BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2414–2422
doi:10.1093/bioinformatics/btu339

Genome analysis

Advance Access publication May 16, 2014

MethylSig: a whole genome DNA methylation analysis pipeline

Yongseok Park[1,*,†], Maria E. Figueroa[2], Laura S. Rozek[3,4] and Maureen A. Sartor[1,5,*]

[1]Department of Computational Medicine and Bioinformatics, [2]Pathology Department, [3]Department of Environmental Health Sciences, [4]Department of Otolaryngology and [5]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: DNA methylation plays critical roles in gene regulation and cellular specification without altering DNA sequences. The wide application of reduced representation bisulfite sequencing (RRBS) and whole genome bisulfite sequencing (bis-seq) opens the door to study DNA methylation at single CpG site resolution. One challenging question is how best to test for significant methylation differences between groups of biological samples in order to minimize false positive findings.

Results: We present a statistical analysis package, methylSig, to analyse genome-wide methylation differences between samples from different treatments or disease groups. MethylSig takes into account both read coverage and biological variation by utilizing a beta-binomial approach across biological samples for a CpG site or region, and identifies relevant differences in CpG methylation. It can also incorporate local information to improve group methylation level and/or variance estimation for experiments with small sample size. A permutation study based on data from enhanced RRBS samples shows that methylSig maintains a well-calibrated type-I error when the number of samples is three or more per group. Our simulations show that methylSig has higher sensitivity compared with several alternative methods. The use of methylSig is illustrated with a comparison of different subtypes of acute leukemia and normal bone marrow samples.

Availability: methylSig is available as an R package at http://sartorlab.ccmb.med.umich.edu/software.

Contact: sartorma@umich.edu

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

Received on February 4, 2014; revised on April 21, 2014; accepted on May 8, 2014

1 INTRODUCTION

DNA methylation (5-methylcytosine) is the most intensively studied and one of the best understood epigenetic marks in mammalian cells, having important roles in imprinting, genome stability and regulation of gene expression without altering the DNA sequence itself (Kulis and Esteller, 2010; Yang *et al*., 2010). In mammalian cells, cytosine methylation occurs almost exclusively at CpG dinucleotides, with the exception of embryonic stem cells, where non-CpG methylation is a frequent occurrence (Lister *et al*., 2009). DNA methylation is essential for normal development and cell differentiation due to its roles in regulating gene expression. For example, unmethylated CpGs in promoter regions can allow binding of specific transcription factors (TFs) while methylated CpGs in these regions can prevent binding (Lim and Maher, 2010). Furthermore, extensive crosstalk occurs between DNA methylation and chromatin modifying histone marks (Izzo and Schneider, 2010; Kassner *et al*., 2013; Shen and Laird, 2013; Vaissière *et al*., 2008). Dysregulation of DNA methylation is a hallmark of cancer, with overall genomic demethylation and gene-specific hypermethylation, most notably in oncogenes and tumor suppressor genes (Sharma *et al*., 2010).

Treatment of DNA with sodium bisulfite deaminates unmethylated cytosines to uracil while methylated cytosines are resistant to this conversion, thus allowing for sequence-specific discrimination between methylated and unmethylated CpG sites (Clark *et al*., 2006). Sodium bisulfite pre-treatment of DNA coupled with next-generation sequencing has facilitated genome-wide quantitative DNA methylation to be studied at single cytosine site resolution (Gu *et al*., 2010; Laird, 2010; Lister and Ecker, 2009). The high cost of whole-genome bisulfite sequencing (bis-seq), and the uneven distribution of CpG sites in the genome motivated the development of modified approaches such as reduced representation bis-seq (RRBS) (Gu *et al*., 2011; Jeddeloh *et al*., 2008; Meissner *et al*., 2005) and enhanced RRBS (ERRBS) (Akalin *et al*., 2012a). These methods have the advantage of requiring fewer sequencing reads by enriching for CpG dense regions of the genome. Bis-seq, RRBS and ERRBS facilitate the study of DNA methylation patterns across the genome in multiple samples and between sample groups.

Recently, ERRBS was used to identify and describe distinct DNA methylation patterns associated with specific forms of acute myeloid leukemias (AML) (Akalin *et al*., 2012a). AML is a highly heterogeneous disease both from the clinical and molecular standpoints, with many distinct molecular subtypes defined by genetic abnormalities; several of these target key epigenetic regulators. An estimated 20–25% of all AMLs are associated with heterozygous somatic mutations of isocitrate dehydrogenase 1 or 2 (*IDH1 or 2*), or ten-eleven translocation 2 (*TET2*) (Patel *et al*., 2011). Any one of these mutations results in an impairment of DNA demethylation pathways and leads to the establishment of a DNA hypermethylation phenotype (Figueroa *et al*., 2010a). A separate class of AMLs, constituting ~15% of all AML cases, are identified by the presence of the t(8;21) translocation giving rise to the AML1/ETO fusion oncoprotein (Petrie and Zelent, 2007).

†Present address: Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

*To whom correspondence should be addressed.

Several methods have been published for analysing genome-wide methylation levels. BSmooth (Hansen *et al.*, 2012), which is targeted to analyse bis-seq data, uses a smoothing approach across the genome within each sample to increase the accuracy of the estimated methylation level for a single CpG site. BSmooth identifies differentially methylated regions (DMRs) by combining top ranked differentially methylated cytosines (DMCs) found using a *t*-statistics approach with either a quantile or direct *t*-statistic cutoff. MethylKit (Akalin *et al.*, 2012b) offers useful annotation features and provides a statistical test by pooling sequencing reads among the individuals in each group. When multiple samples are present, logistic regression with a binary predictor is used, which can be formulated as a binomial-based test. However, methylation levels often vary significantly across individuals, as observed in cancer samples (Hansen *et al.*, 2011). Pooling of reads among individuals may result in inflated type-I error rates when testing for group differences. Non-parametric tests such as the Wilcoxon rank-sum test have also been used to test for DMCs (Nordlund *et al.*, 2013; Wang *et al.*, 2013); however, these tests suffer from a lack of statistical power for experiments with small sample sizes.

