RESEARCH Open Access

# Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions

Yutaka Saito[1,2], Toutai Mituyama[1,2*]

*From* Joint 26th Genome Informatics Workshop and Asia Pacific Bioinformatics Network (APBioNet) 14th International Conference on Bioinformatics (GIW/InCoB2015)
Tokyo, Japan. 9-11 September 2015

## Abstract

**Background:** Detection of differential methylation between biological samples is an important task in bisulfite-seq data analysis. Several studies have attempted de novo finding of differentially methylated regions (DMRs) using hidden Markov models (HMMs). However, there is room for improvement in the design of HMMs, especially on emission functions that evaluate the likelihood of differential methylation at each cytosine site.

**Results:** We describe a new HMM for DMR detection from bisulfite-seq data. Our method utilizes emission functions that combine binomial models for aligned read counts, and beta mixtures for incorporating genome-wide methylation level distributions. We also develop unsupervised learning algorithms to adjust parameters of the beta-binomial models depending on differential methylation types (up, down, and not changed). In experiments on both simulated and real datasets, the new HMM improves DMR detection accuracy compared with HMMs in our previous study. Furthermore, our method achieves better accuracy than other methods using Fisher's exact test and methylation level smoothing.

**Conclusions:** Our method enables accurate DMR detection from bisulfite-seq data. The implementation of our method is named ComMet, and distributed as a part of Bisulfighter package, which is available at http://epigenome.cbrc.jp/bisulfighter.

## Background

Cytosine methylation is an epigenetic modification that affects many biological processes including normal development and pathogenesis [1]. Genome-wide profiling of cytosine methylation is enabled by bisulfite-seq, where unmethylated cytosines are converted and sequenced as thymines [2]. In bisulfite-seq data analysis, a fundamental task is alignment of bisulfite-converted reads to a reference genome, and thus numerous tools have been already developed [3-6]. On the other hand, methods for downstream tasks after read alignment have been relatively limited [7]. Among them, one of the most important is detection of differential methylation between biological samples [8]. Differential methylation analyses can be divided into two categories: those focusing only on pre-specified regions such as known transcription factor binding sites (*e.g.* [9]), and those for de novo finding of differentially methylated regions (DMRs) as novel candidates of regulatory elements (*e.g.* [10]). In this paper, we address the latter case, which is more challenging due to the necessity for determining exact boundaries of DMRs.

* Correspondence: mituyama-toutai@aist.go.jp
[1] Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, 135-0064 Tokyo, Japan
Full list of author information is available at the end of the article

© 2015 Saito and Mituyama This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

DMR detection has been attempted by two-step procedures: first, differentially methylated cytosines (DMCs) are detected by comparison of alignment results between samples; then, DMCs at neighbor positions are grouped as contiguous DMRs by certain distance criteria. Most studies have focused mainly on the first step, and proposed to detect DMCs using Fisher's exact test [10], Student's t-test with methylation level smoothing [11], and logistic regression test [12]. Additionally, many methods have been developed for detecting DMCs based on a variety of probability models [13-15]. In contrast, there have been much less studies on methods for grouping DMCs into DMRs. Although fixed-length distance criteria (*e.g.* sliding windows) have been conventionally used, such strategies depend on the choice of distance parameters (*e.g.* window sizes). Unfortunately, it is difficult to adjust distance parameters empirically because DMR lengths range from hundreds of base pairs as in CpG islands, to millions of base pairs as in cancer aberrations [16].

To address this problem, we have recently proposed a framework for DMR detection based on hidden Markov models (HMMs) [6]. Unlike the two-step procedures, HMMs can integrate detection and grouping of DMCs as joint probability models using emission and transition functions, respectively. Moreover, HMMs enable us to adjust their parameters by well-established learning algorithms so that they incorporate useful information for DMR detection. In particular, we have observed that DMCs exhibit distance distributions distinct from cytosines whose methylation is not changed. Therefore, we have adjusted parameters of transition functions so that they fit these distance distributions. Thanks to this strategy, our method has improved DMR detection accuracy, especially on determining exact boundaries of DMRs. We note that HMM-based DMR detection has also been employed for methylation data other than bisulfite-seq such as Infinium BeadChip [17] and MBDCap-seq [18].

While our previous study has shown the effectiveness of transition functions in HMM-based DMR detection, there is still room for improvement in the design of emission functions. As mentioned above, many studies have proposed various probability models for detecting DMCs [13-15]. An important suggestion from these studies is that DMC detection at individual cytosine sites can be improved by considering probability distributions of methylation levels collected from all genomic cytosine sites. This implies that the information of genome-wide methylation level distributions may also be useful for DMR detection. However, the probability models in [13-15] are specifically developed for DMC detection, and thus cannot be directly applied to the emission functions for HMM-based DMR detection.

In this paper, we describe new emission functions for HMM-based DMR detection from bisulfite-seq data. We show that the emission functions in our previous study [6] have an empirical parameter to represent methylation levels used in binomial models for aligned read counts. From this viewpoint, we propose new emission functions that replace the empirical parameter by beta mixtures for incorporating genome-wide methylation level distributions. We also develop unsupervised learning algorithms to adjust parameters of the beta-binomial models depending on differential methylation types (up, down, and not changed). In experiments on both simulated and real datasets, the new emission functions improve DMR detection accuracy compared with the old ones. Furthermore, our HMM-based method achieves better accuracy than other methods using Fisher's exact test and methylation level smoothing.

## Methods

In this section, we describe a new method for DMR detection from bisulfite-seq data. The method uses new emission functions with an HMM-based framework called ComMet which we developed in our previous study [6]. We first review ComMet, and show that emission functions in our previous study have an empirical parameter to represent methylation levels used in binomial models for aligned read counts. Then, we design new emission functions replacing this empirical parameter by beta mixtures for incorporating genome-wide methylation level distributions. We also present unsupervised learning algorithms to adjust parameters of the beta-binomial models depending on differential methylation types (up, down, and not changed).

