Distribution Agreement

The text below should be reproduced exactly as written, on the Distribution Agreement page. Sign the page on the signature line, and type your name under the signature line. Write the date on the date line.

Distribution Agreement

Signature:

______________________________	4/24/2013______
Hao Feng	Date

Approval Sheet

The approval sheet should be designed according to the plan below. See next page for several comments.

A Hierarchical Bayesian Approach to

Detect Differentially Methylated Loci from

Bisulfite Sequencing Data

By

Hao Feng

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

________Hao Wu, Ph.D & Karen Conneely, Ph.D________

Thesis Advisors

____________Yijuan Hu, Ph.D______________

Reader

Abstract Cover Page

The abstract cover page should be designed according to the plan below. See next page for several comments.

A Hierarchical Bayesian Approach to

Detect Differentially Methylated Loci from

Bisulfite Sequencing Data

By

Hao Feng

Bachelor of Science

University of Science and Technology of China

2011

Thesis Committee Chair: Hao Wu, Ph.D & Karen Conneely, Ph.D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2013

Abstract

The abstract should be designed according to the plan below.

Abstract

A Hierarchical Bayesian Approach to
Detect Differentially Methylated Loci from
Bisulfite Sequencing Data
By Hao Feng

DNA methylation is a central epigenetic modification that has essential roles in cellular processes including genome regulation, development and disease. In studies of DNA methylation, one key task is to identify methylation differences under distinct biological contexts. Recently, as bisulfite sequencing technology (BS-seq) has made it possible to detect methylation in CG loci level, more and more datasets are becoming available to study DNA methylation. A common drawback of datasets in these studies; however, is that the number of sample replicates is usually limited. This can lead to unstable estimation of within group variance, and may subsequently yield unsatisfactory results in differentially methylated loci (DML) detection. Here we propose a new method to apply shrinkage to the variance estimation in an empirical Bayes model. We show that the variance shrinkage in these data can be done by shrinking a dispersion parameter. Simulation results demonstrate the favorable performance of the new methods.

Cover Page
The cover page should be designed according to the plan below. It is almost the same as the abstract cover page. Please see the comments to that page (on page 12).

A Hierarchical Bayesian Approach to

Detect Differentially Methylated Loci from

Bisulfite Sequencing Data

By

Hao Feng

Bachelor of Science

University of Science and Technology of China

2011

Thesis Committee Chair: Hao Wu, Ph.D & Karen Conneely, Ph.D

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2013

Acknowledgments

I would like to thank Dr. Karen Conneely, my advisor of both practicum and thesis, for her help in leading me to this research area and making it possible for me to work on such an attractive and challenging project. I could not make it without her steady support and insightful suggestions.

I would also like to thank Dr. Hao Wu, my indefatigable thesis advisor, for his constant accessibility and assistance. His expert guidance cast light on my research adventure.

Also a special thanks to Dr. Yijuan Hu for taking time to read my thesis.

I am especially grateful for the support from my family and friends during my study at Emory.

Introduction

DNA methylation is an epigenetic modification that plays an important role in normal development and gene regulation[1-3]. It is a biochemical process that mainly occurs on CpG dinucleotides, with very rare cases that happen in CHG and CHH (H= A, T or C)[4]. Typically, by adding a methyl group to the 5-position of a cytosine pyrimidine ring (C), the chemical structure is altered. It has been proposed that the methyl group could physically protect the associated target gene and prevent the transcription process from starting [5, 6]. Consequently, methylation of a cytosine within a gene or its promoter region may affect gene expression.

To understand the biological consequences of DNA methylation status, methylation has traditionally been studied at the gene level. However, methods for assessing whole-genome methylation have improved substantially in the past few years in terms of accuracy, genome coverage, resolution and reduced cost. Current sequencing-based methods for methylation analysis can be classified into two categories: bisulfite conversion–based[7] and enrichment-based methods[8]. The latter category includes MeDIP-seq[9], MBD-seq[10, 11] and methylCap-seq[12], which use different methyl-binding proteins or domains for methyl capture at specific level concentrations of genomic DNA. For example, MeDIP-seq involves antibodies directed against mC or mCG to precipitate methylated DNA fragments, followed by the application of next-generation sequencing to the fragments [8, 9]. On the other hand, bisulfite conversion–based methods include whole-genome bisulfite sequencing (BS-seq or MethylC-seq) [4, 13-19] and reduced representation bisulfite sequencing (RRBS)[20, 21]. To be specific, in the RRBS method, after genomic DNA digestion and size-selection, PCR products are cloned and sequenced. Sequences generated from RRBS libraries are then projected onto genome [20, 22]. Although RRBS provides limited genome coverage (5–10%) and both RRBS and whole-genome bisulfite sequencing can be expensive for large samples, they have become popular because they allow detection at a single-base resolution. Specifically, after treating DNA with sodium bisulfite, unmethylated cytosine is deaminated and changed to uracil, which will be amplified as thymine; while 5-methylcytosine residues are protected by the methyl group and remain unchanged. For a specific CpG site in the genome, using replicates in BS-seq after PCR, we can easily calculate the counts for this specific CpG that is either a thymine or a cytosine. The count of thymine represents the number of sequenced DNA strands that are unmethylated (U) and the count of cytosine represents the number of DNA strands that are methylated (M) at this CpG site. By taking the ratio of methylated number (M) to the total reading number (M+U), the methylation level of a CpG site is calculated as M/(M+U), which is a proportion number that lies strictly between 0 and 1. By this process, genome-wide DNA methylation measurement is achieved at single-base resolution [13].

As quick expansion and rapid increase of data generated by BS-seq techniques becomes available, BS-seq data can be applied to a variety of analyses. Since methylation can influence gene expression, and differences in gene expression can lead to differences in functionality among cells in each kind of tissue, it is possible that reported cell-specific methylation patterns [23] may underlie differences in cell function. Consistent with this, DNA methylation displays distinct signatures for different cell types [24]. For example, in the same region on the genome, malignant tissues could present a hypo-methylated pattern while normal tissues present a hyper-methylated pattern [24-27]. As differential methylation patterns are frequently reported in disease states, one of the key tasks in DNA methylation analyses is to identify and elucidate the role of Differentially Methylated Loci (DML) or Regions (DMR).

Several existing methods including Fisher's exact tests and t-tests have been applied to detect DML/DMR in comparisons between two groups. If there is only one sample per group, Fisher's exact test can be applied to each single CpG site to identify DML [4, 28, 29]. The Fisher's exact test here compares the fraction of methylated Cs in sample 1 and sample 2 in the absence of replicates. When replicates exist, Fisher's exact test would not be appropriate due to the nature of the 2-by-2 contingency table test. When there is more than one sample per group, derivations of a t-test could be applied to compare the proportions in each sample [28, 30]. After computing test statistics, the selected cutoff threshold for identifying DML can be determined using the empirical distribution of test statistics as in [30]. Regions that contain DML can then be claimed as DMR.

