ESTIMATING DNA METHYLATION LEVELS FOR SINGLE-CELL BISULFITE SEQUENCING (BS-SEQ) DATA

by

Yan Jiang

BS in Preventive Medicine, Jilin University, China, 2017

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Yan Jiang

It was defended on

April 12, 2019

and approved by

Committee Member: Ernesto T. A. Marques, Jr., MD, PhD, Associate Professor, Department of Infectious Diseases and Microbiology, Graduate School of Public Health, University of Pittsburgh

Committee Member: Lu Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Yongseok Park, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Yongseok Park, PhD

ESTIMATING DNA METHYLATION LEVELS FOR SINGLE-CELL BISULFITE SEQUENCING (BS-SEQ) DATA

Yan Jiang, MS

University of Pittsburgh, 2019

Abstract

DNA methylation is among the most studied epigenetic marks, with essential influence on biological growths, disease developments and potential public health benefits. Modified from the well-established measuring method bisulfite sequencing (BS-seq), single-cell bisulfite sequencing (scBS-seq) emerged recently to identify DNA methylation status within a single cell to profile and study heterogeneities better. With the unique features of the single-cell DNA methylation data, there is in need of developing a new method to assign the methylation status for each CpG site more accurately and precisely to represent the underlying truth. In this study, we propose a method using Bayes rule and compare its performance with a simple one-third rule method. A simulation study with various settings is conducted to compare the accuracy, precision and bias. The Bayes' method shows an improvement in dimensions of accuracy and bias.

TABLE OF CONTENTS

1.0 INTRODUCTION.. 1

2.0 METHODS 4

2.1 BAYES' RULE METHOD 4

2.2 EVALUATION CRITERIA 6

3.0 SIMULATION 8

3.1 SIMULATION SETTING 8

3.2 SIMULATION RESULTS... 9

3.2.1 DNA METHYLATION DETECTION ACCURACY 9

3.2.2 DNA METHYLATION DETECTION PRECISION 10

3.2.3 DNA METHYLATION DETECTION BIAS 11

3.3 CONCLUSTIONS OF SIMULATIONS 12

4.0 APPLICATION IN REAL DATA ANALYSIS 13

4.1 DATA.. 13

4.2 ERROR RATES ESTIMATION 14

4.3 APPLICATION IN DIFFERENTIALLY METHYLATION ANALYSIS............ 16

5.0 DISCUSSION AND CONCLUSION 17

BIBLIOGRAPHY 18

LIST OF TABLES

Table 1 Correct call proportions of the 2 methods with sample size 100,000 under 3 scenarios. 10

Table 2 Standard deviations of the 2 methods with different sample regions $ under 2 scenarios 10

Table 3 Biases produced by the 4 methods with different sample sizes K.................................. 11

LIST OF FIGURES

Figure 1 Epigenotype as the intermediate phenotype 2

Figure 2 Overall methylation levels and total number of CpG sites in promoter regions 14

Figure 3 Estimated error rates of methylation status among MII cells 15

Figure 4 A two way hierarchical cluster analysis of the relative methylation of promoter regions 16

1.0 INTRODUCTION

DNA methylation is a biochemical process where a methyl group (CH3) is added to the cytosine or adenine. In mammalian cells, the change happens almost exclusively on the 5 position of cytosine (C) when it is followed by guanine (G), called CpG site. This modification of DNA has been found to play important roles in genomic imprinting, genome stability and regulation of gene expression without altering the DNA sequence itself. Dysregulation of DNA methylation has been found to be related to many different diseases, particularly in cancer with overall genomic demethylation and gene specific hypermethylation (Bird, 2002; Kriaucionis & Heintz, 2009; Meissner et al., 2008; Seisenberger et al., 2012; Szulwach et al., 2011; Teschendorff et al., 2010). DNA methylation was the first discovered and remains one of the most studied and best understood epigenetic marks. A recent extensive expansion of our knowledge on epigenetic changes has revealed that the impact of genetic features on the biological phenotype changes is not direct, but through epigenetic changes (Figure 1), which strongly suggests that disease development is better reflected by epigenetic than genetic features. Therefore, epigenetic marks such as DNA methylation can become more conclusive and predictive biomarkers for the detection and diagnosis of many diseases.

