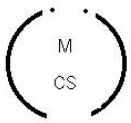
International Journal of Mathematics and
Computer Science, 5(2010), no. 2, -



An Empirical Bayes Approach for Methylation Differentiation at the Single Nucleotide Resolution

Kenneth McCallum, Wenxin Jiang, Ji-Ping Wang

Department of Statistics
Northwestern University
Evanston, IL 60208, USA

e-mail: kennethmccallum2013@u.northwestern.edu, jzwang@northwestern.edu

Abstract

DNA methylation is an important epigenetic phenomenon that is associated with a variety of diseases, particularly cancers. Recent development of high throughput sequencing technology has enabled researchers to investigate the methylation rate at a single nucleotide resolution for any given sample. Testing for methylation rate equality or difference between two samples, however, is challenged by the small sample size observed at many sites across the genome. Fisher's exact test is typically used in this situation; however, it is conservative and it cannot be used to test for specific difference in methylation rate between two samples. In this paper, we propose an empirical Bayes approach that utilizes the genome-wide data as prior information for methylation differentiation between two samples. We show that this new approach is more powerful than Fisher's exact test. In addition, it can be used to test for any specific methylation difference while controlling the false discovery rate (FDR). The new method is applied to a real data set from a colon tumor study.

Key words: 'Empirical Bayes', 'DNA Methylation', Single-nucleotide
AMS subject classification: 62J05, 62J07, 62H35, 62P10
ISSN 1814-0424

1. Introduction

The epigenetic phenomenon of DNA methylation, in which cytosines in CpG dinucleotides are chemically modified by the addition of a methyl group, plays an important role in genetic regulation [1, 2]. Methylation rates are known to change throughout the genome during development in mammals [3, 4]. Furthermore, differential methylation rates are associated with a variety of diseases, including neurodevelopmental disorders [5], and numerous cancers [6, 7, 8, 9].

Most research up to now has focused on methylation rates over large regions of the genome; however, an increasing number of studies attempt to quantify and analyze methylation rates at specific sites [1, 2]. Methods making use of universal bead arrays have been able to detect differential methylation rates at the single nucleotide resolution and have demonstrated that these differences can be used to distinguish normal and cancer tissues [10]. However, the array based methods are limited in terms of the number of sites that can be examined; for example, only 1536 sites were included in the study by Bibikova et al. [10]. Another approach, the one which will be the focus of this study, makes use of high throughput bisulfite sequencing. This method is increasingly common and has the potential to examine hundreds of thousands or millions of sites simultaneously. For example, Laurent et al. [11] and Gu et al. [12] both made use of bisulfite sequencing to generate maps of methylation with single-nucleotide resolution, Han et al. [9] tested for site specific differential methylation in samples taken from subjects with and without lung cancer using bisulfite sequencing, and Houseman et al. [13] used clustering methods to differentiate methylation rates.

The data set produced by Gu et al. [12] is used for illustration in this study. The data was for two tissue samples, a colon tumor and normal colon tissue, both taken from the same donor. Bisulfite sequencing was used to determine methylation status at targeted CpG sites across the genome. At each site (corresponding to the C in a CpG dinucleotide), the data included the number of reads (number of sequenced DNA fragments) that covered the given site, and the number of reads that were positive for methylation at the given site. Figure 1 illustrates the format of the data. A total of 920,441 sites had at least one read for both tissue samples and only these sites were included in the present study. Although a few sites had very large numbers of reads, some with more than 1400, the majority were small, with a median of 10 or fewer across both samples. Summary statistics for the number of reads are given in Table 1.

The central goal of methylation studies is to identify CpG sites or regions that show differential methylation rates between disease or cancer tissue and normal tissue. Given the small number of reads (or sample size) at a given site, Fisher's exact test is often the only choice for testing the equality of methylation rates between two samples. Fisher's exact test, however, is conservative in power. Furthermore, it cannot be used to test for specific difference in the methylation rates between two samples. The latter is particularly important, as in practice, a meaningful difference in methylation is often called only if the methylation rate in one sample is higher/lower than the other by a predefined threshold value (see details

Figure 1: Format of data.

