HMM-Fisher: Identifying differential methylation using a hidden Markov model and Fisher's exact test

Shuying Sun[1*], and Xiaoqing Yu[2]

[1]Department of Mathematics, Texas State University, San Marcos, TX 78666, USA.

[2]Department of Biostatistics, Yale University, New Haven, CT 06511, USA.

[*]Corresponding author's email: ssun5211@yahoo.com

Abstract

DNA methylation is an epigenetic event that plays an important role in regulating gene expression. It is important to study DNA methylation, especially differential methylation patterns between two groups of samples (e.g., patients vs. normal individuals). With next generation sequencing technologies, it is now possible to identify differential methylation patterns by considering methylation at the single CG site level in an entire genome. However, it is challenging to analyze large and complex NGS data. In order to address this difficult question, we have developed a new statistical method using a hidden Markov model and Fisher's exact test (HMM-Fisher) to identify differentially methylated cytosines and regions. We first use a hidden Markov chain to model the methylation signals to infer the methylation state as Not methylated (N), Partly methylated (P), and Fully methylated (F) for each individual sample. We then use Fisher's exact test to identify differentially methylated CG sites. We show the HMM-Fisher method and compare it with commonly cited methods using both simulated data and real sequencing data. The results show that HMM-Fisher outperforms the current available methods to which we have compared. HMM-Fisher is efficient and robust in identifying heterogeneous DM regions.

Software availability and implementation: https://github.com/xxy39/HMM-Fisher.

Key words: differential methylation, bisulfite sequencing, hidden Markov model, Fisher's exact test.

Supplementary information: Available online.

1 INTRODUCTION

DNA methylation is one of the most common molecular changes in cells. In a mammalian cell, it involves the addition of a methyl group (CH_3) to the 5' cytosine (C) of a CG site (cytosine and guanine pair). This epigenetic event plays an important role in the regulation of gene expression, transposon silencing, and transcription factor binding inhibition in both normal and diseased cells (Bell and Felsenfeld, 2000; Clark et al., 1997; Douet et al., 2007; Goll and Bestor, 2005; Hark et al., 2000; Henderson and Jacobsen, 2007; Inoue and Oishi, 2005; Kitazawa et al., 1999; Mancini et al., 1999). Abnormal methylation often occurs across a genome, especially for the regions rich in CG sites. Differential methylation (DM) in diseased cells, when compared with normal cells, can be hypermethylation (i.e., gaining methylation) or hypomethylation (i.e., losing methylation). Recognizing disease methylation events can thus identify important functional regions that play a key role in driving disease phenotypes. In fact, accumulating evidence has shown that DNA methylation is one of the most common molecular modifications in complex diseases (e.g., cancers) (Kim et al., 2010; Lin et al., 2007; Toyota et al., 1999; Zitt and Muller, 2007). Because different DNA methylation events (e.g., *de novo* methylation and demethylation) are common in diseased cells, DNA methylation has been used as biomarkers for the medically important application of early detection of certain diseases such as breast cancer (Suijkerbuijk et al., 2011).

Identifying DNA methylation biomarkers requires us to be able to simultaneously assay the methylation status at each cytosine (C) for a portion of the genome or the entire genome. This is now possible with the bisulfite conversion technique (i.e., converting the un-methylated C to T) and next generation sequencing (NGS) technologies. In fact, several pioneering groups have successfully used the bisulfite-converted methylation sequencing method in both Arabidopsis thaliana and human samples (Cokus et al., 2008; Hansen et al., 2011; Lister et al., 2008; Lister et al., 2009; Lister et al., 2011; Meissner et al., 2008; Sun et al., 2011). From just one sample, NGS technologies can generate hundreds of gigabytes (GBs) of raw methylation sequencing data with complex technical and biological features, raising quality issues. For example, the 3' end of reads may have dramatically low quality, and certain regions may not be efficiently sequenced. It is challenging to preprocess and analyze such large datasets.

These challenges require us to develop new computational and statistical tools. A number of successful efforts in aligning bisulfite-converted reads and in calling methylation already exist, such as BS Seeker (Chen et al., 2010), BRAT (Harris et al., 2010), BSMAP (Xi et al., 2012), and MethylCoder (Pedersen et al., 2011). In addition, a number of packages have been developed for quality assessment, such as TileQC (Dolan and Denver, 2008), NGSQC (Dai et al., 2010), SolexaQA (Cox et al., 2010), PIQA (Martinez-Alcantara et al., 2009), MethyQA (Sun et al., 2013), methylKit (Akalin et al., 2012), SAAP-RRBS (Sun et al., 2012), and BseQC (Lin et al., 2013).

After quality assessment and alignment, one important down-stream analysis is to identify differentially methylated regions, which may be potential DNA methylation biomarkers. In fact, a number of statistical and computational methods currently exist for identifying differential methylation regions or sites using bisulfite-treated methylation sequencing data. Among these methods, QDMR (Zhang et al., 2011) and CpG_MPS (Su et al., 2013) are designed only for DM identification in one group of multiple samples. These methods cannot be used to identify DM between two groups. Fisher's exact test (Becker et al., 2011; Challen et al., 2011; Li et al., 2010; Lister et al., 2009), BEAT (Akman et al., 2014), Bisulfighter (or Commet) (Saito et al., 2014), Methy-Pipe (Jiang et al., 2014), and CpG_MPs (Su et al., 2013) are motived by and/or developed for only a pair of samples. These methods cannot account for the variation of methylation levels between replicates within a group. Some methods are designed for identifying differential methylation between two groups. These methods include: (a) beta-binomial distribution based methods, e.g., BiSeq (Hebestreit et al., 2013), RADmeth (Dolzhenko and Smith, 2014), MethylSig (Park et al., 2014), DSS (Feng et al., 2014), and MOABS (Sun et al., 2014); (b) smoothing-based methods, e.g., BSmooth (Hansen et al., 2012), and BiSeq (Hebestreit et al., 2013); (c) regression-based methods, e.g., RADmeth (Dolzhenko and Smith, 2014), methylKit (Akalin et al., 2012), and eDMR (Li et al., 2013); (d) statistical-test based methods, e.g., DMAP (Stockwell et al., 2014.), adjusted chi-square test (Xu et al., 2013), BSmooth (Hansen et al., 2012), and Fisher's exact test.

