African Journal of Biotechnology Vol. 12(8), pp. 811-819, 20 February, 2013
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB12.2372
ISSN 1684–5315 ©2013 Academic Journals

Full Length Research Paper

Whole-genome methylation caller designed for methyl-DNA immunoprecipitation sequence data

Mengying Ren

[1]Institute of Biostatistics, School of Life Science, Fudan University, Shanghai, People's Republic of China.
E-mail: mandova1005@yahoo.cn.

Accepted 21 December, 2012

DNA methylation is an indispensable epigenetic modification required for regulating the expression of mammalian genomes. Continued efforts have been made to unravel the methylation states genome-wide, featuring the methyl-DNA immunoprecipitation (MeDIP) coupled with next-generation sequencing. Our method uses a single-CpG-resolution, whole-genome methylation caller designed specifically for MeDIP-seq data. It did not require external database for copy number adjustment. Furthermore, it effectively detected genomic regions potentially predisposed to oncogenesis through its prediction of methylation states. The above suggests that our method makes a handy and reliable tool to generate genome-wide methylation profiles. All source codes in PERL language are available upon request of the first author.

Key words: Methyl-DNA immunoprecipitation, next-generation sequencing, Hidden Markov chain.

INTRODUCTION

Methylation of DNA cytosine residues is a common epigenetic mark in many eukaryotes. It is the addition of a methyl group in the fifth carbon position of cytosines found predominantly at the site of CpG, but is also prevalent, though less common, in other sequence context-CHG and CHH (H = A, T, C) (Ramsahoye, 2000; Lister, 2009). DNA methylation is the only known epigenetic system that modifies the DNA molecule itself. It is most fundamental and an indispensable component of the so-called epigenetic mechanism. The phenotype of a cell is primarily determined by its expression profile and its response to environmental cues. Epigenetics provide stability and diversity to the cellular phenotype through chromatin marks that affect local transcriptional potential and are preserved or regenerated during cell division.

Much of the human genome is CpG depleted with the exception of CpG islands which are defined as 200-bp stretches of DNA with a C+G content of 50% and an observed CpG/expected CpG exceeding 0.6(Gardiner-Garden and Frommer, 1987). In the promotor region of genes CpG islands are abundant and non-methylated, but infrequent in highly tissue-specific genes (Bird, 1986). CpG islands tend to remain unmethylated throughout development, with the exception of those islands associated with genes on the inactive X chromosome and those associated with the silent allele of imprinted genes (Yoder et al., 1997). On the contrary, most of CpG dinucleotides outside CpG islands are methylated, especially those found in repeat DNA elements which make up 45% of the genome and contain the majority of 5-methylcytosines (Jordà and Peinado, 2010).

DNA methylation patterns vary in time, space and species. There's a diverse spectrum of animals' methylation levels. DNA methylation is detectable in all stages of *Drosophila melanogaster* development at a level of about one in 1000 to 2000 cytosine residues in adult flies (Gowher et al., 2000), mostly in CpT dinucleotides (Bird, 2002). Up the taxonomy, vertebrate genomes have the highest level of methyl-cytosines. For mice, early embryonic demethylation and following de novo methylation is critical in determining somatic DNA methylation pattern. After fertilization, paternal DNA is actively demethylated and the maternal DNA undergoes passive demethylation. The consequence of this

remodeling of epigenetic marks is the parent-specific pattern of imprinting (reviewed in Carrell, 2012).

The presence of CpG island is perhaps the most striking feature of vertebrate DNA methylation. Of all human genome sequences 0.05% annotated to be located in CpG islands (UCSC table browser, excluding sex chromosomes), many of which remain methylation-free regardless of tissue-specific expression of associated genes. The coincidence of CpG islands and promoter is ubiquitous (Bird, 1986). About 31% of CpG islands are methylation-prone and CpG islands located in promoter regions are seldom methylated (Fan et al., 2010). An active promoter CpG island might occur at the upstream, intron or exon of its associated gene. Hypermethylation of CpG islands located in the promoter regions of tumor suppressor genes is now firmly established as an important mechanism for gene inactivation. CpG island hypermethylation has been described in almost every tumor type (Esteller, 2002). CpG islands differ in their intrinsic susceptibility to de novo methylation, and suggest that the propensity for a CpG island to become aberrantly methylated can be predicted based on its sequence context (Feltus et al., 2003).

There are mainly three kinds of treatment which shall be matched with different analytical steps. Enzyme-based approaches involve digestion of genomic DNA by restriction enzymes which have differential impact on methylated and unmethylated versions of target CpG sites. In these approaches, only particular sequence motifs can be analyzed because specific restriction sites are required to be present (Jacinto et al., 2008). Coupled with either microarrays or capillary sequencing, they have been applied to genome-wide DNA methylation profiling of several organisms but are limited to the analysis of CpG sites located within the enzyme recognition site(s) (Down et al., 2008). Besides, potentially incomplete digestion may cause false positives (Yang et al., 2010).

Being recognized as gold standard of high resolution methylation profiling (Eckhardt et al., 2006), bisulfite (BS) conversion technique is based on the reaction between DNA and sodium bisulfite which converts unmethylated cytosine into uracil and eventually to thymine after amplification, leaving methylated cytosine unchanged. It offers single-CpG resolution (the only one among three methods) and can be coupled with polymerase chain reaction (PCR) (Wang et al., 2008), microarrays or BS-seq. However, following BS conversion, there are so many sequence versions corresponding to a specific genomic region that it is difficult to design enough probes for or accurately map those reads of reduced genomic complexity, posing bioinformatics problems to comprehensive analysis of BS-converted DNA (Down et al., 2008; Xi and Li, 2009; Iraola-Guzmán, 2011). With higher resolution though, BS-coupled methods also require specialized analysis software and a much higher coverage. Besides, BS-seq is currently prohibitively expensive for routine analysis of large genomes, though this will likely not hold in the near future (Reinders et al., 2008). A modified version of BS-seq, reduced representation bisulfite sequencing (RRBS), has recently been developed for efficient profiling of clinical samples (Gu et al., 2010).

