Bioinformatics, 32(10), 2016, 1446–1453
doi: 10.1093/bioinformatics/btw026
Advance Access Publication Date: 27 January 2016
Original Paper

OXFORD

Genome analysis

Differential methylation analysis for BS-seq data under general experimental design

Yongseok Park[1,*] and Hao Wu[2,*]

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA and [2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on 4 August 2015; revised on 7 January 2016; accepted on 13 January 2016

Abstract

Motivation: DNA methylation is an epigenetic modification with important roles in many biological processes and diseases. Bisulfite sequencing (BS-seq) has emerged recently as the technology of choice to profile DNA methylation because of its accuracy, genome coverage and higher resolution. Current statistical methods to identify differential methylation mainly focus on comparing two treatment groups. With an increasing number of experiments performed under a general and multiple-factor design, particularly in reduced representation bisulfite sequencing, there is a need to develop more flexible, powerful and computationally efficient methods.

Results: We present a novel statistical model to detect differentially methylated loci from BS-seq data under general experimental design, based on a beta-binomial regression model with 'arcsine' link function. Parameter estimation is based on transformed data with generalized least square approach without relying on iterative algorithm. Simulation and real data analyses demonstrate that our method is accurate, powerful, robust and computationally efficient.

Availability and implementation: It is available as Bioconductor package DSS.

Contact: yongpark@pitt.edu or hao.wu@emory.edu

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

1 Introduction

DNA methylation is a covalent epigenetic modification of cytosine that plays a critical role in gene regulatory mechanism. It has been well characterized to involve in many biological processes such as cell differentiation, development and aging. It often shows abnormality in many diseases such as cancer (Bird, 2002; Kriaucionis and Heintz, 2009; Meissner *et al.*, 2008; Reik, 2007; Seisenberger *et al.*, 2012; Szulwach *et al.*, 2011b; Teschendorff *et al.*, 2010). Methylated cytosines within promoter regions have been shown to repress gene expression by interfering with DNA-binding proteins that promote transcriptions (Bird and Wolffe, 1999), subsequently leading to phenotypic changes. Studying the dynamics of DNA methylation and their influence has broadened and enhanced our knowledge on many biological mechanisms and disease etiologies, in additional to the genetic components discovered from studying DNA sequences alone.

Technologies for quantifying DNA methylation have evolved rapidly in recent years. Sodium bisulfite treatment of DNA following next-generation sequencing (BS-seq) is a new technology that measures the methylation levels at single-nucleotide resolution. Compared to earlier high-throughput technologies such as methylation microarray (Schumacher *et al.*, 2006) or capture-based sequencing like MeDIP-seq (Weber *et al.*, 2005), BS-seq has improved accuracy, reduced background bias and higher spatial resolution. Thus it has quickly become the technology of choice to study DNA methylation.

One of the most important questions in DNA methylation analysis is to identify genomic loci or regions with different methylation levels among distinct biological conditions. Differentially methylated loci (DML) or regions (DMRs) can be functionally associated with biological processes such as protein bindings and gene expressions to contribute to phenotypic changes. The DML/DMRs could potentially be used as biomarkers for early stage disease diagnosis,

and identifying progression and subtypes of certain disease. DML/DMRs can also provide information for targeted therapies in precision medicine. Developing methods for DML/DMR detection has been and still remains an active epigenomics research topic.

Over the last several years, a number of statistical methods and software tools have been developed for identifying DML/DMRs from BS-seq data (Akalin *et al.*, 2012b; Benoukraf *et al.*, 2013; Feng *et al.*, 2014; Hansen *et al.*, 2012; Park *et al.*, 2014; Saito *et al.*, 2014; Stockwell *et al.*, 2014; Sun *et al.*, 2014). A comprehensive review of the existing methods is provided by Robinson *et al.* (2014). Most of available methods are specifically designed for comparing DNA methylation levels from two treatment groups, partially because most of the existing BS-seq data are generated for two-group comparison due to high cost of the experiment. With rapid advances of sequencing technology and continuous reduction of sequencing cost, an increasing number of experiments are now performed under general, multi-factor design, especially from the reduced representation bisulfite sequencing (RRBS) technology (Jeddeloh *et al.*, 2008; Meissner *et al.*, 2005) and enhanced RRBS (Akalin *et al.*, 2012a). We expect this trend will continue and data with multiple factors/groups and covariates will be common in foreseeable future, even from large-scale population level studies. Flexible and efficient methods are in great demand to comprehensively decipher biological processes in current epigenomics research.

The data from BS-seq can be summarized as numbers of methylated and total reads at each CpG site. The methylated read counts are often modeled by a beta-binomial distribution, which accounts for both the biological and sampling variations (Dolzhenko and Smith, 2014; Feng *et al.*, 2014; Park *et al.*, 2014). A straightforward method to call DML under general design is using a beta-binomial generalized linear model (GLM) at each CpG site. There are currently two methods available for BS-seq under general design: RADMeth (Dolzhenko and Smith, 2014) and BiSeq (Hebestreit *et al.*, 2013). RADMeth directly applies beta-binomial GLM with 'logit' link function. BiSeq first performs a local smoothing within each sample, and then uses a beta regression with 'probit' link on the smoothed methylation levels. The hierarchical testing procedure used in Biseq primarily focuses on identifying DMRs. Another software methylKit (Akalin *et al.*, 2012b) does not explicitly provide any function for multi-factor DML calling. But it internally uses a binomial GLM function for replicated data, which can also be used for general experimental design. All these methods are based on GLM with or without over-dispersion adjustment. Some drawbacks exist using such approach. First it is computationally intensive because model fitting is based on iterative procedure to maximize the likelihood. Given the size of a typical BS-seq dataset (millions of CpG sites), such a procedure is very computationally demanding. This is evidenced by the poor computational performance from both RADMeth and BiSeq. Second, the GLM procedure is numerically unstable when separation or quasi-separation happens in one or more covariates, which frequently occurs in many CpG sites, particularly when methylation levels are close to the boundaries (0 or 1). Such limitations sometimes lead to undesirable results from existing methods.