All of the above methods have some good features, and the authors of these methods have tried to approach this DM identification problem from different perspectives. However, these methods also have

some limitations. For example, some methods are mainly designed for identifying DM regions (i.e., DMRs), not for identifying DM cytosines (i.e., DMCs), or vice versa; some are mainly designed for reduced representative bisulfite sequencing (RRBS) or whole genome bisulfite sequencing (WGBS), but not for both; some have considered the similarity pattern of neighboring CG sites, but most methods do not consider this pattern. Even though a couple of methods (e.g., BSmooth and BiSeq) consider the local similarity feature using "smoothing" methods, the similarity pattern may change across the genome because some neighboring CG sites may have very different methylation levels. Therefore, there is a need to develop a more accurate and effective method. In order to share our understanding of the methylation data and address the challenges mentioned above, our group has developed two hidden Markov model (HMM) (Rabiner, 1989) based methods named HMM-Fisher and HMM-DM. In this article, we introduce the HMM-Fisher method and describe it in detail. We have written another manuscript to introduce HMM-DM (Yu and Sun, 2015a, 2015b). The manuscript of HMM-DM has been submitted and can be found at this web link: https://github.com/xxy39/HMM-DM. The HMM-Fisher method involves using a hidden Markov model to incorporate methylation information from neighboring CG sites. It can capture the unique biological features of methylation sequencing data and remove the impact of sequencing errors. The HMM-Fisher method is designed to identify both DMRs and DMCs, and it is suitable for both RRBS and WGBS data.

In the next several sections, we will give details about the HMM-Fisher method and demonstrate its performance by comparing it with the most commonly cited method BSmooth. We will also briefly show the results of comparing HMM-Fisher with HMM-DM, methylKit, and BiSeq. We have thoroughly compared our two HMM-based methods with BSmooth, methylKit, and BiSeq in another research article (Yu and Sun, 2015). The comparison manuscript can be found at the HMM-Fisher web link https://github.com/xxy39/HMM-Fisher.

2 METHODS

General description.

The DNA methylation levels (or signals) generated by the bisulfite methylation sequencing technique are ratios ranging from 0 to 1. At each CG site the methylation level is calculated as the ratio of the number of sequencing reads with methylated cytosine (or the count of "C") to the total number of reads covering that site (or the count of "C" and "T"). DNA methylation cytosines and regions that differentiate two groups (e.g., between normal and cancerous samples) can be used as methylation biomarkers (Anglim et al., 2008; Levenson, 2010; Li and Tollefsbol, 2010; Lofton-Day et al., 2008; Paluszczak and Baer-Dubowska, 2006; Rawson and Bapat, 2012; Yang et al., 2001). Therefore, it is important to identify differentially methylated sites and regions. In order to do this, we have developed a hidden Markov model-based approach. A hidden Markov model consists of the hidden Markov chain (i.e., a sequence of hidden states) and a sequence of observed data (Baum and Petrie, 1966; Rabiner, 1989). The observed sequence and the hidden sequence are connected by emission distributions.

In this paper, we develop a new approach using a hidden Markov model to first identify the methylation state at each CG site as N (Not methylated), P (Partly methylated), or F (Fully methylated) for each sample. N represents the case when the methylation signal at a CG site is very weak, close to 0; P is when that methylation signal at a CG site is not very weak and not very strong, close to 0.5; F is when the methylation signals at a CG site is very strong, close to 1. There are two reasons of classifying the methylation signal at each CG site into these three categories. First, there are different kinds of unknown noises (errors) in methylation sequencing data. For example, when we check the methylation ratios of two technical replicates of one sample from another publicly available data (Lister et al., 2009), we have found that the correlation between their methylation signals is relatively low. Thus, the methylation ratios obtained from NGS protocols may not be very accurate due to some unknown quality issues. Second, as shown in Figure 1, the methylation signals of neighboring CG sites are very similar in some regions, but they can also be very different. It is therefore important to consider this pattern in our approach. The

Markov structure can model this similarity pattern using a high transition probability; it can also model the situation when neighboring CG sites are not very similar using a low transition probability. Because of the Markov structure's property, we use a hidden Markov model for each sample. After the methylation states at all CG sites of all samples in both groups are inferred, we then use Fisher's exact test to identify DM CG sites. DM CG sites that are close to each other are grouped together as a DMR. DM CG sites that are isolated, or are not close to other DM CG sites, are identified as individual DMCs. As a result, the HMM-Fisher can identify and report both DMRs and DMCs. Next we describe the key features of HMM-Fisher in detail.

Key features of the HMM-Fisher method

The hidden state at each CG site is N, P, or F as described above. We let the initial probability be 1/3 for all three states. The transition probabilities between any states h_i and h_{i+1} (where i=1, …, K-1), P (h_{i+1}| h_i), are given in Table 1. According to the biological meaning of the three hidden states (N, P, F), the means of the emission distributions of these states are likely to be 0, 0.5, and 1. Hence, we use the following truncated normal distributions with means 0, 0.5, and 1 as the emission distributions for each sample:

$$o_i|h_i = \begin{cases} Tnormal(0,\ \sigma_N^2) & if\ h_i = N \\ Tnormal\left(\frac{1}{2},\ \sigma_P^2\right) & if\ h_i = P \\ Tnormal(1,\ \sigma_F^2) & if\ h_i = F \end{cases}$$

For these distributions, the lower and upper bounds are set to be 0 and 1 because the methylation ratios range from 0 to 1. For each sample (or individual), the joint probability of the hidden Markov model including the hidden state sequence (H= {h_1, … , h_K}) and the observed methylation ratio sequence (O = {o_1, … , o_K}) at K CG sites is: $P(O,H) = P(O|H)P(H) = P(h_1)P(o_1|h_1)\prod_{i=2}^{K} P(h_i|h_{i-1})P(o_i|h_i)$.