The problem and challenge of DML/DMR detection is that it typically lacks enough replicates, which could make the estimated values and test results unstable. For example, in one dataset of DNA methylation of early mammalian embryo, only 2-5 replicates were included [31]. When performing a two-sample t-test in each of the CpG site, low number of replicates in each sample would violate the assumption of sufficiently large number of replicates in central limit theorem, making the asymptotic distribution do not hold anymore. Moreover, we also need to correctly account for the variation in our measurements of methylation. When measuring methylation levels, there are two sources of variation: technical variation, which reflects the measurement error resulting from the technology, and biological variation, which reflects the heterogeneity among samples we get from the same group or population [23, 32, 33]. Since the methylation process can be interpreted as a stochastic process, replicates in the same group generally will not have exactly the same methylation level. Existing methods can capture the overall variation, but we want to distinguish biological variation and technical variation in our result, because our goal is to identify CpG sites and regions that exhibit consistent differences even when taking biological variation into account.

To solve this problem, it is natural to think about borrowing information from other CpG sites to achieve better estimation of variation and subsequently improve our test results. Previous studies in microarray and RNA sequencing showed better estimation and significant improvement after using shrinkage estimators [32, 34-36]. These methods proposed to use the negative binomial (NB) model, which is a gamma-Poisson mixture, to fit the gene expression data and correctly capture biological variance. Here, we propose a beta-binomial hierarchical Bayesian model to capture the biological variation and apply shrinkage in estimating it. This model can be interpreted as follows: the beta distribution models the unobserved true methylation levels in each CpG site across replicates in each group; and conditioning on the true methylation level as the probability, the counts of the methylated cytosine follow a binomial distribution. Here, the biological variation is captured by the beta distribution and the technical variation is captured by the binomial distribution. The biological variation is captured in the beta model by mean μ and dispersion $\emptyset$:

$$var = \mu(1-\mu)\emptyset \tag{1.1}$$

Note that dispersion $\emptyset$ represents the variation of a CpG site's methylation level relative to its mean. Each CpG site within a single condition (e.g. within cases, or within controls) will have its own dispersion under our model. After verifying with real data, a prior is established that $\emptyset$ follows a log-normal distribution across the whole genome. Then estimating and shrinking can be done for each $\emptyset$ using a Bayesian approach.

After correctly accounting for dispersion and improving our estimator of within-group biological variance, a simple two-condition comparison can be framed as a Wald test or likelihood ratio test. Using simulations based on real data, we demonstrate that the proposed method yields improved DML detection in the top discovery rate (TDR). The rest of the paper is organized as follows. In the Methods Section, we present the hierarchical model, estimation, and testing procedure. In the Results Section, we present the results from our simulations comparing DML detection using our proposed method versus alternative ones. In the Discussion Section, we discuss the interpretation of the dispersion, the connection with other studies, and future directions.

Methods

Here, we denote that at the *i*th CpG site, *j*th group and *k*th replicate, X_{ijk} is the number of reads that show methylation, N_{ijk} is the total number of reads that cover this position, and $p_{ijk}$is the underlying methylation level. At the *i*th CpG site, the overall methylation level is denoted by μ_{ij} and the dispersion is $\emptyset_i$. Then the following beta-binomial hierarchical model is proposed:

$$X_{ijk}|p_{ijk}, N_{ijk} \sim Binomial\ (N_{ijk}, p_{ijk})$$

$$p_{ijk} \sim Beta\left(\mu_{ij}, \emptyset_{ij}\right)$$

$$prior: \emptyset_{ij} \sim log - normal\ (\ m_{0j}, r_{oj}^2)$$

$m_{0j}\ and\ r_{oj}^2$ are parameters from a common prior which can be estimated from the data. In a two group comparison setting, *j* is equal to 1 or 2. A method of moment estimator (MME) is used to estimate the prior parameters.

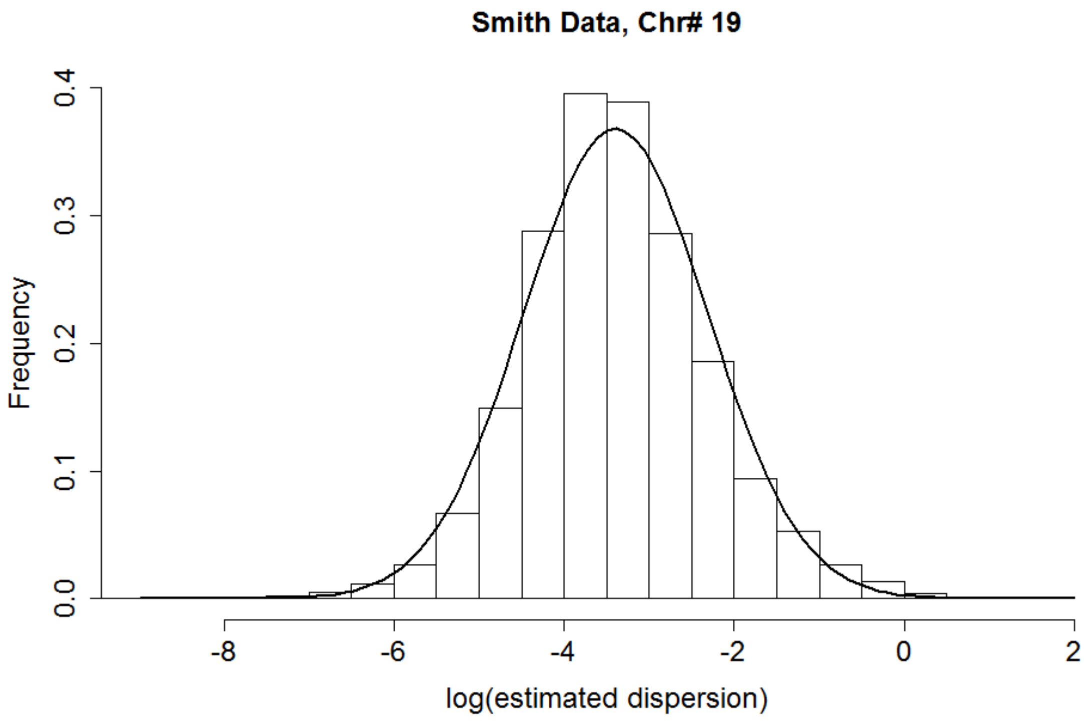

Fig. 1. Histogram of the logarithm of estimated CpG-specific dispersion ($\emptyset_{ij}$, estimated by MME) from mouse embryogenesis data [31] for one chromosome. The solid line is the density for normal distribution with parameters estimated from log($\emptyset_{ij}$). $\emptyset_{ij}$ can be approximately modeled as a log-normal distribution.

Using data on DNA methylation from mouse embryogenesis [31], the distribution of the logarithm of estimated dispersion, i.e. log($\emptyset$), is approximately Gaussian as shown in Figure 1. From the histogram we can see that the dispersion can be approximated as a log-normal distribution. This Gaussian-like data structure in Figure 1 confirms that it is appropriate to use the log-normal assumption as we proposed.

p_{ijk} represents the underlying true methylation proportion which is restricted in [0,1]. It can vary across replicates even for the same CpG site (same i) and same group (same j). However, within the same CpG site *i* and same group *j*, our model assumes that p_{ijk} comes from the same beta distribution for all replicates. The beta distribution can be re-parameterized as a function of the mean u_{ij} and dispersion parameter $\emptyset_{ij}$. Compared with the traditional form of the beta (α, β) distribution, the parameters have the following relationship:

$$\mu = \frac{\alpha}{\alpha + \beta}$$

$$\emptyset = \frac{1}{\alpha + \beta + 1}$$

Thus,

$$var(\mathrm{p}) = \left(\frac{\alpha}{\alpha+\beta}\right)\left(1-\frac{\alpha}{\alpha+\beta}\right)\left(\frac{1}{\alpha+\beta+1}\right) = \mu(1-\mu)\emptyset$$

After re-parametrization, μ captures the underlying true methylation level and $\emptyset$ captures the dispersion or biological variation in our model.