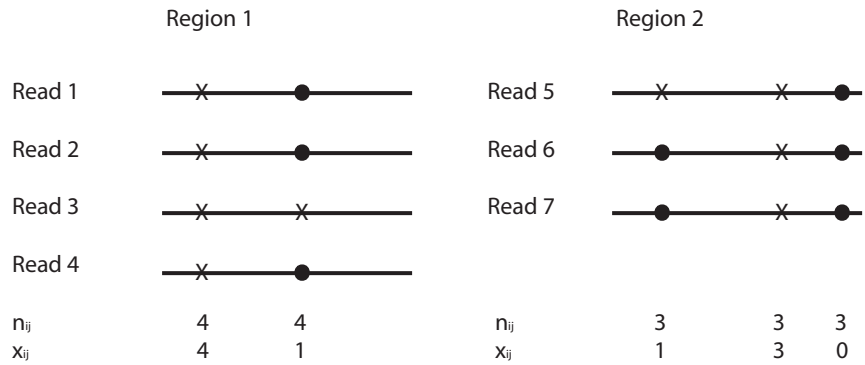


Figure 1 shows two regions of equal length that contain the targeted CpG sites for methylation examination. Bisulfite sequencing generated two groups of short reads of identical length, each of which covered one of these regions. For example, in region 1, four reads were generated. Within each read, the methylation status at the sites were identified as positive or negative. In the figure, n is used to denote the total number of reads observed at a given site, and x the number of methylations (positives) observed across reads.

Table 1: Quantiles for number of reads per site (n_{ij}).

	Minimum	Q1	Median	Q2	Maximum
Normal	1	2	6	14	14043
Tumor	1	3	10	25	14361

below). These two limitations motivate us to seek an alternative approach.

In this paper, we proposed an empirical Bayes (eB) approach, in which we utilize the large amount of data observed from the entire genome to construct a prior distribution for the methylation rate. Based on the posterior distribution of the methylation rate at each site, we test for difference of methylation rate while controlling false discovery rate. We show this new approach has improved power compared to Fisher's exact test. In addition, it can be used to test any specific difference of methylation rates between two samples.

2. Methods & Results

2.1. The Model

Let x_{ij} be the observed methylations out of a total of n_{ij} reads at site i from sample j for $i = 1, ..., M$ and $j = 1, 2$, and θ_{ij} be the true, unobserved, methylation rate. Let

$\mathbf{N} = (\mathbf{N}_1, \mathbf{N}_2)$, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{N}_j = \{n_{ij} : i = 1, ..., M\}$ and $\mathbf{X}_j = \{x_{ij} : i = 1, ..., M\}$. We assume that the methylation rates are independent across samples and sites, and have a common distribution within each sample,

$$\theta_{ij} \sim Beta(\gamma_j, \lambda_j).$$

Note that although the model allows for the possibility that the samples may differ with respect to the hyper-parameters γ and λ, results given below show that in practice a single set of hyper-parameters can be used if the samples are similar. We further assume that each read is a random observation from the population, i.e., the entire underlying tissue. Then x_{ij} follows a binomial distribution

$$x_{ij}|(n_{ij}, \theta_{ij}) \sim Binomial(n_{ij}, \theta_{ij}).$$

The posterior probability for the methylation rate given the reads data is then

$$\theta_{ij}|(n_{ij}, x_{ij}) \sim Beta(\gamma_j + x_{ij}, \lambda_j + n_{ij} - x_{ij}).$$

To estimate the hyper-parameters, observe that the likelihood function is

$$L(\gamma_j, \lambda_j; \mathbf{N}_j, \mathbf{X}_j) = \Pi_{i=1}^{M} \int_0^1 P[X_{ij} = x_{ij}|\theta_{ij}, n_{ij}] p(\theta_{ij}|\gamma_j, \lambda_j) d\theta_{ij}.$$

Under the model,

$$P[X_{ij} = x_{ij}|\theta_{ij}, n_{ij}] = \binom{n_{ij}}{x_{ij}} \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{n_{ij} - x_{ij}}$$

and

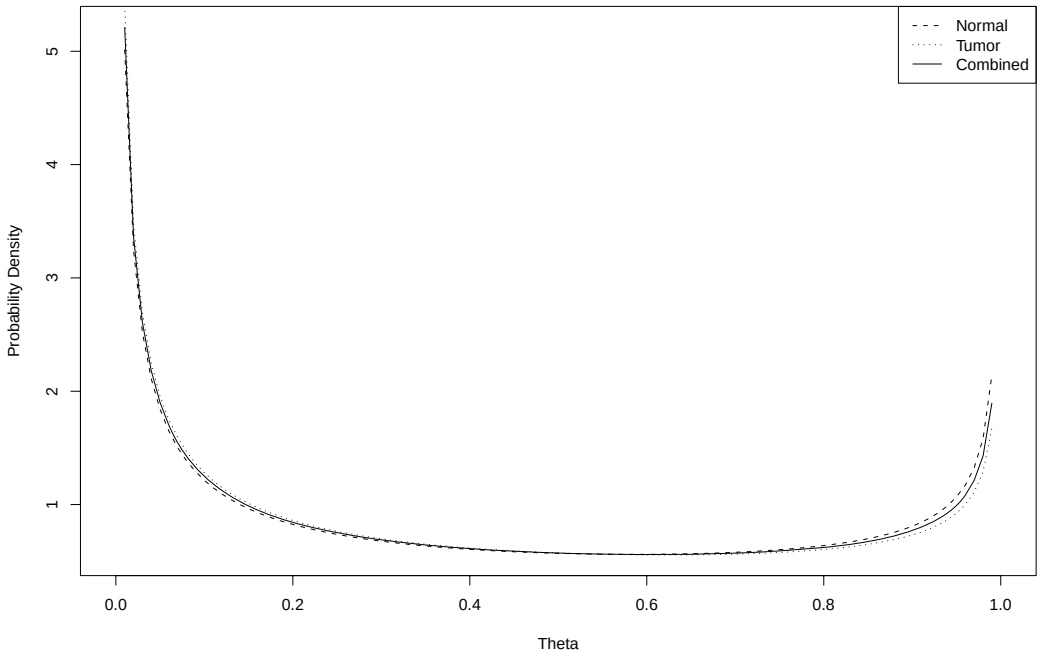
$$p(\theta_{ij}|\gamma_j, \lambda_j) = B^{-1}(\gamma_j, \lambda_j) \theta_{ij}^{\gamma_j - 1} (1 - \theta_{ij})^{\lambda_j - 1},$$