After obtaining the methylation states using the above HMM approach, we conduct a Fisher's exact test to first identify differentially methylated CG sites. If the distance of two CG sites is less than 100 bases, we combine them while conducting Fisher's exact test in order to reduce the impact of small sample size and sequencing errors. For example, in Table 2 we list the count of the methylation states for

two consecutive CG sites at all eight samples; Fisher's exact test p-value is 0.0006, which means that these two groups have significant methylation difference. For the CG sites that are significantly differentially methylated, if the average methylation level in one group is higher than the one in the other group, it is defined as hypermethylation (Hyper), otherwise, it is defined as hypomethylation (Hypo). For the sites that are not significantly differentially methylated, it is defined as equally methylated (EM).

Differentially methylated sites are summarized as two types. The first type is DMCs, that is, non-consecutive singleton CG sites that have strong differential methylation signals. The second type is DMRs, that is, regions with consecutive CG sites that have strong differential methylation signals. In order to summarize and obtain DMCs and DMRs, we use the following metrics obtained from our model: p-value, DM status (or state), and distance between CG sites. The algorithm we have developed to do the summarization has been provided in the Supplemental file.

Parameter setting and methylation state estimation

As for the transition probabilities of each individual, we let the counts of the transitions $\{y_{j1}, y_{j2}, y_{j3}\}$ (that is, the changes from state j to 1, 2, and 3 respectively, for j =1, 2, 3) follow a multinomial distribution with parameters $\{n_j= y_{j1}+ y_{j2} +y_{j3,} t_{j1}, t_{j2}, t_{j3}\}$. We let each $\{t_{j1}, t_{j2}, t_{j3}\}$ follow a Dirichlet distribution, Dirichlet (1,1,1), then the posterior distribution of $\{t_{j1}, t_{j2}, t_{j3}\}$ follows a Dirichlet distribution, Dirichlet $(y_{j1}+1, y_{j2}+1, y_{j3}+1)$. As for the variances of the emission distributions, we use the empirical estimates. Using different cutoff values, we have studied the methylation variance for three types of CG sites: the CG sites with no methylation or very low methylation signals (e.g., methylation level <0.1, or <0.2), CG sites with partial methylation (e.g., 0.3< methylation level <0.7, or 0.2< methylation level <0.8), and CG sites with full methylation or very high methylation signals (e.g., methylation level >0.8, or >0.9). We then estimate the variances of these three types of CG sites based on these estimates. In particular, the default variance for the three methylation states (N, P, F) are 0.12^2, 0.15^2, and 0.13^2.

We have implemented the above HMM-Fisher method using the statistical language R. More specifically, we have developed a software package named HMM-Fisher. All the code files, example data, example scripts, and user manual can be found at the following web page: https://github.com/xxy39/HMM-Fisher.

Breast cancer dataset

To illustrate our method, we use a publicly available RRBS dataset of eight breast cancer cell lines (Sun et al., 2011). There are four estrogen receptor positive (ER+) samples (BT474, MCF7, ZR751, and T47D) and four negative (ER-) samples (BT20, MCF10A, MDAMB231, and MDAMB468). Different from other cell lines, MCF10A is a non-tumorigenic cell line. For better biological results, it is better not to include this cell line. However, we include it in our analysis because the sample size is small and our goal of using the publicly available data is to demonstrate our new statistical method. The sequenced reads of each sample are first trimmed with the BRAT (Harris et al., 2010) trimming function to remove low quality bases at both ends and then aligned afterwards. Methylation signals (or levels) are later obtained for each CG site using the BRAT *acgt-count* function. After removing CG sites with abnormally low coverage, 77,822 CG sites on chromosome 1 are used for further analysis.

Simulated dataset

To best mimic the complex patterns of DNA methylation, we simulate a dataset with known DMRs based on the first 10,000 CG sites from the breast cancer data described above. The four ER+ samples are used as the control group. In order to obtain the test group, we simulate DM regions by inserting differential methylation regions with different lengths and variation types into the control group. We simulate methylation sequencing data this way in order to ensure that the simulated DM regions can preserve the natural changes of methylation levels across CG sites and among samples. The simulation is conducted in three steps. First, the 10,000 CG sites are grouped into different types of regions based on their methylation levels and the variation across different samples in the control group. There are four types of

regions: H, High-methylation regions with small cross-sample variation; L, low-methylation regions with small cross-sample variation; M-H, High-methylation regions with large cross-sample variation; and M-L, low-methylation regions with large cross-sample variation. This step generates 2,459 regions. Second, from these regions, we randomly choose 80 regions covering 929 CG sites to insert methylation differences. These DMRs are designed to have different lengths (1~76 CGs) and methylation patterns. Third, within these selected regions, the methylation levels of the test group are sampled from uniform distributions (Table 3). To create methylation differences, for H and M-H regions that have high methylation level in the control group, the test group is simulated with relatively low methylation level; similarly for L and M-L regions, the test group is simulated with relatively high methylation levels. In addition, to ensure a true difference in regions with large cross-sample variation and small size (< 3 CGs), more stringent uniform distributions are used for these regions. For further description of the simulation, see the Supplemental file.

3 RESULTS

Simulated dataset

We apply HMM-Fisher to the simulated dataset described in the Methods section. The DM CG sites are identified with different p-value thresholds. With the p-value threshold set at 0.05, HMM-Fisher identifies 1,171 DM CG sites, achieving a sensitivity of 97.20% and a specificity of 97.01%. Out of the 80 selected DMRs, 55 regions are completely identified, 16 regions are partially identified, and 9 singletons are not identified. Figure 2 shows a simulated DMR that is successfully identified by HMM-Fisher. Within the 14-CG region (Figure 2, upper panel), the test group exhibits a consistently low methylation level in all four blue samples. While in the control group, three samples have relatively high methylation, but one sample has fairly low methylation (Figure 2, upper panel). This large cross-sample methylation variation may make it difficult for the traditional methods to identify such a DMR. However, HMM-Fisher detects

this region as hypermethylated with a small p-value, suggesting that our method can easily handle the CG sites with large variation among samples. The lower panel of Figure 2 shows the estimated state for each CG by the HMM step. For the simulated DM CGs (shaded box), the estimated states accurately reflect the raw methylation levels, that is, N state for the test group and N/P/F states for the control group. Then in the Fisher's exact test step, the neighboring CG sites are combined together, yielding p-values that are less than 0.007 for all 14 DM CG sites.