In this paper, we propose a novel statistical method to detect DML for BS-seq data under general experimental design. At each CpG site, the count data are modeled by a beta-binomial regression with 'arcsine' link function. Model fitting is based on the transformed methylation levels by applying generalized least square (GLS) procedure, which provides estimates for regression coefficients and their covariance. Hypothesis testing is achieved using a Wald test at each CpG site, and the significance levels can be used for DML calling. We perform extensive simulation studies and real data analyses to show the advantages of our proposed method over existing ones. The method, termed '**DSS-general**', is now implemented as a part of the Bioconductor R package DSS.

2 Methods

2.1 The data model

Suppose the input data include N CpG sites and D samples. For CpG site i ($i = 1, 2, \ldots, N$) in dataset d ($d = 1, 2, \ldots, D$), let Y_{id} and m_{id} be the methylated and total read counts respectively. To account for both sampling and biological variation, Y_{id} is modeled using a beta-binomial distribution. To be specific, let p_{id} be the underlying methylation level for CpG site i in sample d, then Y_{id} is assumed to follow a binomial distribution with total trials m_{id} and probability p_{id}. To account for biological variation, p_{id} is assumed to follow a beta distribution with mean π_{id} and dispersion parameter φ_i. Then Y_{id} follows beta-binomial distribution, i.e. $Y_{id} \sim \text{beta-bin}(m_{id}, \pi_{id}, \varphi_i)$.

Let $\boldsymbol{X}$ be a design matrix of dimension $D \times p$, which contains group information and other continuous or discrete covariates. Under generalized linear model framework, we consider π_{id} associated with the experimental design through a linear function: $g(\pi_{id}) = \mathbf{x}_d \boldsymbol{\beta}_i$. Here, $\mathbf{x}_d$ is the d^{th} row of the design matrix $\boldsymbol{X}$, and β_i is a p-dimensional vector of coefficients for CpG site i. $g(\cdot)$ is called link function. We propose to use the following 'arcsine' link function, which was originally considered in Yu (2009) for variance stabilization of binomial data: $g(x) = \arcsin(2x - 1)$. Our final model for BS-seq data under general design is:

$$\begin{aligned} Y_{id} &\sim \text{beta-bin}(m_{id}, \pi_{id}, \varphi_i) \\ \arcsin(2\pi_{id} - 1) &= \mathbf{x}_d \beta_i. \end{aligned} \tag{1}$$

DML detection for any experimental factor/covariate can be conducted by hypothesis test: $H_{0j} : \beta_{ij} = 0$. Furthermore, any linear combination of them can be formulated with a general hypothesis test: $H_0 : \mathbf{C}^T \beta_i = 0$, where $\mathbf{C}$ is a real-valued P-dimensional vector.

2.2 The choice of 'arcsine' link function

The use of 'arcsine' link function is a key component in our method since this link function leads to an efficient and stable estimation procedure. Usually, the selection of link function depends on investigator's belief on the linear relationship between the response and covariates. Commonly used link functions include the 'logit' link $\text{logit}(t) = \log(t/(1-t))$, and 'probit' link $\text{probit}(t) = \Phi^{-1}(t)$, which is the inverse standard normal cumulative density function. However, the parameter estimation with these link functions has no closed form and requires computationally intensive iterative procedures. Furthermore, when the methylation levels are near the boundaries (0 or 1), the parameter estimation becomes unstable or there are no valid parameter estimates when separation or quasi-separation happens. This is a particular concern in the genome-wide DNA methylation analysis because a majority of the CpG sites have methylation levels close to the boundaries.

The 'arcsine' link provides satisfactory solutions to both problems. After the 'arcsine' transformation, the variance of transformed data approximately only depends on dispersion but not the mean (shown below in Equation 2). This provides possibility to apply GLS method for parameter estimation (see next section for details). In contrast, if one chooses to use other link functions such as 'logit' or 'probit' and apply transformation, the variance of transformed data will still be a function of mean, and so iterative procedure is needed for parameter estimation from regression model. Moreover, unlike

logistic regression or beta-binomial GLM, which will fail when separation or quasi-separation happens in at least one covariate, the GLS procedure is very stable even when most of the observed methylation levels in most samples are close to 0 or 1.

2.3 Parameter estimation procedure

The crucial first step is to estimate the linear model coefficient β_i at each CpG site. Due to incomplete coverage of BS-seq, many CpG sites contain missing data, i.e. they are not covered from all samples. The DML detection will be performed for a CpG site as long as there is enough residual degree of freedoms, i.e. it is not required that data at the CpG site are fully complete. In general, the parameter estimation can be achieved by a beta-binomial GLM, which involves maximizing the whole likelihood from data. However, such an approach relies on computationally intensive iterative procedure. Given the size of BS-seq data (~30 million CpG sites from whole genome BS-seq and 3–4 millions from RRBS), the GLM procedure will be very computationally demanding. To circumvent this problem, we consider the following approximation procedure.