where B stands for the Beta function. Therefore,

$$\begin{aligned} L(\gamma_j, \lambda_j; \mathbf{N}_j, \mathbf{X}_j) &= \Pi_{i=1}^{M} \int_0^1 \binom{n_{ij}}{x_{ij}} B^{-1}(\gamma_j, \lambda_j) \theta_{ij}^{\gamma_j + x_{ij} - 1} (1 - \theta_{ij})^{\lambda_j + n_{ij} - x_{ij} - 1} d\theta_{ij} \\ &= \Pi_{i=1}^{M} \binom{n_{ij}}{x_{ij}} B^{-1}(\gamma_j, \lambda_j) B(\gamma_j + x_{ij}, \lambda_j + n_{ij} - x_{ij}), \end{aligned}$$

The maximum likelihood estimates of γ_j and λ_j, denoted $\hat{\gamma}_j$ and $\hat{\lambda}_j$, can easily be found using a method such as the Newton-Raphson algorithm. For the tumor and normal colon tissue data from [12], we fitted the Beta-binomial model for each sample separately, and then for the combined data. Results are summarized in Table 2. The MLEs for γ are approximately equal across samples while the MLEs for λ show a greater difference, with the tumor tissue having a λ value approximately 10% greater than the normal tissue. Despite this small discrepancy in λ, the density curves, shown in Figure 2, appear almost identical. This suggests that little would be gained by specifying separate priors for the two samples in this case.

Figure 2: Prior Distribution Densities.



The densities of the fitted prior distributions are given. These priors are assumed to be i.i.d. across all sites in the data set used to fit them.

To verify the fit of the model, based on the empirical distribution of n_{ij}, we calculated the expected empirical distribution of the observed methylation rates, defined as x_{ij}/n_{ij}, based on the fitted beta model from the joint data, treating x_{ij} as a random variable. For a given n_{ij}

$$P(X_{ij} = x_{ij}|n_{ij}, \hat{\gamma}, \hat{\lambda}) = \int_0^1 P[X_{ij} = x_{ij}|n_{ij}, \theta_{ij}]p(\theta_{ij}|\hat{\gamma}, \hat{\lambda})d\theta.$$

The probability for observing methylation rate $q \equiv x_{ij}/n_{ij}$ is then found by taking the weighted average over all pairs (x, n) such that $x/n = q$. That is, if $S_q = \{(x, n) : x/n = q\}$, then the expected probability to observe q in the sample given the empirical distribution of n_{ij} is

$$\Sigma_{S_q} P(X_{ij} = x_{ij}|n_{ij}, \hat{\gamma}, \hat{\lambda})P(N_{ij} = n_{ij})$$