The abilities of HMM-Fisher to detect differentially methylated regions with different sizes and different cross-sample variation types are described in Figure 3. The 80 simulated DMRs are separated into two types based on their sizes (number of CG sites included), long regions with at least 20 CG sites and short regions with less than 20 CG sites. DMRs are also classified according to their variation across samples, regions with small variation (H and L regions) and regions with large variation (M-H and M-L regions). For CG sites within each region type, the sensitivity is calculated for different thresholds of p-value. In general, HMM-Fisher shows high sensitivities for both long and short regions for all p-value thresholds. For long regions, almost 100% of simulated DM CG sites are identified (Figure 3, green line). For short regions, the sensitivities are slightly lower but are still $\geq$ 95% for most thresholds (Figure 3, blue line), suggesting that HMM-Fisher is powerful in detecting differential methylation signals that occur in small clusters. A similar pattern is seen in regions with different types of cross-sample variation. The small-variation regions have a high sensitivity as we expected. HMM-Fisher also shows a high sensitivity for regions with large variation, indicating that our method can accurately identify DM CG sites even with large cross-sample variation.

To evaluate our method, we compare HMM-Fisher with the commonly cited method BSmooth (Hansen et al., 2012) using the simulated dataset. We set parameters for the smoothing step in BSmooth to be comparable to HMM-Fisher. The smoothing windows of BSmooth are required to have at least 1 CG and to be at least 5 bp in length. The maximum distance between any two consecutive CG sites to break the smoothing window is set at 100 bp. In the modified t-test step, all 10,000 CG sites are tested for

group differences, and the variance is estimated for the control group. Any CG site with a t-statistic beyond a certain threshold is defined as differentially methylated.

Table 4 shows the sensitivity and false positive rate (FPR) for both HMM-Fisher and BSmooth using different thresholds of their statistics. In general, HMM-Fisher (Table 4A) achieves much higher sensitivity and much lower FPR compared with BSmooth (Table 4B). Then for each method we choose the threshold that shows the best performance and then compare them in detail in Table 5. With the p-value cutoff of 0.03, HMM-Fisher yields a sensitivity of 97.20% and a FPR of 2.44%, while the t-statistic threshold of 2.5 in BSmooth yields a much lower sensitivity of 78.33% and a much higher FPR of 10.12%. Moreover, HMM-Fisher is more powerful in identifying DMRs with small sizes and large cross-sample variations. For regions with less than 20 CG sites, HMM-Fisher detects 95.73% CGs while BSmooth identifies 77.02%; for regions with large cross-sample variation, BSmooth only detects 42.14% while HMM-Fisher achieves a significantly higher sensitivity of 92.50%.

Breast cancer dataset

To illustrate the application of our method, we apply HMM-Fisher to a real breast cancer dataset to identify the differentially methylated CGs and regions between the ER+ and ER- groups. To ensure biological meaning for differential methylation, we require that only CG sites with a mean difference of methylation level between the two groups ≥ 0.3 can be defined as differentially methylated. The DM CG sites where the ER+ has a higher methylation level than ER- are classified as hypermethylated DM, and the DM CG sites where the ER- has a higher methylation level are classified as hypomethylated DM. With the p-value threshold set at 0.05, HMM-Fisher identifies 1,917 DM CG sites, forming 805 DMRs with a median length of 8 bp (minimum 1 bp; maximum 305 bp). The majority (73.03%) of the differentially methylated CG sites are hypomethylated in ER+ group, while only 26.95% are hypermethylated. In addition, to examine the performance of HMM-Fisher in regions with different cross-sample variations, we classify all identified DM CGs based on their variation status. The CG sites where the four samples in one group all have methylation levels ≥ 0.6 or all have methylation levels ≤ 0.4 are

defined as small-variation in that group; otherwise, the CG sites will be defined as large-variation in that group. Of the 1,917 DM CG sites identified by HMM-Fisher, a majority of them have large variation in either one group (48.4%) or both groups (40.1%), while only a small proportion (16.5%) have small variations in both groups. This finding further confirms that HMM-Fisher can handle the CG sites with large cross-sample variations. Moreover, 1,348 of the identified DM CG sites are associated with a total of 227 genes, and these DM sites are located in the gene body (1,119 CGs) and/or promoter regions (287 CGs) of these 227 genes. The 10 most frequent genes are listed in Table 6. Most of these genes are highly methylated in ER- compared to ER+.

In the Introduction section, we mention that we have developed another HMM-based method HMM-DM (Yu and Sun, 2015a, 2015b) for the identification of differential methylation. In order to have a better understanding of the performance of our HMM-based methods, in addition to BSmooth, we have thoroughly compared HMM-Fisher and HMM-DM with two other commonly cited methods, methylKit (Akalin et al., 2012) and BiSeq. The results of comparing these 5 methods are documented in detail in another manuscript that we have submitted (Yu and Sun, 2015). Our 5-method comparison results show that our HMM-methods have better performance as shown below. First, HMM-based methods have higher sensitivity: HMM-DM (97.74%), HMM-Fisher (97.20%), BiSeq (96.23%), methylKit (88.27%), BSmooth (66.63%). Second, HMM-based methods have lower false discovery rate: HMM-DM (1.77%), HMM-Fisher (2.44%), BiSeq (3.75%), methylKit (4.27%), and BSmooth (5.13%). Finally, HMM-based methods are more robust in identifying DM regions that are short and/or have large methylation variation.

4 DISCUSSION

In the Methods section, we have introduced the statistical features of the HMM-Fisher and this method is implemented using R scripts as a software package named HMM-Fisher. There are other useful features of HMM-Fisher as a software package. First, users can conduct quality control to select the CG sites with a certain percentage of observations (e.g., 80%) to reduce the impact of missing values. Second, as a

Markov model-based method, it is important to check the convergence of the hidden Markov chain. Therefore, we have provided an R function to check the joint probability of the hidden Markov model and plot the joint probability of different iterations. In fact, the hidden Markov model used in the HMM-Fisher converges after a few iterations. Third, one important feature of HMM-Fisher is that it can generate both the differentially methylated cytosine singletons DMCs and DMRs. When running the HMM-Fisher, the default setting is to produce both DMCs and DMRs in one output file. If users only want to generate output files for DMRs, they can achieve this simply by setting the "singleton" option to be "FALSE" or remove the DM regions with length equal to 1 (that is, the DMCs) from the output file. Fourth, in order to make it easy for users to further analyze and interpret the DM cytosines and regions they obtain, we have also provided the R function to add genetic annotations for each DM site. That is, users can obtain the gene names with which their significant DM sites are associated. For detailed information about the above features, users may check section 4 and 5 of the HMM-Fisher User Manual, which can be found at HMM-Fisher's web page.