Also, if we combine the two pieces of the hierarchical model together, we can see that the biological variation is captured by the beta distribution and the remaining variation is captured by the binomial distribution.

Previous studies have shown that the methylation levels for consecutive CpG sites are autocorrelated across the genome [37, 38], suggesting that μ_{ij} may vary smoothly along the genome [30]. Therefore, the package *BSmooth* [30] can be implemented in estimating μ_{ij}. In this situation, μ_{ij} is assumed to be a smooth function with respect to genomic location l_i :

$$\mu_{ij} = f_j(l_i)$$

$f_j(l_i)$ can be estimated with a local-likelihood smoother [39]. To do this, the data from all replicates is collapsed and combined to improve the coverage depth and estimation. The estimated methylation level $\hat{\mu}_{ij}$ for the *i*th CpG site is then obtained as estimated profile $\hat{f}_j(l_i)$. Notably, since smoothing is highly dependent on the assumption of strong autocorrelation of methylation levels across the genome, it is possible that in situations where this assumption is violated, smoothing could yield imprecise estimation. In this case, a more powerful strategy may be to simply use the maximum likelihood estimator (MLE) of the beta distribution to get the point estimate of μ_{ij} in each *i*th CpG site and *j*th group.

Our next goal is to estimate $\emptyset_{ij}$in our model and apply an empirical Bayes method to our estimation. When putting the prior and distribution together as follows:

$$\mathrm{p}\left(\emptyset_{ij} \middle| x_{ijk}, N_{ijk}, \mu_{ij}\right) \propto f\left(\emptyset_{ij}\right) \prod_{k} P\left(x_{ijk} \,\middle| N_{ijk}, \mu_{ij}, \emptyset_{ij}\right)$$

The posterior point estimate of $\emptyset$ then satisfies:

$$\log(\mathrm{p}\left(\emptyset_{ij} \middle| x_{ijk}, N_{ijk}, \mu_{ij}\right)) \propto$$

$$\sum_k \varphi(x_{ijk} + (\emptyset_{ij}^{-1} - 1)\mu_{ij}) + \sum_k \varphi(N_{ijk} - x_{ijk} + (\emptyset_{ij}^{-1} - 1)(1 - \mu_{ij}))$$

$$- \sum_k \varphi(N_{ijk} + (\emptyset_{ij}^{-1} - 1)) -\mathrm{n}\ \varphi((\emptyset_{ij}^{-1} - 1)\mu_{ij}) -\mathrm{n}\ \varphi(\left(\emptyset_{ij}^{-1} - 1\right)(1 - \mu_{ij})) +$$

$$n\ \varphi\left(\emptyset_{ij}^{-1} - 1\right) - \log\left(\emptyset_{ij}\right) - \log\left(r_{0j}\right) - \frac{(\log(\emptyset_{ij}) - m_{0j})^2}{2{r_{0j}}^2}$$

In practice, we can get the point estimate of $\emptyset$ by maximizing the above equation using the Newton-Raphson method after plugging in the hyper-parameters (prior) m_{0j} and r_{oj}^2 . We also plugged in the estimated μ_{ij} with or without smoothing as described. We maximized the likelihood using the “optimize” function in R (2.15.2). Because we use a prior with estimated $m_{0j}\ and\ r_{oj}^2$, the estimated $\emptyset_{ij}$ is therefore an empirical Bayes estimate shrunken toward the common prior. Also notable is that the last line of the above equation includes the penalty function $-\log(\emptyset_{ij}) - \log(r_{0j}) - \frac{(\log(\emptyset_{ij}) - m_{0j})^2}{2{r_{0j}}^2}$, which will penalize large extreme $\emptyset_{ij}$ in our estimation.

After finishing the parameters estimation described above for each *j*th group, hypothesis tests can be done at each *i*th CG loci to compare methylation at group *j*=1 and group *j*=2. In order to detect DML in a Wald test framework, it is necessary to correctly account for the variance of each CpG site of each group. Here our point estimate for μ in the *i*th CpG site and *j*th group is $\hat{\mu}_{ij} = \frac{\sum_k X_{ijk}}{\sum_k N_{ijk}}$ by combining data from all replicates together. Since we have assumed the following distribution:

$$X_{ijk} \sim Beta - binom\ (\ N_{ijk}, \mu_{ij}, \emptyset_{ij}\ )$$

The variance of X_{ijk} should be:

$$var(X_{ijk}) = N_{ijk}\mu_{ij}(1-\mu_{ij})[1 + (N_{ijk}-1)\emptyset_{ij}]$$

Then we can get:

$$var(\hat{\mu}_{ij}) = var\left(\frac{\sum_k X_{ijk}}{\sum_k N_{ijk}}\right)$$

$$= (\frac{1}{\sum_k N_{ijk}})^2 \sum_k \{N_{ijk}\mu_{ij}(1-\mu_{ij})[1 + (N_{ijk}-1)\emptyset_{ij}]\} \qquad (\mathbf{1.2})$$

After estimating the dispersion parameter $\emptyset_{ij}$ and methylation level μ_{ij} for each CpG site *i* and group j, we can calculate the estimated variance in each CpG site *i* of each group *j* by plugging in estimated values to the equation (**1.2**). For two-group comparisons, we can estimate variances for each group by plugging in the corresponding estimates of the dispersion parameter $\emptyset_{ij}$ and methylation level μ_{ij} for each group. If we have methylation data from both cancer patients and control subjects, then we can denote the cancer patient group as group # 1 and the control group as group # 2. For a specific CpG site, the estimated variance of methylation for the cancer group 1 is $\widehat{var}_{i1}$ and the estimated variance for the control group is $\widehat{var}_{i2}$.

Hypothesis tests for the comparison of two groups can then be performed in the Wald test framework. To be specific, the Wald test for the two-group comparison of the *i*th CpG site is constructed as:

$$t_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\widehat{var}_{i1} + \widehat{var}_{i2}}}$$

where $\widehat{var}_{ij}$ $(j = 1,2)$ is the estimated variance for group 1 or 2 as described above.

Likewise, a likelihood ratio test (LRT) can also be constructed using the estimated parameters. If the probability mass function (PMF) of a beta-binomial distribution is $g(.)$, then the LRT at the *i*th CpG site is:

$$LRT_i = -2(\sum_j \sum_k \log(g_0(X_{ijk}, N_{ijk})) - \sum_j \sum_k \log(g_j(X_{ijk}, N_{ijk})))$$

Here g_0 is the PMF of the beta-binomial distribution under the null hypothesis and g_1 and g_2 are the PMF under the alternative hypothesis for group 1 and group 2. g_0, g_1 and g_2 are obtained by plugging in the corresponding estimated $\hat{\mu}$ and $\hat{\emptyset}$ to the PMF of beta-binomial distribution under the null and alternative models. The test statistics of the likelihood ratio test follow a χ^2 distribution with two degrees of freedom.