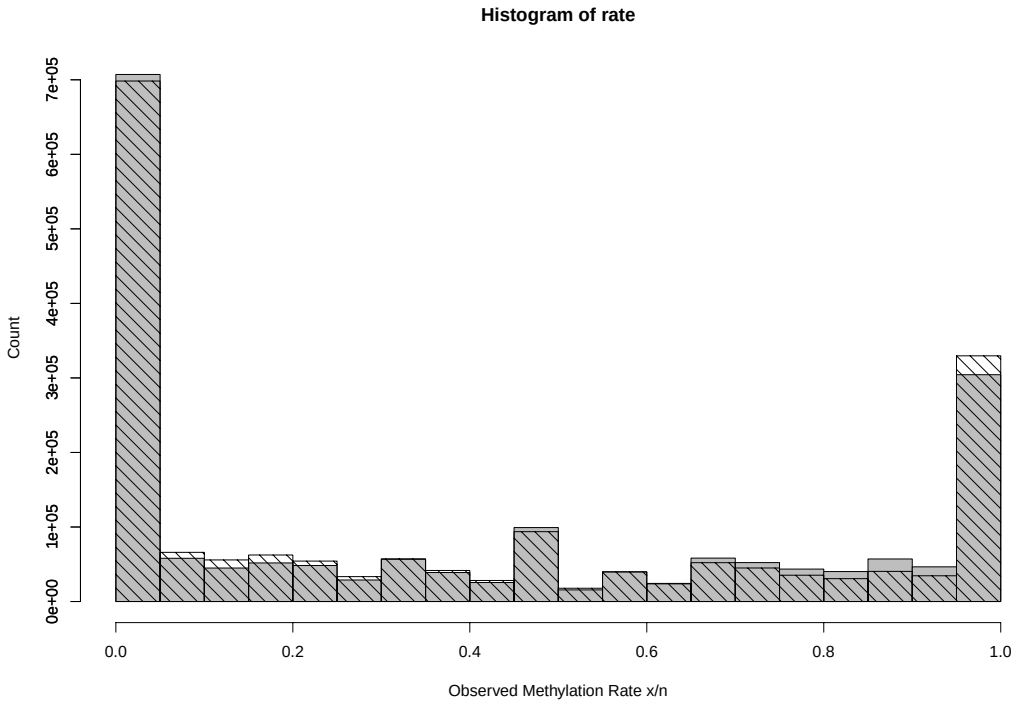
where $P(N_{ij} = n_{ij})$ is the probability of a site having n_{ij} reads based on the empirical distribution. This is then compared to the observed distribution of methylation rates. Figure 3 gives the shape of the distributions. The shape is similar to that of the curves in Figure

Table 2: Maximum likelihood estimates of hyperparameter values.

Parameter	Normal	Tumor	Combined
γ	0.365518	0.359081	0.362389
λ	0.550387	0.614117	0.582426

Estimates of the parameters for the prior distribution are given for normal colon tissue data, tumor colon tissue data, and the combined data set.

Figure 3: Theoretical and Observed Methylation Rates.



Proportion of positive reads out of total reads (x_{ij}/n_{ij}) is given on the x-axis. Number of sites matching a given proportion is shown on the y-axis. The grey bars represent the observed data while the bars with black diagonal stripes indicate the theoretical number given the prior distribution for the underlying methylation rate and the empirical distribution for the number of reads.

2, though it reflects the fact that the distribution of observed rates is discrete. The spikes that occur near 0 and 1 in the plot are partially due to the large number of sites with small

numbers of reads, which are highly constrained in terms of values they can take on. Overall, the evidence shows that the model is a very good fit for the data.

2.2. Hypothesis Tests

Two different sets of hypotheses are considered. The first is a simple test of equality

$$H_0 : \theta_{i1} = \theta_{i2} \text{ vs. } H_A : \theta_{i1} \neq \theta_{i2}.$$

The second is a test of difference of rates given by

$$H_0' : |\theta_{i1} - \theta_{i2}| \leq c \text{ vs. } H_A' : |\theta_{i1} - \theta_{i2}| > c$$

for some constant c. The second hypothesis is particularly interesting, as in practice, differential methylation is often called when the difference is substantial, e.g., c=0.2 [11].

Given $\hat{\gamma}_j$, $\hat{\lambda}_j$ the posterior distribution of the methylation rate is

$$\theta_{ij}|(x_{ij}, n_{ij}) \sim Beta(\hat{\gamma}_j + x_{ij}, \hat{\lambda}_j + n_{ij} - x_{ij}).$$

For convenience, we shall denote the posterior distribution as $\pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(\theta_{ij})$ in the following context. For testing $H_0 : \theta_{i1} = \theta_{i2}$ versus $H_1 : \theta_{i1} \neq \theta_{i2}$, we define the posterior log odds as follows:

$$\Delta_i \equiv log\big[\frac{\pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(\theta_{i1} > \theta_{i2})}{1 - \pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(\theta_{i1} > \theta_{i2})}\big].$$

We reject H_0 if $|\Delta_i| > \delta_\alpha$, where δ_α is the cutoff value corresponding to level α.

Given the prior distribution and the number of reads at a site for each sample, it is possible to calculate the level (α) and power of the test for a given critical value (δ_α) analytically. However, doing so for every combination of number of reads appearing in the sample would be extremely computationally intensive. Here we estimate it using Monte Carlo simulations. We first generate Monte-Carlo samples as follows:

1. Sample (n_{i1}, n_{i2}) pairs with replacement from the observed data. We sample the pairs instead of individual n_{ij}'s to account for possible dependence of reads count between samples due to various factors including DNA sequence features.