### HMM-based DMR detection from bisulfite-seq data

In our previous study [6], we developed ComMet, an HMM-based framework for DMR detection from bisulfite-seq data (Figure 1). The motivation for employing HMMs came from our observation of real data where DMCs showed distance distributions distinct from CpGs whose methylation was not changed. We incorporated these distributions into transition functions of HMMs. ComMet uses the state transition diagram shown in Figure 1a where transition probabilities among Up, Down, and NoCh states represent distinct distance distributions among DMCs. ComMet adjusts transition probabilities for each dataset to be analyzed using expectation-maximization algorithms. ComMet detects DMRs by dynamic programming algorithms that maximize log-likelihood ratio scores $\log \frac{P(\text{region},\ dir)}{P(\text{region},\ \text{NoCh})}$, where *dir* (= Up or Down) is the direction of differential methylation. The output of ComMet is a list of DMRs ranked by their log-likelihood ratio scores.

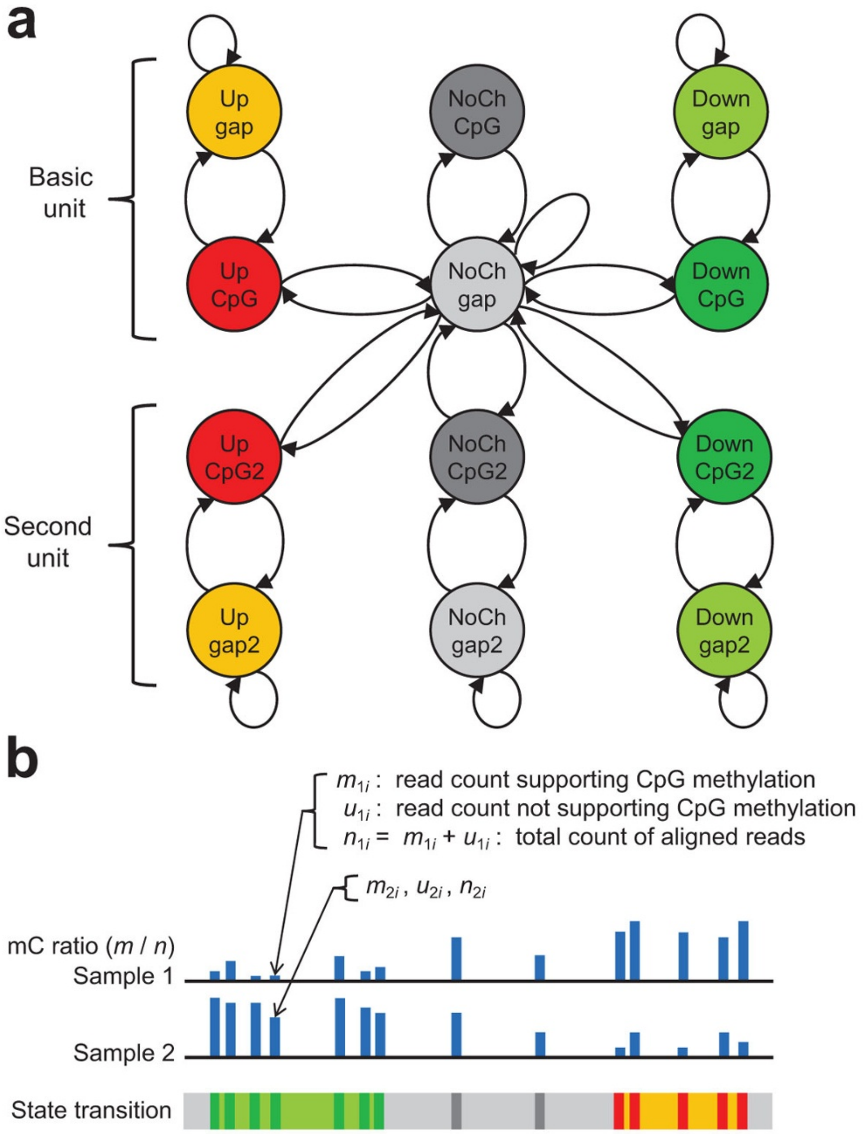

a
Up gap
NoCh CpG
Down gap
Basic unit
Up CpG
NoCh gap
Down CpG
Up CpG2
NoCh CpG2
Down CpG2
Second unit
Up gap2
NoCh gap2
Down gap2
b
$m_{1i}$ : read count supporting CpG methylation
$u_{1i}$ : read count not supporting CpG methylation
$n_{1i} = m_{1i} + u_{1i}$ : total count of aligned reads
$m_{2i}$ , $u_{2i}$ , $n_{2i}$
mC ratio ($m$ / $n$)
Sample 1
Sample 2
State transition

**Figure 1 ComMet: an HMM-based framework for DMR detection from bisulfite-seq data**. (a) HMM architecture. The HMM has pairs of states for CpG positions and their interval positions (named *gap*), each of which has three types of differential methylation: hypermethylation (Up), hypomethylation (Down), and no change (NoCh). Transition probabilities among Up, Down, and NoCh states represent distinct distance distributions among DMCs. Throughout this study, we use the dual architecture consisting of the basic and second units since it can achieve better accuracy than the basic unit only [6]. (b) Example of input bisulfite-seq data and corresponding state transitions. Colors in the state transition track correspond to those in (a).

