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 Otalaryngology, 5 Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA,

1 Current location: Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

Associate Editor: Dr. 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 analyze 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 to 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

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 & 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 while methylated CpGs in these regions can prevent binding (Lim & Maher, 2010). Furthermore, extensive cross-talk occurs between DNA methylation and chromatin modifying histone marks (Vaissière et al., 2008; Shen & Laird, 2013; Izzo & Schneider, 2010; Kassner et al., 2013). 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 (Lister & Ecker, 2009; Gu et al., 2010; Laird, 2010). 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 bisulfite sequencing (RRBS) (Meissner et al., 2005; Jeddeloh et al., 2008; Gu et al., 2011) 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 approximately 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 & Zelent, 2007).

Several methods have been published for analyzing genome-wide methylation levels. BSmooth (Hansen et al., 2012), which is targeted to analyze bis-seq data, uses a smoothing approach across the
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

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

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

\[ P(X = x) = \binom{n}{x} \frac{\Gamma(x + \alpha) \Gamma(n + \beta - x) \Gamma(\alpha + \beta)}{\Gamma(n + \alpha + \beta)} \frac{(\mu x + \alpha)}{(\mu x + \alpha + (n-x) \phi)}, \]

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

Let \( X_{i,j,k} \sim \text{Beta-Bin}([\mu x_i\theta_i, (1 - \mu x_i)\theta_i], n_{i,j}) \), \( g = 1, 2, j = 1, 2, \ldots, J_g, \) where \( X_{i,j,k} \) and \( n_{i,j,k} \) are number of methylated cytosine reads and coverage at site of the \( j^{th} \) sample in group \( q \), respectively. Our goal is to test the hypothesis \( H_{0,i} : \mu x_i = \mu x_j \) against \( H_{1,i} : \mu x_i \neq \mu x_j \).

There are several methods to estimate the parameters \( \mu x_i \) and \( \theta_i \), 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 \( \mu x_i \) and \( \theta_i \) simultaneously would be computationally demanding. We propose an approximation method.

The parameter \( \theta_i \) is a simple function of the over-dispersion parameter. Since our goal is to compare means from different groups, \( \theta_i \) 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 \( \theta_i \) and use the estimated \( \theta_i \) to conduct the likelihood ratio test.

Let \( \ell(\mu x_1, \mu x_2, \theta_i) \) be the log joint likelihood function. We estimate \( \theta_i \) by setting \( d\ell(\cdot)/d\theta_i = 0 \) (see supplement for formulas) while using the observed group mean \( \bar{\mu x}_{g,i,j} \) as the estimate for \( \mu x_g \). Here we restrict \( \theta_i \geq 0 \).

Given \( \theta_i = \hat{\theta}_i \), the MLE for \( \mu x_{1,g} \) is \( \mu x_{1,g} = \hat{\mu} \), and \( \mu x_{2,g} = \hat{\mu} \). To simplify the estimation process, we use \( \hat{\mu} \) instead of \( \bar{\mu} \).

The likelihood ratio test then becomes

\[ D = 2\ell(\mu x_1 = \mu x_2, \hat{\theta}_i) - \ell(\mu x_1 = \mu x_2, \hat{\theta}_i). \]

The distribution of D asymptotically approaches a \( \chi^2 \) distribution as sample size increases. However, the distribution of D has significantly heavier tails than \( \chi^2 \) for small sample sizes, since it is based on the estimated nuisance parameter \( \theta_i \). Motivated by the two-sample t-test, which can be considered a likelihood ratio test conditional on the estimated variance, we propose the approximation \( D \sim t^2 \), where \( t^2 \) 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 \( \chi^2 \) distribution (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 (Table S1), we generated data based on the properties of data from chromosome 1 of the five IDH1/2 mutated (cancer) and 4 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 15-30% (weaker signal) or 25-40% (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 estimate 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 200-300bp windows (Figure S2).