Methyl-DNA immunoprecipitation (MeDIP), introduced in 2005, is based on the direct immunoprecipitation of methylated DNA (Weber et al., 2005). Firstly, genomic DNA purified by standard procedures is sheared by sonication to produce random fragments ranging in size of 300 to 600 bp, which is a key to guaranteeing efficient immonoprecipitation and a reasonable level of resolution. After that, DNA must be denatured at 95°C to yield single-stranded DNA fragments. The rest of the assay is a standard immonoprecipitation protocol followed by incubation with anti-5-methylcytidine antibody. The immunoprecipitated DNA can be hybridized with microarrays or sequenced. MBD-isolated Genome Sequencing (MiGS) combines precipitation of methylated DNA by recombinant methyl-CpG binding domain of MBD2 protein and sequencing of the isolated DNA by a massively parallel sequencer (Serre et al., 2010; Lan et al., 2011). Another method “MIRA” uses a different combination of proteins to recover CpG islands, obtaining a resolution that is similar to bisulfate sequencing (Rauch and Pfeifer, 2010). In MethylCap-seq (Brinkman et al., 2010), captured DNA is washed and eluted in a step-wise manner using increasing salt concentrations to obtain genome stratification with reduced complexity. The efficiency of immunoprecipitation in MeDIP depends on the density of methylated CpG sites, which vary greatly within any given mammalian genome, making it difficult to distinguish variations in enrichment from confounding CpG density effects (Weber et al., 2007). MeDIP combined with next generation sequencing (MeDIP-seq) have a great potential to become the most cost-effective and unbiased method in whole-genome methylome profiling.

The key step in MeDIP-seq analysis is the identification and quantification of methylated regions. Batman (Down et al., 2008), short for “Bayesian tool for methylation analysis”; can estimate absolute DNA methylation levels, across a wide range of CpG densities, from MeDIP-based experiments. Until then, it had not been possible to estimate absolute methylation levels from MeDIP, and analysis of regions with low CpG density has been assumed to be problematic (Weber et al., 2007). The work is also the first MeDIP-seq data to represent a high-resolution whole-genome DNA methylation profile of a mammalian genome. Yang et al. (2010) used peak search (widely used in CHiP-seq data to find regions of high read density) based on Poisson model to identify methylated regions on a whole-genome scale, to deal with single-sample cases. As of multiple sample analysis, a recent study (Ruike et al., 2010) obtained DNA methylation profiles for 8 human breast cancer cell lines

and 1 normal human mammary epithelial cells. This study classified regions as hyper-, hypo- and not differentially methylated groups by pairwise comparisons of MeDIP-seq depth. A similar categorization based on Batman DNAm score was used by Feber et al., 2011 to perform global analysis to identify directional changes in DNAm. Bismark (Krueger and Andrews, 2011) is a flexible tool for the analysis of bisulfite sequencing data which performs both read mapping and methylation calling in a single convenient step. Its mapping scheme aims to find a unique alignment by running four alignment processes corresponding to four sequence identity simultaneously which enables Bismark to uniquely determine the strand origin of a bisulfite read. Methylation calls in Bismark take the surrounding sequence context into consideration and discriminate between cytosines in CpG, CHG and CHH context. MeQA (Huang et al., 2011) is a pipeline for pre-processing, data quality assessment and distribution of sequences reads and estimation of DNA methylation levels of MeDIP-seq datasets. Inspired by the valuable concept of Batman's coupling factor, MEDIPS (Chavez et al., 2010) weighs the raw MeDIP-seq signals with respect to the estimated coupling factor-dependent normalization parameters. It is a time-efficient statistical method for normalizing and analyzing MeDIP-seq data.

MATERIALS AND METHODS

A detail of the data is described in the work of Feber et al. (2011). Three samples, each containing a pool of no more than 10 individuals representing normal, neurofibroma (NF) and malignant nerve sheath tumor (MPNST), respectively, are sequenced and aligned to human genome NCBI build 36. Only sequences with mapping quality ≥ 10 are used, totaling 104.6, 104.8 and 103.0 million 50-mer reads. Reads are counted for each CpG site in the hg18 reference genome (set to 0 if not present in retrieved reads).

Hidden Markov model (HMM) is a stochastic method which has been used in various applications like speech processing, signal processing and character recognition. Apart from gene finding and annotation (Krogh, 1997; Zhu et al., 2006a) (early works reviewed in Durbin et al., 1998; Birney, 2001), its application in biological sequence analysis includes genome segmentation by introducing macros-states (Melodelima and Gautier, 2007), modeling length distribution of sequences (Zhu et al., 2006b) and splicing sites recognition (Dong and Sun, 2007). In our HMM model, methylation can be predicted at single CpGs. Each CpG site has a "hidden" state of being un-methylated non-CGI (UN, 1), methylated non-CGI (MN, 2), un-methylated CGI (UC, 3) or methylated CGI (MC, 4). The states of successive CpG sites are assumed to follow a Markov process. Standard Baum-Welch algorithm (Baum et al., 1970) is used with slight modification which is described below. Poisson emission probabilities are assumed for unmethylated states UN and UC, while normal distribution are applied within class MN and MC.