While transition functions of ComMet incorporated distance distributions of DMCs, the design of emission functions was not well established in our previous study. Given alignment results of bisulfite-converted reads, we can observe the counts of reads supporting CpG methylation as the number of C-C matches, and the counts of reads not supporting CpG methylation as the number of C-T mismatches (Figure 1b). Let us denote the count of reads supporting methylation by $m_{si}$, the count of reads not supporting methylation by $u_{si}$, and the total count of aligned reads by $n_{si}$, for each CpG site $i$ and each sample $s$ = 1, 2. If a CpG site is differentially methylated, the counts can be considered to be taken from separate probability distributions reflecting the difference of methylation levels between two samples. On the other hand, if a CpG site is not differentially methylated, the counts should be the consequence of the common methylation level. Therefore, in our previous study, emission functions for CpG states were designed as follows:

$$
\begin{aligned}
e_i^{\mathrm{U}} &= \mathrm{Binom}(m_{1i}|n_{1i},\ \theta_{1\,i}^{\mathrm{U}})\mathrm{Binom}(m_{2\,i}|n_{2\,i},\ \theta_{2\,i}^{\mathrm{U}}),\\
e_i^{\mathrm{D}} &= \mathrm{Binom}(m_{1i}|n_{1i},\ \theta_{1i}^{\mathrm{D}})\mathrm{Binom}(m_{2i}|n_{2i},\ \theta_{2i}^{\mathrm{D}}),\\
e_i^{\mathrm{N}} &= \mathrm{Binom}(m_{1\,i}|n_{1\,i},\ \theta_{0i}^{\mathrm{N}})\mathrm{Binom}(m_{2i}|n_{2i},\ \theta_{0i}^{\mathrm{N}}),
\end{aligned}
$$

where U, D, and N represent Up, Down, and NoCh states, respectively, Binom() is a binomial distribution, and $\theta_i^{\cdot}$ is the occurrence probability of reads supporting CpG methylation at the $i$-th CpG site for each differential methylation state. (Note that we use common emission functions between the basic and second units in Figure 1a, and no emission function for gap states.) The problem here is how to model $\theta_i^{\cdot}$ depending on differential methylation states. One may consider to use $\theta_{1i}^{\mathrm{U}} = \theta_{1i}^{\mathrm{D}} = m_{1i}/n_{1i}$, $\theta_{2i}^{\mathrm{U}} = \theta_{2i}^{\mathrm{D}} = m_{2i}/n_{2i}$, and $\theta_{0i}^{\mathrm{N}} = (m_{1i} + m_{2i})/(n_{1i} + n_{2i})$. However, this cannot discriminate the direction of differential methylation due to $\theta_i^{\mathrm{U}} = \theta_i^{\mathrm{D}}$, and thus is not a suitable choice. In our previous study, we resorted to introduce an empirical parameter *pseudo*, resulting in

$$
\begin{aligned}
\theta_{1i}^{\mathrm{U}} &= \frac{m_{1i} + pseudo}{n_{1i} + pseudo},\\
\theta_{2i}^{\mathrm{U}} &= \frac{m_{2i}}{n_{2i} + pseudo},\\
\theta_{1i}^{\mathrm{D}} &= \frac{m_{1i}}{m_{1i} + pseudo},\\
\theta_{2i}^{\mathrm{D}} &= \frac{m_{2i} + pseudo}{n_{2i} + pseudo},\\
\theta_{0i}^{\mathrm{N}} &= \frac{m_{1i} + m_{2i}}{n_{1i} + n_{2i}},
\end{aligned}
$$

We note that *pseudo* can be regarded as a pseudocount added to actual read counts, playing a role to represent state-dependent methylation levels. For example, if a CpG site has the differential methylation state of Up, we expect that the methylation level is high in the sample 1 and low in the sample 2 (Figure 1b). Accordingly, *pseudo* is added to $m_{1i}$ (supporting CpG methylation in the sample 1) and to $u_{2i}$ (not supporting CpG methylation in the sample 2).

Although the empirical parameter partially solved the problem of designing emission functions, our previous study did not address how to adjust it. The optimal value of *pseudo* depends on the magnitude of read counts *m* and *n* (i.e. sequencing depth). Moreover, it may also depend on underlying biological processes between samples such as normal development and pathogenesis. In fact, as will be shown in the "Results and discussion" section, ComMet with the above emission functions may result in poor accuracy of DMR detection depending on the value of *pseudo*.

### New emission functions and learning algorithms

To design new emission functions for ComMet, we recall that the empirical parameter in our previous study had a role to represent state-dependent methylation levels. This viewpoint leads to the idea that the empirical parameters can be replaced by utilizing genome-wide methylation level distributions observed from real data. To formulate this intuition, we propose new emission functions in the following form:

$$e_i^{\mathrm{U}} = \int_0^1 \int_0^1 \mathrm{Binom}(m_{1i}|n_{1i},\ \theta_1)\mathrm{Binom}(m_{2i}|n_{2i},\ \theta_2)p(\theta_1,\ \theta_2|\mathrm{U})d\theta_1 d\theta_2, \qquad (1)$$

$$e_i^{\mathrm{D}} = \int_0^1 \int_0^1 \mathrm{Binom}(m_{1i}|n_{1i},\ \theta_1)\mathrm{Binom}(m_{2i}|n_{2i},\ \theta_2)p(\theta_1,\ \theta_2|\mathrm{D})d\theta_1 d\theta_2, \qquad (2)$$

$$e_i^{\mathrm{N}} = \int_0^1 \mathrm{Binom}(m_{1i}|n_{1i},\ \theta_0)\mathrm{Binom}(m_{2i}|n_{2i},\ \theta_0)p(\theta_0|\mathrm{N})d\theta_0. \qquad (3)$$

Note that new emission functions use probability distributions of state-dependent methylation levels $p(\theta|\cdot)$. This is in contrast to the emission functions in our previous study using fixed values $\theta_i^{\cdot}$.