For CpG site \( i \), let \( S_i = \{ k : -R \leq k - i \leq R \} \), 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 $\theta_i$ is the solution of the equation:

$$\sum_{k \in S} H(k - i) \frac{df}{d\theta_i} (\mu_{11}, \mu_{12}, \theta_i) = 0.$$

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, $\hat{\mu}_{1g}$ and $\hat{\mu}_i$ can be obtained using

$$\sum_{k \in S} H(k - i) \frac{df}{d\mu_{1g}} (\mu_{11}, \mu_{12}, \hat{\theta}_i) = 0,$$

and

$$\sum_{k \in S} H(k - i) \frac{df}{d\mu_i} (\mu_{11} = \mu_1, \mu_{12} = \mu_2, \hat{\theta}_i) = 0.$$

### 2.4 Identifying enriched or differentially methylated transcription factors

For each transcription factor (TF) in a given data base such as ENCODE uniform TF (http://genome.ucsc.edu/ENCODE2), 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 $N_{\text{total}}$ be the total number of compared CpG sites that can be annotated into TFs and among these, let $N_{\text{dmc}}$ be the total number of identified DMCs. For the TF $i$, let $n_{T}^i$ be the number of CpG sites and $n_{D}^i$ be the number of DMCs within this TF $i$. We test $H_{0i}: p_T^i = p_D^i$ vs. $H_{1i}: p_T^i \neq p_D^i$, where $p_T^i = n_T^i/N_{\text{total}}$ and $p_D^i = n_D^i/N_{\text{dmc}}$. Since $N_{\text{total}} \gg N_{\text{dmc}}$, we use likelihood ratio test based on binomial distribution for $p_T^i$ and treat $p_D^i$ 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 to 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 data to annotating and visualizing results (Figure 1). It accepts methylation data defining the number of Cs and Ts at each CpG site for multiple samples that are assigned to one or two or more groups (e.g., myCpG methylation call files from Bismark software, Krueger & 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 $t^2$ (default) or a $\chi^2$ 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 (table S1) and tested for DMCs. Sample libraries were prepared using the ERRBS method (Akalin et al., 2012a), and CpG sites with $\geq 10^6$ but $\leq 500$X coverage were included. At each CpG site, there were $\leq 21$ pairs of percent methylation level and coverage, since 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 to the expected $p$-values from the binomial model. We analyzed CpG sites covered by at least two samples in both groups. Results are displayed in QQ-plots of the $-\log_{10}(p$-values) (Figure 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 (Figure S3), with $p$-values as low as $1.25 \times 10^{-177}$ with four covered samples and $2.7 \times 10^{-285}$ with 21 samples, and 41% of CpG sites satisfying FDR $< 0.05$. In contrast, the minimum FDR for methylSig was 0.066 among the 2.46 million CpG sites tested.

#### 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 BSmiSh. Data were generated to mimic the observed properties of ERRBS data, including coverage levels and variation in methylation levels among samples (see Methods 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 15-30% (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 (\(R = 4, 000\)) with \(~16\) CpG sites in each DMR. In the higher level, DMCs were assigned to 50 different regions (\(R = 1, 000\)) 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) (Figure 3).

Because not all methods being compared have an appropriate type-I error rate, and researchers are often interested in the top ranked 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) (Figure 3(c)-(f)), but not for independent DMCs (Figure 3(a)-(b)). For DMCs that occur independently across CpG sites, BSmooth has strikingly low detection power (Figure 3(a)-(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 (Figure 3(f)). In practice, however, our site-specific test is potentially more robust since it does not depend on the assumption of correlated dispersion parameters.

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 \(\sim 65\) CpG sites per DMR. (e and f) DMCs with a high level of clustering \(\sim 65\) CpG sites per DMR. A total of 65,284 CpG sites were used, and 100 simulations were conducted for each of the six cases.

Fig. 2: QQ-plots of observed versus expected \(\log_2(p\text{-values})\) for the binomial and beta-binomial models show that methylSig has a well-calibrated type-I error rate when each group has \(\geq 3\) samples (\(n = 6\)). 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.
3.4 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 (Figure S2(a)). Currently, each significant tiled region is considered a separate DMR, and 25bp is the default size of the tiles. While 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, since 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 25bp of each other) (Figure S2(b)).