$$e_{\pi_i}(x_i) = e^{-\alpha[\pi_i]}\frac{\alpha^{x_i}}{x_i!} \quad (1)$$

$$e_{\pi_j}(x_i) = \frac{1}{\sqrt{2\pi\beta}} e^{-\frac{(x_i-\alpha)^2}{2\beta}} \quad (2)$$

Where, $\pi_i \in \{1,3\}$ $\pi_j \in \{2,4\}$ is the inferred state at site i/j, $\alpha/\alpha,\beta$ are the Poisson/normal parameters for corresponding state π_i , x_i denotes the read count at the ith CpG site. Two sets of parameters are to be estimated: 24 transition probabilities (including four starting and four ending transitions), six Poisson/normal emission distribution parameters for four states, respectively. The initials are picked randomly under constrained conditions below: states 2 and 3 are mostly likely to remain in themselves; while *a*[4][4] is also relatively high given the density of CpG sites once this region becomes methylated. All transitions to states 2 and 3 are high. Transitions between states 4 and 1 are rare. The initial $\alpha(\beta)'s$ are chosen through trial and error. To get rid of serious underflow and overflow problem, scaling parameters (Rabiner, 1990) to both forward $\tilde{f}'s$ and backward parameters $\tilde{b}'s$ are initially used and then merged:

$$\tilde{f}_l(i+1) = \frac{1}{s_{i+1}} e_l(x_{i+1}) \sum_k \tilde{f}_k(i) a_{kl} \quad (3)$$

(similar for $\tilde{b}'s$) such that $\sum_l \tilde{f}_l(i) = 1, \sum_l \tilde{b}_l(i) = 1$. The geometric averages of two sets of scaling parameters are used as final scaling parameters and $\tilde{f}'s$ and $\tilde{b}'s$ are re-calculated afterwards. Standard forward and backward f's and b's are calculated for each of 22 training sequences (chromosomes) (Durbin et al., 1998), transition parameters (a's) are updated as follows:

$$A_{kl} = \sum_j \frac{\sum_i \tilde{f}_k^j(i) a_{kl} e_l(x_{i+1})^j \tilde{b}_l^j(i+1)}{\tilde{P}(x^j)}$$

(4a)

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad (4b)$$

i=1,2,... L^j (length of sequence j), j=1,…22, A_{kl} is the expected number of occurrences of k-to-l transition. And using point estimation, α,β are updated according to:

$$\alpha[k] = \frac{\sum_{ij} x(i)^j \tilde{f}_k^j(i)\tilde{b}_k^j(i) st^j(i)/\tilde{P}(X^j)}{\sum_{ij} \tilde{f}_k^j(i)\tilde{b}_k^j(i) st^j(i)/\tilde{P}(X^j)}$$

(5)

$$\beta[k] = \frac{\sum_{ij} (x(i)^j)^2 \tilde{f}_k^j(i)\tilde{b}_k^j(i) st^j(i)/\tilde{P}(X^j)}{\sum_{ij} \tilde{f}_k^j(i)\tilde{b}_k^j(i) st^j(i)/\tilde{P}(X^j)} - \left(\frac{\sum_{ij} x(i)^j \tilde{f}_k^j(i)\tilde{b}_k^j(i) st^j(i)/\tilde{P}(X^j)}{\sum_{ij} \tilde{f}_k^j(i)\tilde{b}_k^j(i) st^j(i)/\tilde{P}(X^j)}\right)^2$$

(6)

Where, $k \in \{1,2,3,4\}$, *st*(*i*)'s are the new scaling parameters.

Table 1. Comparison statistics regarding HEP.

Percentage	Global correlation	Overscored	Underscored
Batman	46.5	1.87	4.28
This study	40.5	1.41	14.72

Underflow problem caused by ultra-high read counts are dealt with, replacing original read count with a jittered number around the mean of state MN or MC. State path for each chromosome is the summary of repeated Viterbi (Durbin et al., 1998) inferences starting from different initials.

Algorithm

Initialization

Pick initial model parameters according to criterion described above.

Recurrence

For each sequence (chromosome) j = 1…22, Calculate $\tilde{f}_k^j(i)$ i = 1… L^j and s_i using Equation 3, where,

$$s_{i+1} = \sum_l e_l(x_{i+1}) \sum_k \tilde{f}_k(i) a_{kl}$$

Calculate $\tilde{b}_k^j(i)$, $i = L^j \ldots 1$ and t_i (similar to *si*). Re-scale $\tilde{f}_k^j(i)$ and $\tilde{b}_k^j(i)$ using $st_i = \sqrt{s_i \square t_i}$ calculate $p(X_j) = \sum_k \tilde{f}_k(L^j) a_{k0}$; where, X_j stands for the jth sequence (chromosome). Add contribution to A_{kl} in (4a) and to the denominator and numerator in (5) and (6). Calculate new parameters using Equations 4b, 5 and 6.

Termination

Stop when iteration times exceed a predefined threshold.

Decoding

Use Viterbi algorithm to infer state at each site for each sequence (chromosome). Methylation states are summarized according to state votes. All actual calculations are log-scaled. It has been previously shown that MeDIP-derived data need to be corrected for local CpG densities in order to compute unbiased methylation levels (Down et al., 2008; Pelizzola et al., 2008), so global methylation score is up-adjusted according to whether or not the CpG site is predicted to be in a CGI.