Let $Z_{id} = \arcsin(2Y_{id}/m_{id} - 1)$. Then we have

$$E[Z_{id}] \approx \arcsin(2E[Y_{id}]/m_{id} - 1) = \arcsin(2\pi_{id} - 1) = \mathbf{x}_d\beta_i.$$

The variance of Z_{id} is (detailed derivation in Supplementary Materials)

$$\mathrm{var}(Z_{id}) \approx \frac{1 + (m_{id} - 1)\varphi_i}{m_{id}}. \quad (2)$$

It is interesting to note that the variance of Z_{id} does not depend on mean structure approximately. Given dispersion parameter φ_i, GLS can be applied to estimate the regression coefficients β_i. To be specific, define the following covariance matrix:

$$\mathbf{V}_i = \mathrm{diag}\left(\frac{1 + (m_{id} - 1)\varphi_i}{m_{id}}\right),$$

then

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\mathbf{V}_i^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_i^{-1}\mathbf{Z}.$$

For beta-binomial model, there are several ways to estimate dispersion parameter such as maximizing likelihood, and using Pearson χ^2 or deviance statistics. Here we propose to use Pearson χ^2 statistics based on transformed linear model to estimate φ_i because it is less computationally demanding and has relatively good property. We first let $\varphi_i = 0$ and the initial covariance matrix is $\mathbf{V}_{i0} = \mathrm{diag}(1/m_{id})$. The parameter estimator from GLS with covariance matrix $\mathbf{V}_{i0}$ is: $\boldsymbol{\beta}_i^{(0)} = (\mathbf{X}^T\mathbf{V}_{i0}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_{i0}^{-1}\mathbf{Z}$.

Consider Pearson chi-square statistics $\chi_i^2 = \sum_d m_{id}(Z_{id} - \mathbf{x}_d\beta_i^0)^2$. Let $\hat{\sigma}_i^2 = \chi_i^2/(D - p)$, an estimator for φ_i is obtained as below (detailed derivations in Supplementary Materials):

$$\hat{\varphi}_i = \frac{D(\hat{\sigma}_i^2 - 1)}{\sum_d (m_{id} - 1)}. \quad (3)$$

Our model is based on beta-binomial distribution and hence $0 < \varphi_i < 1$, which requires $1 < \hat{\sigma}_i^2 < \frac{\sum_d (m_{id} - 1)}{D} + 1$. However, because of random variation and approximation bias, it is possible that $\hat{\sigma}_i^2$ does not satisfy the constraints. To avoid this, we restrict $\hat{\varphi}_i$ to be bounded by 0.001 and 0.999.

Given estimated dispersion, the estimate of variance structure is now

$$\hat{\mathbf{V}}_i = \mathrm{diag}\left(\frac{1 + (m_{id} - 1)\hat{\varphi}_i}{m_{id}}\right).$$

GLS procedure is applied once more based on $\hat{V}_i$, and the updated estimates for regression coefficients and covariance matrix are obtained as

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\hat{\mathbf{V}}_i^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}_i^{-1}\mathbf{Z},$$

and

$$\hat{\boldsymbol{\Sigma}}_i \equiv \widehat{\mathrm{var}(\hat{\boldsymbol{\beta}}_i)} = (\mathbf{X}^T\hat{\mathbf{V}}_i^{-1}\mathbf{X})^{-1}.$$

The estimation procedure utilizes two GLS for each CpG site without relying on intensive iterative algorithm. It has profound connection with a beta-binomial GLM, in which the initial regression coefficients are estimated from logistic regression and using Pearson χ^2 statistics to estimate dispersion parameter similar to Equation (3) (Hinde and Demétrio, 1998). The covariance structure $\mathbf{V}_i$ is diagonal matrix. Thus our GLS procedure is also a wighted least square, where weight for sample d is $V_{id}^{-1/2}$.

2.4 Hypothesis testing

Hypothesis testing for differential methylation at CpG site i can be formulated as: $H_0 : \mathbf{C}^T\beta_i = 0$. The procedure is very general and can test any linear combination of the effects. For example, to test the effect of factor k, $\mathbf{C}$ is a vector having 1 at the k^{th} element and 0's in all others. With point estimation and estimated covariance matrix, the null hypothesis is tested through a standard Wald test procedure. The Wald test statistics is calculated as:

$$t_i = \mathbf{C}^T\frac{\hat{\boldsymbol{\beta}}_i}{\sqrt{\mathbf{C}^T\hat{\boldsymbol{\Sigma}}_i\mathbf{C}}}.$$

The Wald test statistics approximately follow normal distribution, and the P-values can be obtained accordingly. False discovery rate (FDR) can be estimated using established procedures such as Benjamini–Hochberg's method (Benjamini and Hochberg, 1995).

2.5 Simulation settings

In all simulations, data are generated semi-parametrically. The counts are generated from the data model described in Equation (1) with model parameters estimated from the human lung adenocarcinoma dataset (described later). The model is 2 × 2 factorial design with 20 000 CpG sites for 3 and 10 replicates in each condition group (12 and 40 datasets in total). In order to mimic real DNA methylation level patterns, we first estimate smooth methylation levels over CpG basepair locations from real data and selected 20 000 CpG sites. The methylation levels from these selected CpG sites are used as baseline. The regression coefficients β_g (g = 0, 1, 2) are simulated in the following way: (1) β_0 (the intercept) is used the smooth estimates from selected CpG sites; (2) Randomly select n_b number of starting index from 20 000 CpG sites and randomly draw the DMR size from uniform(5, 50). (3) β_1 and β_2 for a block of CpG sites within the same DMR regions are sampled from Uniform(0.1, 0.7), which related to 0.2 − 12% change of methylation level at 0% (or 100%), or 5 − 32% change at 50% methylation level; (3) The dispersion parameter φ_i' s are independently generated from log-normal distribution with mean -3 and standard deviation 0.7, which are similar to the real data estimates. The simulations are repeated

for $n_b = 40, 160, 400$, which roughly provides proportion of DML at 5%, 20% and 50% of total number of CpG sites. The DM statuses for two factors are independently generated.

