{Uniform}[u_3, v_3] \text{ for } h_i=\text{Hypo}, & p=1, 2, \ldots, M\\ \alpha_{i6}\sim\text{Uniform}[u_4, v_4] \text{ for } h_i=\text{Hypo}, & p=M+1, M+2, \ldots, P\end{array}$$

Instead of using fixed values, we assign hyperpriors to the parameters of the distribution of θ_i^e. For EM states, all hyperpriors are set to ensure no limitation on the shape of emission β distribution, which allows us to model EM states with various methylation levels. For Hypo and Hyper states, u and v are set to limit the shape of β-distribution within a certain range, such that two groups have different methylation levels. In particular, for Hyper states, where the samples in group 1 have higher methylation levels than group 2, u_1 and v_1 will be set to ensure a relatively higher mean in $\beta(\alpha_{i1}, 1)$, and u_2, v_2 will be set to ensure a relatively lower mean in β $(1, \alpha_{i2})$. A similar strategy is applied to Hypo states to ensure a lower mean for $\beta(1, \alpha_{i5})$ and a higher mean for $\beta(\alpha_{i6}, 1)$. With the above prior distributions, the conditional probabilities of θ_i^e given observed data and hidden states can be derived. Slice sampling (Neal, 2003) is then used to sample θ_i^e from the conditional probabilities.

2.5 Estimating differential methylation states

The differential methylation states of HMM are estimated by a Markov chain Monte Carlo (MCMC) method. In particular, for a given CG site i, the Gibbs sampler (Gelfand and Smith, 1990) is used to sample the three hidden states from their conditional probability distribution:

$$P(h_i=k|O, h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_L, \theta^e, \theta^t)\propto P(O, h_1, \ldots, h_{i-1}, h_i=k, h_{i+1}, \ldots, h_L, \theta^e, \theta^t)$$

$$\propto\left[\prod_{p=1}^{P}P(O_{pi}|h_i=k,\theta^e)\right]\cdot\left[P(h_i)\cdot\prod_{j=2}^{L}P(h_j|h_{j-1},\theta^t)\right]\cdot P(\theta^e)\cdot P(\theta^t)$$

$$\propto\begin{cases}\left[\prod_{p=1}^{P}P(O_{pi}|h_i=k,\theta^e)\right]\cdot\left[P(h_i=k)\cdot P(h_2|h_1=k,\theta^t)\right] & i=1\\ \left[\prod_{p=1}^{P}P(O_{pi}|h_i=k,\theta^e)\right]\cdot\left[P(h_{i=k}|h_{i-1},\theta^t)\cdot P(h_{i+1}|h_i,\theta^t)\right] & 1<i<L\\ \left[\prod_{p=1}^{P}P(O_{pi}|h_i=k,\theta^e)\right]\cdot P(h_{L=k}|h_{L-1},\theta^t) & i=L\end{cases}$$

where k=Hyper, Hypo, EM.

For each CG site, a posterior probability is provided for each of three possible states, showing the possibility of the CG site being one of the states. The state with highest posterior probability is assigned as the state of CG site *i*. To call a Hyper or Hypo CG site, we require the difference of mean methylation levels in two groups (mean difference) to be larger than a certain threshold. This setting is to make sure the identified differentially methylated CG sites are biologically meaningful, that is, there is a measureable difference between the two groups.

2.6 Identifying differentially methylated regions

Each CG site is inferred by our HMM as equally methylated, hypermethylated, or hypomethylated. The hypermethylated and hypomethylated CG sites are called as DM. We then summarize these DM CG sites into either single site or regions with at least two CG sites. The adjacent DM CG sites are grouped into regions considering their differential methylation state, distance, and sequencing coverage (see the Supplementary file posted on the HMM-DM web page). Only CG sites with the same states can be included in the same region. Therefore, this method generates two types of DM regions, hypermethylated regions and hypomethylated regions.