RESULTS

Compatibility with the gold standard

After counting reads for each CG site and adding sites of zero counts, there are on average around 71.2% CpG sites with positive read counts for normal sample. The methylation call confirmed the reported bimodality: 79.9% CpG sites displayed hypomethylation (methylation score <0.3), 18.5% were hypermethylated (methylation score ≥0.7) and 1.58% heterogeneously methylated. Of all sites 68% are within CGIs, consistent with previous estimation of around 80% (Eckhardt et al., 2006). Correlation of methylation calls (averaged over 100-bp window to enable comparability) with Batman's m scores are 32.4, 28.4 and 36.6% for normal, NF and MPNST samples.

Our inferred single-CpG methylation score has a 46.5% correlation with Human Epigenome Project (HEP) (Eckhardt et al., 2006) for all tissues pooled (Batman with HEP, 40.5% on 100bp-window base). Table 1 displays the performance of two methods, where "overscore" is defined as a site whose HEP methylation score minus predicted score exceeds 0.4, and vice versa for "underscore". The tendency to underscore in our method suggests its being overly prudent and the possible inadequacy of CpG density adjustment. Since considerable between-tissue variation was recognized (Rakyan et al., 2004), tissue-wise comparison was performed and demonstrated in Figure 1. All comparison is based/ converted to hg18 reference sequences (Zhang et al., 2000). Our methylation correlation distinguished tissues better with a much larger between-tissue variation; predictably low correlation between HEP and sperm sample is evident to be seen.

Careful inspection of the primer sequences reveal that the bisulfate primers used in HEP studies did not consider the CpG sites of the reverse strand, that is, the design of alleged no-CpG-containing, BS-treated-DNA-specific primers assumed non-methylation status of the minus strand. Of all 31704 CpG sites implicated only 65 were located on the reverse strand and 290 unmapped in both strands according to reference genome hg18. As a result we sought other ways to further validate our method which was derived from strand-insensitive MeDIP-seq data.

Comparison with independent gene expression data

Promoter methylation is suspected as playing an important role in the pathogenesis of MPNST (Kawaguchi et al., 2005). We analyzed the promoter regions of 55 genes previously reported to be associated with neurofibromatosis tumors (Miller et al., 2009) (Tables 2 and 3). Putative promoter positions are extracted according to Ensembl annotation.

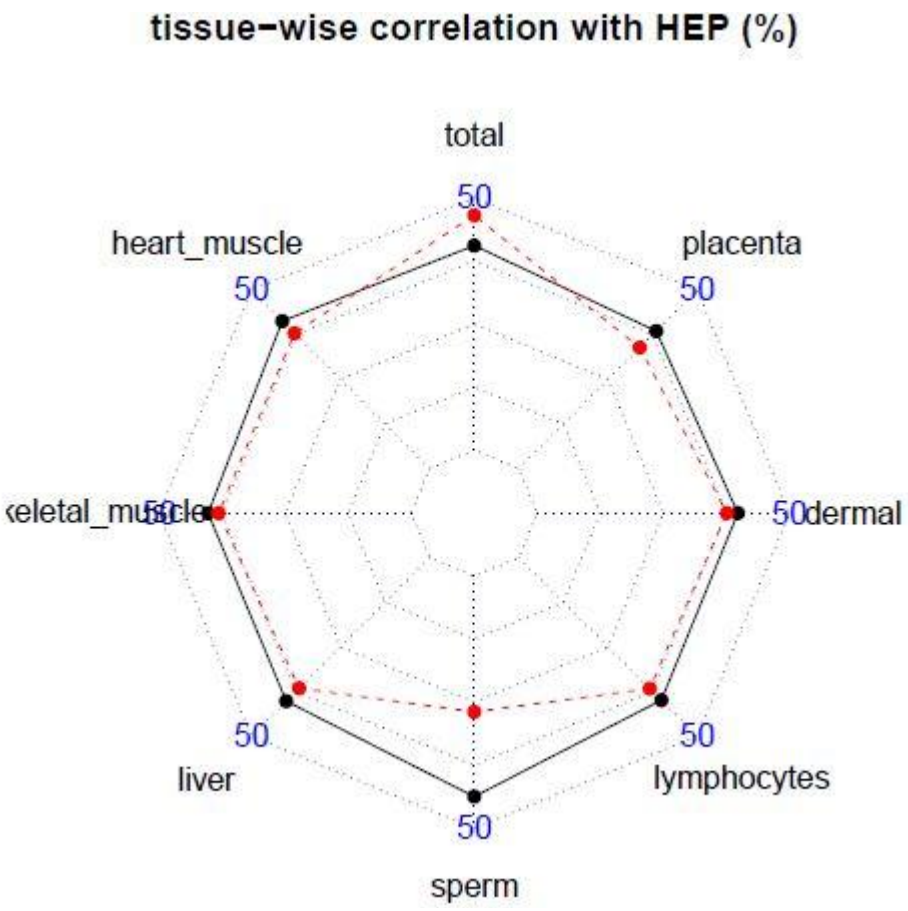

Figure 1. The radar chart for tissue-wise correlation with HEP data. Batman scores (represented by black line) has a 6% lower overall correlation with HEP, though both are less than 50%. Our methylation correlation (red) distinguished tissues better with a much larger between-tissue variation (25.0) than Batman (16.0).

From Table 2, our methods distinguishes differentially methylated region (DMR) with significantly lower false positive/negative rates. However, for up-regulated genes (benign to malignant), our method has a slightly higher false positive rate, partly due to the fact that the predominant role for DNA methylation is down-regulation and the resemblance of normal to benign tissues overrides that of benign to malignant. Figure 2 gives a summary for DMR recognition. Our method outperforms Batman except that in “benign to malignant” case, more up-regulated genes are called to have an increment in methylation score in our method than in Batman.