Results

Simulation

We used simulation data to test our proposed method and compare the results with existing methods. All simulations are based on the original distribution of RBBS data from a study on mouse embryogenesis [31]. For each simulation data, a number of 20,000 CpG sites were simulated. The true methylation levels are drawn from a pool of hundreds of pieces of smoothed methylated curves, generated by applying *BSmooth* [30] to the original data. The number of replicates per CpG site is usually 2-5 as is typical due to the expense of collecting BS-seq data. The underlying percentage of true differentially methylated CGs among all CGs is set to be within 5%-10%, as it best represents the proportion of DML across different stages of development. In some cases, each of the dispersion parameters $\emptyset_{ij}$ is drawn from a log-normal distribution with parameters based on published data [31]. In other cases, each of the dispersion parameters $\emptyset_{ij}$ is drawn from a Gamma distribution for comparison.

By applying shrinkage to the data, we obtained got the shrunken estimated value of dispersions. The scatterplots of true dispersions versus the estimated dispersions are shown in **Figure 2**. From the plot we see some over-shrinkage due to small number of coverage, where the prior will dominate in that case. **Figure 2** also shows a boxplot comparison of the biases of the shrinkage method vs. naïve method. Our proposed method successfully avoids extreme values and thus obtains improved precision.

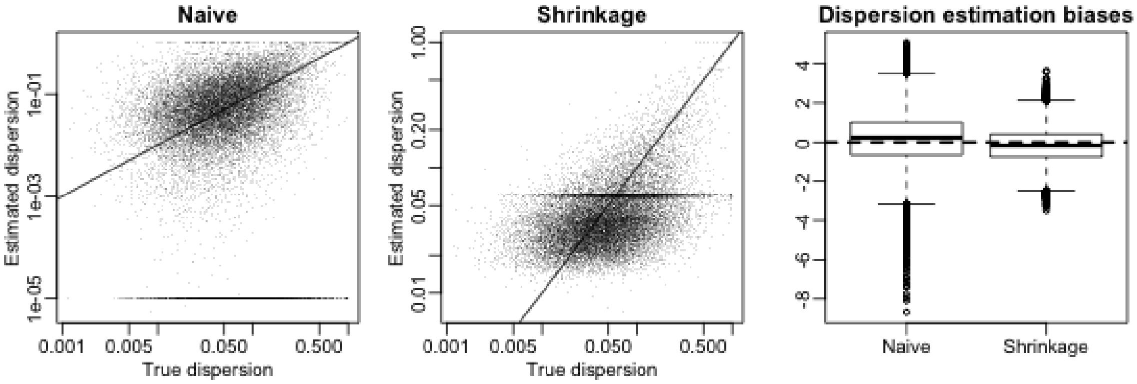

Fig. 2. Scatterplot of true dispersion vs. estimated shrunken dispersion from a naïve method (left) and shrinkage method (middle). Boxplot of the bias for multiple CG sites (right). Our method showed improved precision by avoiding extreme values. Some over-shrinkage estimations are observed because of the small coverage. All plots are for 20,000 CpG sites.

When comparing the MSE of point estimation of dispersion $\emptyset_{ij}$ our method has consistently lower MSE than the naïve method of moment estimator (MME) (**Figure 3**).

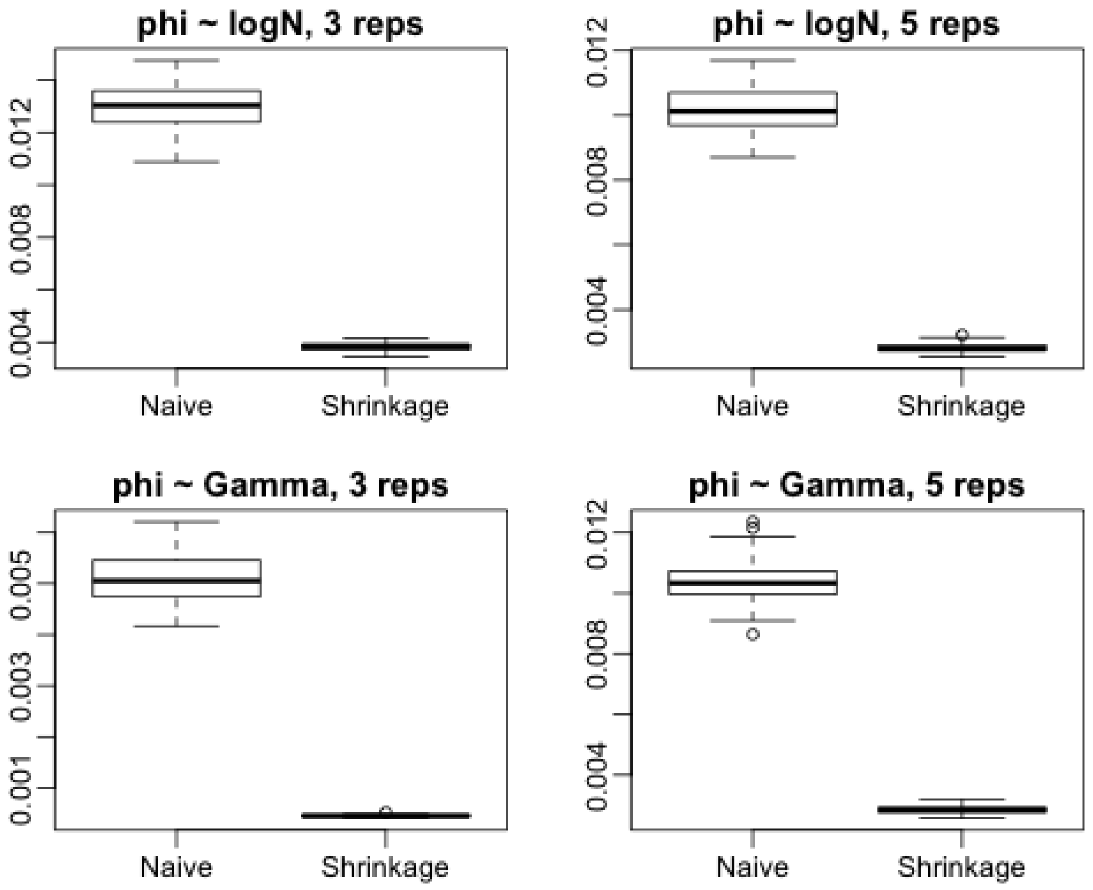

Fig. 3. Boxplots comparing the distribution of MSE for dispersion $\emptyset_i$ estimators from our proposed method and naïve method of moment estimators (MME) over 30 simulations. Dispersion $\emptyset_i$ is randomly generated from the log-normal distribution (top row) and Gamma distribution (bottom row). Each group contains either 3 replicates (left column) or 5 replicates (right column). All simulations are for 20,000 CpG sites.

After estimating all parameters, we applied our proposed Wald test method with both shrinkage dispersion and naïve dispersion to identify DML between groups 1 and 2 in the simulated data. For comparison, we implemented classic t-tests and Fisher's exact tests, as they are also very popular existing methods. We also implemented a newly developed adjusted Chi-square based method[40] to compare the performance.