3 Results

3.1 Simulated data

We apply the HMM-DM method to the simulated dataset described in the Methods section. With the cutoff of posterior probability set as ≥0.8, we obtain 1068 identified DM CG sites, yielding a sensitivity of 97.74% and a specificity of 98.24%. Out of the 80 selected DM regions, 68 are completely identified, seven are partially identified, and five singletons are not identified. An example of an identified DM region is illustrated in Figure 2. In this region, although the test group has a generally lower methylation level than the control group, large variation across the four control samples makes it difficult to identify by traditional methods. But HMM-DM successfully identifies all 15 CG sites within this 150 bp-long region as hypomethylated with posterior probabilities ≥0.9, suggesting a high confidence in calling DM regions that have large within group variation. As for the computation time, it takes HMM-DM 0.3 CPU hour (dual quad-core 2.66 Ghz Xeon E5430 processor) to detect DM regions for 10,000 CG sites.

To illustrate the efficiency of HMM-DM in identifying DM CG sites, we first evaluate its overall performance under different cutoffs of the posterior probabilities. As higher posterior probability corresponds to a

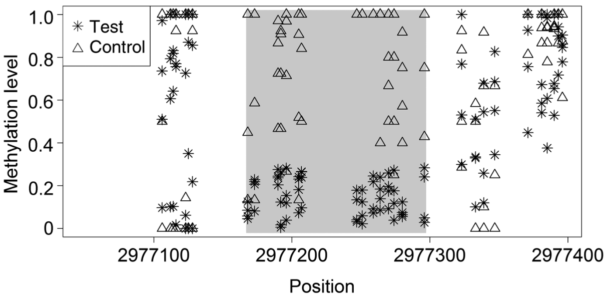

Figure 2: A typical DM region.
The methylation levels of control and test groups are represented in triangles and stars, respectively. The shaded box represents a simulated DM region identified by HMM-DM. All CG sites within this DM region are identified as hypo (hypomethylated in test group) and the background CG sites are identified as EM (equally methylated).

higher level of confidence, we expect that applying a relatively higher cutoff of posterior probability can filter out false positives and CG sites with weak DM signals. As shown in Table 2, the false positive rate decreases from 3.29% to 1.77% with a cutoff changing from 0.4 to 0.8. Over 90% of the identified true positive CG sites have posterior probabilities higher than 0.95 (data not shown). Therefore, when we filter the results with different cutoffs, the sensitivity stays as high as 99% and drops slightly to 92% when only DM CG sites with posterior probability ≥0.95 are considered as positives. This indicates that HMM-DM has a generally high sensitivity and accuracy in identifying DM CG sites.

We further examine HMM-DM's sensitivity in detecting DM regions with different lengths (number of CG sites included in each region) and different levels of within-group variation (Figure 3). The 80 designed DM regions are separated into three categories based on their sizes: long DM regions with >20 CG sites, median DM regions with 3–20 CG sites, and short DM regions with ≤2 CG sites. For the CG sites within each of the 10 long DM regions, almost 100% of designed DM CG sites are identified with high posterior probabilities. Only 1.8% of CG sites are filtered out when the cutoff is set as high as ≥0.95 (Figure 3, squares). For the median length DM regions, the sensitivity is 99.45% without any filtering. This number drops slightly for different cutoffs of posterior probability, but is still higher than 90% when the cutoff is increased to 0.95 (Figure 3, stars). As for the short DM regions, although HMM-DM shows a lower sensitivity than in the longer regions, over 80% of DM CG sites are identified with a posterior probability cutoff of ≥0.8 (Figure 3, triangles), suggesting that HMM-DM is capable of detecting DM regions even when the differential signal occurs in rather small clusters. A similar pattern is shown in the analysis of DM regions with different variation levels: both small-variation DM regions (H and L regions) and large-variation DM regions (M-H and M-L regions) show high sensitivities in all cutoff values, with a slightly lower sensitivity in large-variation DM regions when the cutoff increases to 0.9 (Figure 3, crosses and diamonds). All above results indicate that although the variation levels and DM region sizes may influence HMM-DM to some extent, our method can accurately identify DM CG sites that occur in small clusters with the presence of large within-group variation.