DISCUSSION

In this paper we presented a methylation calling model which has the intuition of our current knowledge of the phenomenon. The efficiency of immunoprecipitation in MeDIP depends on the density of methylated CpG sites, which vary greatly within any given mammalian genome (Weber et al., 2007). Regions with dense cytosine methylation are least affected and yield a relatively strong signal (Reinders et al., 2008). This makes it difficult to distinguish variations in enrichment from confounding CpG density effects, calling for a thorough model that deals with inherent sequence bias and allows for local fitting of hypothetical distributions. Given the relatively low resolution of MeDIP-seq data, we are able to generate whole-genome, single-resolution estimation of the methylation status of each CpG site. High-resolution methylation profiles for both DNA strands, which require much higher computational power, have yet to come, and MeDIP-seq strategy will continue to make desirable tool since current BS-seq technology could not afford simultaneous detection of myriad versions regarding a certain DNA fragment containing CpG sites at both strands. Our method, being a naive methylation caller, ignores chimeric reads that indicate potential SNP effects obscured by altered methylation level. Very recent studies (He et al., 2011; Ito et al., 2011) urges researchers to come out with more elaborate methods to distinguish not only methylation in multiple sequence contexts including CHG and CHH but also subtypes or

Table 2. Methylation of suspected promoter regions of genes associated with NF/MPNST: normal to benign.

Name	Expres.	Met	This study	Batman	Name	Expres.	Met	This study	Batman
EMP2	↓	↑	NA	↑0.015	EN2	↑	↓	↑0	↓0.016
EPB41L3	↓	↑	↑0.001	↑0.01		↑	↓	↓0.001	↓0.025
	↓	↑	NA	↑0.01		↑	↓	↓0.011	↓0.006
	↓	↑	NA	↓0.042	HGF	↑	↓	↓0.017	↑0.158
GFAP	↓	↑	NA	↑0.038	MDK	↑	↓	↓0.007	↓0.008
	↓	↑	NA	↑0.063		↑	↓	NA	↑0.038
HLA-DQB1	↓	↑	↑0.122	↑0.029	PAX6	↑	↓	↓0.004	↑0.013
KLK6	↓	↑	↑0.215	↑0.065		↑	↓	NA	↓0.013
LGI1	↓	↑	↑0.215	↑0.132		↑	↓	↓0.005	↓0.018
MBP	↓	↑	NA	↑0.006		↑	↓	↑0.028	↑0.088
	↓	↑	↑0.231	↑0.285	SMAD3	↑	↓	NA	↑0.029
	↓	↑	↑0.096	↑0.088	WT1	↑	↓	NA	↑0.048
	↓	↑	↑0.023	↑0.088		↑	↓	↓0.359	↓0.029
NGFR	↓	↑	↑0.023	↓0.013		↑	↓	↓0.016	↑0.001
	↓	↑	↑0.021	↑0.023		↑	↓	↑0.257	↑0.069
CDKN2A	↓	↑	NA	↑0.002		↑	↓	NA	↑0.107
	↓	↑	NA	↑0.008	APOD	↑	↓	↑0.039	↑0.036
	↓	↑	NA	↑0.017		↑	↓	↓0.094	↓0.003
	↓	↑	NA	↓0.007	CASP1	↑	↓	NA	↓0.04
CTSD	↓	↑	↑0.008	↑0.051	CD36	↑	↓	↑0.052	↑0.14
	↓	↑	↑0.008	↓0.167	EGFR	↑	↓	NA	↑0.013
	↓	↑	NA	↑0.003		↑	↓	NA	↑0.004
GNAI2	↓	↑	NA	↓0.005		↑	↓	↓0.004	↓0.021
	↓	↑	NA	↑0.026	KIT	↑	↓	↓0.018	↑0.012
HPCAL1	↓	↑	NA	↑0.026	LEPR	↑	↓	NA	↑0.015
	↓	↑	NA	↑0.015		↑	↓	NA	↑0.016
MFI2	↓	↑	↑0.003	↑0.015	MME	↑	↓	NA	↓0.004
NES	↓	↑	NA	↓0.023	SOCS3	↑	↓	NA	↑0.013
NFKB1	↓	↑	NA	↓0.011		↑	↓	↓0.016	↓0.011
BCL2	↓	↑	↑0.002	↓0.026	ADM	↑	↓	NA	↑0.026
BCL2L2	↓	↑	NA	↓0.026		↑	↓	↓0.002	↑0.007
EDNRB	↓	↑	↑0.034	↑0.105	CAPN1	↑	↓	NA	↓0.008
ERBB3	↓	↑	NA	↓0.002	FBN2	↑	↓	↓0.016	↓0
	↓	↑	NA	↑0.016	IGFBP3	↑	↓	NA	↓0.025
	↓	↑	NA	↓0.029	PDGFRA	↑	↓	↑0.008	↑0.003
MPZ	↓	↑	NA	↑0.159	PIAS3	↑	↓	NA	↓0
PDGFA	↓	↑	↑0.093	↓0.109		↑	↓	NA	↑0.022
	↓	↑	↑0.093	↓0.109	PLAU	↑	↓	NA	↑0.03
S100B	↓	↑	↑0.073	↓0.109	PTGES	↑	↓	↓0.075	↑0.035
SOX5	↓	↑	NA	↑0.13	PTGS2	↑	↓	↓0.096	↓0.02
SOX2	↓	↑	↑0.142	↑0.047	TFPI	↑	↓	↑0.086	↓0.013
SOX2-OT	↓	↑	NA	↑0.047	TWIST1	↑	↓	NA	↑0.004
	↓	↑	↑0.142	↑0.047	SOX9	↑	↓	↑0.196	↑0.119
	↓	↑	NA	↑0.032		↑	↓	↑0.134	↑0.041
SOX8	↓	↑	↑0.074	↑0.089	SOX11	↑	↓	↑0.017	↑0.006
	↓	↑	↑0.251	↑0.123		↑	↓	NA	↑0.016
SOX10	↓	↑	↑0.442	↑0.354		↑	↓	↑0.061	↓0
SOX13	↓	↑	NA	↑0.354					