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 $<100kb$) where at most 500 CpG sites have data reads, users can visualize sample-specific coverage levels and % methylation at each site, together with group averages, significance levels and a number of genomic annotations (Figure 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, UTR, intron, exon, noncoding RNA and intergenic regions), and TFs. This is important when interpreting DNA methylation differences, since 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 (Figure 4b).

Fig. 4: Data visualization. (a) When the chromosome range is large (>1 million bp), 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 (<1mbp 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.

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 25bp 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 $\leq 0.05$ and estimated methylation difference $\geq 25\%$. 
3.7.2 Site specific analysis Both the beta-binomial and binomial models found substantial more DMCs using the binomial model than beta-binomial model (Figure 5(a)). 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 to AMLs with TET2 (Figueroa et al., 2010a). Indeed, we identified ~6X more DMCs in the IDH1/2 group than in the TET2 group with methylSig. In contrast, the binomial model identified more (~1.7X) DMCs in the TET2 group. Both groups have been observed to have a strong bias towards hypermethylation. Using methylSig, 98% of IDH1/2 and 88% of TET2 DMCs were hypermethylated compared to 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) (Figure 5(a)). 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 % methylation change than binomial model, such that the small p-value is more highly associated with large % methylation change using beta-binomial model than it is with binomial model (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 8,521 (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 to NBM from Figure 6(a), about 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. 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, CDS (intron and exon), and noncoding RNA when using RefGene models (Figure 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^-100) (Figure 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 1.7 × 10^-102, respectively), and for Brf2 and Pol3 sites in the AML1/ETO subtype (Figure S5).

![Figure 5](http://bioinformatics.oxfordjournals.org/) 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.

3.7.3 Tiled analysis For the tiled analysis, we divided the genome into regions of continuous non-overlapping 25bp 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 (Figure 5(b)) identified more DMRs for all subtypes of AMLs compared to the site-specific analyses. Again, the great majority, 97% and 88.0%, of DMRs were hypermethylated for IDH1/2 and TET2 compared to 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.

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
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 to all CpG sites used to identify DMCs.

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”.

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 BSsmooth 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 % methylation, but are too heterogeneous or have very low 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, since they lead to abnormally high levels of 2-hydroxyglutarate, which is a direct inhibitor of the TET DNA demethylases required to convert 5mC to 5hmC (Xu et al., 2011). Similarly, loss-of-function mutations in TET2 result in impaired DNA demethylation (Figueras et al., 2010b; Ko et al., 2010). While 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 % methylation change. Recently, more
sophisticated methods based on Bayesian hierarchical models have been published (Sun et al., 2014; Feng 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 (COPA) 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 analyze alternative cytosine methylation contexts. Finally, the methylSig R package provides a unique visualization approach and useful annotation functions.

ACKNOWLEDGEMENT

Funding: Funding for this work was provided by NCI Grant # R01CA158286-01A1 and NIEHS grant # P30 ES017885-01A1.

REFERENCES

Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Xi, Y., Tong, P., Meong, M., Goodell, M. A. & Li, Y. (2014) 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 (COPA) 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 analyze alternative cytosine methylation contexts. Finally, the methylSig R package provides a unique visualization approach and useful annotation functions.

ACKNOWLEDGEMENT

Funding: Funding for this work was provided by NCI Grant # R01CA158286-01A1 and NIEHS grant # P30 ES017885-01A1.

REFERENCES

Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Xi, Y., Tong, P., Meong, M., Goodell, M. A. & Li, Y. (2014) 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 (COPA) 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 analyze alternative cytosine methylation contexts. Finally, the methylSig R package provides a unique visualization approach and useful annotation functions.

ACKNOWLEDGEMENT

Funding: Funding for this work was provided by NCI Grant # R01CA158286-01A1 and NIEHS grant # P30 ES017885-01A1.

REFERENCES

Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Xi, Y., Tong, P., Meong, M., Goodell, M. A. & Li, Y. (2014) 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 (COPA) 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 analyze alternative cytosine methylation contexts. Finally, the methylSig R package provides a unique visualization approach and useful annotation functions.