Next, we present unsupervised learning algorithms for estimatingthese distributions for each dataset to be analyzed. Figure 2 shows the overview of the algorithms. As shown in Figure 2a, we exploit that methylation levels *m/n* collected from all genomic CpG sites form a distribution with two modes of high and low methylation. Such bimodal methylation level distributions are a common feature observed in many real datasets, and have also been reported by other researchers (*e.g.* 9, 10, 16, 19). Moreover, recent studies have suggested that detection of DMCs at individual cytosine sites can be improved by considering genome-wide methylation level distributions [13-15]. We propose to utilize this information for HMM-based DMR detection. We model genome-wide methylation level distributions by using beta mixtures as follows:

$$
\begin{aligned}
p(\theta_1) &= p_{\mathrm{H}}(\theta_1) + p_{\mathrm{L}}(\theta_1) + \mathrm{uniform}\\
&= w_{\mathrm{H1}}\mathrm{Beta}(\theta_1|\alpha_{\mathrm{H1}},\ \beta_{\mathrm{H1}}) + w_{\mathrm{L1}}\mathrm{Beta}(\theta_1|\alpha_{\mathrm{L1}},\ \beta_{L1}) + w_{\mathrm{unif1}},\\
p(\theta_2) &= p_{\mathrm{H}}(\theta_2) + p_{\mathrm{L}}(\theta_2) + \mathrm{uniform}\\
&= w_{\mathrm{H2}}\mathrm{Beta}(\theta_2|\alpha_{\mathrm{H2}},\ \beta_{\mathrm{H2}}) + w_{\mathrm{L2}}\mathrm{Beta}(\theta_2|\alpha_{\mathrm{L2}},\ \beta_{L2}) + w_{\mathrm{unif2}},
\end{aligned}
$$

where H and L represent two modes of high and low methylation, each of which is modeled by a beta distribution, and unif represents a background ground methylation level modeled by an uniform distribution (Figure 2b). Using these component distributions, we

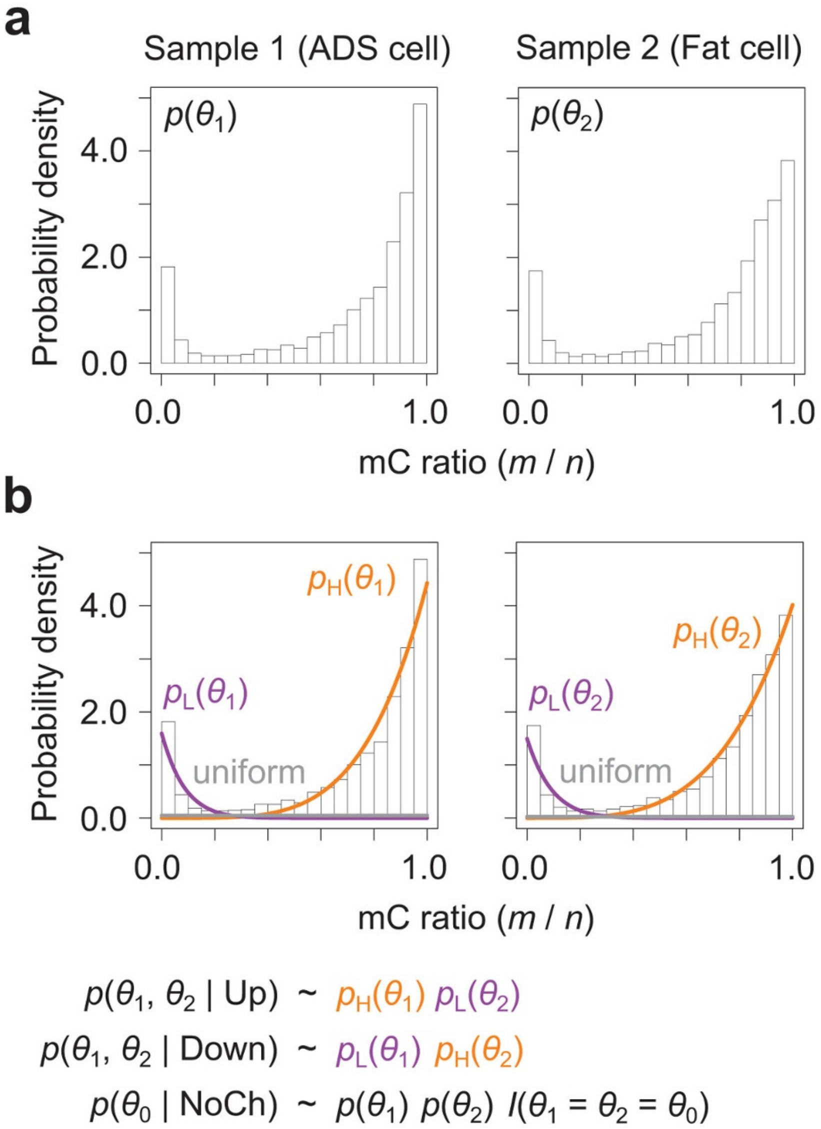


**Figure 2 Learning procedures for new emission functions**. (a) Bimodality of genome-wide methylation level distributions. Real bisulfite-seq data from [10] are shown. (b) Learning parameters for beta mixtures. Genome-wide methylation level distributions are fitted by beta mixtures, and differential methylation states are represented as alterations between two mixture components. The histograms are depicted using all genomic CpG sites, while the probability distributions are estimated using 10000 randomly selected CpG sites.

represent probability distributions of state-dependent methylation levels as follows:

$$p(\theta_1,\ \theta_2|\mathrm{U}) = \mathrm{Beta}(\theta_1|\alpha_{\mathrm{H1}}, \beta_{\mathrm{H1}})\mathrm{Beta}(\theta_2|\alpha_{\mathrm{L2}},\ \beta_{\mathrm{L2}}), \quad (4)$$

$$p(\theta_1,\ \theta_2|\mathrm{D}) = \mathrm{Beta}(\theta_1|\alpha_{\mathrm{L1}}, \beta_{\mathrm{L1}})\mathrm{Beta}(\theta_2|\alpha_{\mathrm{H2}},\ \beta_{\mathrm{H2}}), \quad (5)$$

$$\begin{aligned} p(\theta_0|\mathrm{N}) = {} & (w_{\mathrm{H1}}\mathrm{Beta}(\theta_1|\alpha_{\mathrm{H1}}, \beta_{\mathrm{H1}}) + w_{L1}\mathrm{Beta}(\theta_1|\alpha_{\mathrm{L1}},\ \beta_{\mathrm{L1}}) + w_{\mathrm{unif1}}) \\ & (w_{\mathrm{H2}}\mathrm{Beta}(\theta_2|\alpha_{\mathrm{H2}}, \beta_{\mathrm{H2}}) + w_{L2}\mathrm{Beta}(\theta_2|\alpha_{\mathrm{L2}},\ \beta_{\mathrm{L2}}) + w_{\mathrm{unif2}}) \quad (6) \\ & I(\theta_1 = \theta_2 = \theta_0), \end{aligned}$$