The left half are down-regulated genes and the right, up-regulated. Columns from left to right are: name of the gene, expected expression, expected methylation pattern in its promoter, our estimated difference in methylation score, difference in Batman's score. False negative means for down-regulated genes (i.e. expected up-rise in methylation score), methylation callers predicts a decrease; and vice versa for false positive.

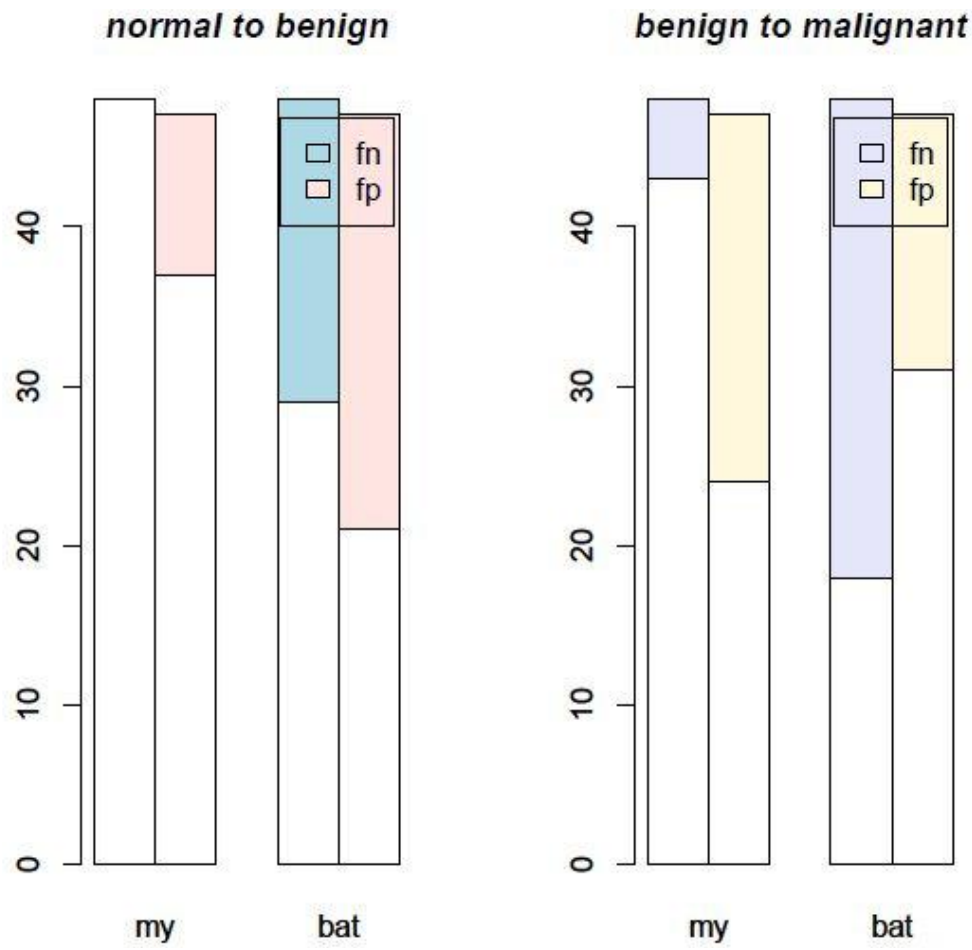

Figure 2. Statistics for the differentially methylated region (DMR) classification. Colored bars on the vertical axis denotes number of mispredicted DMRs (in “normal to benign” case, we predicted with zero false negative rate). Our method outperforms Batman except that in “benign to malignant” case, more up-regulated genes are called to have an increment in methylation score in our method than in Batman.

Table 3. Methylation of suspected promoter regions of genes associated with NF/MPNST: benign to malignant.

Name	Expres.	Met	This study	Batman	Name	Expres.	Met	This study	Batman
EMP2	↓	↑	NA	↓0.044	EN2	↑	↓	↑0.344	↑0.091
EPB41L3	↓	↑	↓0.013	↓0.026		↑	↓	↑0.223	↑0.107
	↓	↑	NA	↓0.039		↑	↓	↑0.418	↑0.127
	↓	↑	↑0.047	↓0	HGF	↑	↓	↓0	↓0.046
GFAP	↓	↑	NA	↓0.077	MDK	↑	↓	NA	↓0.035
	↓	↑	↑1	↑0.299		↑	↓	↑0.015	↓0.034
HLA-DQB1	↓	↑	↓0.085	↓0.034	PAX6	↑	↓	↑0.27	↑0.095
KLK6	↓	↑	↓0.063	↑0.186		↑	↓	↑0.866	↑0.522
LGI1	↓	↑	↑0.054	↓0.043		↑	↓	↑0.092	↑0.048
MBP	↓	↑	NA	↑0.005		↑	↓	↑0.971	↑0.357
	↓	↑	↑0.129	↓0.037	SMAD3	↑	↓	↑0.002	↓0.026
	↓	↑	↑0.033	↓0.02	WT1	↑	↓	↑1	↑0.365
	↓	↑	↑0.031	↑0.004		↑	↓	↑0.928	↑0.527
NGFR	↓	↑	NA	↓0.017		↑	↓	↑0.319	↑0.224
	↓	↑	NA	↓0.2		↑	↓	↑0.431	↑0.509

Table 3. Continued.