3 Results

3.1 Simulation

Comprehensive simulation studies are conducted to evaluate the performance of DSS-general from several different aspects.

3.1.1 DML detection accuracy

We first compare the DML detection accuracies from several methods, including DSS-general (version 2.10.0), RADMeth, BiSeq (version 1.10.0) and a binomial GLM with 'logit' link in R (version 3.2.2) for comparison. All methods are applied with default settings. Here, the proportion of true positives among a given number of top-ranked CpGs is used as criterion. This refers to true discovery rate (TDR) hereafter. Higher TDR is expected from better method. This criterion is also referred to precision–recall analysis, which is a commonly used and considered as better measurement of the accuracy for genome-wide differential analysis than receiver operating characteristic (ROC) (Davis and Goadrich, 2006).

Each simulation is repeated 50 times to obtain the average TDR estimates. Figure 1 shows the TDR curves by varying the number of replicates in each group and the percent of DML CpG when data are simulated using 'arcsine' link function. It can be seen that DSS-general and RADMeth are very similar and outperform other two methods for all top ranked CpG sites. For example, among top 200 ranked CpGs with 3 replicates per group and ~ 5% DML from DSS-general and RADMeth, 79% are true DML, whereas the percentages are 70%, and 59% from BiSeq and binomial GLM respectively. The performance difference becomes less significant when sample sizes increase (right versus left panels in Fig. 1) and when the proportion of DML increases (Panels from different rows in Fig. 1). It is clearly seen that the performance from Binomial GLM is not acceptable even with increased sample sizes. This is not surprising because Binomial GML does not adjust for over-dispersion as other three methods are designed based on beta-binomial model.

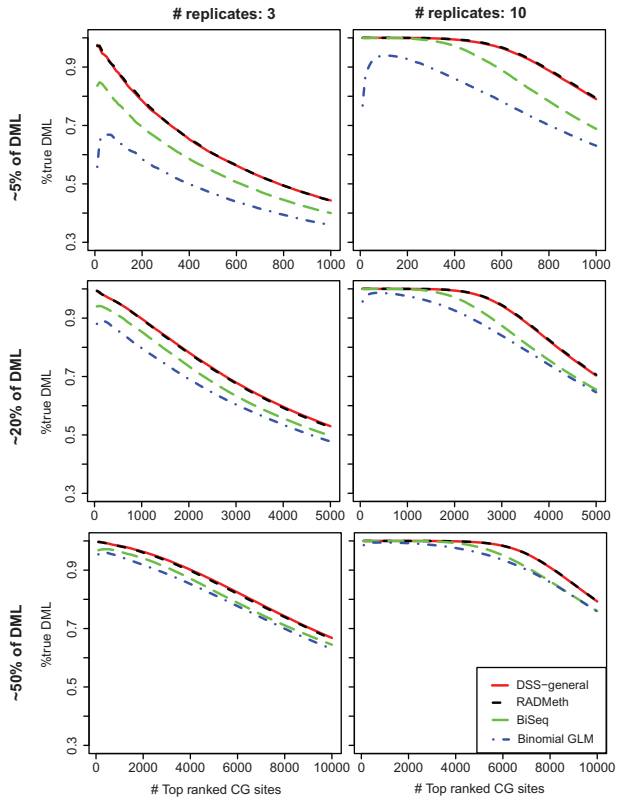

Fig. 1. Comparison of DML detection accuracies from simulations. The proportions of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. Data are simulated based on the model in Equation (1). As can be seen from left to right panels, when sample sizes are increasing, the detecting accuracy is also increasing. We also notice that when proportion of DML increases, the detecting accuracy also becomes better and the difference between different methods tends to be smaller

To assess the robustness to model mis-specification of DSS-general, we conduct another simulation with the same setting as above except for changing the link function to 'logit', i.e. the linear relationship shown in Equation (1) is $\text{logit}(\pi_{id}) = \mathbf{x}_d\beta_i$. In this case, since the scales of the coefficients under 'logit' link are greater than those from 'arcsine' link for the same data (by a ratio of approximately 2.3), we multiply 2.3 for β_g' s for all simulations using 'logit' link. As seen in Supplementary Figure S1, the TDRs from all methods have roughly similar patterns and DSS-general and RADMeth still provide the best accuracy among all methods.

3.1.2 Statistical inference

Accurate Statistical inference is another important indicator of a good statistical method. Figure 2(A) displays the normal QQ-plot of the Wald statistics when testing the first factor. The middle part of the distribution matches normal quantiles very well, while the heavy tails in both ends represent DML. This suggests that the good approximation of using normal distribution to obtain *P*-values. Figure 2(B) shows the histogram of the nominal p-values from DSS-general. It has desired shape of a *P*-value distribution with a spike close to 0, which is enriched by DML, and reasonably flat for the rest of the region. For comparison purpose, we also show the *P*-value distributions from other methods in Figure 2(C–E). Those p-values are not quite uniformly distributed for non-DML CpG sites, suggesting better statistical inference of DSS-general.