where $I()$ is an indicator function that takes 1 or 0 depending on whether the condition is true or false. This corresponds to represent differential methylation states as alterations between two modes. For example, Up state $p(\theta_1, \theta_2|\mathrm{U})$ is represented as high methylation in the sample 1, $p_{\mathrm{H}}(\theta_1)$, and low methylation in the sample 2, $p_{\mathrm{L}}(\theta_2)$. By substituting the equations 4-6 into the equations 1-3, the new emission functions are finally written as

$$e_i^{\mathrm{U}} = \binom{n_{1i}}{m_{1i}}\binom{n_{2i}}{m_{2i}}\frac{B(m_{1i} + \alpha_{\mathrm{H1}},\ u_{1i} + \beta_{\mathrm{H1}})}{B(\alpha_{\mathrm{H1}},\ \beta_{\mathrm{H1}})}\frac{B(m_{2i} + \alpha_{\mathrm{L2}},\ u_{2i} + \beta_{\mathrm{L2}})}{B(\alpha_{\mathrm{L2}},\ \beta_{\mathrm{L2}})}, \quad (7)$$

$$e_i^{\mathrm{D}} = \binom{n_{1i}}{m_{1i}}\binom{n_{2i}}{m_{2i}}\frac{B(m_{1i}+\alpha_{\mathrm{L1}},\ u_{1i}+\beta_{\mathrm{L1}})}{B(\alpha_{\mathrm{L1}},\ \beta_{\mathrm{L1}})}\frac{B(m_{2i}+\alpha_{\mathrm{H2}},\ u_{2i}+\beta_{\mathrm{H2}})}{B(\alpha_{\mathrm{H2}},\ \beta_{\mathrm{H2}})}, \tag{8}$$

$$e_i^{\mathrm{N}} = \frac{1}{Z}\binom{n_{1i}}{m_{1i}}\binom{n_{2i}}{m_{2i}}\sum_{s\in\{\mathrm{H1,\ L1,\ unif1}\}}\ \sum_{t\in\{\mathrm{H2,\ L2,\ unif2}\}} w_s w_t \frac{B(m_{1i}+m_{2i}+\alpha_s+\ \alpha_t-1,\ u_{1i}+u_{2i}+\beta_s+\beta_{t-1})}{B(\alpha_s,\ \beta_s)B(\alpha_t,\ \beta_t)}, \tag{9}$$

$$Z = \sum_{s\in\{\mathrm{H1,\ L1,\ unif1}\}}\ \sum_{t\in\{\mathrm{H2,\ L2,\ unif2}\}} w_s w_t \frac{B(\alpha_s+\alpha_t-1,\ \beta_s+\beta_t-1)}{B(\alpha_s,\ \beta_s)B(\alpha_t,\ \beta_t)}, \tag{10}$$

where $B()$ is a beta function, and $\alpha_{\mathrm{unif1}} = \beta_{\mathrm{unif1}} = \alpha_{\mathrm{unif2}} = \beta_{\mathrm{unif2}} = 1$ by definition of uniform distribution.

The parameter estimation of $w.$, $\alpha.$, and $\beta.$ involves several technical issues. First, we perform maximum likelihood estimation that maximizes the likelihood of read counts $m$ and $n$, rather than methylation levels $m/n$. Read counts preserve the information of sequencing depth (i.e. the magnitude of read counts), which is cancelled in methylation levels. Therefore, this enables to incorporate the information of sequencing depth into parameter values, thereby to overcome the drawback of the previous emission functions where the optimal value of *pseudo* depends on sequencing depth. The estimation problem is regarded as maximum likelihood estimation for beta-binomial mixtures, and thus can be solved as a simpler case of well-established maximum likelihood estimation for Dirichlet-multinomial mixtures described in [20]. Second, we can reduce the computational cost by only using read counts from a small number of randomly selected CpG sites. As shown in Figure 2, the histograms depicted using all genomic CpG sites are well fitted by the probability distributions estimated from 10000 CpG sites. Thus, we use 10000 CpG sites also for other datasets throughout this study. Third, we need to restrict the ranges of parameter values so that the integral in the equation 3 is tractable, and each beta component distribution corresponds to exactly one mode of methylation levels. Accordingly, parameter estimation is performed under the constraints of $\alpha.$, $\beta. \geq 1$ and $\beta_{\mathrm{H1}} = \beta_{\mathrm{H2}} = \alpha_{\mathrm{L1}} = \alpha_{\mathrm{L2}} = 1$.

## Results and discussion

To evaluate DMR detection accuracy, we conducted experiments on both simulated and real datasets. Unfortunately, there is no database of gold standards for benchmarking DMR detection (*i.e.* true biological DMRs). Therefore, we employ multilateral evaluation using a series of simulated and real datasets. The overall protocols are similar to those used in [6]. In experiments on simulated data, detected DMRs were evaluated for their overlap with simulated true DMRs. In experiments on real data, detected DMRs were evaluated for agreement with gene expression and DNase I hypersensitivity.