To overcome current limitations, we developed methylSig, a genome-wide DNA methylation analysis pipeline for use with bis-seq, RRBS or ERRBS data. MethylSig utilizes a beta-binomial model across the samples in each defined group for each CpG site or region to identify either DMCs or DMRs. It can also incorporate local information across a chromosome to improve estimates of variances and/or methylation levels. MethylSig offers annotation functions to map DMCs/DMRs to gene structures and a unique data visualization approach to view several aspects of the data simultaneously across the genome, providing a platform from which to process, analyse and visualize genome-wide methylation data.

2 METHODS

2.1 Beta-binomial approach

If X follows a binomial distribution with number of trials n and the probability of 'success' P, the distribution function of X is

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,\ldots,n.$$

In the case that p varies between sets of trials and $p\sim$ Beta (α,β), we say X has a beta-binomial distribution with probability mass function

$$P(X=x)=\binom{n}{x}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)}$$

$$=\binom{n}{x}\frac{\Gamma(\mu\theta+x)\Gamma((1-\mu)\theta+n-x)\Gamma(\theta)}{\Gamma(n+\theta)\Gamma(\mu\theta)\Gamma((1-\mu)\theta)},$$

where $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}\,dt$. The mean and variance of X are $n\mu$ and $n\mu(1-\mu)\{1+(n-1)\phi\}$, where mean $\mu=\alpha/(\alpha+\beta)$. Let $\theta=\alpha+\beta$, then the over-dispersion parameter $\phi=1/(1+\theta)$. To simplify the estimation process, we use θ instead of ϕ.

Let $X_{igj}\sim$Beta-Bin$(\mu_{ig}\theta_i,(1-\mu_{ig})\theta_i,n_{igj})$, $g=1,2$, $j=1,2,\ldots,J_g$, where X_{igj} and n_{igj} are number of methylated cytosine reads and coverage at site i of the j^{th} sample in group g, respectively. Our goal is to test the hypothesis $H_{i0}:\mu_{i1}=\mu_{i2}$ against $H_{i1}:\mu_{i1}\neq\mu_{i2}$.

There are several methods to estimate the parameters μ_{ig} and θ_i, such as estimation by moments and maximum likelihood estimator (MLE) (Griffiths, 1973), and an alternative method discussed by Tripathi *et al.* (1994). Although simple, the estimator from the method of moments is not from maximizing the likelihood; thus, it is not useful when applying the likelihood ratio test. Because there is no closed form for the MLE, obtaining estimates of μ_{ig} and θ_i simultaneously would be computationally demanding. We propose an approximation method.

The parameter θ_i is a simple function of the over-dispersion parameter. Because our goal is to compare means from different groups, θ_i is a nuisance parameter. Based on the Wilks phenomenon results on generalized likelihood ratio test (Fan *et al.*, 2001), we can first estimate the nuisance parameter θ_i and use the estimated $\hat{\theta}_i$ to conduct the likelihood ratio test.

Let $\ell(\mu_{i1},\mu_{i2},\theta_i)$ be the log joint likelihood function. We estimate $\hat{\theta}_i$ by setting $d\ell(\cdot)/d\theta_i=0$ (see supplement for formulas) while using the observed group mean $\sum_{j=1}^{J_g}x_{igj}/\sum_{j=1}^{J_g}n_{igj}$ as the estimate for μ_g. Here we restrict $\hat{\theta}_i\geq 0$.

Given $\theta_i=\hat{\theta}_i$, the MLE for μ_{ig}, $g=1,2$ is the solution of the equation

$$\frac{d\ell}{d\mu_{ig}}(\mu_{i1},\mu_{i2},\theta_i=\hat{\theta}_i)=0,\ g=1,2.$$

Under the null hypothesis $H_{i0}:\mu_{i1}=\mu_{i2}=\mu_i$, the MLE for μ_i is the solution of the equation

$$\frac{d\ell}{d\mu_i}(\mu_{i1}=\mu_i,\mu_{i2}=\mu_i,\theta_i=\hat{\theta}_i)=0.$$

The likelihood ratio test then becomes

$$D=2\{\ell(\mu_{ig}=\hat{\mu}_{ig},\theta_i=\hat{\theta}_i)-\ell(\mu_{ig}=\hat{\mu}_i,\theta_i=\hat{\theta}_i)\}.$$

The distribution of D asymptotically approaches a χ_1^2 distribution as sample size increases. However, the distribution of D has significantly heavier tails than χ_1^2 for small sample sizes, because it is based on the estimated nuisance parameter $\hat{\theta}$. Motivated by the two-sample *t*-test, which can be considered a likelihood ratio test conditional on the estimated variance, we propose the approximation $D\sim t_p^2$, where t_p is the Student *t* distribution with p degrees of freedom. Our permutation results in Section 3.2 show empirically that this is a close approximation, with the *P*-values from our proposed method closely matching the expected *P*-values under the null. In contrast, the results are noticeably anti-conservative using the χ_1^2 distribution (Supplementary Figure S1).

2.2 Simulations

To assess the ability of methylSig to identify true positive DMCs, we conducted a set of simulations, and compared three versions of methylSig with four other methods: the binomial-based test of MethylKit, BSmooth, a standard *t*-test and Wilcoxon rank test. Using ERRBS data from AML samples (Supplementary Table S1), we generated data based on the properties of data from chromosome 1 of the five IDH1/2 mutated (cancer) and four normal bone marrow (NBM) samples. This resulted in 65 284 CpG sites for which at least three samples in each group satisfied the required coverage level. When generating data, we preserved the actual CpG locations and coverage levels for each sample. To produce data similar to a real situation, we first estimated the site-specific dispersion parameters using beta-binomial approach and group mean methylation levels for the normal group using BSmooth. We then used a beta-binomial distribution with the estimated dispersion parameters and group methylation levels to generate data for the normal

samples and non-DMCs for the cancer samples. For cancer samples at DMC sites, we randomly chose methylation differences uniformly between 15% and 30% (weaker signal) or 25% and 40% (stronger signal). We also simulated three levels of clustering of DMCs into regions, resulting in six scenarios. We performed 100 simulations for each scenario. We used the percent estimates of methylation to perform the standard *t*-test and Wilcoxon rank test. We select top ranked CpG sites based on *t*-statistics provided by BSmooth after smoothing.

2.3 Options for combining local information

To incorporate local information when estimating group mean methylation levels and the dispersion parameters, we also provide a local MLE using triangular Kernel weights. This method is particularly useful for small sample sizes and when the group methylation levels or dispersion parameters are locally similar or highly correlated, as we observed they are for up to 200–300 bp windows (Supplementary Figure S2).