Further, we compare the type I error rates from different methods. Data are generated from null models where none of the CpG site are DML, and four methods are applied to obtain *P*-values for

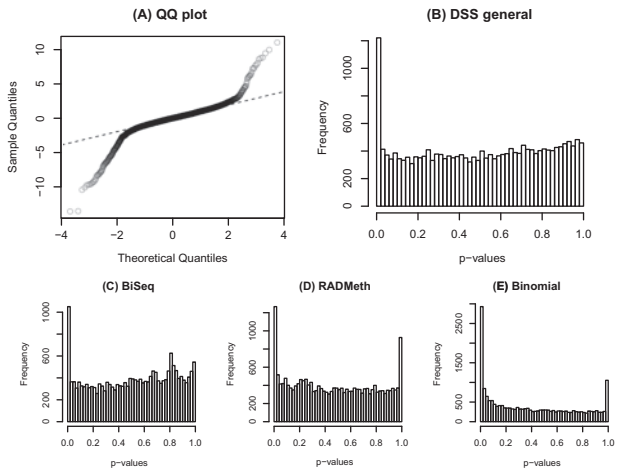

Fig. 2. Properties of test statistics and *P*-value distribution based on 10 replicates per group and ~ 5% DML proportion. (A) Normal QQ-plot of test statistics; (B)–(E) Histogram of *P*-values estimated from simulation using DSS-general, BiSeq, RADMeth and Binomial GLM respectively

all CpG sites. The type I error rates are then obtained under *P*-value threshold of 0.05. Again, the simulation is repeated for 50 times. Figure 3 show the boxplots of type I error rates from different methods and the average type I error rates are further summarized in Table 1. BiSeq provides very accurate type-I error rates. Both RADMeth and DSS-general have slightly inflated type-I rate and the inflation is reduced with increases of sample sizes. When data are simulated from 'arcsine' link with 10 replicates per group (Fig. 3B), DSS-general gives very accurate type-I error rate (0.053 for nominal level of 0.05). Binomial GLM's type-I error rate appears not acceptable and does not improve even when sample sizes become larger.

3.2 Analyses of real data

3.2.1 Human lung adenocarcinoma

This dataset is obtained from GEO database under accession number GSE52140. It contains eight RRBS experiments for two non-small cell lung cancer (NSCLC) cell lines A549 and HTB56 under two conditions: normal and highly metastatic (Hascher *et al.*, 2014). Metastasis often happens in late stages of caner when it spreads from one part of the body to another. It is the leading cause for cancer mortality. One primary goal in this study is understanding the dynamics of DNA methylation during tumor metastasis. This dataset is a typical 2 × 2 crossed design. There are two biological replicates in each cell × condition combination. On average, this dataset covers 3.4 million CpG sites with the average sequencing depth of 22.

There are 5 200 129 CpG sites covered by at least one experiment. A full model including cell, condition and cell × condition interaction is fit using DSS-general, BiSeq, RADMeth. Most of those 5 million CpG sites are not covered by all experiments, which leads to the results from part of the CpG sites. On average, for testing different factors, DSS-general covers around 60% of the CpG sites, compared to RADMeth's 55% and BiSeq's only 27%. This discrepancy is caused by different data filtering missing data processing algorithms from different software. DSS-general works as long as there is enough residual degree of freedoms without requiring fully complete data from all experiments. Therefore, it can provide results for more CpG sites than other methods.

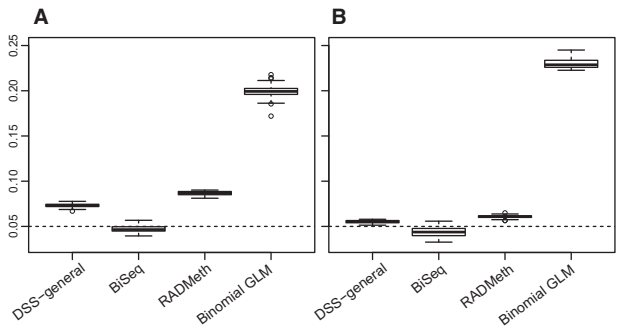

Fig. 3. Boxplots of type I errors from simulation. Data are generated from null model (no DML) with 'arcsine' link with (A) 3 replicates per group and (B) 10 replicates per group. Type I error rates are obtained based on p-value threshold of 0.05. Box-plot results are from 50 simulations

Table 1. Type I errors for DML detection, from simulation

# replicates	DSS-general	BiSeq	RADMeth	binomial GLM
3	0.063	0.052	0.081	0.151
10	0.053	0.046	0.061	0.175

The results from testing condition effect (normal versus metastasis) are compared. To have a fair comparison, we only select CpG sites covered by all three methods. In total, there are 1 400 281 such CpG sites, about 27% of all from the merged data. The *P*-value distributions for these CpG sites are shown in Figure 4. Since those p-values are from mixed DML and non-DML, we expect *P*-value distribution with a spike close to 0, which is enriched by DML, while close to uniformly distributed for CpG sites with larger *P*-values from non-DML. DSS-general shows the most desirable distribution, indicating better model fitting and statistical inference. Further, we look at the Spearman's rank correlation of reported *P*-values from different methods. The correlation is very high between DSS-general and RADMeth at 0.94 while it is 0.65 between DSS-general and BiSeq. The low correlation between BiSeq and the other two methods is likely from the smoothing procedure implemented by BiSeq only. Pairwise scatterplots of the p-values are shown in Supplementary Figure S3.