Table 2: Sensitivity and FPR (%) of HMM-DM with different cutoffs of posterior probability.

Cutoffs	**0.4**	**0.5**	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**	**0.99**
Sensitivity	99.25	99.25	99.14	99.14	98.82	98.49	98.06	97.74	96.88	96.23	92.79	88.05
FPR	3.29	3.26	2.92	2.75	2.45	2.29	1.91	1.77	1.48	1.40	1.05	0.78

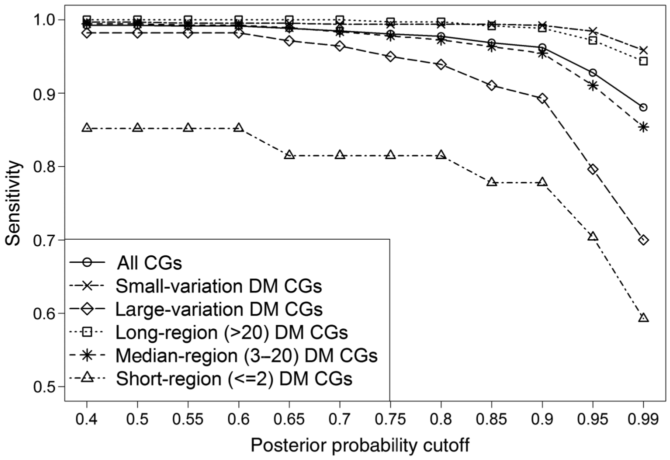

Figure 3: Sensitivity of HMM-DM.
Shown are the overall sensitivity (circles) and the sensitivities in DM regions with different sizes and different levels of within group variation, under different cutoffs of posterior probability. The larger the DM region size, and the lower the variation level, the higher the probability to identify a CG site.

Table 3: Sensitivity and FPR (%) of BSmooth with different cutoffs of *t*-test.

Thresholds	**1.5**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.6**	**6**	**12**
Sensitivity	94.29	91.47	85.79	82.18	78.47	72.47	70.72	68.78	66.63	62.40	50.00
FPR	18.60	17.48	15.53	14.25	10.21	8.21	6.93	5.99	5.14	4.20	3.33

In order to further demonstrate our method, we compare it with the most commonly used and cited method BSmooth (Hansen et al., 2012) using the simulated dataset. The parameters for the smoothing step of BSmooth are set to be comparable to HMM-DM: the minimum number of methylation loci in a smoothing window is set as 1, the minimum length of a smoothing window is set as 5, and the maximum gap between two methylation loci (before the smoothing is broken across the gap) is set as 100 bp. In the modified *t*-test step, all the 10,000 simulated CG sites are tested for methylation differences; the variance is estimated for the control group. Any CG with modified *t*-statistics beyond a certain threshold is identified as a DM CG.

For all different *t*-statistics thresholds, the sensitivity and FPR are calculated for BSmooth (Table 3). Comparing HMM-DM with BSmooth, we find that HMM-DM achieves higher sensitivity than BSmooth and it has a much smaller false positive rate. In particular, in Table 4 we compare the HMM-DM results with posterior probability ≥0.8 to the BSmooth results with a *t*-statistic threshold of 2.5. The HMM-DM result yields a sensitivity of 97.74% (908 true positives) and a false positive rate of 1.77% (160 false positives), while the *t*-statistics threshold of 2.5 in BSmooth yields a much smaller sensitivity of 79.47% (729 true positives) and a much higher false positive rate of 10.12% (926 false positives). In addition, HMM-DM is more accurate in detecting DM regions with shorter length and larger within-group variation. For the DM regions with 3–20 CG sites, and with no more than 2 CG sites, HMM-DM achieves a sensitivity of 97.26% and 81.48%, respectively, while BSmooth identifies 76.64% and 77.77%, respectively. For the DM regions with relatively larger variation (M-H and M-L regions), HMM-DM detects 93.93% of the CG sites and BSmooth detects only 42.14%. In summary, HMM-DM is more powerful than BSmooth in identifying DM CG sites, without sacrificing the specificities.