For CpG site i, let $S=\{k:-R\le k-i\le R\}$, where R is a predefined range in base pairs to combine local information. For CpG sites k in S, the weight function is $H((k-i)/(R+1))$. The default function is $H(u)=(1-u^2)^3$, which can be redefined.

The $\hat{\theta}_i$ is the solution of the equation:

$$\sum_{k\in S} H(k-i)\frac{d\ell}{d\theta_i}(\mu_{k1},\mu_{k2},\theta_i)=0.$$

The degrees of freedom $p=\sum_{k\in S}H(k-i)(J_{k1}+J_{k2}-2)$ when using the squared *t*-distribution approximation for the likelihood ratio statistics. Note that only sites with data for at least two samples in each group are used. Similarly, $\hat{\mu}_{ig}$ and $\hat{\mu}_i$ can be obtained using

$$\sum_{k\in S} H(k-i)\frac{d\ell}{d\mu_{ig}}(\mu_{i1},\mu_{i2},\theta_i=\hat{\theta}_i)=0,$$

and

$$\sum_{k\in S} H(k-i)\frac{d\ell}{d\mu_i}(\mu_{i1}=\mu_i,\mu_{i2}=\mu_i,\theta_i=\hat{\theta}_i)=0.$$

2.4 Identifying enriched or differentially methylated TFs

For each TF in a given database such as ENCODE uniform TF (http://genome.ucsc.edu/ENCODE/), our goal is to identify which TFs are enriched, that is, which TFs' binding sites have a significantly larger proportion of DMCs than the overall proportion of DMCs. Let N_{total} be the total number of compared CpG sites that can be annotated into TFs and among these, let N_{dmc} be the total number of identified DMCs. For the TF i, let n_i^T be the number of CpG sites and n_i^D be the number of DMCs within this TF i. We test $H_{i0}: p_i^T=p_i^D$ versus $H_{i1}: p_i^T\neq p_i^D$, where $\hat{p}^T=n_i^T/N_{\text{total}}$ and $\hat{p}^D=n_i^D/N_{\text{dmc}}$. Because $N_{\text{total}}\gg N_{\text{dmc}}$, we use likelihood ratio test based on binomial distribution for p_i^D and treat $\hat{p}^T$ as the known true proportion.

Alternatively, we can ask which TFs have a significant level of hypermethylation or hypomethylation across their binding sites, which could indicate whether the TF is having a weaker or stronger regulatory effect, respectively. To address this, for each sample we first tile all reads from regions to which a particular TF is predicted to bind. We then apply our beta-binomial model to the data for each TF to identify TFs with hyper- or hypo-methylated binding sites. This performs a self-contained hypothesis test, in that the level of differential methylation is compared with the null hypothesis of no differential methylation, as opposed to the level of differential methylation outside of the TF binding sites.

3 RESULTS

3.1 Overview of methylSig

The methylSig workflow proceeds through several steps, from reading in data to annotating and visualizing results (Fig. 1). It accepts methylation data defining the number of Cs and Ts at each CpG site for multiple samples that are assigned to one of two or more groups (e.g. myCpG methylation call files from Bismark software, Krueger and Andrews, 2011). If coverage is low or the aim is to examine methylation trends, the data may be tiled within regions of specified width or local information may be incorporated to increase power. MethylSig utilizes a beta-binomial model to compare methylation levels at each CpG site or tiled region between two groups of samples, as defined by the user. Parameter estimation includes two stages. In the first stage, the dispersion parameter is estimated at each CpG site or region; this parameter accounts for the biological variation among samples within the same group. A weighted likelihood can be used to incorporate information from nearby CpG sites or regions. In the second stage, the group methylation level at each CpG site or region is calculated using the estimated dispersion parameter; again information can be incorporated from nearby CpG sites. A statistic based on the likelihood ratio test is used to evaluate the significance level of the difference in methylation. *P*-values and *q*-values are calculated based on either a t_p^2 (default) or a χ_1^2 approximation (recommended for large sample sizes). Finally, the methylSig package provides data visualization and annotation functions as well as functions to identify enriched TFs.

3.2 Evaluation of type-I error using permutations

To evaluate the type-I error rate of methylSig under the null hypothesis of no signal, and compare it with that of the binomial model, we permuted 21 AML samples (Supplementary Table S1) and tested for DMCs. Sample libraries were prepared using the ERRBS method (Akalin *et al.*, 2012a), and CpG sites with ≥10× but ≤500× coverage were included. At each CpG site, there were ≤21 pairs of percent methylation level and coverage, because

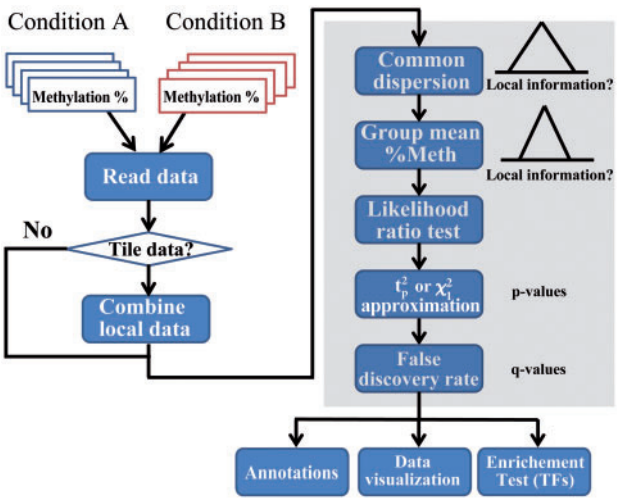

Fig. 1. Workflow of methylSig package

some sites were not covered in some samples; these pairs were permuted and randomly assigned to samples in two groups (11 'treatment' and 10 'control' samples) to generate groups lacking meaningful methylation differences.

The binomial and beta-binomial tests were applied to calculate *P*-values, which were compared with the expected *P*-values under the null hypothesis of no differential methylation. We analysed CpG sites covered by at least two samples in both groups. Results are displayed in QQ-plots of the $-\log_{10}(P\text{-values})$ (Fig. 2). For sites covered by at least six samples, methylSig closely follows the expected *P*-value distribution resulting in a well-calibrated type-I error rate; however, the *P*-values from the binomial model are anti-conservative (Supplementary Figure S3), with *P*-values as low as 1.25×10^{-177} with four covered samples and 2.7×10^{-285} with 21 samples, and 41% of CpG sites satisfying False Discovery Rate (FDR) < 0.05. In contrast, the minimum FDR for methylSig was 0.066 among the 2.46 million CpG sites tested.