Next, we examine DML from different methods. Both BiSeq and RADMeth are providing adjusted *P*-values, however, those adjustments are used to identify DMRs in which RADMeth uses fixed width regions and BiSeq identifies variable size regions. To evaluate site specific performance of different methods by adjusting multiple comparisons, we apply Bonferroni correction with critical value of 0.05 to call DML. DML Numbers are 2670 from DSS-general, 958 from BiSeq and 0 from RADMeth, showing that DSS-general is more sensitive. We further explore the potential biological impact of DML on genes by considering those with at least five DML. There are 44 such genes from DSS-general, one from BiSeq (PRDM16) and none from RADMeth. A complete list of the genes detected from DSS-general is provided in Supplementary Table S1. Among those 44 genes reported from DSS-general, many of them have been previously reported to be related to lung cancer, for example, KLF6 (Spinola *et al.*, 2007), KLF2 (Xie *et al.*, 2011), KLC1 (Togashi *et al.*, 2012), PCSK6 (Hawes *et al.*, 2010), MAFK (Boutros *et al.*, 2009). To have more objective evaluation of biologically meaningful genes, we jointly search the name of each gene and 'lung cancer' on PubMed. We deem the gene meaningful if a search returns published paper. Out of 44 genes, 23 (52%) come with at least one published paper. As control, we randomly sample 44 from all genes 100 times and at each time, we conduct the same search. Only about 7/44 (16%) returns at least one paper from these random genes, suggesting that DSS-general indeed detect biologically meaningful candidate genes. In order to find some genes from other methods, we lower significance level to *P*-values of 5×10^{-5} without multiple testing correction. DSS-general reports 340 genes compared to BiSeq's 16 and RADMethd's 12. Larger number and meaningful genes detected from DSS-general provide more potential candidates for further functional studies.

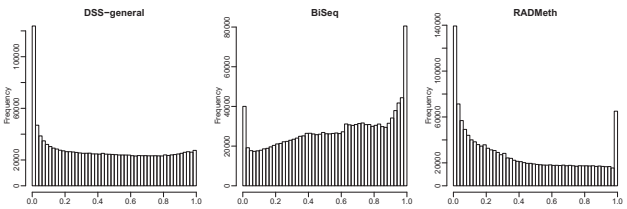

Fig. 4. *P*-values from human lung adenocarcinoma data. Histograms of *P*-values for testing the conditions effect (normal versus metastasis), from DSS-general, BiSeq and RADMeth

Furthermore, the ranking of CpG sites from different method are compared in order to evaluate the performance of top ranked DML. Top 2000 CpG sites are selected from different methods respectively, and Venn diagram is generated (Supplementary Fig. S4). There exist some discrepancies between BiSeq and other methods, which is consistent with the findings from p-value correlations. Agreement of rankings from DSS-general and RADMeth is much higher. We also select genes with at least five DML, based on the top 2000 CpG sites. A better method is likely to find more genes with multiple DML. DSS-general selects 20 genes, compared to BiSeq's 3 and RADMeth's 16, indicating the good performance of DSS-general even with the ranks of the CpG sites alone. A venn diagram for the overlaps of genes is shown in Supplementary Figure S5. We perform more in-depth comparison of the genes reported from above procedure by doing the same PubMed search as mentioned before. Twelve out of 20 genes (60%) reported from DSS-general are associated with at least one published paper from PubMed search with 5.8 average number of citations for these genes. For RADMeth genes, 8/16 (50%) has been reported, and the average citation number is 4.3. The three genes detected by BiSeq all have result from PubMed search, but the average citation is as low as 1.7. Overall, these results show not only DSS-general is more sensitive and providing more candidate genes, but also the genes are also more biologically meaningful compared to other methods.

As the major strength of general design is to identify more complicated effects, we next look at the cell type × condition interactions. The *P*-value distributions are shown in Supplementary Figure S6, and again the results from DSS-general have the most desirable shape. Again Bonferroni corrected p-values with threshold of 0.05 is used to identify DML. Interestingly we find that BiSeq detects 12 738 DML, while DSS-general finds 7165 and RADMeth none. The number of genes with at least five DML are 147 from DSS-general, 320 from BiSeq and 0 from RADMeth. To have a more fair comparison, we again look at the top ranked CpG sites from different methods. For those genes with at least five DML from top 2000 DML, DSS-general detects 41 genes, BiSeq 6, and RADMeth 38, again suggesting that DSS-general find more consistent DML. Overlaps of these genes are shown in Supplementary Figure S7.

Figure 5 shows the methylation profiles of KLF11 gene, one reported only by DSS-general from top 2000 DML. Hypomethylation pattern exists in metastatic HTB56 cells exon–intron junction region only. Additional methylation profiles of a few other candidate genes are shown in Supplementary Figures S8–S11 for genes PAX9, KLF2, FLJ12825 and RBPMS2. They exhibit different types of cell × condition interactions. For example, PAX9 have hypomethylation in HTB56 cell in metastatic state only, while KLF2 show hypomethylation in metastatic A549 cells, and FLJ12825 and RBPMS2 show hypermethylation in normal HTB56 cells. Lacking of proper tools for detecting DML of interaction effect, these interesting findings are not reported in the original paper (Hascher *et al.*, 2014). DSS-general can potentially fill this gap for more in-depth BS-seq data analyses to provide novel candidates for further functional analysis for cancer biologists.

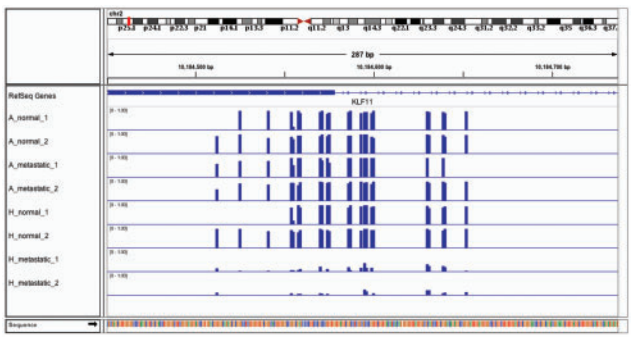

Fig. 5. Methylation profiles for KLF11, from human lung adenocarcinoma data. This demonstrates the cell type by condition interaction effect for the KLF11 gene

3.2.2 Mouse pronuclear DNA in the zygote

Data is obtained from GEO database under accession number GSE56650. The goal of this research is to understand the epigenetic dynamics during embryogenesis (Guo *et al.*, 2014). DNA methylation levels were profiled for male and female pronuclear DNA in the mouse zygote. There are 13 RRBS samples in total. For female, there are three wild type, two Tet3 knock-out, and two Aphidicolin treated pronuclei samples. Samples for male include two wild type, two Tet3 knock-out and two Aphidicolin treated. This is a 2 × 3 design with two genders (male and female) and three conditions (wild type, Tet3 knock-out and Aphidicolin treated). An approach similar to the human lung adenocarcinoma data is used. BiSeq has limitation as it is not able to test the difference of two levels when there are three or more in one factor. In this example, BiSeq can only test whether all three conditions have no difference, and it is not possible to compare wild type versus Tet3 knock-out. To compare all three methods, we only consider gender effect in this section for illustration purpose.