Since DML detection is often used as a hypothesis generating tool, and the goal is to have as many true positives as possible in the top-ranked CpG sites, we compute the proportion of true DML (i.e. 1 − false discovery proportion) in the top-ranked CpGs up to the top 1000. In these 5 methods, top CpG sites are ranked either based on the increasing order of p-values. In the true discovery rate (TDR) plot, our method shows a high proportion of true positives among the top ranked CpG sites [**Figure 4**]. The proposed Wald test outperformed the classic t-test and Fisher's exact test as the Wald tests consistently have higher true positive rates. We tried our method using different prior distribution for the dispersion parameter $\emptyset$, including the log-normal distribution, Gamma distribution and empirical distribution from real data. The TDR plot showed consistent better performance for the Wald test with shrunk variance compared to other approaches. Under different prior distributions for the dispersion, our method showed consistently better performance, which demonstrates the robustness of our method. In cases where the replicate numbers are small, the improvement of Wald test over other approaches in greatest; this is partly because the variance is difficult to estimate based on 2 replicates in each group using the naïve method. When the replicate number is large, the naïve dispersion estimates are closer to those of the shrinkage method. In this case the Wald test with a naïve dispersion estimator is also a good choice since it is computationally less intensive.

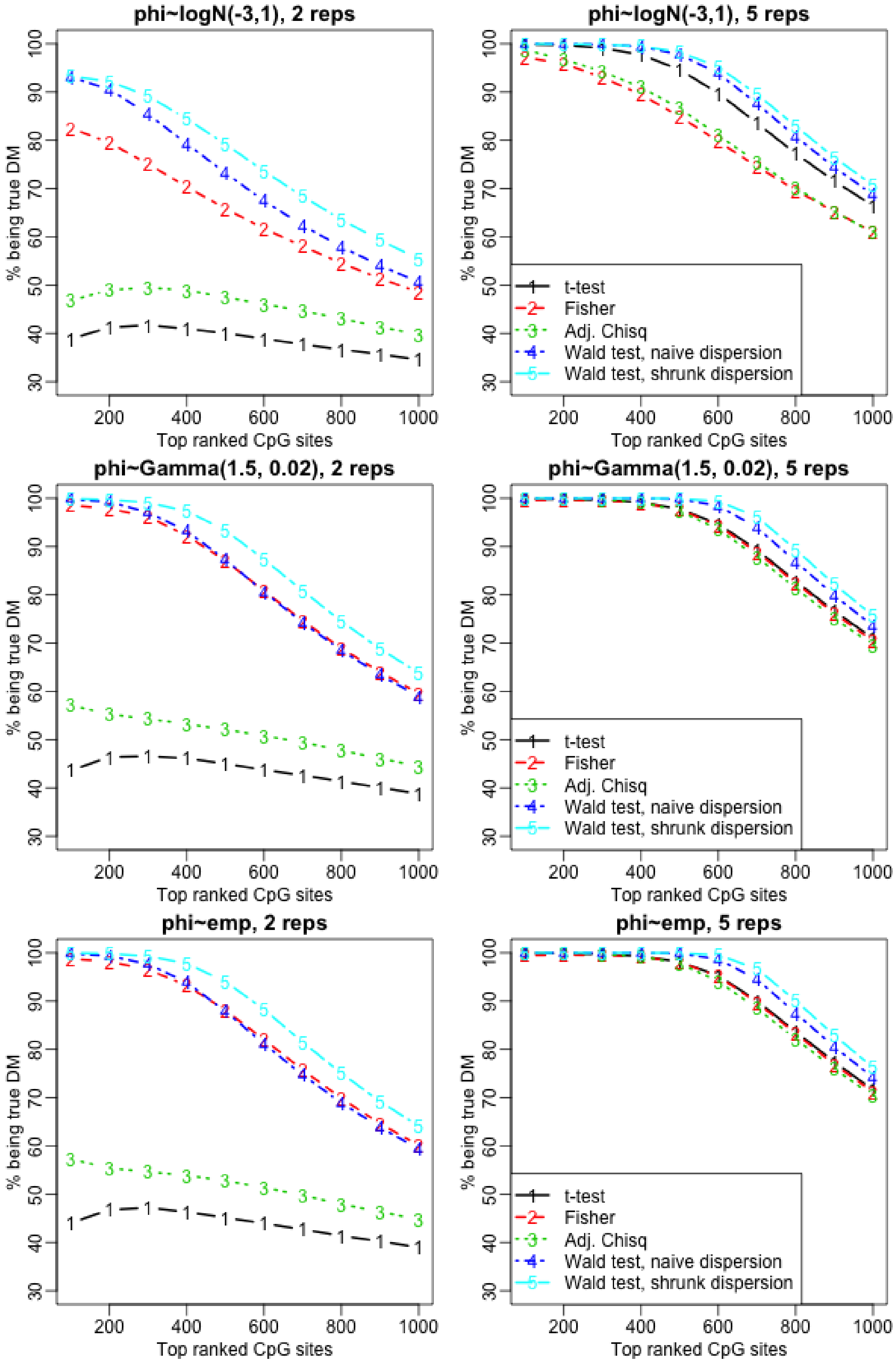
phi~logN(-3,1), 2 reps
phi~logN(-3,1), 5 reps
phi~Gamma(1.5, 0.02), 2 reps
phi~Gamma(1.5, 0.02), 5 reps
phi~emp, 2 reps
phi~emp, 5 reps
% being true DM
Top ranked CpG sites
30
40
50
60
70
80
90
100
200
400
600
800
1000
t-test
Fisher
Adj. Chisq
Wald test, naive dispersion
Wald test, shrunk dispersion

Fig. 4. The True Discovery Rate (TDR) plot of DML detection accuracy from our proposed method for top 1000 CpG sites. CpG sites are ranked by p-values from low to high. The proportion of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. The dispersion parameter $\emptyset$ comes from the log-normal distribution (top row), Gamma distribution (middle row), or empirical distribution from real data (bottom row). In situations where replicate numbers are low (left column), our method improved the most since it solved the difficulty of estimating variance from 2 replicates per group. Our proposed Wald test method showed consistent improvement.

Since the Wald test and likelihood ratio test (LRT) are asymptotically equivalent, we also implemented our estimated shrunk dispersion parameter into the likelihood ratio test. **Figure 5** shows that the performance of these two tests is comparable. It is thus justifiable to use either the Wald test or the LRT for a two group comparison. To explore more details, the test statistics and quantile-quantile plot (QQ plot) of Wald test are shown in **Figure 6**. The Wald test statistics showed good normality; and QQ plot showed good combination of the existence of null and alternative hypothesis.