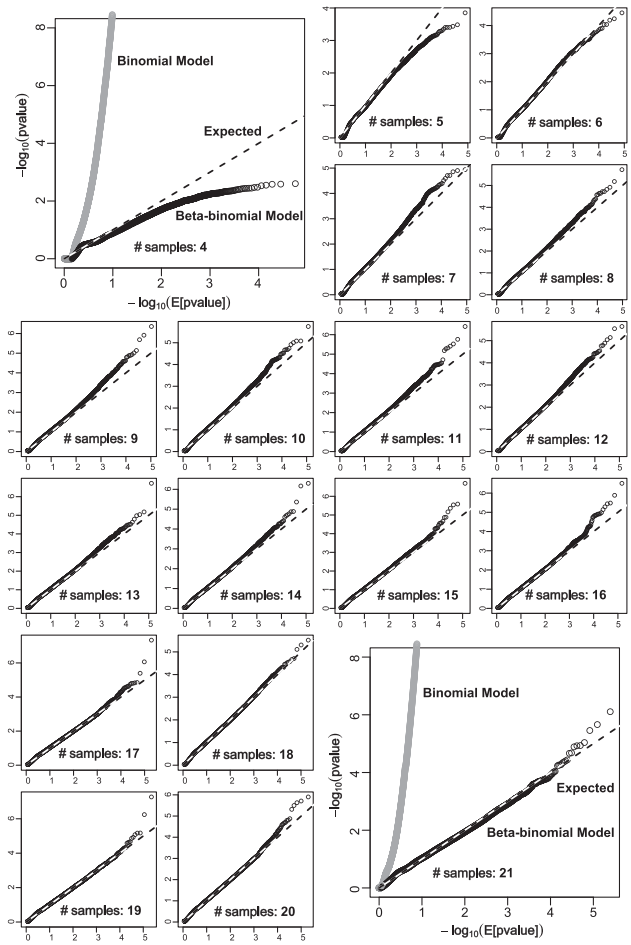

Fig. 2. QQ-plots of observed versus expected $\log_{10}(P\text{-values})$ for the binomial and beta-binomial models show that methylSig has a well-calibrated type-I error rate when each group has ≥ 3 samples (number of samples = 6). Results from the binomial model consistently show an inflated type-I error rate (gray; shown for four and 21 samples). Data from 21 samples were permuted among groups for each CpG site

3.3 Simulation

To assess methylSig's ability to identify true positive DMCs, we conducted simulations comparing methylSig with four other methods: a standard *t*-test, Wilcoxon rank sum test, the binomial test, and BSmooth. Data were generated to mimic the observed properties of ERRBS data, including coverage levels and variation in methylation levels among samples (see Section 2 for details).

Simulations were performed varying two main parameters, the range of differences in percent methylation and the level of DMC clustering into DMRs. The difference in percent methylation for DMCs or DMRs for a simulation was either randomly chosen to be between 15% and 30% (to simulate weaker signals) or 25–40% (stronger signals). We simulated three levels of DMC clustering. First, we simulated DMCs independently, randomly choosing 5% of the CpG sites to be DMCs. This simulates situation that DMCs uncorrelated with neighboring sites. Next DMCs are simulated to be clustered into regions (DMRs) by dividing the chromosome into *R* regions, and randomly selecting 5% of the regions to be DMRs. In the lower level of clustering, DMCs were assigned to 200 different regions ($R = 4000$) with ~16 CpG sites in each DMR. In the higher level, DMCs were assigned to 50 different regions ($R = 1000$) with ~65 CpG sites in each DMR. In both cases, the methylation difference for each CpG site in a DMR was set to be equal.

We compared seven methods: BSmooth, a standard *t*-test, the binomial-based test, the Wilcoxon rank test, our site-specific beta-binomial test, and our beta-binomial test incorporating local information (dispersion and methylation levels separately) (Fig. 3). Because not all methods being compared have a appropriate type-I error rate, and researchers are often interested in the top ranked declared DMCs, we compared the proportion of true DMCs among the total declared DMCs for varying cut-offs for each method. In all cases, our methods performed as well or better than the alternatives, with methylSig using local information for dispersion estimation outperforming methylSig without use of local information. This is not surprising given the observed high correlation of dispersion parameters within a 200–300 bp range. MethylSig using local information to estimate methylation levels outperformed methylSig without use of local information, for lower and high levels of DMC clustering (narrow and broad DMRs) (Fig. 3c–f), but not for independent DMCs (Fig. 3a and b). For DMCs that occur independently across CpG sites, BSmooth has strikingly low detection power (Fig. 3a and b). The performance of BSmooth improves dramatically for clustered DMCs, with performance similar to the site-specific version of methylSig for broad DMRs with 25–40% differences in methylation (Fig. 3f). In practice, however, our site-specific test is potentially more robust because it does not depend on the assumption of correlated dispersion parameters.

3.4 Site specific versus tiled analysis

Although a change in methylation at a single CpG site may disrupt gene regulation, in some experiments, the average coverage per cytosine site may be low, making tiling data a desirable option to increase power. Much biological regulation occurs over a wider region covering multiple CpGs. Thus tiling nearby data may provide complementary insights about the

effect of DNA methylation on biological phenotypes. Using the 21 AML samples and four NBM controls, we assessed the auto-correlation among nearby sites using a weighted Pearson correlation coefficient to account for the varying levels of coverage. We found that on average, adjacent CpG sites had high correlation ($R > 0.9$) with correlation dropping rapidly when distance >200–300 bp (Supplementary Figure S2a). Currently, each significant tiled region is considered a separate DMR, and 25 bp is the default size of the tiles. Although we observed this to work well for RRBS data, users will likely want to increase the tiling width for bis-seq data.