CDKN2A	↓	↑	NA	↓0.015		↑	↓	NA	↓0.058
	↓	↑	NA	↓0.044	APOD	↑	↓	↑0.193	↑0.032
	↓	↑	NA	↓0.051		↑	↓	↑0.244	↑0.057
	↓	↑	NA	↓0.036	CASP1	↑	↓	NA	↑0.024
CTSD	↓	↑	↑0.23	↑0.143	CD36	↑	↓	↑0.591	↓0.033
	↓	↑	↑0.182	↓0.105	EGFR	↑	↓	NA	↓0.04
	↓	↑	NA	↑0.001		↑	↓	NA	↓0.016
GNAI2	↓	↑	NA	↓0.014		↑	↓	↑0.042	↓0.051
	↓	↑	NA	↓0.023	KIT	↑	↓	↑0.008	↓0.042
HPCAL1	↓	↑	↑0.069	↑0.006	LEPR	↑	↓	NA	↓0.036
	↓	↑	↑0.069	↓0.027		↑	↓	NA	↓0.026
MFI2	↓	↑	↑0.006	↑0.027	MME	↑	↓	NA	↓0.003
NES	↓	↑	NA	↓0.012	SOCS3	↑	↓	NA	↓0.015
NFKB1	↓	↑	NA	↓0.016		↑	↓	↑0.002	↓0.021
BCL2	↓	↑	↑0.05	↓0.034	ADM	↑	↓	NA	↓0.046
BCL2L2	↓	↑	↑0.002	↑0.021		↑	↓	NA	↓0.064
EDNRB	↓	↑	↑0.154	↓0.075	CAPN1	↑	↓	NA	↓0.007
ERBB3	↓	↑	↑0.065	↓0.014	FBN2	↑	↓	↓0.028	↓0.038
	↓	↑	↑0.004	↓0.023	IGFBP3	↑	↓	NA	↓0.021
	↓	↑	NA	↓0.017	PDGFRA	↑	↓	↓0.002	↓0.017
MPZ	↓	↑	↑0.25	↑0.214	PIAS3	↑	↓	NA	↓0.024
PDGFA	↓	↑	↑0.109	↓0.347		↑	↓	NA	↓0.019
	↓	↑	↑0.075	↑0.009	PLAU	↑	↓	NA	↓0.055
S100B	↓	↑	↑0.804	↑0.336	PTGES	↑	↓	↑0.016	↓0.029
SOX5	↓	↑	↑0.29	↑0.278	PTGS2	↑	↓	NA	↓0.058
SOX2	↓	↑	↓0.019	↓0.045	TFPI	↑	↓	↓0.017	↓0.017
SOX2-OT	↓	↑	↑0.475	↑0.085	TWIST1	↑	↓	↑0.037	↓0.041
	↓	↑	↓0.019	↓0.045	SOX9	↑	↓	↑0.738	↑0.168
	↓	↑	NA	↓0.057		↑	↓	↑0.163	↑0.039
SOX8	↓	↑	↑0.154	↑0.021	SOX11	↑	↓	↓0.017	↓0.006
	↓	↑	↑0.729	↑0.275		↑	↓	NA	↓0.02
SOX10	↓	↑	↑0.456	↑0.158		↑	↓	NA	↓0.075
SOX13	↓	↑	↑0.059	↑0.008					

variations of “methylation” such as 5CaC, 5fC converted from 5meC. Although, this model is copy number indifferent, we expect more conserved treatment of ultra-high read counts. We believe that as sequencing technology continuously lowers the threshold for obtaining higher-quality, longer reads, future-generation methylation analysis tools will demand more informed models intertwined with complex bioinformatics techniques.

REFERENCES

Baum LM, Petrie T, Soules G, Weiss N (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math.Statist. 41(1):164–171.

Bird A (1986). CpG-rich islands and the function of DNA methylation. Nature 321:209–213.

Bird A (2002). DNA methylation patterns and epigenetic memory. Genes Dev. 16(1):6–21.

Birney E (2001). Hidden Markov models in biological sequence analysis. IBM J. Res. Dev. 45(3-4):449-454.

Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG (2010). Whole-genome DNA methylation profiling using MethylCap-seq. Methods. 52(3):232–236.

Carrell DT (2012). Epigenetics of the male gamete. Fertil. Steril. 97(2):267–274.

Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, Herwig R, Adjaye J (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome Res. 20(10):1441–1450.

Dong C, Sun Y (2007). Donor recognition synthesis method base on simulate anneal. In Proceedings of the Life system modeling and simulation 2007 international conference on Bio-Inspired computational intelligence and applications, Springer-Verlag, Berlin, Heidelberg.

Down TA, Rakyan VK, Turner DJ, Flicek P (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat. Biotechnol. 26(7):191–203.