In addition to the DM identification methods reviewed in the Introduction section, there are some methods developed for identifying DM between two groups, e.g., ABCD-DNA (Robinson et al., 2012), COHCAP (Warden et al., 2013), and M&M algorithm (Zhang et al., 2013). These methods integrate with other types of genetic or epigenetic data (e.g., gene expression, copy number variation, or other types of sequencing data). There are also some methods developed for epigenome-wide association studies (EWAS) (Liu et al., 2013; Zou et al., 2014). In addition, other DMR identification methods, such as, IMA (Wang et al., 2012), A-clustering (Sofer et al., 2013), Minfi (Aryee et al., 2014), MethyAnalysis (Du and Bourgon, 2014), DMRCate (Peters et al., 2015), Probe Lasso (Butcher and Beck, 2015), and BumpHunter (Jaffe et al., 2012) are motivated by and/or designed for Illumina methylation array data. These methods may be applied directly to or modified for bisulfite sequencing data. Their performance on bisulfite sequencing data has not been studied yet. Because HMM-Fisher is developed for DM identification between two groups using bisulfite sequencing data, it is not designed

for Illumina array data, EWAS, or data integration. Therefore, we have not reviewed these papers in the Introduction section. However, these methods do have good ideas and advantages that are worth reviewing.

5 CONCLUSIONS

In this paper, we have developed a new statistical approach, HMM-Fisher, to identify differentially methylated regions. This approach is developed based on a hidden Markov model and Fisher's exact test. We have shown this approach by comparing it with currently available methods using both simulated data and real data. Our results show that this new approach has several advantages. First, it is a data-driven method that identifies both DMRs and DMCs in an entire genome, both inside and outside of predefined CpG islands or genes. Second, it incorporates neighboring CG site methylation information and thus reduces the impact of sequencing errors. Hence, it is a robust and efficient method. Third, it is suitable for the methylation sequencing data generated from both the whole genome sequencing and target-region sequencing.

AUTHORS' CONTRIBUTIONS

SS developed the statistical method and wrote the original R script. XY revised and improved the R script to implement the method. Both authors have been involved in writing the manuscript and approved the final document.

ACKNOWLEDGEMENTS

Xiaoqing Yu's stipend was provided by the Case Comprehensive Cancer Center when she was a graduate student at Case Western Reserve University. This work was also 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 Biology, 13, R87.
- Akman, K., T. Haaf, S. Gravina, J. Vijg and A. Tresch (2014): "Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data," Bioinformatics, 30, 1933-1934.
- Anglim, P. P., T. A. Alonzo and I. A. Laird-Offringa (2008): "DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update," Mol Cancer, 7, 81.
- Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry (2014): "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays," Bioinformatics, 30, 1363-1369.
- Baum, L. E. and T. Petrie (1966): "Statistical inference for probabilistic functions of finite state Markov chains," Annals of Mathematical Statistics 37, 1554-1563.
- Becker, C., J. Hagmann, J. Muller, D. Koenig, O. Stegle, K. Borgwardt and D. Weigel (2011): "Spontaneous epigenetic variation in the Arabidopsis thaliana methylome," Nature, 480, 245-249.
- Bell, A. C. and G. Felsenfeld (2000): "Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene," Nature, 405, 482-485.
- Butcher, L. M. and S. Beck (2015): "Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data," Methods (San Diego, Calif.), 72, 21-28.
- Challen, G. A., D. Sun, M. Jeong, M. Luo, J. Jelinek, J. S. Berg, C. Bock, A. Vasanthakumar, H. Gu, Y. Xi, S. Liang, Y. Lu, G. J. Darlington, A. Meissner, J.-P. J. Issa, L. A. Godley, W. Li and M. A. Goodell (2011): "Dnmt3a is essential for hematopoietic stem cell differentiation," Nature genetics, 44, 23-31.
- Chen, P. Y., S. J. Cokus and M. Pellegrini (2010): "BS Seeker: precise mapping for bisulfite sequencing," BMC Bioinformatics, 11, 203.
- Clark, S. J., J. Harrison and P. L. Molloy (1997): "Sp1 binding is inhibited by (m)Cp(m)CpG methylation," Gene, 195, 67-71.
- Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008): "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning," Nature, 452, 215-219.
- Cox, M. P., D. A. Peterson and P. J. Biggs (2010): "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data," BMC Bioinformatics, 11, 485.
- Dai, M., R. C. Thompson, C. Maher, R. Contreras-Galindo, M. H. Kaplan, D. M. Markovitz, G. Omenn and F. Meng (2010): "NGSQC: cross-platform quality analysis pipeline for deep sequencing data," BMC Genomics, 11 Suppl 4, S7.
- Dolan, P. C. and D. R. Denver (2008): "TileQC: a system for tile-based quality control of Solexa data," BMC Bioinformatics, 9, 250.
- 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.
- Douet, V., M. B. Heller and O. Le Saux (2007): "DNA methylation and Sp1 binding determine the tissue-specific transcriptional activity of the mouse Abcc6 promoter," Biochem Biophys Res Commun, 354, 66-71.
- Du, P. and R. Bourgon (2014): "methyAnalysis: DNA methylation data analysis and visualization," R package version 1.10.0.
- Feng, H., K. N. Conneely and H. Wu (2014): "A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data," Nucleic Acids Research, 42, e69-e69.
- Goll, M. G. and T. H. Bestor (2005): "Eukaryotic cytosine methyltransferases," Annu Rev Biochem, 74, 481-514.