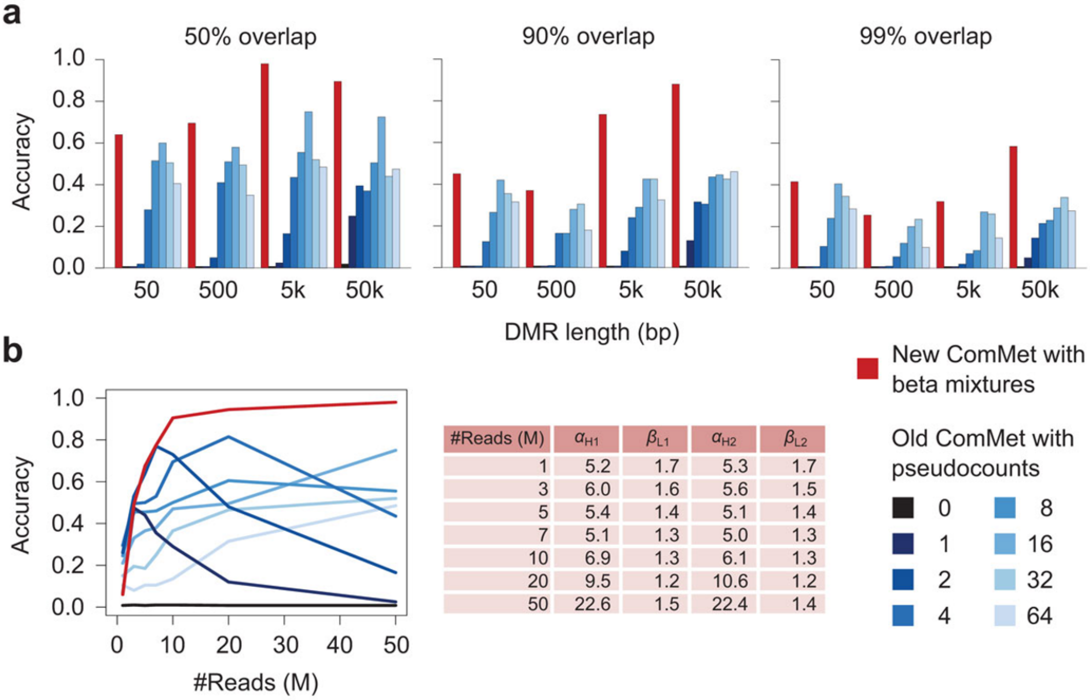
We compared DMR detection accuracy between ComMet using the new emission functions and that using the old ones. In addition, we also compared new ComMet with other methods using Fisher's exact test [10] and methylation level smoothing [11]. We used LAST [5] to align bisulfite-converted reads to reference sequences. The alignment results were used as the common input for each DMR detection method.

### Experiments on simulated data

We simulated bisulfite-converted reads using DNemulator [5]. The human chromosome × (chrX) was used as a reference. Methylation levels were assigned for all CpG sites in the chrX. 87-bp sinlgle-end reads were generated from random loci in the chrX with cytosines converted to thymines according to their methylation levels. Quality values were attached to reads according to SRR094461 in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra), which is bisulfite-seq data produced by the Illumina's platform. These reads were treated as the dataset for the sample 1. Next, 100 random regions were defined as DMRs for Up or Down, and methylation levels of all CpG sites in these regions were changed to the maximum or the minimum, respectively. Reads were again generated, and treated as the dataset for the sample 2. To test the effects of sequencing depth, we varied the number of generated reads for each dataset from 1 to 50 million (M). We also varied the length of simulated DMRs by preparing four versions of datasets: 50 bp, 500 bp, 5 kbp, and 50 kbp.

We evaluated DMR detection accuracy using the rate of correct predictions in the top 100 DMRs detected by each method. A correct prediction was defined as a simulated true DMR reciprocally overlapped with a detected DMR in a certain proportion of their lengths. For example, a correct prediction with 50% reciprocal overlap was counted only if the length of the overlapping region was larger than half the length of the simulated true DMR, and half the length of the detected DMR. Similarly, we also defined correct predictions for 90% and 99% reciprocal overlaps.

Figures 3 and 4, and Figure S1 in Additional file 1 show the experimental results. ComMet using the new emission functions achieved better accuracy than that using the old ones with various values of the *pseudo* parameter (Figure 3a and Figure S1 in Additional file 1). It should be noted that, while the old emission functions attained comparable accuracy to the new ones when used with the optimal value of *pseudo*, it is difficult to



| #Reads (M) | $\alpha_{H1}$ | $\beta_{L1}$ | $\alpha_{H2}$ | $\beta_{L2}$ |
|---|---|---|---|---|
| 1 | 5.2 | 1.7 | 5.3 | 1.7 |
| 3 | 6.0 | 1.6 | 5.6 | 1.5 |
| 5 | 5.4 | 1.4 | 5.1 | 1.4 |
| 7 | 5.1 | 1.3 | 5.0 | 1.3 |
| 10 | 6.9 | 1.3 | 6.1 | 1.3 |
| 20 | 9.5 | 1.2 | 10.6 | 1.2 |
| 50 | 22.6 | 1.5 | 22.4 | 1.4 |

**Figure 3 Comparison between new and old ComMet evaluated on simulated data**. (a) For various DMR lengths and the fixed sequencing depth of 50M reads, accuracy evaluated with 50% (left), 90% (center), or 99% (right) reciprocal overlap is shown. (b) For varying sequencing depths and the fixed DMR length of 5000 bp, accuracy evaluated with 50% reciprocal overlap is shown. Also shown are estimated values of parameters in new emission functions.

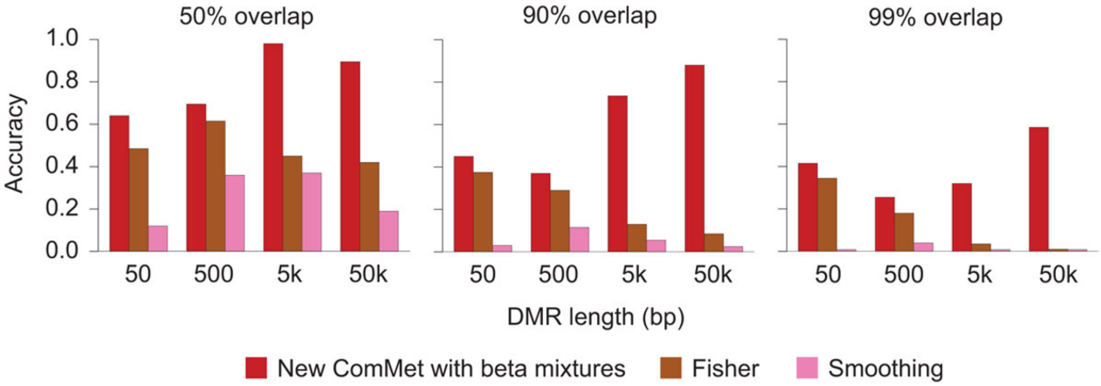


**Figure 4 Comparison between new ComMet and other methods evaluated on simulated data**. For various DMR lengths and the fixed sequencing depth of 50M reads, accuracy evaluated with 50% (left), 90% (center), or 99% (right) reciprocal overlap is shown.

find such optimal values in a practical situation where accuracy cannot be systematically evaluated. In fact, as explained in the "Methods" section, the optimal value of *pseudo* critically depends on sequencing depth, while the parameters in the new emission functions were successfully adjusted by our learning algorithms (Figure 3b and Figure S1 in Additional file 1). ComMet with the new emission functions also achieved better accuracy than Fisher's exact test and the smoothing method (Figure 4).