3.5 Basing variance estimation on normal samples only

DNA methylation levels are often more heterogeneous among cancer samples than among normal samples of the same tissue (Hansen *et al.*, 2011). This potentially reduces the power to identify methylation differences between diseased and normal samples when the variance is estimated using data from both groups. An alternative approach, with a slightly different interpretation of results, is to calculate variances from normal samples only. This is particularly useful when relevant DNA methylation changes occur in only a subset of the disease samples. Such a subset of cancer samples may harbor a common mutation or have a similar clinical outcome. Thus, we offer the ability to base variance estimates on just one of the groups in methylSig. However, because the number of normal samples is sometimes relatively small, as in our AML example, this approach also increases the uncertainty of the variance estimates. Therefore, in these cases and in our implementation below, we recommend combining information from nearby CpG sites to obtain more robust variance estimation. This is also motivated by the observation that not only methylation levels, but also dispersion estimates are correlated among nearby CpG sites ($R = 0.68$ on average for adjacent CpG sites within 25 bp of each other) (Supplementary Figure S2b).

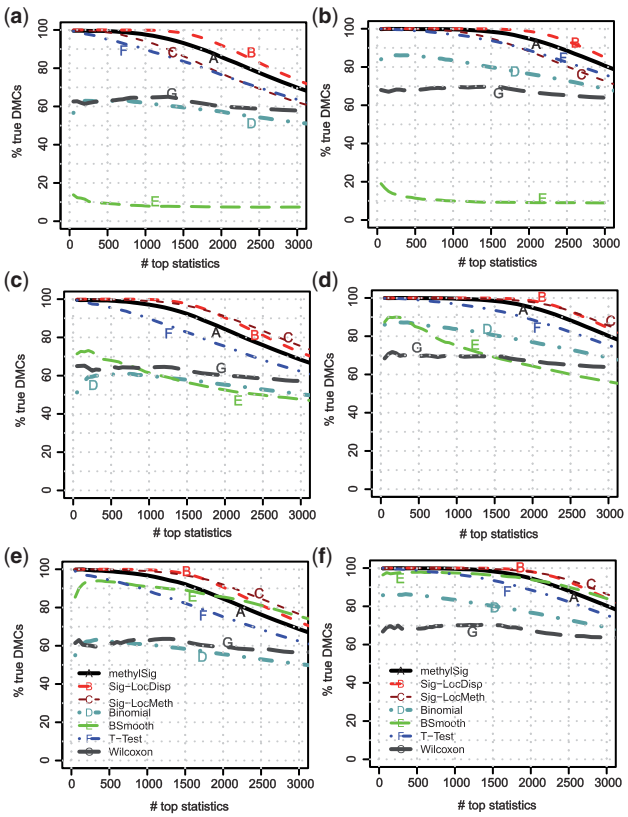

Fig. 3. Simulation study comparing methylSig to alternative methods. Each plot shows the percent of true DMCs among increasing numbers of top ranked DMCs. The *x*-axis value at which a line falls below 0.95 provides the number of sites resulting in a 5% false positive rate. Left column: 15–30% methylation differences. Right column: 25–40% methylation differences. (a and b) uncorrelated DMCs. (c and d) DMCs with a low level of clustering (16 CPG sites per DMR. (e and f) DMCs with a high level of clustering (~65 CpG sites per DMR). A total of 65 284 CpG sites were used, and 100 simulations were conducted for each of the six cases

3.6 Data visualization

MethylSig offers a unique two-tiered visualization of the methylation data depending on the zoom level. For narrow regions (recommended for <100 kb) where at most 500 CpG sites have data reads, users can visualize sample-specific coverage levels and percent methylation at each site, together with group averages, significance levels and a number of genomic annotations (Fig. 4a). To assess the potential biological impact and differential methylation events, we also annotate these CpG sites to the genomic context, including CpG islands and shores, RefGene information (promoter, untranslated region (UTR), intron, exon, non-coding RNA and intergenic regions), and TFs. This is important when interpreting DNA methylation differences, because the presence of methylcytosine may have different effects on gene regulation depending on the context of the genomic region (van Vlodrop *et al.*, 2011). For broad regions, the visualization is simplified to the locations of multiple genomic features (e.g. CpG islands, enhancer regions, etc) along with $\log_{10}$-scale significance levels of the DMCs/DMRs (Fig. 4b).

3.7 Implementation of methylSig with ERRBS data from multiple subtypes of AML

3.7.1 Subtypes of AML Among the 21 ERRBS samples used in our permutation analysis, we classified 13 samples into the following three subgroups: *IDH1/2* mutated ($n = 5$), *TET2* mutated ($n = 4$) and AML1/ETO fusion ($n = 4$). An additional four samples correspond to NBM specimens obtained from healthy donors. Figueroa *et al.* (2010a) previously showed that these AML subtypes display distinct DNA methylation patterns.

We applied the beta-binomial and binomial models to identify DMCs or DMRs (using 25 bp windows) characterizing the AML subtypes and normal samples for CpG sites covered by at least three samples in each group. Significant DMCs/DMRs were defined as those CpG sites or regions with FDR ≤ 0.05 and estimated methylation difference $\geq 25\%$.

3.7.2 Site-specific analysis Both the beta-binomial and binomial models found substantial numbers of DMCs. However, as expected based on the anti-conservative nature of the binomial model, we found substantially more DMCs using the binomial model than beta-binomial model (Fig. 5a). With beta-binomial

model we identified largely different numbers of DMCs when comparing *IDH1/2* mutation, *TET2* mutation, or AML1/ETO gene fusion to NBM samples. Samples harboring *IDH1/2* mutations have previously been reported to have broader and more severe effects on DNA methylation compared with AMLs with *TET2* (Figueroa *et al*., 2010a). Indeed, we identified ~6× more DMCs in the *IDH1/2* group than in the *TET2* group with methylSig. In contrast, the binomial model identified more (~1.7×) DMCs in the *TET2* group. Both groups have been observed to have a strong bias toward hypermethylation. Using methylSig, 98% of *IDH1/2* and 88% of *TET2* DMCs were hypermethylated compared with NBM. Strong effects on DNA methylation were also observed in samples harboring the AML1/ETO gene fusion, although approximately equal numbers of hyper- and hypo-methylated sites were observed in this subtype (Figueroa *et al*., 2010a) (Fig. 5a). For both of these AML subtypes, methylSig identified several thousand DMCs. In contrast, Binomial model identified more similar numbers of DMCs between each of the AML subtypes and NBM samples, with the fewest DMCs found for *IDH1/2*, and only 87% and 63% of the DMCs were hypermethylated in *IDH1/2* and *TET2* samples, respectively. The beta-binomial model also tends to identify DMCs with a larger percent methylation change than binomial model, such that the small *P*-value is more highly associated with large percent methylation change using beta-binomial model than it is with binomial model (Supplementary Figure S4). For example, removing the minimum methylation difference of 25% criteria increases the number (%) of DMCs for *IDH1/2* from 116 605 (7.8%) to 607 745 (40.4%) using the binomial model, but only increases the number from 8521 (0.57%) to 11 018 (0.73%) using our beta-binomial method.