In this section, we have compared HMM-DM with the algorithm BSmooth, which is commonly cited. In another manuscript (Yu and Sun, 2016), we have compared HMM-DM and another HMM-based method (HMM-Fisher) developed by our group with multiple algorithms (methylKit, BSmooth, BiSeq) that are developed based on different models and assumptions. The brief comparison results of sensitivity and false positive rates using simulation data are shown below. Sensitivity levels of the five methods are: 97.74% (HMM-DM), 97.20% (HMM-Fisher), 96.23% (BiSeq), 88.27% (methylKit), and 66.63% (BSmooth). False discovery rates of the five methods are: 1.77% (HMM-DM), 2.44% (HMM-Fisher), 3.75% (BiSeq), 4.27% (methylKit), and 5.13% (BSmooth). Among these five methods, HMM-DM has the highest sensitivity and lowest false discovery rate.

Table 4: Comparing the performance of HMM-DM and BSmooth.

	HMM-DM	**BSmooth**
Sensitivity, all DM regions	97.74%	78.47%
FPR, all DM regions	1.77%	10.21%
Sensitivity, DM region >20	99.72%	81.35%
Sensitivity, DM region	97.26%	76.64%
Sensitivity, DM region ≤2	81.48%	77.77%
Sensitivity, small-variation DM region	99.38%	89.98%
Sensitivity, large-variation DM region	93.93%	42.14%

Shown are comparison results of HMM-DM with posterior probability ≥0.8 and BSmooth with modified *t*-statistics threshold of 2.5. Seven metrics are considered: the sensitivity and FPR for all simulated DM regions, sensitivity for DM regions with >20 CG sites, DM regions with 3–20 CG sites, DM regions with ≤2 CG sites, as well as sensitivity for DM regions with small and large variation (see the first column).

3.2 Breast cancer data

To illustrate the application of our method, we apply HMM-DM to detect the DM CG sites between ER^+ and ER^- groups in a breast cancer sequencing dataset (Sun et al., 2011). In chromosome 1, a total of 77,822 CG sites have been considered. To ensure that the detected DM CG sites have biological meaning rather than statistical significance alone, only CG sites with a mean difference (between the two groups) ≥0.3 can be identified as DM. CG sites in which ER^- has higher methylation levels than ER^+ are defined as hypermethylated, and CG sites in which ER^- has lower methylation levels are defined as hypomethylated. We identify 2326 DM CG sites, forming 898 DM regions. The median length of these DM regions is 8 bp (minimum is 1 bp and maximum is 305 bp). 76.91% (1789) of the detected DM CG sites are hypermethylated in the ER^- group, while 23.09% (537) are hypomethylated. In addition, all identified DM CG sites are categorized by their variation status within ER^- and ER^+ groups, respectively. The CG sites in which the four samples in one group all have methylation levels ≤0.4, or all have methylation level ≥0.6, are classified as small-variation; otherwise, the CG sites are classified as large-variation. The majority of identified DM CG sites have large variation either in one group (67%) or in both groups (31%), and only 2% of DM CG sites have small variation in both groups, suggesting HMM-DM is capable of identifying DM CG sites with various degrees of within-group variation. Moreover, out of the 2326 DM CG sites detected by HMM-DM, there are 1577 CG sites covering a total of 236 genes either in the gene body (1296 CG sites) or in the promoter region (343 CG sites). Table 5 lists the top 10 frequent genes, the majority of which show higher methylation levels in the ER^- group, a breast cancer type that is more difficult to treat. In particular, there are five genes located in the 1p36 tumor suppressor region, suggesting a possible mechanism for the severity of the ER^- condition.