2. Sample θ_{ij} values for each site in each sample from the fitted prior distribution, i.e., $Beta(\hat{\gamma}, \hat{\lambda})$ from the combined data or $Beta(\hat{\gamma}_j, \hat{\lambda}_j)$ from separate samples.

3. Generate x_{ij} from $Binomial(n_{ij}, \theta_{ij})$

Two simulated data sets of size equal to the original data are generated. In one set, we use $Beta(\hat{\gamma}, \hat{\lambda})$ to generate the θ_{ij} values for both samples. For the second set, $Beta(\hat{\gamma}, \hat{\lambda})$ is used only for the sites with equal θ_{ij} values while the remaining sites are simulated using the separate estimates from the normal and tumor tissues (i.e., $Beta(\hat{\gamma}_j, \hat{\lambda}_j)$). In both cases, the first 100,000 sites are set so that $\theta_{i1} = \theta_{i2}$ while the remaining ones are allowed to vary.

Table 3: Test of Equality

	Simulation 1			Simulation 2		
	Critical	level	power	Critical	level	power
1 prior	2.5	0.1	0.518	2.5	0.1	0.522
1 prior	3.15	0.05	0.438	3.15	0.05	0.442
2 priors	2.5	0.1	0.524	2.5	0.1	0.522
2 priors	3.15	0.05	0.452	3.21	0.05	0.441
Fisher's Exact		.1	0.356		0.1	0.355
Fisher's Exact		0.05	0.317		0.05	0.316

Simulation 1 used a single prior from the combined data set to generate the methylation rates. Simulation 2 used the prior from the combined data set to generate methylation rates for the subset of the simulated data where the rates were set equal across tissue samples, and used each sample's individually calculated prior for the remaining data points. The designations of 1 prior and 2 prior refer to whether the combined data estimates of the parameters or the individual tissue sample estimates were used in calculating the log odds.

After the simulated data set is complete, the posterior log odds can be calculated for each site. A suitable critical value can then be selected for a level α test by setting δ_α equal to the $100(1-\alpha)th$ percentile of the absolute values of the log odds for the subset of sites with $\theta_{i1} = \theta_{i2}$. Similarly, power can be estimated by taking the proportion of sites with $\theta_{i1} \neq \theta_{i2}$ with posterior log odds that have absolute values less than δ_α. In implementing this test for the simulated data sets, rather than refitting the values of $\hat{\gamma}$ and $\hat{\lambda}$, the values used in the simulation were reused in calculating the posteriors. This is justified by the large size of the data sets and the resultant accuracy and precision of the MLE.

Two versions of the test are conducted on each of the simulated data sets. The first uses the combined estimates of the hyper-parameters, while the second uses the separate estimates of the hyper-parameters for each of the two data sets. The results in this case indicate that it makes little or no difference which way the priors are specified. This is not unexpected since the two prior distributions were so similar. However, this might not generalize to all cases if two specimens are markedly different in their methylation patterns. Table 3 shows the approximate critical values for level 0.1 and 0.05 tests, and estimated power. Results for Fisher's exact test are also given for comparison. It should be cautioned that the critical values given here depends on both the hyper-parameters and the distribution of read counts, and hence are specific to this data set and should not be taken to be generally applicable.

For the test of difference, we define the posterior log odds that $|\theta_{i1} - \theta_{i2}| > c$ as follows,

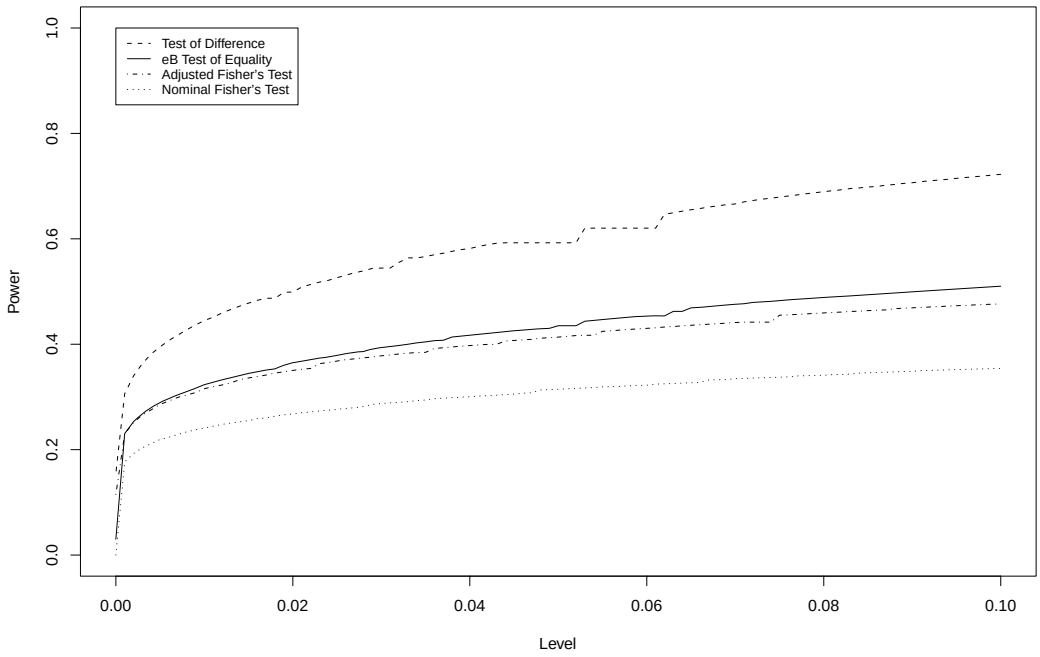
$$\Delta_i^c = log\big[\frac{\pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(|\theta_{i1} - \theta_{i2}| > c)}{1 - \pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(|\theta_{i1} - \theta_{i2}| > c)}\big].$$