Mapping CpG sites to CpG islands and gene bodies using RefGene models, as can be seen for *IDH1/2* compared with NBM from Figure 6a, ~49% of CpG sites covered by the ERRBS assay are annotated to a CpG island; this percent increased to 62% among identified DMCs. Conversely, the percentage of CpG sites from inter-CpG island regions decreased from 33% covered by ERRBS overall to 20% among DMCs.

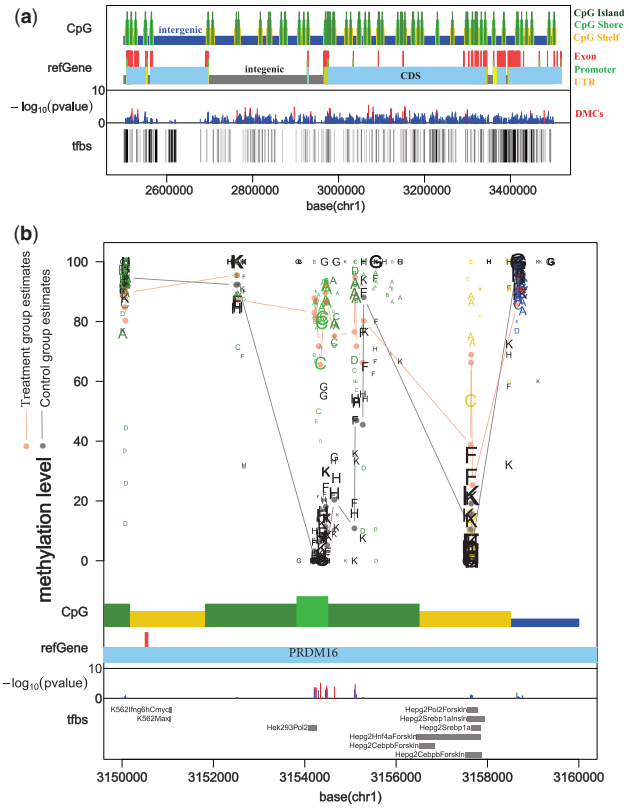

Fig. 4. Data visualization. (a) When the chromosome range is large (>1 mb), the visualization function does not show data. It is helpful to identify regions where DMCs or DMRs exist, and then zoom in to a particular region of interest. (b) When zoomed in to a small region (< 1 mb or at most 500 CpG sites with data), users can visualize the coverage levels, percent methylation levels and group mean methylation estimates at each CpG site. Different letters represent different samples, and the size of each letter represents the read coverage at that CpG site for the related sample. Letter with color (consistent with CpG island annotation) presents treatment group (IDH1/2 mutation) and black for control group (normal). Red and black dots and lines represent the group estimates for treatment and control groups, respectively. CpG islands and shores, intergenic regions, $-\log_{10}$(*P*-values) and protein-DNA binding sites are shown in the annotation tracks below the methylation data

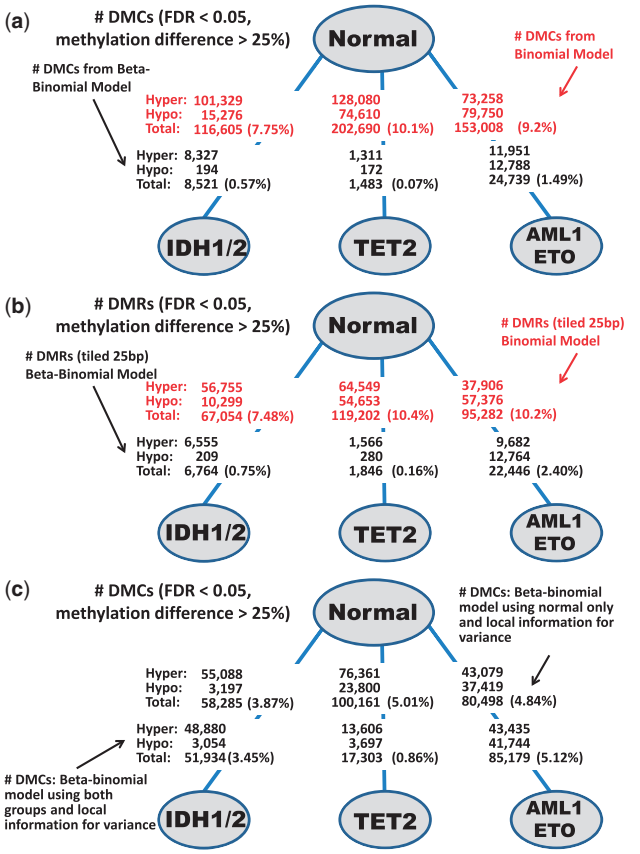

Fig. 5. Number of DMCs identified using methylSig (beta-binomial model) and methylKit (Binomial model) when comparing AML subtypes with *IDH1/2* mutation, *TET2* mutation, another gene mutation and AML1/ETO to NBM samples, respectively. (a) Site specific analysis (DMCs). (b) Tiled analysis (DMRs). (c) Estimating variance with using local information to obtain variances

CpG shores and shelves were neither enriched nor depleted with DMCs. This result is consistent with previous observations regarding the patterns of DMCs in IDH-mutant AMLs (Akalin *et al*., 2012a). We did not observe enrichment of DMCs in any regions defined by gene structure, including promoter, 5′ or 3′ UTR, coding sequence (CDS) (intron and exon), and non-coding RNA when using RefGene models (Fig. 6b) . Using methylSig to test for TFs with potentially enriched DMCs in their binding sites, we identified an extremely strong enrichment of Suz12 sites (*P*-value: 2.4×10^{-190}) (Fig. 6c). Suz12 is a component of the Polycomb Repressive Complex 2 (PRC2), which trimethylates lysine 27 of histone 3. DNA regions normally harboring this histone mark in embryonic stem cells have been observed to be hypermethylated in multiple types of cancer (Widschwendter *et al*., 2006). This suggests that PRC2 target sites also have a strong tendency toward hypermethylation in *IDH1/2* mutant AML samples. An enrichment of Suz12 sites was also observed for the *TET2* and AML1/ETO subtypes (*P*-values: 4.7×10^{-9} and .., respectively), and for Brf2 and Pol3 sites in the AML1/ETO subtype (Supplementary Figure S5).