4 Discussion

Our method has the following advantages. First, HMM-DM is not limited to any specific BS protocol. It is suitable for detecting DM regions using data generated from both WGBS and targeted BS. Second, HMM-DM has a finer resolution compared to BSmooth. Because it is a CG site-based approach, the changes of differential methylation patterns over short distances can be captured. Therefore, HMM-DM allows users to fully benefit from the single-CG resolution of methylation measurement provided by the BS technologies. Third, variation within the same biological group is considered. β-Distributions can model methylation levels with different variation easily with different shape parameters. This property is particularly beneficial when dealing with cancer samples, where the between-sample variation is usually large. Fourth,

Table 5: Top 10 frequent genes that include identified DM CG sites.

Location	Gene name	CG sites in gene body	CG sites in promoter
1p36.32	*AJAP1*	67 (67/–)	39 (39/–)
1p34.3	*GRIK3*	57 (37/20)	44 (44/–)
1p36.31	*CAMTA1*	43 (25/18)	–
1p36.23-p33	*PRDM16*	33 (23/10)	–
1p21	*LOC100129620*	25 (25/–)	–
1p21-22	*NTRK1*	24 (21/3)	–
1p13.2	*C1orf183*	18 (18/–)	5 (5/–)
1p36.3	*GABRD*	21 (19/2)	–
1p36.33	*RNF223*	20 (20/–)	–
1q42.13	*OBSCN*	19 (18/1)	–

Shown are the location, gene name, number of CG sites covering the gene body and promoter region for the top 10 frequent genes. Number of hyper-methylated and hypo-methylated CG sites are shown in brackets, separated by slash. "–" Indicates that no hyper or hypo DM CG sites are identified in the specific genomic regions.

HMM-DM has a simple parameter setting for users, because all key parameters (e.g. the priors for transition and emission probabilities) are estimated from the data directly. After obtaining the raw results, users can choose the desire thresholds for posterior probability and mean difference between groups for further filtering. Fifth, the HMM used by HMM-DM can well handle genomic regions with either dense or sparse CG sites. This is because the emission probability distribution of an HMM plays a dominant role compared to the transition probability distribution, so that the gaps in RRBS data may have little effect on the estimated DM status.

Our method also has a limitation. For example, the current method does not directly incorporate coverage in the model. However, this limitation can be made up for by performing quality control on coverage when preparing for the input data. For example, before applying the HMM-DM to a dataset, CG sites with extremely low coverage are removed.

It has been shown that methylation levels of neighboring CG sites may be spatially correlated (Eckhardt et al., 2006). However, in the whole-genome methylation sequencing data, the spatial correlation levels at different genomic regions are still unknown, and these correlations levels may vary along a genome depending on the length and CG density of different regions. In our model, we consider the spatial correlation by modeling the transition probabilities using the HMM, which borrows information from neighboring sites and avoids over-smoothing issues at the same time.

Our simulation data preserve the features of real BS data. The simulated DM regions are not chosen randomly, but are based on the natural blocks of methylation status in real data. In real BS data, adjacent CG sites tend to have similar methylation levels and a similar differential methylation status. It is relatively less frequent to observe dramatic changes within hundreds of base pairs. Therefore, it is only proper to follow the natural change of the methylation status, instead of creating DM blocks by arbitrary settings. As for the choice of simulating DM regions in the test group, besides uniform distributions we have also simulated another set of data using β-distributions. As HMM-DM performs similarly in both settings, we only report results from the uniform distribution in this paper.

5 Conclusion

In this paper we have developed a HMM based approach HMM-DM to detect DM regions from BS data generated from different protocols, such as WGBS and reduced representative BS. The HMM-DM method is illustrated using both simulated and real datasets from breast cancer samples. The application to simulated data shows an increased power compared to the currently most commonly cited method BSmooth, especially in DM regions that are short and have large within-group variation.

Software and supplementary information: see https://github.com/xxy39/HMM-DM.

Acknowledgments: This work was supported by Dr. Shuying Sun's start-up funds and the Research Enhancement Program provided by Texas State University.