Table 4: Test of Differences

Critical	Level	Power
0.857	0.1	0.741
1.658	0.05	0.620

Critical values, level, and power for the test of differences of methylation rates are reported with a $c = 0.2$ as the null hypothesized largest absolute difference. Test were done with a single prior on simulated data using the prior fitted to the combined data.

Figure 4: Power versus Level.



Power is shown on the y-axis and level on the x-axis. Values are estimates based on simulated data.

As with the test of equality, a critical value for a given level α, and the corresponding power, can be determined by simulations. A difference threshold of $c = 0.2$ was chosen for the test, which corresponds to the bin width for categorizing methylation rates used in other studies (eg. Laurent et al., 2010, [11]). Simulations indicate greater power for the test of differences than for the test of equality at a given level. Results are summarized in Table 4. The power of all three tests is plotted against the level in Figure 4.

Table 5: Hypotheses and True Values

	Test Negative	Test Positive	Total
H_0	M_{00}	M_{01}	$M_{0\cdot}$
H_1	M_{10}	M_{11}	$M_{1\cdot}$
Total	$M_{\cdot 0}$	$M_{\cdot 1}$	$M_{\cdot\cdot}$

The number of true negative, $M_{0\cdot}$, and true positives, $M_{1\cdot}$, compared to the number testing as negative, $M_{\cdot 0}$, and testing as positive, $M_{\cdot 1}$. Only $M_{\cdot 0}$, $M_{\cdot 1}$, and $M_{\cdot\cdot}$ are directly observed.

2.3. False Discovery Rate

Using the estimate of level, α, and power, β, from the simulations, the false discovery rate (FDR) can be estimated for the original data set. Let $M_{\cdot\cdot}$ be the number of sites, $M_{0\cdot}$ and $M_{1\cdot}$ be the total number of true null and alternative hypotheses respectively. Let $M_{\cdot 0}$ and $M_{\cdot 1}$ be the numbers of claimed negatives and positives. Table 5 tabulates four different incidents incurred in hypothesis testing: true negatives (M_{00}), false natives (M_{10}), true positives (M_{11}), and false positives (M_{01}). Then

$$E[M_{\cdot 1}] = \alpha M_{0\cdot} + \beta M_{1\cdot} = \alpha M_{0\cdot} + \beta(M_{\cdot\cdot} - M_{0\cdot}).$$

This implies that $M_{0\cdot}$ can be estimated by

$$\hat{M}_{0\cdot} = [M_{\cdot 1} - \beta M_{\cdot\cdot}]/[\alpha - \beta]$$

The FDR is then estimated by

$$FDR = \frac{\alpha \hat{M}_{0\cdot}}{M_{\cdot 1}}.$$

Since M_{00}, M_{01}, M_{10}, and M_{11} are all functions of the specified level α, estimation of $M_{0\cdot}$ and FDR requires an appropriate choice of α. For the real data, using the estimated α and β from the simulation studies presented in Figure 4, we calculated $\hat{M}_{0\cdot}$ for α ranging from 0.1 to 0.0001. Interestingly, $\hat{M}_{0\cdot}$ increased monotonically from around 740,000 to over 870,000 as the type I error level decreased from 0.1 to less than 0.0001. To determine which value of α can lead to a most accurate estimate of $M_{0\cdot}$, we simulated data sets containing 800,000 true nulls and 192,000 true alternatives where the methylation rate θ_{ij} followed the prior distribution fitted from the eB approach. The monotonicity, however, was not observed; and $M_{0\cdot}$ was estimated very accurately for any α value used in the same range. This likely indicates some violations of model assumptions in the real data. We leave this as an open question for future investigation.