- Hansen, K., B. Langmead and R. Irizarry (2012): "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions," Genome Biology, 13, R83.
- Hansen, K. D., W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry and A. P. Feinberg (2011): "Increased methylation variation in epigenetic domains across cancer types," Nat Genet, 43, 768-775.
- Hark, A. T., C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse and S. M. Tilghman (2000): "CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus," Nature, 405, 486-489.
- 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, 1647–1653.
- Henderson, I. R. and S. E. Jacobsen (2007): "Epigenetic inheritance in plants," Nature, 447, 418-424.
- Inoue, S. and M. Oishi (2005): "Effects of methylation of non-CpG sequence in the promoter region on the expression of human synaptotagmin XI (syt11)," Gene, 348, 123-134.
- 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," International Journal of Epidemiology, 41, 200-209.
- Jiang, P., K. Sun, F. M. F. Lun, A. M. Guo, H. Wang, K. C. A. Chan, R. W. K. Chiu, Y. M. D. Lo and H. Sun (2014): "Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis," PLoS ONE, 9, e100360.
- Kim, M. S., J. Lee and D. Sidransky (2010): "DNA methylation markers in colorectal cancer," Cancer Metastasis Rev, 29, 181-206.
- Kitazawa, S., R. Kitazawa and S. Maeda (1999): "Transcriptional regulation of rat cyclin D1 gene by CpG methylation status in promoter region," J Biol Chem, 274, 28787-28793.
- Levenson, V. V. (2010): "DNA methylation as a universal biomarker," Expert Rev Mol Diagn, 10, 481-488.
- 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.
- Li, Y. and T. O. Tollefsbol (2010): "Impact on DNA methylation in cancer prevention and therapy by bioactive dietary components," Curr Med Chem, 17, 2141-2151.
- 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.
- Lin, J., M. Lai, Q. Huang, Y. Ma, J. Cui and W. Ruan (2007): "Methylation patterns of IGFBP7 in colon cancer cell lines are associated with levels of gene expression," J Pathol, 212, 83-90.
- Lin, X., D. Sun, B. Rodriguez, Q. Zhao, H. Sun, Y. Zhang and W. Li (2013): "BSeQC: quality control of bisulfite sequencing experiments," Bioinformatics, 29, 3227-3229.
- Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker (2008): "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," Cell, 133, 523-536.
- 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. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M.

Evans and J. R. Ecker (2011): "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells," Nature, 471, 68-73.
- Liu, Y., M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, K. Shchetynsky, A. Scheynius, J. Kere, L. Alfredsson, L. Klareskog, T. J. Ekström and A. P. Feinberg (2013): "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in Rheumatoid Arthritis," Nature biotechnology, 31, 142-147.
- Lofton-Day, C., F. Model, T. Devos, R. Tetzner, J. Distler, M. Schuster, X. Song, R. Lesche, V. Liebenberg, M. Ebert, B. Molnar, R. Grutzmann, C. Pilarsky and A. Sledziewski (2008): "DNA methylation biomarkers for blood-based colorectal cancer screening," Clin Chem, 54, 414-423.
- Mancini, D. N., S. M. Singh, T. K. Archer and D. I. Rodenhiser (1999): "Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors," Oncogene, 18, 4108-4119.
- Martinez-Alcantara, A., E. Ballesteros, C. Feng, M. Rojas, H. Koshinsky, V. Y. Fofanov, P. Havlak and Y. Fofanov (2009): "PIQA: pipeline for Illumina G1 genome analyzer data quality assessment," Bioinformatics, 25, 2438-2439.
- 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.
- Paluszczak, J. and W. Baer-Dubowska (2006): "Epigenetic diagnostics of cancer--the application of DNA methylation markers," J Appl Genet, 47, 365-375.
- 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.
- Pedersen, B., T. F. Hsieh, C. Ibarra and R. L. Fischer (2011): "MethylCoder: software pipeline for bisulfite-treated sequences," Bioinformatics, 27, 2435-2436.
- Peters, T. J., M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V Lord, S. J. Clark and P. L. Molloy (2015): "De novo identification of differentially methylated regions in the human genome," Epigenetics & Chromatin, 8, 6.
- Rabiner, L. R. (1989): "A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition," P Ieee, 77, 257-286.
- Rawson, J. B. and B. Bapat (2012): "Epigenetic biomarkers in colorectal cancer diagnostics," Expert Rev Mol Diagn, 12, 499-509.
- Robinson, M. D., D. Strbenac, C. Stirzaker, A. L. Statham, J. Song, T. P. Speed and S. J. Clark (2012): "Copy-number-aware differential analysis of quantitative DNA sequencing data," Genome Research, 22, 2489-2496.
- Saito, Y., J. Tsuji and T. Mituyama (2014): "Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions," Nucleic Acids Research, 42, e45.
- Sofer, T., E. D. Schifano, J. A. Hoppin, L. Hou and A. A. Baccarelli (2013): "A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure," Bioinformatics, 29, 2884-2891.
- Stockwell, P. A., A. Chatterjee, E. J. Rodger and I. M. Morison (2014.): "DMAP: Differential Methylation Analysis Package for RRBS and WGBS data," Bioinformatics advanced online publication, doi:10.1093/bioinformatics/btu1126.
- Su, J., H. Yan, Y. Wei, H. Liu, H. Liu, F. Wang, J. Lv, Q. Wu and Y. Zhang (2013): "CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data," Nucleic Acids Research, 41, e4-e4.
- Suijkerbuijk, K. P., P. J. van Diest and E. van der Wall (2011): "Improving early breast cancer detection: focus on methylation," Ann Oncol, 22, 24-29.
- Sun, D., Y. Xi, B. Rodriguez, H. Park, P. Tong, M. Meong, M. Goodell and W. Li (2014): "MOABS: model based analysis of bisulfite sequencing data," Genome Biology, 15, R38.