References

Akalin, A., M. Kormaksson, S. Li, F. Garrett-Bakelman, M. Figueroa, A. Melnick and C. Mason (2012): "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles," Genome Biol., 13, R87.

Bock, C., I. Beerman, W.-H. Lien, Z. D. Smith, H. Gu, P. Boyle, A. Gnirke, E. Fuchs, D. J. Rossi and A. Meissner (2012): "DNA methylation dynamics during in vivo differentiation of blood and skin stem cells," Mol. Cell, 47, 633–647.

Campagna, D., A. Telatin, C. Forcato, N. Vitulo and G. Valle (2013): "PASS-bis: a bisulfite aligner suitable for whole methylome analysis of Illumina and SOLiD reads," Bioinformatics, 29, 268–270.

Chen, P., S. Cokus and M. Pellegrini (2010): "BS Seeker: precise mapping for bisulfite sequencing," BMC Bioinforma, 11, 203.

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

Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin and S. Beck (2006): “DNA methylation profiling of human chromosomes 6, 20 and 22,” Nat. Genet., 38, 1378–1385.

Feng, H., K. N. Conneely and H. Wu (2014): “A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data,” Nuc. Acids Res., 42, e69.

Gelfand, A. and A. Smith (1990): “Sampling-based approaches to calculating marginal densities,” J. Am. Stat. Assoc., 85, 398–409.

Gu, H., C. Bock, T. S. Mikkelsen, N. Jager, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander and A. Meissner (2010): “Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution,” Nat. Methods, 7, 133–136.

Gu, H., Z. D. Smith, C. Bock, P. Boyle, A. Gnirke and A. Meissner (2011): “Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling,” Nat. Protoc., 6, 468–481.

Hansen, K., B. Langmead and R. Irizarry (2012): “BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions,” Genome Biol., 13, R83.

Harris, E. Y., N. Ponts, A. Levchuk, K. L. Roch and S. Lonardi (2010): “BRAT: bisulfite-treated reads analysis tool,” Bioinformatics, 26, 572–573.

Hebestreit, K., M. Dugas and H.-U. Klein (2013): “Detection of significantly differentially methylated regions in targeted bisulfite sequencing data,” Bioinformatics, 29, 1647–1653.

Jaffe, A. E., P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg and R. A. Irizarry (2012): “Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies,” Int. J. Epidemiol., 41, 200–209.

Jayanth, N. and M. Puranik (2011): “Methylation stabilizes the imino tautomer of dAMP and amino tautomer of dCMP in solution,” J. Phys. Chem. B, 115, 6234–6242.

Krueger, F. and S. Andrews (2011): “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications,” Bioinformatics, 27, 1571–1572.

Krueger, F., B. Kreck, A. Franke and S. Andrews (2012): “DNA methylome analysis using short bisulfite sequencing data,” Nat. Methods, 9, 145–151.

Laurent, L., E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring and C.-L. Wei (2010): “Dynamic changes in the human methylome during differentiation,” Genome Res., 20, 320–331.

Li, Y., J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, Y. Huang, H. Cao, X. Zhou, S. Guo, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck and X. Zhang (2010): “The DNA methylome of human peripheral blood mononuclear cells,” PLoS Biology, 8, e1000533.

Li, S., F. Garrett-Bakelman, A. Akalin, P. Zumbo, R. Levine, B. To, I. Lewis, A. Brown, R. D’Andrea, A. Melnick and C. Mason (2013): “An optimized algorithm for detecting and annotating regional differential methylation,” BMC Bioinformatics, 14, S10.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren and J. R. Ecker (2009): “Human DNA methylomes at base resolution show widespread epigenomic differences,” Nature, 462, 315–322.

Lister, R., M. Pelizzola, Y. Kida, R. Hawkins, J. Nery, G. Hon, J. Antosiewicz-Bourget, R. O’Malley, R. Castanon and S. Klugman (2011): “Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells,” Nature, 471, 68–73.

Meissner, A., T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch and E. S. Lander (2008): “Genome-scale DNA methylation maps of pluripotent and differentiated cells,” Nature, 454, 766–770.