Durbin R, Eddy SR, Krogh A, Mitchison G (1998). Biological Sequence

Analysis - Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefligera C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. 38:1378–1385.
Esteller M (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene 21(35):5427–5440.
Fan SC, Zou JX, Xu HB, Zhang XG (2010). Predicted methylation landscape of all CpG islands on the human genome. Chin. Sci. Bull. 55(22):2353–2358.
Feber A, Wilson GA, Zhang L, Presneau N, Idowu B (2011). Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. Genome Res. 21:515–524.
Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM (2003). Predicting aberrant CpG island methylation. PNAS 100(21):12253–12258.
Gardiner-Garden M, Frommer M (1987). CpG islands in vertebrate genomes. J. Mol. Biol. 196(2):261–282.
Gowher H, Leismann O, Jeltsch A (2000). DNA of *Drosophila melanogaster* contains 5-methylcytosine. EMBO J. 19:6918–6923.
Gu H, Bock C, Mikkelsen TS, J¨ager N, Smith ZD (2010). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Methods 7(2):133–138.
He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, Sun Y, Li X, Dai Q, Song CX, Zhang K, He C, Xu GL (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 333(6047):1303–1307.
Huang J, Renault V, Sengen`es J, Touleimat N, Michel S, Lathrop M, Tost J (2011). MeQA: A pipeline for MeDIP-seq data quality assessment and analysis. Bioinformatics.
Iraola-Guzmán S, Estivill X, Rabionet R (2011). DNA methylation in neurodegenerative disorders: a missing link between genome and environment? Clin. Genet. 80(1):1–14.
Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333(6047):1300–1303.
Jacinto FV, Ballestar E, Esteller M (2008). Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome. Biotechniques 44(1):35–43.
Jordà M, Peinado MA (2010). Methods for DNA methylation analysis and applications in colon cancer. Mutat. Res./Fund. Mol. M. 693(1-2):84–93.
Kawaguchi K, Oda Y, Saito T, Takahira T, Yamamoto H, Tamiya S, Iwamoto Y, Tsuneyoshi M (2005). Genetic and epigenetic alterations of the PTEN gene in soft tissue sarcomas. Hum. Pathol. 36(4):357–363.
Krogh A (1997). Two methods for improving performance of a HMM and their application for gene finding. In Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, pp. 179–186. AAAI Press.
Krueger F, Andrews SR (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics.
Lan X, Adams C, Landers M, Dudas M, Krissinger D, Marnellos G, Bonneville R, Xu M, Wang J, Huang THM, Meredith G, Jin VX (2011). High resolution detection and analysis of CpG dinucleotides methylation using mbd-seq technology. PLoS ONE. 6(7):e22226.
Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(2).
Melodelima C, Gautier C (2007). A Markovian approach for the segmentation of chimpanzee genome. In Proceedings of the 1st international conference on Bioinformatics research and development, 251–262, Springer-Verlag, Berlin, Heidelberg.
Miller SJ, Jessen WJ, Mehta T, Hardiman A, Sites E, Kaiser S, Jegga AG, Li H, Upadhyaya M, Giovannini M, Muir D, Wallace MR, Lopez E, Serra E, Nielsen GP, Lazaro C, Stemmer-Rachamimov A, Page G, Aronow BJ, Ratner N (2009). Integrative genomic analyses of neurofibromatosis tumours identify SOX9 as a biomarker and survival gene. EMBO Mol. M. 1(4):236–248.
Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AM (2008). MEDME: An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome Res. 18(10):1652–1659.
Rabiner LR (1990). Readings in speech recognition. chapter A tutorial on hidden Markov models and selected applications in speech recognition, 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, Andrews TD, Howe KL, Otto T, Olek A, Fischer J, Gut IG, Berlin K, Beck S (2004). DNA methylation profiling of the human major histocompatibility complex: A pilot study for the human epigenome project. PLoS Biol. 2(12):e405, 11.
Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V (2000). Non-CpG methylation is prevalent in embryonic stem cells andmaybe mediated by DNA methyltransferase 3a. PNAS. 97(10):5237–5242.
Rauch TA, Pfeifer GP (2010). DNA methylation profiling using the methylated-CpG island recovery assay (MIRA). Methods 52(3):213–217.
Reinders J, Vivier CD, Theiler G, Chollet D, Descombes P, Paszkowski J (2008). Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. Genome Res. 18(3):469–476.
Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G (2010). Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics 11(137).
Serre D, Lee BH, Ting AH. MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res. 38(2):391–399.
Wang G, Hu X, Lu C, Su C, Luo S, Luo ZW (2008). Promoter-hypermethylation associated defective expression of Ecadherin in primary non-small cell lung cancer. Lung Cancer 62(2):162–172.
Weber M, Davies JJ, Wittig D, Oakeley EJ (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. 37(8):853–862.
Weber M, Hellmann I, Stadler MB, Ramos L (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. 39(4):457–466.
Xi Y, Li W (2009). BSMAP: whole genome bisulfite sequence mapping program. BMC Bioinformatics 10(232).
Yang Y, Wang W, Li Y, Tu J, Bai Y (2010). Identification of methylated regions with peak search based on poisson model from massively parallel methylated DNA immunoprecipitation-sequencing data. Electrophoresis 31:3537–3544.
Yoder JA, Walsh CP, Bestor TH (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 13(8):335–340.
Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7(1-2):203–214.
Zhu H, Wang J, Yang Z, Song Y (2006a). Interpolated hidden markov models estimated using conditional ML for eukaryotic gene annotation. In Proceedings of the 2006 international conference on Computational Intelligence and Bioinformatics - Volume Part III, 267–274, Springer-Verlag, Berlin, Heidelberg..
Zhu H, Wang J, Yang Z, Song Y (2006b). A method to design standard HMMs with desired length distribution for biological sequence analysis. In Proceedings of the 6th international conference on Algorithms in Bioinformatics, 24–31, Springer-Verlag, Berlin, Heidelberg.