3.7.3 Tiled analysis For the tiled analysis, we divided the genome into regions of continuous non-overlapping 25 bp windows and tiled (pooled) data within each region. Similar to the site-specific analysis, both the binomial and beta-binomial methods identified substantial numbers of DMRs. As expected due to the increase in statistical power from pooling data from nearby sites, methylSig (Fig. 5b) identified more DMRs for all subtypes of AMLs compared with the site-specific analyses. Again, the great majority, 97% and 88.0%, of DMRs were hypermethylated for *IDH1/2* and *TET2* compared with NBM, respectively. In contrast, fewer significant changes were identified using tiled analysis with the binomial model, with the *IDH1/2* subtype again having the fewest significant changes, and 85% and 64% of DMRs hypermethylated for *IDH1/2* and *TET2* subtypes, respectively.

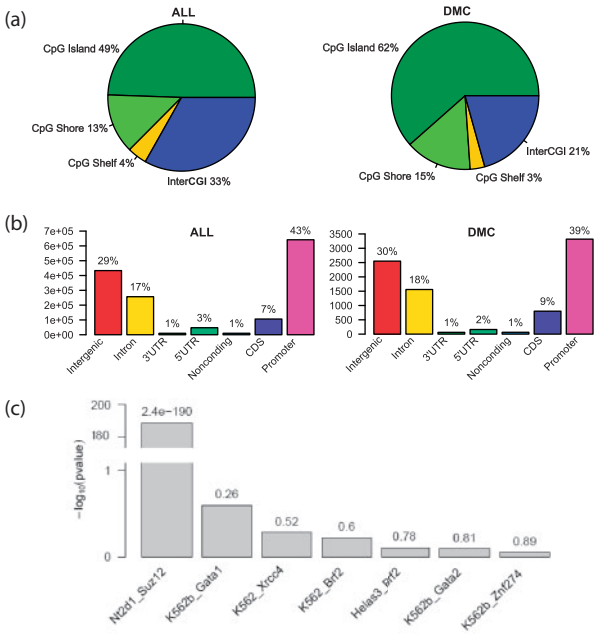

Fig. 6. Annotation results from comparing *IDH1/2* mutant to NBM samples. (a) When annotating CpG sites to CpG island, 13% more CpG sites (from 49% to 62%) are annotated to CpG islands in DMCs while the percentage of CpG sites from inter CpG island regions decreased 13% (from 33% to 20%). (b) No noticeable change is observed when annotating CpG sites to gene bodies using refGene Model. (c) When using ENCODE uniform TF information, Suz12 is the only one TF that is significantly enriched in DMCs compared with all CpG sites used to identify DMCs

3.7.4 Basing variance estimation on normal samples only and using local information When estimating variance from normal samples, many CpG sites with severe heterogeneities among cancer patients are identified as DMCs. To show the utility of this approach we combined the 13 AML samples, simulating a situation where there is no prior knowledge about molecular subtype, and compared this heterogeneous group to the NBM samples using both variance estimation methods (all samples and normal-only samples) with methylSig. We then determined which sites were identified as DMCs in the 'normal-only' analysis to help understand and explore the possibility of cancer subtypes. Clustering the 13 AMLs and four normal samples based on these DMCs separated the AML subtypes and normal samples, with only one *IDH1/2* outlier (Fig. 7). Clustering the samples based on DMCs using all samples to estimate variance was not able to separate the subtypes (Supplementary Figure S6).

When estimating variance with local information, as expected, many more DMCs were identified (3.5–12×, Fig. 5a and c, bottom results). This option of using local information is useful when nearby CpG sites have similar variation across samples within each group. We suggested using 200–300 bp window in each direction based on our observations (Supplementary Figure S2). The patterns of the results are similar with or without local information. We found that most DMCs are hypermethylated when comparing *IDH1/2* or *TET2* mutation group to normal samples, while there was no significant difference between percent hyper- and hypomethylated when comparing

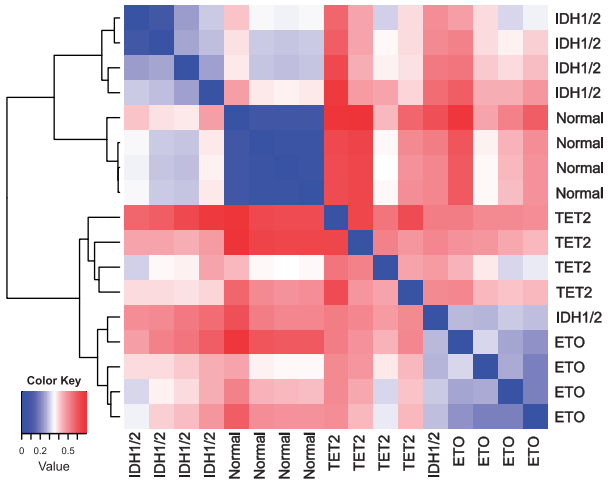

Fig. 7. Clustering AML and NBM samples using CpG sites significant using only NBM samples to estimate the variance is able to separate the molecular subtypes without prior information. The distance measure used was 1 - pairwise correlation. Hierarchical clustering method was 'ward'

AML1/ETO samples to normal samples. When combining local information with using normal samples only to estimate variance option, interestingly, we found the number of identified DMCs when comparing *TET2* to normal samples changed from the least to the most (Fig. 5c). This suggests the possible stronger methylation heterogeneity across samples in *TET2* mutation group.