### Experiments on real data

We conducted experiments that evaluate agreement between detected DMRs and changes in gene expression. Note that similar experiments have been employed also

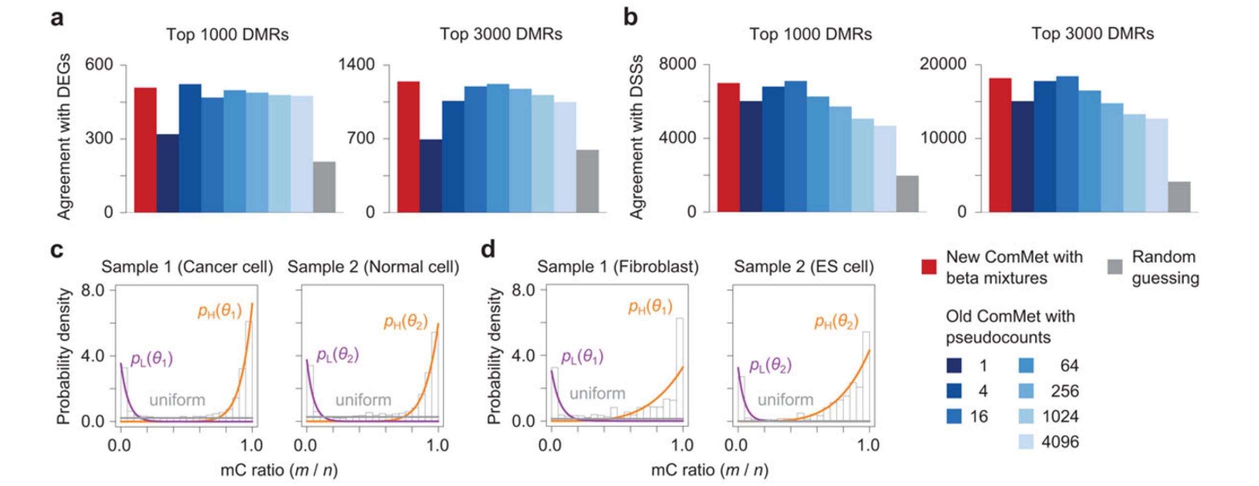


**Figure 5 Comparison between new and old ComMet evaluated on real data**. (a) Agreement between detected DMRs and changes in gene expression. (b) Agreement between detected DMRs and changes in DNase I hypersensitivity. See the main text for details. (c,d) The bimodal distributions of methylation levels observed in real data, and the fitted beta mixtures.

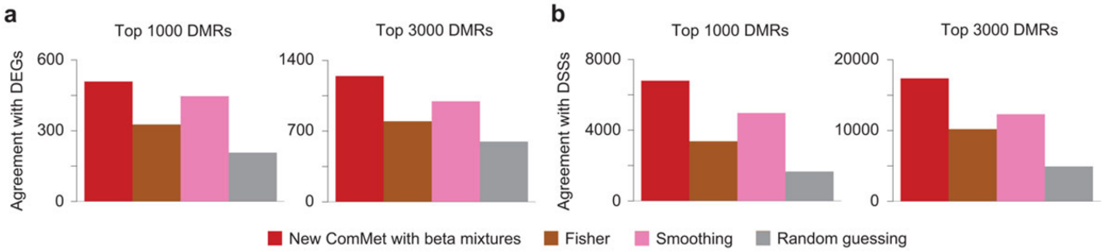


**Figure 6 Comparison between new ComMet and other methods evaluated on real data**. (a) Agreement between detected DMRs and changes in gene expression. (b) Agreement between detected DMRs and changes in DNase I hypersensitivity. See the main text for details.

in previous studies [6,11]. We collected data from [16], where human breast cancer and normal breast are measured by both RNA-seq and bisulfite-seq. We aligned RNA-seq reads to the human genome using TopHat [21]. Gene expression was measured by fragments per kilobase of transcript per million mapped reads (FPKM) using Cufflinks [22]. Differentially expressed genes (DEGs) were determined by the threshold of five-fold FPKM change. We evaluated agreement between DEGs and detected DMRs according to the previous study [6,11]. We focused on DEGs whose ±5 kbp regions around transcription start sites (TSSs) contained detected DMRs. The numbers of DEGs were counted for the top 1000 and 3000 DMRs detected by each method. We used these counts as a measure of the agreement. For the base-line of accuracy, we calculated the expected number of DEGs when DMRs were randomly placed in the TSS windows (denoted by *random guessing*).

In addition, we evaluated agreement between detected DMRs and changes in DNase I hypersensitivity as conducted in [6]. We collected data from [19], where human foreskin fibroblasts and embryonic stem cells are measured by bisulfite-seq. For these cell types, we obtained DNase I hypersensitivity data from the ENCODE project http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks/. The data for each cell type contain the set of 150-bp regions that show the local maxima of DNase I hypersensitivity with false discovery rate (FDR) less than 1%. We defined “differentially sensitive sites” (DSSs) as those 150-bp regions present in either one of the two cell types. The agreement between DSSs and detected DMRs was evaluated similarly to the experiment for DEGs. We focused on DSSs whose ±5 kbp regions around the midpoints contained detected DMRs. The numbers of DSSs were counted for the top 1000 and 3000 DMRs detected by each method. We used these counts as a measure of the agreement.