- Sun, S., A. Noviski and X. Yu (2013): "MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment," BMC Bioinformatics, 14, 259.
- 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.
- Toyota, M., N. Ahuja, M. Ohe-Toyota, J. G. Herman, S. B. Baylin and J. P. Issa (1999): "CpG island methylator phenotype in colorectal cancer," Proc Natl Acad Sci U S A, 96, 8681-8686.
- Wang, D., L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia and S. Liu (2012): "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data," Bioinformatics, 28, 729-730.
- 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," Nucleic Acids Research, 41, e117-e117.
- 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.
- 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.
- Yang, X., L. Yan and N. E. Davidson (2001): "DNA methylation in breast cancer," Endocr Relat Cancer, 8, 115-127.
- Yu, X. and S. Sun (2015): "Comparing five statistical methods of differential methylation identification using bisulfite sequencing data," Manuscript submitted for publication.
- Yu, X. and S. Sun (2015a): "HMM-DM: identifying differentially methylated regions using a hidden Markov model," Manuscript submitted for publication.
- Yu, X. and S. Sun (2015b): "HMM-DM," GitHub repository, https://github.com/xxy39/HMM-DM.
- Zhang, B., Y. Zhou, N. Lin, R. F. Lowdon, C. Hong, R. P. Nagarajan, J. B. Cheng, D. Li, M. Stevens, H. J. Lee, X. Xing, J. Zhou, V. Sundaram, G. Elliott, J. Gu, T. Shi, P. Gascard, M. Sigaroudinia, T. D. Tlsty, T. Kadlecek, A. Weiss, H. O'Geen, P. J. Farnham, C. L. Maire, K. L. Ligon, P. A. F. Madden, A. Tam, R. Moore, M. Hirst, M. A. Marra, B. Zhang, J. F. Costello and T. Wang (2013): "Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm," Genome Research, 23, 1522-1540.
- Zhang, Y., H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, X. Li, Q. Wu, F. Wang and Y. Cui (2011): "QDMR: a quantitative method for identification of differentially methylated regions by entropy," Nucleic Acids Research, 39, e58-e58.
- Zitt, M. and H. M. Muller (2007): "DNA methylation in colorectal cancer--impact on screening and therapy monitoring modalities?," Dis Markers, 23, 51-71.
- Zou, J., C. Lippert, D. Heckerman, M. Aryee and J. Listgarten (2014): "Epigenome-wide association studies without the need for cell-type composition," Nat Meth, 11, 309-311.

Tables

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

For any j =1, 2, and 3, $t_{j1}+t_{j2}+t_{j3} = 1$.

$H_i \rightarrow h_{i+1}$	N	P	F
N (Not methylated)	t_{11}	t_{12}	t_{13}
P (Partly methylated)	t_{21}	t_{22}	t_{23}
F (Fully methylated)	t_{31}	t_{32}	t_{33}

Table 2. A contingency table example for Fisher's exact test at two adjacent CG sites.

	N	P	F
ER+	0	3	5
ER-	7	1	0

Table 3. Uniform distributions that are used to simulate the DMRs for the test group.

	> 3 CGs	≤ 3 CGs
H DMRs	Uniform (0, 0.4)	Uniform (0, 0.2)
L DMRs	Uniform (0.6, 1)	Uniform (0.8, 1)
M-H DMRs	Uniform (0, 0.3)	Uniform (0, 0.2)
M-L DMRs	Uniform (0.7, 1)	Uniform (0.8, 1)

Table 4. Sensitivity and FPR (%) of HMM-Fisher and BSmooth

A. HMM-Fisher with different p-value thresholds

Threshold of p-value	0.01	0.03	0.05	0.08	0.1	0.15	0.2
Sensitivity	91.93	97.20	97.20	97.85	97.85	98.92	98.92
FPR	1.44	2.44	2.99	2.99	5.36	7.21	9.28

B. BSmooth with different t-statistics thresholds

t-statistics threshold	**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	78.33	72.47	70.72	68.78	66.63	62.40	50.00
FPR	18.60	17.48	15.53	14.25	10.12	8.21	6.93	5.99	5.14	4.20	3.33

Table 5. Comparing the performance of HMM-Fisher and BSmooth.

HMM-Fisher p-value cutoff is 0.05 and BSmooth modified t-statistics threshold is 2.5. Six metrics are considered: the sensitivity and FPR for all simulated DMRs, sensitivity for DMRs with ≥ 20 CGs, DMRs with < 20 CGs, DMRs with small, and DMRs with large variation (see the first column).

	HMM-Fisher	BSmooth
Sensitivity , all DMRs	97.20%	78.33%
FPR, all DMRs	2.99%	10.12%
Sensitivity, DMR ≥ 20	99.03%	81.35%
Sensitivity, DMR < 20	95.73%	77.02%
Sensitivity, small-variation DMR	99.23%	89.98%
Sensitivity, large-variation DMR	92.50%	42.14%

Table 6. Top 10 genes with identified DM CGs.

This table shows chromosome location, gene name, number of CGs located in the gene body and promoter region for the top 10 frequent genes. Number of hypermethylated and hypomethylated CGs are shown in brackets, separated by slash. "-" indicates that no hyper or hypo DM CGs are identified in that specific genomic regions.

Location	**Gene name**	**CGs in gene body (hyper/hypo)**	**CGs in Promoter (hyper/hypo)**
1p36.32	AJAP1	68 (-/67)	24 (-/24)
1p34.3	GRIK3	49 (9/40)	43 (-/43)
1p36.31	CAMTA1	48 (7/41)	-
1p36.23-p33	PRDM16	34 (10/24)	-
1p21-22	NTRK1	22 (-/22)	7 (2/5)
1p21	LOC100129620	28 (-/28)	-
1p21.3	LPPR5	17 (-/17)	11 (-/11)
1p13.1	NHLH2	27 (-/27)	-
1p34	HIVEP3	22 (14/8)	-
1p36.33	RNF223	17 (-/17)	-

Figures

Figure 1. The methylation patterns of two samples.