4 DISCUSSION

Statistical analysis of genome-wide bis-seq experiments with multiple biological samples per group is challenging due to heterogeneous read coverage and methylation levels, relatively small sample sizes and the large number of tests required for a site-specific analysis. Appropriately applying a statistical method to adjust for biological variation, and using information from nearby CpG sites to improve estimation, can significantly reduce the false positive findings while maximizing power. Here, we introduced methylSig, a genome-wide DNA methylation analysis pipeline for bis-seq or RRBS/ERRBS data that estimates biological variation using a beta-binomial approach and maintains a well-calibrated type-I error rate. MethylSig is able to incorporate local information to improve the estimation of variances and group methylation levels. Our simulation results show that in terms of sensitivity, methylSig outperforms both standard statistical tests such as the *t*-test and Wilcoxon rank test, as well as BSmooth and the binomial test of MethylKit under a variety of realistic scenarios. Our method reduces false positive findings by filtering out sites that have a high difference in percent methylation, but are too heterogeneous or have very low of coverage in many samples to be of biological relevance.

Our analyses using ERRBS methylation data from AML and NBM samples showed that methylSig can lead to statistically and biologically relevant results. One subgroup of AML samples had mutations in either *IDH1* or *IDH2*. Nearly all of the DMCs in cancers with neomorphic *IDH1/2* mutations are expected to be hypermethylated, because they lead to abnormally high levels of 2-hydroxyglutarate, which is a direct inhibitor of the TET DNA demethylases required to convert 5 mC to 5 hmC (Xu *et al*., 2011). Similarly, loss-of-function mutations in *TET2* result in impaired DNA demethylation (Figueroa *et al*., 2010b; Ko *et al*., 2010). Although most of the DMCs in the *TET2* mutant subgroup are also expected to be hypermethylated, this occurs to a lesser degree than for the *IDH1/2* group due to the possible partial compensation by other TET proteins. The great majority of DMCs identified by methylSig in these two subgroups were hypermethylated.

Compared to MethylKit, which uses a binomial model without adjusting for biological variation, we demonstrated that our approach performs favorably in terms of type-I error rate, and correlation between significance and percent methylation change. Recently, more sophisticated methods based on Bayesian hierarchical models have been published (Feng *et al*., 2014; Sun *et al*., 2014). Although these methods also make use of a beta-binomial model, they differ from methylSig by using priors for the methylation levels (Sun *et al*., 2014) or dispersion parameters (Feng *et al*., 2014) based on the entire data; methylSig, in contrast, incorporates local information for methylation levels and/or dispersion parameters.

For sites that are highly heterogeneous in both groups, methylSig may not be able to identify regions that are important for a subgroup of samples. This is similar to other traditional statistical methods that test for significant group differences. In this case, a test for a significant subgroup, such as cancer outlier profile analysis used for expression data and implemented in Oncomine (Tomlins *et al*., 2005), may be more appropriate. In cases where one of the groups (e.g. controls) is more homogeneous, users may estimate variance based only on that group in order to identify regions potentially important in a subset of the diseased samples, which are likely more heterogeneous overall.

MethylSig can also be used to analyse alternative cytosine methylation contexts. Finally, the methylSig R package provides a unique visualization approach and useful annotation functions.

Funding: Funding for this work was provided by National Cancer Institute Grant No. R01CA158286-01A1 and National Institute of Environmental Health Sciences Grant No. P30 ES017885-01A1.

Conflict of Interest: none declared.

REFERENCES

Akalin,A. *et al.* (2012a) Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet*., **8**, e1002781.

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

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

Fan,J. *et al.* (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Ann Stat*., **29**, 153–193.

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

Figueroa,M.E. *et al.* (2010a) Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell*, **18**, 553–567.

Figueroa,M.E. *et al.* (2010b) DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*, **17**, 13–27.

Griffiths,D.A. (1973) Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, **29**, 637–648.

Gu,H. *et al.* (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods*, **7**, 133–136.

Gu,H. *et al.* (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc*., **6**, 468–481.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet*., **43**, 768–775.

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

Izzo,A. and Schneider,R. (2010) Chatting histone modifications in mammals. *Brief. Funct. Genomics*, **9**, 429–443.

Jeddeloh,J. *et al.* (2008) Reduced-representation methylation mapping. *Genome Biol*., **9**, 231.

Kassner,I. *et al.* (2013) Crosstalk between SET7/9-dependent methylation and ARTD1-mediated ADP-ribosylation of histone H1.4. *Epigenetics Chromatin*, **6**, 1.

Ko,M. *et al.* (2010) Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*, **468**, 839–843.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.
Kulis,M. and Esteller,M. (2010) DNA methylation and cancer. *Adv. Genet.*, **70**, 27–56.
Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
Lim,D.H. and Maher,E.R. (2010) DNA methylation: a form of epigenetic control of gene expression. *Obstet. Gynaecol.*, **12**, 37–42.
Lister,R. and Ecker,J.R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.*, **19**, 959–966.
Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
Meissner,A. *et al.* (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*, **33**, 5868–5877.
Nordlund,J. *et al.* (2013) Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.*, **14**, r105.
Patel,K.P. *et al.* (2011) Acute myeloid leukemia with IDH1 or IDH2 mutation frequency and clinicopathologic features. *Am. J. Clin. Pathol.*, **135**, 35–45.
Petrie,K. and Zelent,A. (2007) AML1/ETO, a promiscuous fusion oncoprotein. *Blood*, **109**, 4109–4110.
Sharma,S. *et al.* (2010) Epigenetics in cancer. *Carcinogenesis*, **31**, 27–36.
Shen,H. and Laird,P.W. (2013) Interplay between the cancer genome and epigenome. *Cell*, **153**, 38–55.
Sun,D. *et al.* (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.
Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
Tripathi,R.C. *et al.* (1994) Estimation of parameters in the beta binomial model. *Ann. Inst. Stat. Math.*, **46**, 317–331.
Vaissière,T. *et al.* (2008) Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutat. Res.*, **659**, 40–48.
van Vlodrop,I.J.H. *et al.* (2011) Analysis of promoter CpG island hypermethylation in cancer: location, location, location! *Clin. Cancer Res.*, **17**, 4225–4231.
Wang,T. *et al.* (2013) RRBS-Analyser: a comprehensive web server for reduced representation bisulfite sequencing data analysis. *Hum. Mutat.*, **34**, 1606–1610.
Widschwendter,M. *et al.* (2006) Epigenetic stem cell signature in cancer. *Nat. Genet.*, **39**, 157–158.
Xu,W. *et al.* (2011) Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell*, **19**, 17–30.
Yang,X. *et al.* (2010) Targeting DNA methylation for epigenetic therapy. *Trends Pharmacol. Sci.*, **31**, 536–546.