Figures 5 and 6 show the experimental results. The advantage of the new emission functions over the old ones was also validated on real data, achieving better agreement of detected DMRs with DEGs (Figure 5a) and DSSs (Figure 5b). We again emphasize that the performance of the old emission functions critically depends on the choice of *pseudo* values, while the optimal value is difficult to find empirically. In contrast, as shown in Figure 5cd, the bimodal distributions of methylation levels were observed in real datasets, and our learning algorithms successfully fitted the beta mixtures. The new ComMet achieved better accuracy than Fisher's exact test and the smoothing method also for real datasets (Figure 6).

## Conclusions

In this paper, we described the new emission functions for HMM-based DMR detection from bisulfite-seq data. We proposed to incorporate the information of genome-wide methylation level distributions into emission functions, replacing the empirical parameter used in our previous study. ComMet with the new emission functions successfully improved DMR detection accuracy compared to the previous version. Recent studies suggest that detection of DMCs at individual cytosine sites can be improved by considering genome-wide methylation level distributions [13-15]. Therefore, our results have shown that such information is useful not only for detecting DMCs, but also for DMR detection. Furthermore, our HMM-based method achieves better accuracy than other methods using Fisher's exact test and methylation level smoothing. The implementation of ComMet is distributed as a part of Bisulfighter package, which is available at http://epigenome.cbrc.jp/bisulfighter.

## Additional material

**Additional file 1: Figure S1**. Benchmark for DMR detection at varying sequencing depth. For each DMR length, accuracy evaluated with 50% reciprocal overlap is shown. Also shown are estimated values of parameters in new emission functions.

### List of abbreviations

DMR: differentially methylated region; DMC: differentially methylated cytosine; HMM: hidden Markov model; DEG: differentially expressed gene; DSS: differentially sensitive site.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

YS developed the method, wrote the program, performed the experiments, and drafted the manuscript. TM conceived of the study, and coordinated the project. All authors have read and approved the final manuscript.

### Acknowledgements

This study was supported by Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST), New Energy and Industrial Technology Development Organization (NEDO), and Grant-in-Aid for Young Scientists (B) Grant Number 15K16089 from Japan Society for the Promotion of Science (JSPS).

### Declarations

The publication charges for this article have been funded by Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST).
This article has been published as part of *BMC Genomics* Volume 16 Supplement 12, 2015: Joint 26th Genome Informatics Workshop and 14th International Conference on Bioinformatics: Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S12.

### Authors' details

[1]Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, 135-0064 Tokyo, Japan. [2]Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency (JST), 4-1-8 Honcho, Kawaguchi, 332-0012 Saitama, Japan.

Published: 9 December 2015

### References

1. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond.** *Nat. Rev. Genet* 2012, **13(7)**:484-492.
2. Laird PW: **Principles and challenges of genomewide DNA methylation analysis.** *Nat. Rev. Genet* 2010, **11(3)**:191-203.
3. Xi Y, Li W: **BSMAP: whole genome bisulfite sequence MAPping program.** *BMC Bioinformatics* 2009, **10**:232.
4. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics* 2011, **27(11)**:1571-1572.
5. Frith MC, Mori R, Asai K: **A mostly traditional approach improves alignment of bisulfite-converted DNA.** *Nucleic Acids Res* 2012, **40(13)**:100.
6. Saito Y, Tsuji J, Mituyama T: **Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions.** *Nucleic Acids Res* 2014, **42(6)**:45.
7. Bock C: **Analysing and interpreting DNA methylation data.** *Nat. Rev. Genet* 2012, **13(10)**:705-719.
8. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X: **Statistical methods for detecting differentially methylated loci and regions.** *Front Genet* 2014, **5**:324.
9. Takada H, Mituyama T, Wei Z, Yoshihara E, Jacinto S, Downes M, Evans RM: **Methylome, transcriptome, and PPARγ cistrome analyses reveal two epigenetic transitions in fat cells.** *Epigenetics* 2014, **9(9)**:1195-1206.
10. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM, Ecker JR: **Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.** *Nature* 2011, **471(7336)**:68-73.
11. Hansen KD, Langmead B, Irizarry RA: **BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.** *Genome Biol* 2012, **13(10)**:83.
12. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE: **methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.** *Genome Biol* 2012, **13(10)**:87.
13. Dolzhenko E, Smith AD: **Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments.** *BMC Bioinformatics* 2014, **15(1)**:215..
14. Feng H, Conneely KN, Wu H: **A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data.** *Nucleic Acids Res* 2014, **42(8)**:69.
15. Raineri E, Dabad M, Heath S: **A note on exact differences between beta distributions in genomic (Methylation) studies.** *PLoS ONE* 2014, **9(5)**:97349.
16. Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, Kuan S, Edsall LE, Camargo AA, Stevenson BJ, Ecker JR, Bafna V, Strausberg RL, Simpson AJ, Ren B: **Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.** *Genome Res* 2012, **22(2)**:246-258.

17. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S: **ChAMP: 450k Chip Analysis Methylation Pipeline.** *Bioinformatics* 2014, **30(3)**:428-430.
18. Mao Z, Ma C, Huang TH, Chen Y, Huang Y: **BIMMER: a novel algorithm for detecting differential DNA methylation regions from MBDCap-seq data.** *BMC Bioinformatics* 2014, **15(Suppl 12)**:6.
19. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL: **Dynamic changes in the human methylome during differentiation.** *Genome Res* 2010, **20(3)**:320-331.
20. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D: **Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.** *Comput. Appl. Biosci* 1996, **12(4)**:327-345.
21. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25(9)**:1105-1111.
22. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: **Improving RNA-Seq expression estimates by correcting for fragment bias.** *Genome Biol* 2011, **12(3)**:22.

doi:10.1186/1471-2164-16-S12-S3
**Cite this article as:** Saito and Mituyama: **Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions.** *BMC Genomics* 2015 **16**(Suppl 12):S3.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central