In the absence of a reliable estimate of $M_{0\cdot}$, a precise estimate of FDR cannot be calculated. The most conservative estimate of FDR can be obtained by substituting $M_{\cdot\cdot}$ for $\hat{M}_{0\cdot}$ into the FDR formula. An FDR of 0.05 is achieved by setting the level as $\alpha = 0.00092$, at which $M_{\cdot 1} = 16,976$ sites were identified as differentially methylated. In contrast, only

5,003 sites were identified as differentially methylated at the same FDR using Fisher's exact test (the FDR was controlled by requiring the q-value of each individual hypothesis to be ≤ 0.05 using the QVALUE R package downloaded from http://www.bioconductor.org). The eB test clearly shows improved power over Fisher's test, however, the majority of the true positive sites remain un-identified due to the limitation of small sample size (n_{ij}).

The same method was applied to the test of difference ($H_0' : |\theta_{i1} - \theta_{i2}| \leq c$ vs. $H_A' : |\theta_{i1} - \theta_{i2}| > c$ at $c = 0.20$). An FDR≤ 0.05 was achieved at level 0.000088. A total of 1,630 sites were identified to have significantly pronounced difference (≥ 0.2) in methylation rates between the two samples.

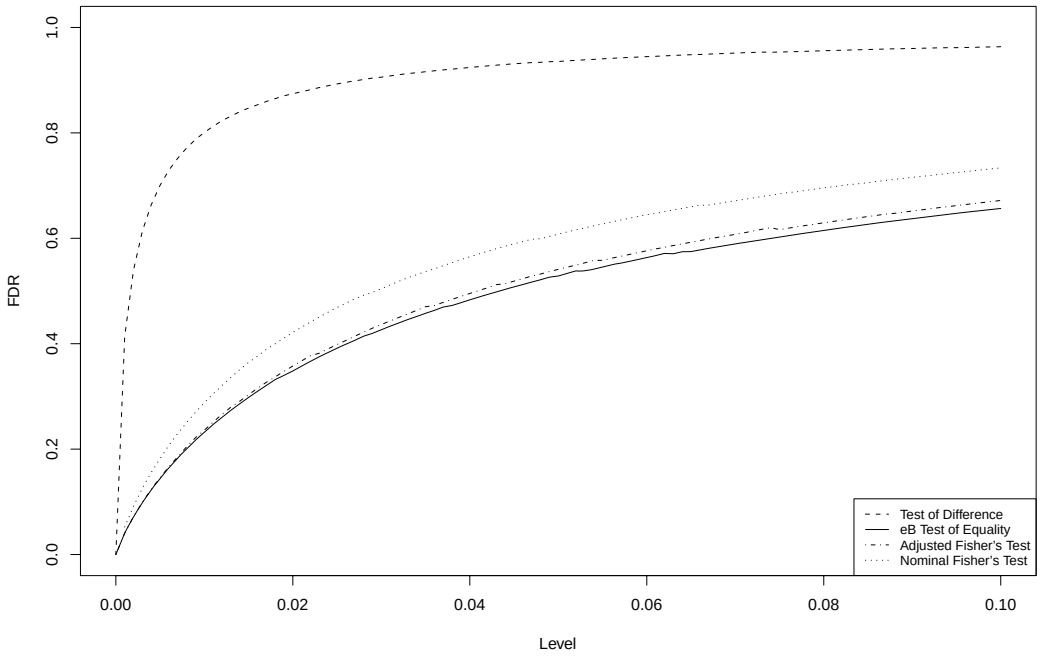
To gain further insights into the FDR behavior, in Figure 5, we plotted the FDR as a function of the level summarized from the simulation studies described above. The proposed eB method has a lower FDR than Fisher's exact test at all levels less than 0.1. Since the actual level of Fisher's exact test is typically lower than the nominal level, we also plotted the FDR vs actual level of Fisher's exact test. The actual level was assessed in the same way as for the eB approach, by finding the actual type I error rate in the simulated data at each given nominal level. The posterior log odds test has a uniformly lower FDR than the Fisher's exact test even after the adjustment for the difference in nominal and actual levels. In practice, as Fisher's test is always performed under the nominal level while the true level is never known, a comparison of the power or FDR under the nominal level is more meaningful. A spreadsheet with the locations on the genome that tested as positive is available at http://bioinfo.stats.northwestern.edu/∼jzwang/.