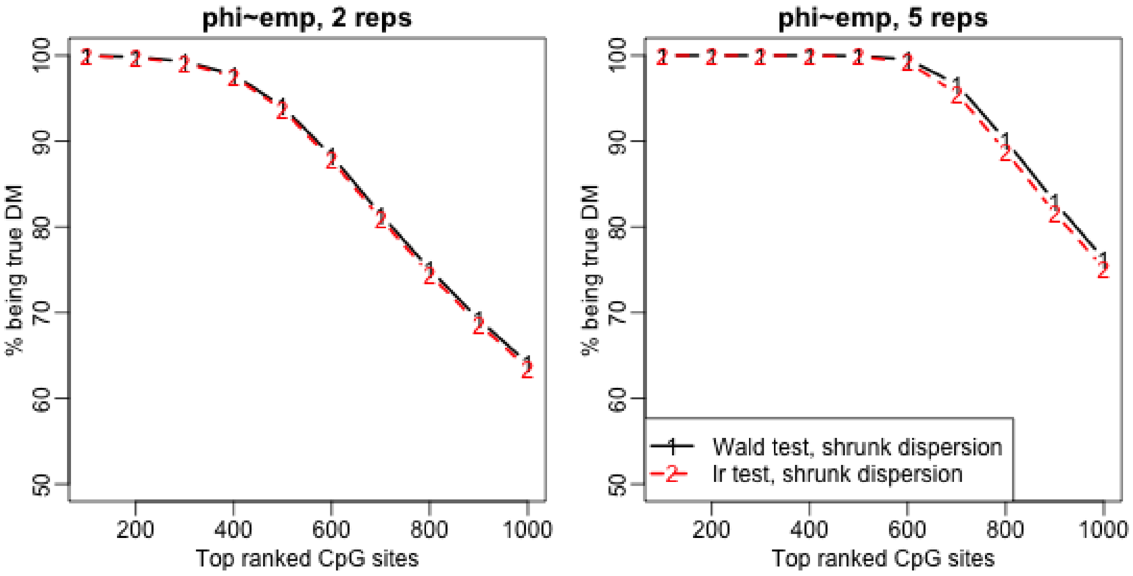

Fig. 5. True Discovery Rate (TDR) plot of DML detection accuracy using Wald test and likelihood ratio test (LRT) for top 1000 detected CG loci. Detected CG loci are ranked by p-values from low to high. The proportion of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. In cases where that the replicates number is small (left) vs. large (right), the two methods still have comparable similar performance.

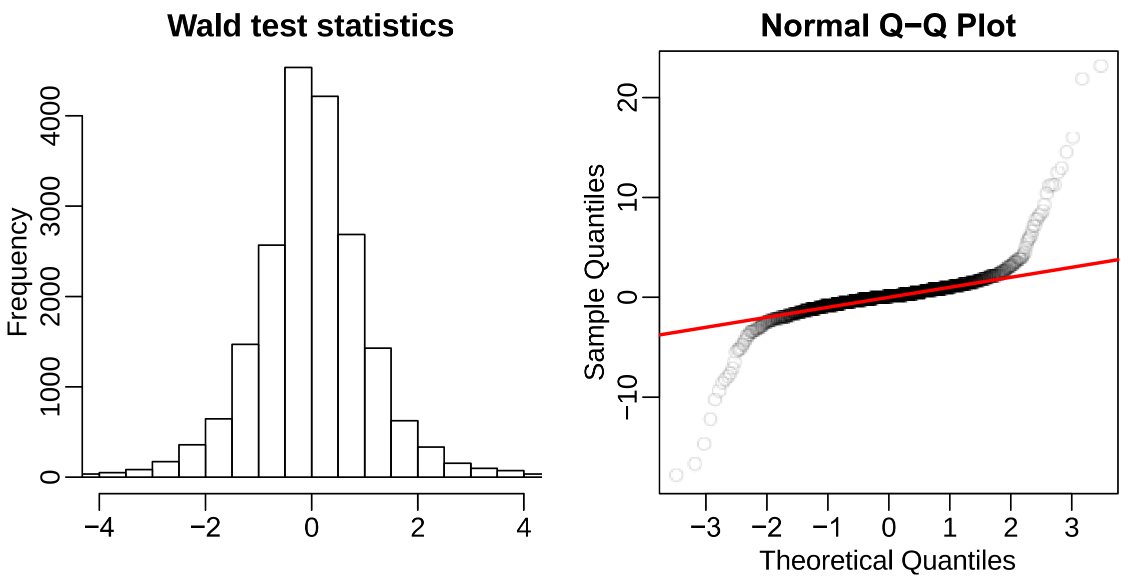

Fig. 6. The histogram (left) of test statistics from Wald test and the quantile-quantile (QQ) plot (right) of Wald test statistics.

Although it is justifiable to use either the Wald test or the likelihood ratio test, the Wald test can be applied to more situations where the researcher want to choose their own alternative hypothesis. To be specific, according the nature of the variance in the beta-binomial distribution, when the methylation levels are close to the boundaries of 0 or 1, the estimated variances are smaller than those near the middle. Consequently, after we plug in the variance to the Wald test, it would lead to a more anti-conservative p-value when the underlying methylation levels are close the boundaries [**Figure 7**]. For example, for a methylation level change of 0.04 at a fixed dispersion, the methylation level change from 0.95 to 0.99 is much more significant than the change from 0.48 to 0.52. For this scenario we propose an alternative approach to handle the issue similar to *MATS* in RNA-seq data [41]. We test the hypothesis that the difference in the methylation level for a given CpG site between group 1 and group 2 is above a user-defined cutoff c, i.e. $|\mu_1 - \mu_2| > c$ [**Figure 7**]. The cutoff c is a user-defined parameter that represents the extent of methylation change that the researcher wants to identify. By implementing this method, we can avoid the discrepancy of p-values of same amount methylation change under different methylation level.

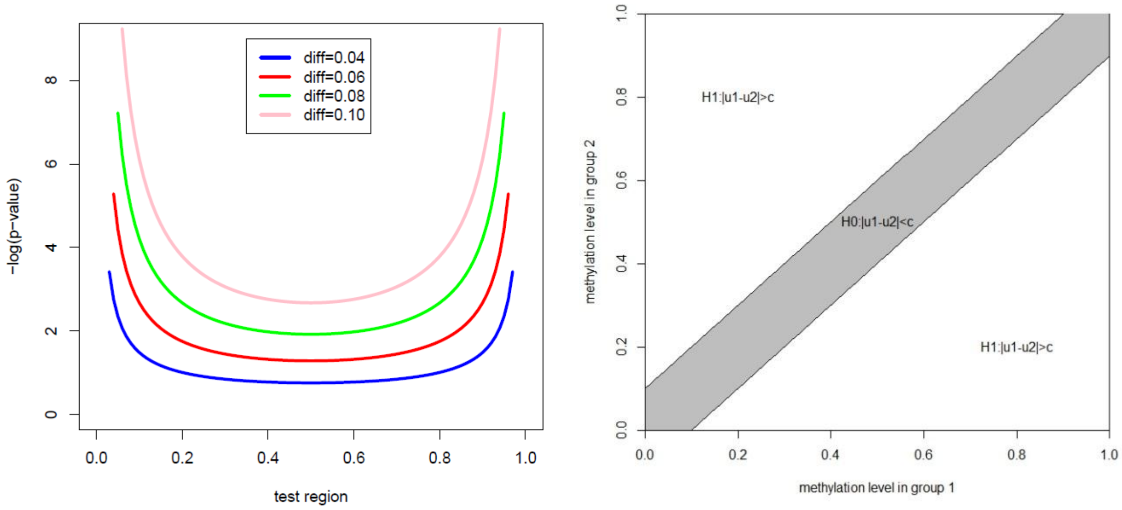

Fig. 7. (left) For a same level of difference (4 possible showed here), within the same level of change in methylation level, test results tend to be more significant as the p-values get smaller near the boundary. (right) Our proposed null and alternative hypotheses. The H_0 null hypothesis is that the difference in methylation levels between two groups is below the user-defined cutoff c (the gray area). The H_1 alternative hypothesis is that the difference is above the user-defined cutoff c (the white area).

For example, if a researcher is interested in identifying DML with at least 0.1 changes in methylation level, the cutoff c should be set to 0.1. If we calculate the probability of

$|\mu_1 - \mu_2| > c$ from the BS-seq counts, i.e. $P(|\mu_{1i} - \mu_{2i}| > c|data)$, we can rank CpG sites based on the probability for each CpG site *i* to improve DML detection. Simulation results show our method has higher or equal TDR in the top ranked CpG sites [**Figure 8**].