Neal, R. M. (2003): “Slice sampling,” Ann. Stat., 31, 705–767.

Park, Y., M. E. Figueroa, L. S. Rozek and M. A. Sartor (2014): “MethylSig: a whole genome DNA methylation analysis pipeline,” Bioinformatics, 30, 2414–2422.

Robinson, M. D., A. Kahraman, C. W. Law, H. Lindsay, M. Nowicka, L. M. Weber and X. Zhou (2014): “Statistical methods for detecting differentially methylated loci and regions,” Front. Genet., 5.

Rohde, C., Y. Zhang, R. Reinhardt and A. Jeltsch (2010): “BISMA – Fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences,” BMC Bioinformatics, 11, 230.

Saito, Y., J. Tsuji and T. Mituyama (2014): “Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions,” Nuc. Acids Res., 42, e45.

Song, Q., B. Decato, E. E. Hong, M. Zhou, F. Fang, J. Qu, T. Garvin, M. Kessler, J. Zhou and A. D. Smith (2013): “A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics,” PLoS One, 8, e81148.

Stockwell, P. A., A. Chatterjee, E. J. Rodger and I. M. Morison (2014): “DMAP: differential methylation analysis package for RRBS and WGBS data,” Bioinformatics, 30, 1814–1822.

Strathdee, G. and R. Brown (2002): “Aberrant DNA methylation in cancer: potential clinical interventions,” Expert. Rev. Mol. Med., 4, 1–17.

Sun, D., Y. Xi, B. Rodriguez, H. J. Park, P. Tong, M. Meong, M. A. Goodell and W. Li (2014): “MOABS: model based analysis of bisulfite sequencing data,” Genome Biol., 15, R38.

Sun, S. and X. Yu (2016a): "HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test," Statistical Application in Genetics and Molecular Biology, in press.

Sun, S. and X. Yu (2016b): "HMM-Fisher," GitHub Repository, https://github.com/xxy39/HMM-Fisher.

Sun, Z., Y. W. Asmann, K. R. Kalari, B. Bot, J. E. Eckel-Passow, T. R. Baker, J. M. Carr, I. Khrebtukova, S. Luo, L. Zhang, G. P. Schroth, E. A. Perez and E. A. Thompson (2011): "Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing," PLoS One, 6, e17490.

Sun, Z., S. Baheti, S. Middha, R. Kanwar, Y. Zhang, X. Li, A. S. Beutler, E. Klee, Y. W. Asmann, E. A. Thompson and J.-P. A. Kocher (2012): "SAAP-RRBS: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing," Bioinformatics, 28, 2180–2181.

Wang, H.-Q., L. Tuominen and C.-J. Tsai (2011): "SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures," Bioinformatics, 27, 225–231.

Warden, C. D., H. Lee, J. D. Tompkins, X. Li, C. Wang, A. D. Riggs, H. Yu, R. Jove and Y.-C. Yuan (2013): "COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis," Nuc. Acids Res., 41, e117.

Wei, S., R. Brown and T. Huang (2003): "Aberrant DNA methylation in ovarian cancer: is there an epigenetic predisposition to drug response?," Ann. NY Acad. Sci., 983, 243–250.

Xi, Y., C. Bock, F. Müller, D. Sun, A. Meissner and W. Li (2012): "RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing," Bioinformatics, 28, 430–432.

Xi, Y. and W. Li (2009): "BSMAP: whole genome bisulfite sequence MAPping program," BMC Bioinformatics, 10, 232.

Xu, H., R. H. Podolsky, D. Ryu, X. Wang, S. Su, H. Shi and V. George (2013): "A method to detect differentially methylated loci with next-generation sequencing," Genetic Epidemiology, 37, 377–382.

Yu, X. and S. Sun (2016): "Comparing five statistical methods of differential methylation identification using bisulfite sequencing data," Statistical Application in Genetics and Molecular Biology, in press.