3. Discussion

In this paper, we showed the two advantages of proposed empirical Bayes approach over Fisher's exact test in methylation differentiation studies. This method is particularly useful when the number of reads at each site (or sample size) is small, while genome-wide data can provide rich information regarding the methylation rates across sites. Indeed, as shown in Figure 2, the fitted beta distribution has majority of probability mass concentrated around 0 and 1. This suggests that most sites have either very high or very low methylation rates. This prior information tends to shrink the posterior distribution of θ_{ij} towards the two ends. For example, if the observed $x_{ij} = 0$ and $n_{ij} = 2$, then it is highly likely that this site has a low methylation rate regardless of the small sample size, and vice versa. This strong prior information forms the basis for the power improvement when using posterior log odds as the test statistics.

Several possibilities exist for generalizations or refinements of this approach. Firstly, the bias issue in estimation of $\hat{M}_{0.}$ needs further investigation. It is not clear to us whether there is a causal relationship between the type I error level α and the bias. Secondly, currently all sites are treated as independent. In a real genome, it is possible that sites nearby may be correlated in methylation rate. Characterizing such dependence may help further improve the power of the eB approach. Finally, only two tissue samples were used to generate the data for this study; however, it will often be desirable to incorporate multiple specimens for each

Figure 5: FDR versus Level.



FDR is shown on the y-axis and level on the x-axis. The two curves for Fisher's exact test differ due to the overly conservative nature of the test. Values are estimates based on simulated data. The abbreviation "eB" stands for empirical Bayes.

condition. If methylation rates across specimens can be considered to be independent, then the density of the vector of methylation rates will be a product of beta densities. Similarly, the vector of positive reads will have probability mass given by the product of independent binomial pmfs. Because of independence, the posterior density for the vector of methylation rates will then be the product of beta densities, with the beta densities being the same as the posteriors would be if each specimen were treated separately. Once this distribution is known, the distribution of weighted averages of the methylation rates can be easily obtained.

Acknowledgements

This work is supported by NCI grant U54CA143869.

References

[1] J. Tost, DNA Methylation: An Introduction to the Biology and Disease-Associated Changes of a Promising Biomarker, Molecular Biotechnology, 44 (2010), 71-81.

[2] A. Bird, DNA methylation patterns in epigenetic memory, Genes and Development, 16 (2002), 6-21.

[3] S. Guibert, T. Forne, and M. Weber, Dynamic regulation of DNA methylation during mammalian development, Epigenetics, 1 (2009), 81-98.

[4] W. Reik, W. Dean, and J. Walter, Epigenetic reprogramming in mammalian development, Science, 293 (2001), 1089-1093.

[5] K. D. Robertson, DNA methylation and human disease, Nature Reviews Genetics, 6 (2005), 597-610.

[6] G. Heller, C. C. Zielinski, and S. Zöchbauer-Müller, Lung Cancer: From single-gene methylation to methylome profiling, Cancer Metastasis Review, 29 (2010), 95-107.

[7] B. C. Christensen, C. J. Marsit, A. E. Houseman, J. J. Godleski, J. L. Longacker, S. Zeng, R.-F. Yeh, M. R. Wrensch, J. L. Wiemels, M. R. Karagas, R. Bueno, D. J. Sugarbaker, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, Differentiation of Lung Adenocarcinoma, Pleural Mesothelioma, and Nonmalignant Pulmonary Tissues Using DNA Methylation Profiles, Cancer Research, 69 (2009), 6315-6321.

[8] D. T. Hsiung, C. J. Marsit, E. A. Houseman, K. Eddy, C. S. Furniss, M. D. McClean, and K. T. Kelsey, Global DNA Methylation Level in Whole Blood as a Biomarker in Head and Neck Squamous Cell Carcinoma, Cancer Epidemiology, Biomarkers and Prevention, 16 (2007), 108-114.

[9] W. Han, T. Wang, A. A. Reilly, S. M. Keller, and S. D. Spivack, Gene promoter methylation assayed in exhaled breath, with differences in smokers and lung cancer patients, Respiratory Research, 10 (2009), 86.

[10] M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E. W. Garcia, B. Wu, D. Doucet, N. J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D. L. Barker, M. S. Chee, J. Floros, and Jian-Bing Fan, High-throughput DNA methylation profiling using universal bead arrays, Genome Research, 16 (2006), 383-393.

[11] L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. K. Sung, I. Rigoutsos, J. Loring, and C. L. Wei, Dynamic changes in the human methylome during differentiation, Genome Research, 20 (2010), 320-331.

[12] H. Gu, C. Bock, T. S. Mikkelsen, N. Jäger, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander, and A. Meissner, Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution, Nature Methods, 7 (2010), 133-136.

[13] E. A. Houseman, B. C. Christensen, R.-F. Yeh, C. J. Marsit, M. R. Karagas, M. Wrensch, H. H. Nelson, J. Wiemels, S. Zeng, J. K. Wiencke, and K. T. Kelsey, Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions, BMC Bioinformatics, 9 (2008), 365.