The CpG coverage from this dataset is very sparse. Only 987 855 sites are covered by at least one sample. To compare gender effect, 184 148 sites have common results from all three methods. Supplementary Figure S12 shows the *P*-value distributions. Similar to the previous findings, results from DSS-general is the most desirable. Numbers of DML called (with significance level of 0.05 after Bonferroni correction) are 21 215, 1456, 2059 respectively from DSS-general, BiSeq and RADMeth, indicating higher power from DSS-general. We again look at the top ranked 2000 CpG sites and look for genes with at least 5 DML. The numbers of genes detected are 12 from DSS-general, 4 from BiSeq and 7 from RADMeth respectively. Overlaps of the genes are shown in Supplementary Figure S13. Although it is difficult to evaluate these genes without further functional analyses, providing more candidate genes with the same number of DML suggests more consistent and sensitive results from DSS-general.

3.3 Software and computational performance

DSS-general is implemented as a part of DSS Bioconductor package. It has exceptional computational performance. For a simulation with 20 000 CpG sites and 12 samples, it takes about 5.6 s for one factor on a MacBook pro laptop with i7 2.7 GHz CPU and 16G RAM. In contrast, analysis of the same data takes 3.6 min from RADMeth and 15.5 min from BiSeq. Based on the fact that the computational times from all methods are approximately linear to the number of CpG sites, for a typical RRBS dataset with 5 million CpG sites, it is around 23 min for DSS-general, 15.2 h for RADMeth and 64.6 h for BiSeq, i.e. DSS-general is 40 and 166 times faster than RADMeth and BiSeq respectively.

In addition, from the implementation of software program point of view, both RADMeth and BiSeq perform model fitting and hypothesis testing in one function. If a user wants to test for other different factors, the time consuming model fitting procedure has to be

repeated. In contrast, DSS-general separates the model fitting and hypothesis testing steps. After the computationally intensive model fitting, different hypothesis tests can be conducted using the same parameter estimates, which makes DSS-general even more computationally efficient when multiple hypothesis tests are to be performed.

4 Discussion

Considering the large number of tests required, developing statistical method to analyze BS-seq data with multiple factors or covariates is challenging due to complicated model settings and computationally demanding parameter estimation procedure. Beta-binomial model is commonly used for count data from BS-seq experiment due to the possibility of considering both biological and sampling variations. Even though beta-binomial GLM is a straightforward solution, computationally inefficient and numerically unstable estimation procedures are major obstacles for BS-seq data analysis in practice. We developed a novel statistical method, termed 'DSS-general', to detect DML from BS-seq data under general experimental design. The method is based on transformation and approximation, which significantly improves computational efficiency. DSS-general characterizes the count data using a beta-binomial regression with an 'arcsine' link function. The regression coefficients are estimated from transformed methylation levels using GLS, where the covariance matrices are determined by sequencing depth and biological variances. The estimator of covariance matrix for the parameter estimations is carefully derived. A Wald test procedure is then applied for hypothesis testing to identify DML. Similar approaches that operate on transformed data have been developed for other sequencing data such as RNA-seq (Law *et al.*, 2014) and ChIP-seq (Chen *et al.*, 2015), and provide good performances.

An important distinction of DSS-general is the use of 'arcsine' link function instead of the more frequently used 'logit' or' probit'. This is mainly due to the statistical convenience because the stabilization of the variances after transformation makes the least square procedure possible. As shown from comparisons of computational performance with existing methods, DSS-general is 40 and 166 times more efficient than RADMeth and BiSeq respectively. Along with compatible or better accuracy and statistical property, DSS-general is a more desirable tool for the analyses of BS-seq data under general experimental design. Although RADMeth also provides similarly good results and computationally more efficient than BiSeq, unlike DSS-general and BiSeq, users cannot apply continuous covariates or factors with more than two values. Compared with 'logit' link, a disadvantage of the 'arcsine' link is that the coefficient is more difficult to interpret. However, since the main goal of DML detection is to detect and rank CpG sites and the detection/rankings are often based on quantities such as test statistics, *P*-values or FDR, the magnitudes of the coefficients are often not of interest. Thus the interpretability of regression coefficients will not be a major limitation of the method.

The algorithm of DSS-general contains a step for dispersion estimation. We evaluated the performance of dispersion estimation on DML detection accuracy. We compared the results from DSS-general when using the estimated dispersions versus using the true dispersions. Supplementary Figure S2 shows the TDR curves from the comparison. As expected, using true dispersions gives slightly better TDRs, particularly when sample sizes are small. However, the difference becomes very small when sample sizes increase to 10 replicates per group, indicating good performance of the proposed dispersion estimation procedure. It is also important to note that some dispersion estimation methods from other sequencing data are based on empirical Bayes (EB) procedure, which borrows information across all genes or CpG sites and achieves a 'shrinkage' effect (Feng *et al.*, 2014; Love *et al.*, 2014; Robinson and Smyth, 2007; Wu *et al.*, 2013). Under our setting, the EB-type of shrinkage method cannot be easily adopted because of the complicated relationship between the dispersion and data distribution, i.e. it is difficult to express the data likelihood as a function of the dispersion. Furthermore, the shrinkage is often based on maximizing an objective function such as penalized likelihood. The computational efficiency becomes a major issue without closed form of shrinkage estimator. In DSS-general, the bounded estimation of dispersions can be considered as a type of shrinkage that imposes some biological variance and stabilizes the results.