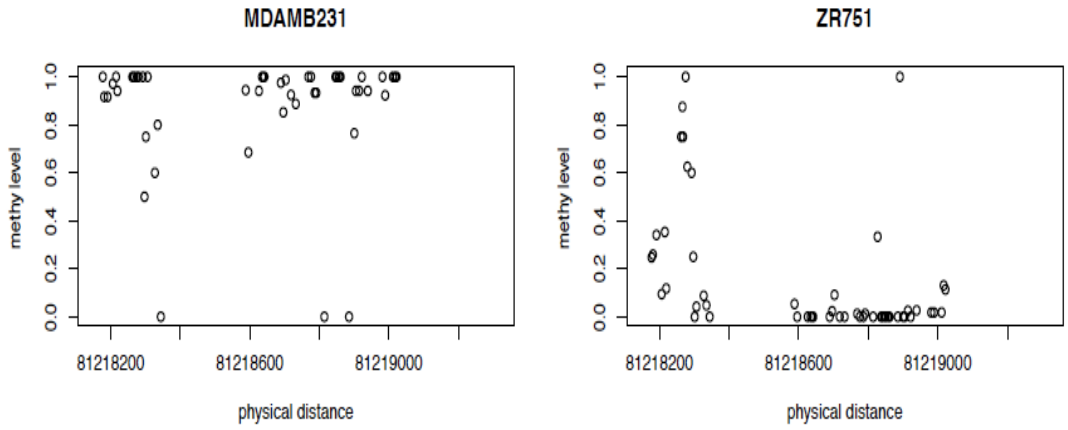

Figure 1

Figure 2. A simulated differentially methylated region identified by HMM-Fisher.

The upper panel shows the raw methylation levels in the control group (red) and the test group (blue), respectively. The lower panel shows the estimated state (N, P, F) for each CG in the control group (red) and the test group (blue). The shaded box indicates the simulated differentially methylated region.

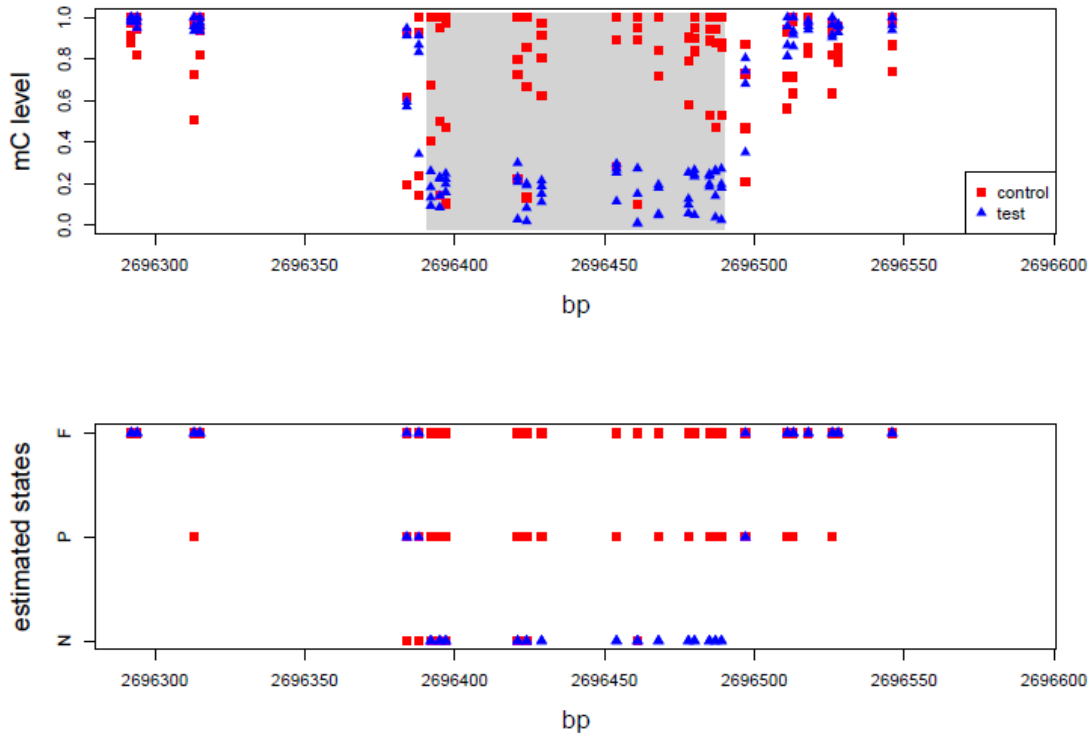

Figure 2

Figure 3. Sensitivity of HMM-Fisher to detect DMRs with different sizes and variation types.

Shown are the overall sensitivity (black), as well as sensitivity for DMRs with small across-sample variation (red), large across-sample variation (purple), long regions with at least 20 CGs (green), and short regions with less than 20 CGs (blue).

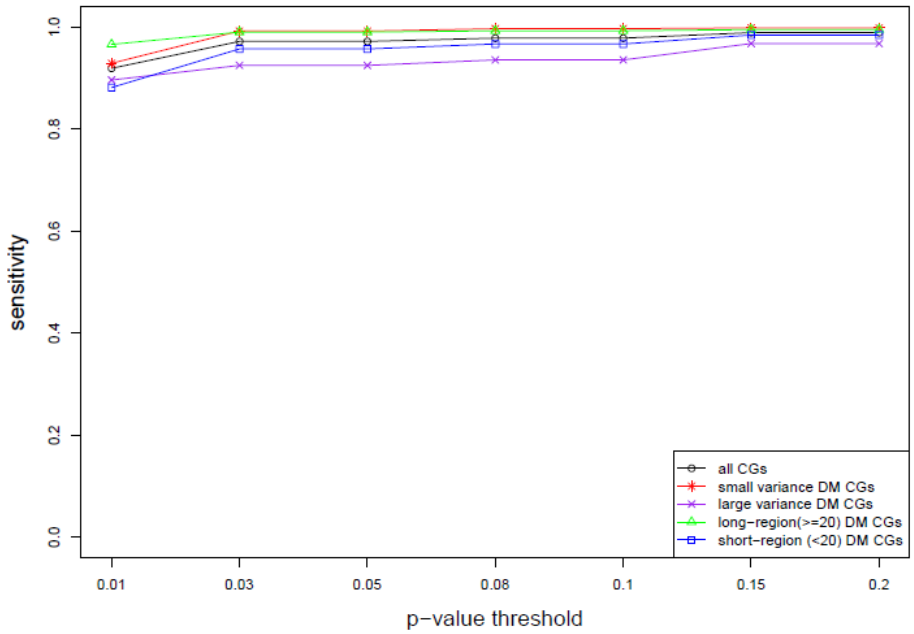

Figure 3