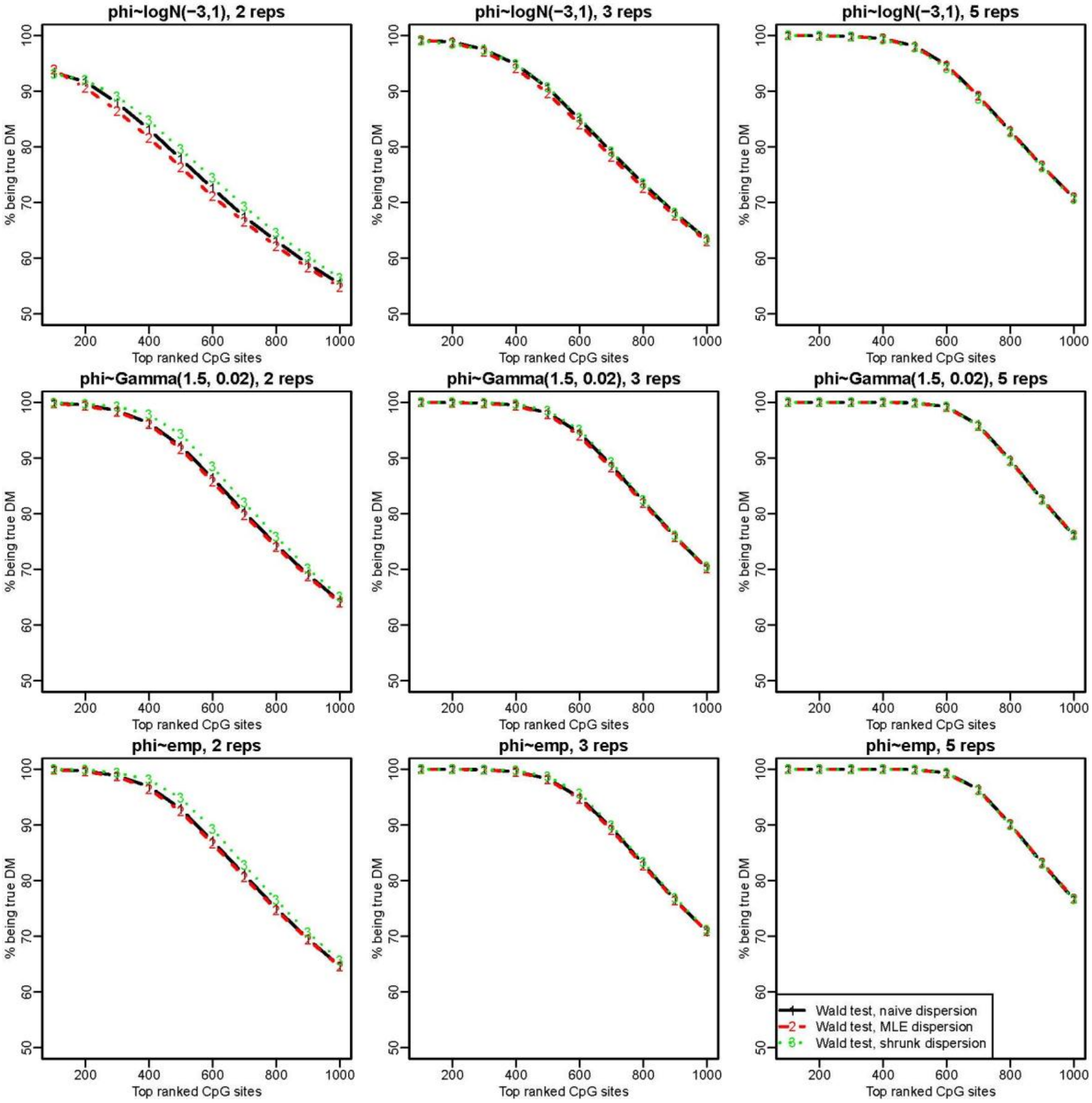

Fig. 8. The True Discovery Rate (TDR) plot of DML detection accuracy from our proposed method for top 1000 CpG sites using user-defined methylation level change cutoff at 0.1. CpG sites are ranked by p-values from low to high. The proportion of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. The dispersion parameter $\emptyset$ comes from the log-normal distribution (top row), Gamma distribution (middle row), or empirical distribution from real data (bottom row). In situations where replicate numbers are low (left column), our method improved the most since it solved the difficulty of estimating variance from 2 replicates per group. By using the Wald test we can easily convert test into testing $P(|\mu_{1i} - \mu_{2i}| > c|data)$ form, which avoid discrepancy p-values for a same amount of

methylation change under different methylation level. Here c is the user-defined cutoff that user desire to detect.

Discussion

In this study, we presented a method of estimating the dispersion parameter in a beta-binomial model for DNA methylation data. We showed that the new model could capture the biological variance in a CpG-specific manner and, as a result, lead to better detection of DML compared to existing methods. Therefore, we showed that a good estimation of dispersion $\emptyset_{ij}$ is important to DML detection when a small number of biological replicates are used. Accounting for biological variance has turned out to be very successful in other fields. For example, our approach is similar to the negative-binomial model, i.e. the Gamma-Poisson mixture, that is widely used in RNA-seq data [32, 34, 42, 43]. In that case, the Gamma distribution models the biological variation and the Poisson distribution models the variation due to the counting process in sequencing.

It is widely accepted that CpG sites have different biological variance, and thus we should not expect a constant $\emptyset$ for all CpG sites. In experiments with a small number of replicates, an empirical Bayes method that combines the information from the observed sample and shrinks toward a common prior of $\emptyset_{ij}$ helps stabilize the estimate of $\emptyset_{ij}$. Also, without any assumptions on the relationship of dispersion $\emptyset_{ij}$ with methylation level μ_{ij}, we estimated dispersion $\emptyset_{ij}$ independently of methylation level μ_{ij}.

As previous studies have shown that the methylation levels μ_{ij} are autocorrelated across the genome [37, 38], implementing smoothing in estimating methylation level is increasing popular. However, we should notice that if the assumption is violated, using smoothing could hurt the result and lead to bias in point estimation and detection [**Figure 9**]. For example, if the differentially methylated CpG sites are sparsely distributed, or if detecting DML rather than DMR is desired, or Reduced Representation Bisulfite Sequencing (RRBS) data is being used in certain case, researchers should carefully consider choose whether or not to use smoothing.