Although the results shown in this paper are based on CpG methylation, our method can also be applied to CpH methylation from whole-genome BS-seq. In addition, other types of base resolution methylation sequencing data such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Song *et al.*, 2013; Szulwach *et al.*, 2011a) can also be compared using DSS-general. Though there is no such dataset under general experimental design so far, they could become popular with the continuous advances of technologies.

DSS-general is specifically developed to test differential methylation at single CpG site level. It has been reported that the methylation levels are spatially correlated across the genome (Lister *et al.*, 2009), and smoothing-based methods were developed to account for the correlation (Hansen *et al.*, 2012; Park *et al.*, 2014). Although we considered spatial correlation when generating data in simulation study, we did not develop method to take into account this information to improve DML calling. This is partly because that all currently available data with general design are from RRBS, where the CpG coverage is sparse and spatial correlation is more difficult to be modeled. Nonetheless, incorporating the spatial information into our method will be a research topic in the near future. Furthermore, DSS-general is designed for data under linear model with fixed effects. The method cannot be directly applied to analyzing data from more complexed designs such as paired or clustered, which has to be modeled by a mixed effects model. Developing method for analyzing such data is also in our short term plan.

Acknowledgement

The authors thank Mr. Tianlei Xu for helping identify the RRBS datasets used in the real data analyses.

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.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, 289–300.

Benoukraf,T. *et al.* (2013) Gbsa: a comprehensive software for analysing whole genome bisulfite sequencing data. *Nucleic Acids Res.*, **41**, e55–e55.

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

Bird,A.P. and Wolffe,A.P. (1999) Methylation-induced repression: belts, braces, and chromatin. *Cell*, **99**, 451–454.

Boutros,P.C. *et al*. (2009) Prognostic gene signatures for non-small-cell lung cancer. *Proc. Natl. Acad. Sci.*, **106**, 2824–2828.

Chen,L. *et al*. (2015) A novel statistical method for quantitative comparison of multiple chip-seq datasets. *Bioinformatics*, btv094.

Davis,J. and Goadrich,M. (2006). The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd international conference on Machine learning*, ACM. pp. 233–240.

Dolzhenko,E. and Smith,A.D. (2014) Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, **15**, 215.

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–e69.

Guo,F. *et al*. (2014) Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell*, **15**, 447–458.

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

Hascher,A. *et al*. (2014) DNA methyltransferase inhibition reverses epigenetically embedded phenotypes in lung cancer preferentially affecting polycomb target genes. *Clinical Cancer Res.*, **20**, 814–826.

Hawes,S.E. *et al*. (2010) DNA hypermethylation of tumors from non-small cell lung cancer (NSCLC) patients is associated with gender and histologic type. *Lung Cancer*, **69**, 172–179.

Hebestreit,K. *et al*. (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, **29**, 1647–1653.

Hinde,J. and Demétrio,C.G. (1998) Overdispersion: models and estimation. *Comput. Stat. Data Anal.*, **27**, 151–170.

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

Kriaucionis,S. and Heintz,N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science*, **324**, 929–930.

Law,C.W. *et al*. (2014). Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, **15**, R29.

Lister,R. *et al*. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.*, **15**, 550.

Meissner,A. *et al*. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*, **33**, 5868–5877.

Meissner,A. *et al*. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.

Park,Y. *et al*. (2014) methylsig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, **30**, 2414–2422.

Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.

Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Robinson,M.D. *et al*. (2014) Statistical methods for detecting differentially methylated loci and regions. *Front. Genet.*, **5**, 324.

Saito,Y. *et al*. (2014) Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Res.*, **42**, e45.

Schumacher,A. *et al*. (2006) Microarray-based dna methylation profiling: technology and applications. *Nucleic Acids Res.*, **34**, 528–542.

Seisenberger,S. *et al*. (2012) The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell*, **48**, 849–862.

Song,C.X. *et al*. (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, **153**, 678–691.

Spinola,M. *et al*. (2007) Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the klf6 gene. *Cancer Letters*, **251**, 311–316.

Stockwell,P.A. *et al*. (2014) Dmap: differential methylation analysis package for rrbs and wgbs data. *Bioinformatics*, btu126.

Sun,D. *et al*. (2014) Moabs: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.

Szulwach,K.E. *et al*. (2011a) 5-hmc-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.*, **14**, 1607–1616.

Szulwach,K.E. *et al*. (2011b) Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet.*, **7**, e1002154

Teschendorff,A.E. *et al*. (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.*, **20**, 440–446.

Togashi,Y. *et al*. (2012) Klc1-alk: a novel fusion in lung cancer identified using a formalin-fixed paraffin-embedded tissue only. *PLoS One*, **7**, e31323.

Weber,M. *et al*. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.

Wu,H. *et al*. (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232–243.

Xie,P. *et al*. (2011) Smurf1 ubiquitin ligase targets kruppel-like factor klf2 for ubiquitination and degradation in human lung cancer h1299 cells. *Biochem. Biophys. Res. Commun.*, **407**, 254–259.

Yu,G. (2009) Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Stat. Probab. Lett.*, **79**, 1621–1629.