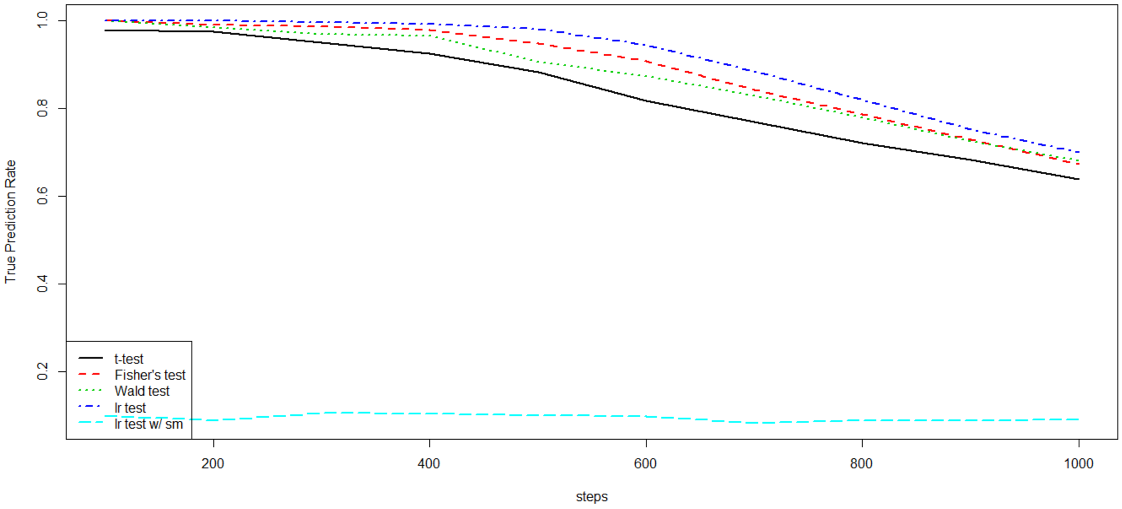

Fig. 9. TDR plot showing that smoothing can reduce TDR if the assumption of autocorrelation is violated. Here all the differentially methylated CGs are sparsely distributed across the genome and do not cluster in specific regions. In this case using smoothing (indigo) would smooth out the true point difference and lead to reduced power compared to direct point estimation. Here the indigo color line represents the TDR achieved when smoothing. In this case it is the least powerful method. 20,000 CpG sites were simulated, 10% of then are true DML.

Future work could include detecting differentially methylated loci in region level rather than in CpG site level. This would be more meaningful since DMR could be connected to functional DNA regions that either regulate genes or are genes themselves. In most cases using smoothing could improve the point value estimation and shrinkage could improve the variance estimation. Use a combination of smoothing and shrinkage in the test is also a study that deserves to be further explored. It would also be possible to extend our method to a more complex experimental design setting such as multiple groups' comparison in a GLM framework, or studies with continuous outcome variables. Finally, this method could be applied not only towards estimating the dispersion of DNA methylation data but also other beta-binomial family problems like estimating heritabilities of binary traits or spatial heterogeneity of disease in pathology studies [44].

Reference

1. Bestor TH: **The DNA methyltransferases of mammals.** *Hum Mol Genet* 2000, **9:**2395-2402.
2. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes Dev* 2002, **16:**6-21.
3. Reik W: **Stability and flexibility of epigenetic gene regulation in mammalian development.** *Nature* 2007, **447:**425-432.

4. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462:**315-322.

5. Bird AP, Wolffe AP: **Methylation-induced repression - Belts, braces, and chromatin.** *Cell* 1999, **99:**451-454.

6. Hendrich B, Bird A: **Identification and characterization of a family of mammalian methyl-CpG binding proteins.** *Mol Cell Biol* 1998, **18:**6538-6547.

7. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong CB, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao YJ, et al: **Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications.** *Nature Biotechnology* 2010, **28:**1097-U1194.

8. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, Beck S, Butcher LM: **Methylome analysis using MeDIP-seq with low DNA concentrations.** *Nature Protocols* 2012, **7:**617-636.

9. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, et al: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nature Biotechnology* 2008, **26:**779-785.

10. Serre D, Lee BH, Ting AH: **MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome.** *Nucleic Acids Research* 2010, **38:**391-399.

11. Rauch TA, Pfeifer GP: **DNA methylation profiling using the methylated-CpG island recovery assay (MIRA).** *Methods* 2010, **52:**213-217.

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

13. Cokus SJ, Feng SH, Zhang XY, Chen ZG, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.** *Nature* 2008, **452:**215-219.

14. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D: **Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.** *Nat Genet* 2005, **37:**853-862.

15. Gu H, Bock C, Mikkelsen TS, Jager N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A: **Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution.** *Nature Methods* 2010, **7:**133-U169.

16. Adorjan P, Distler J, Lipscher E, Model F, Muller J, Pelet C, Braun A, Florl AR, Gutig D, Grabs G, et al: **Tumour class prediction and discovery by microarray-based DNA methylation analysis.** *Nucleic Acids Res* 2002, **30:**e21.

17. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133:**523-536.

18. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Sung KWK, Rigoutsos I, Loring J, Wei CL: **Dynamic changes in the human methylome during differentiation.** *Genome Research* 2010, **20:**320-331.

19. Li YR, Zhu JD, Tian G, Li N, Li QB, Ye MZ, Zheng HC, Yu JA, Wu HL, Sun JH, et al: **The DNA Methylome of Human Peripheral Blood Mononuclear Cells.** *Plos Biology* 2010, **8**.

20. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454:**766-770.

21. Smith ZD, Gu HC, Bock C, Gnirke A, Meissner A: **High-throughput bisulfite sequencing in mammalian genomes.** *Methods* 2009, **48:**226-232.

22. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.** *Nucleic Acids Res* 2005, **33:**5868-5877.

23. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT: **Significance analysis and statistical dissection of variably methylated regions.** *Biostatistics* 2012, **13:**166-178.

24. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo M, Tan AC, Galm O, et al: **A DNA methylation fingerprint of 1628 human samples.** *Genome Research* 2012, **22:**407-419.

25. Feinberg AP: **The epigenetics of cancer etiology.** *Semin Cancer Biol* 2004, **14:**427-432.

26. Jones PA, Baylin SB: **The epigenomics of cancer.** *Cell* 2007, **128:**683-692.

27. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer.** *Nat Rev Genet* 2002, **3:**415-428.

28. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE: **methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.** *Genome Biol* 2012, **13:**R87.

29. Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, Milne TA, Huang Y, Biswas D, Hess JL, et al: **Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia.** *PLoS Genet* 2012, **8:**e1002781.

30. Hansen KD, Langmead B, Irizarry RA: **BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.** *Genome Biol* 2012, **13:**R83.

31. Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, Meissner A: **A unique regulatory phase of DNA methylation in the early mammalian embryo.** *Nature* 2012, **484:**339-344.

32. Wu H, Wang C, Wu Z: **A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data.** *Biostatistics* 2012.

33. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, et al: **Increased methylation variation in epigenetic domains across cancer types.** *Nat Genet* 2011, **43:**768-775.

34. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, **23:**2881-2887.

35. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

36. McCarthy DJ, Chen YS, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.** *Nucleic Acids Research* 2012, **40:**4288-4297.

37. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, et al: **DNA methylation profiling of human chromosomes 6, 20 and 22.** *Nat Genet* 2006, **38:**1378-1385.

38. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, Wen B, Feinberg AP: **Comprehensive high-throughput arrays for relative methylation (CHARM).** *Genome Res* 2008, **18:**780-790.

39. Loader C: *Local regression and likelihood.* New York: Springer; 1999.

40. Xu HY, Varghese G: **A Method to Detect Differentially Methylated Loci With Next Generation Sequencing.** *Genetic Epidemiology* 2012, **36:**723-724.

41. Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, Carstens RP, Xing Y: **MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data.** *Nucleic Acids Res* 2012, **40:**e61.

42. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.

43. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.** *BMC Bioinformatics* 2010, **11:**422.

44. Madden LV, Hughes G: **Plant disease incidence: distributions, heterogeneity, and temporal analysis.** *Annu Rev Phytopathol* 1995, **33:**529-564.