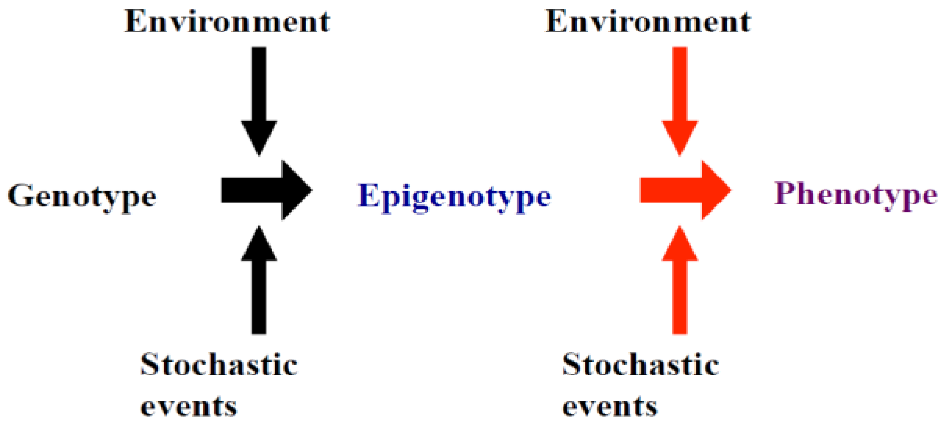

Figure 1 Epigenotype as the intermediate phenotype

Detection technologies of DNA methylation have been shifted from mass spectrometry and array-based assays to sequencing based technologies. Treatment of DNA with sodium bisulfite causes the conversion of unmethylated cytosines to uracils while methylated cytosines remain protected from this conversion. The consequence of this process is that methylated and unmethylated CpG sites can be identified using next generation sequencing (NGS). Over the past decade, protocols for sodium bisulfite treatment of DNA coupled with NGS (BS-seq) have become the gold standard approach for assessment of genome-wide DNA methylation. This trend is partially due to high level of fidelity and reproducibility of BS-seq, with rates well above 99% generally being achieved for conversion of unmethylated cytosine residues to uracil. Series of key findings and technological advances over the past years have led to our current ability for quantitative query of the methylation status across the whole genome. BS-seq has brought revolutionary basepair resolution to study genome wide DNA methylation. Whole genome bisulfite sequencing (WGBS) provides the first and still remains the only method to obtain accurate, quantitative estimates of the percent of cells in a population that are methylated at each of the millions of CpG sites across the entire genome.

DNA methylation is heterogeneous even among the same type of tissue within the same individuals. This heterogeneity of DNA methylation patterns may be partially responsible for the heterogeneity of the cell populations. The recently development of single cell Bisulfite sequencing (scBS-seq) technology has opened the door to study cell specific methylation patterns (Farlik et al., 2015; Smallwood et al., 2014). Each dataset generated by scBS-seq provides methylation information for a single cell. Although the short length of sequencing reads generated by scBS-seq is not a major concern, the low genomic coverage (~20% of CpGs) presents a major statistical challenge in characterizing cell specific information. Another challenge in the analysis of these types of data in diploid organisms such as human is the presence of allele specific methylation patterns. These challenges and others must be addressed to facilitate the identification of global and local methylation levels along with spatial methylation patterns within each cell.

Unlike a regular bisulfite sequencing dataset composed of a mixed cell population which is epigenetically heterogeneous, a diploid single-cell dataset is expected to display methylation levels only of 0%, 50% or 100% at an individual CpG site. Some observed data in reality can deviate from these percentages caused by technical artifacts such as sequencing and mapping errors. When a CpG site is mapped with multiple reads, one simple way to call the CpG methylation status is using one-third rule, in which we set 1/3 and 2/3 as cutoffs and assign data not higher than 1/3 to 0%, not lower than 2/3 to 100%, and any level in between to 50%. In this thesis, we address this problem by proposing a statistical method based on Bayes' rule to better identify the underlying methylation status for each of the millions of CpG sites. We will compare the performance of this method to the 1/3 rule using a simulation study and perform a real data analysis by applying this method.

2.0 METHODS

2.1 BAYES' RULE METHOD

Single cell DNA methylation data only include methylation status at CpG sites that are covered by some sequence reads. Let n_{si} and y_{si} be the total number of reads and methylated reads in sample s at CpG site i, then if this CpG site is methylated $n_{si} = y_{si}$ and if it is unmethylated $y_{si} = 0$. For diploid organisms, methylation status could be different for each allele. Therefore, the methylation status in each of the CpG sites can only be 0%, 50%, or 100%. However, due to sequencing and mapping errors, and potential insufficient bisulfite conversion for one of both alleles, we frequently observe any methylated proportion of reads. One simple method is to separate these three statuses with so called one-third rule: 0 if $y_{si}/n_{si} \leq 1/3$, ½ if $\frac{1}{3} < y_{si}/n_{si} < 2/3$, and 1 if if $y_{si}/n_{si} \geq 2/3$. Then we combine this methylation status to make a summary data.

Another more rigorous way is to estimate the methylation status by incorporating more information. Let s_{si} be the underlying methylation status for CpG site i in sample s. Then as we discussed above, it can only be 0, ½ or 1. Let p_{si} be the true proportion of methylated reads by considering the error rates, in which $p_{si} \leq p_a$ if $s_{si} = 0, p_b \leq p_{si} \leq p_c$ if $s_{si} =$ ½ and $p_{si} \geq p_d$ if $s_{si} = 1$.

Here $y_{si}|p_{si} \sim Bin(n_{si}, p_{si})$, and $P(p_{si}|y_{si}, n_{si}) = P(y_{si}|p_{si}, n_{si})P(p_{si}|n_{si})/P(y_{si}|n_{si})$ by following Bayes' Rule. Our goal is to impute the status of methylation level for CpG site i given data (y_{si}, n_{si}).

Our proposed method is based on restrictedly maximizing posterior likelihood function $P(p_{si}|y_{si}, n_{si})$. I.e. we compare the restricted maximums of likelihood when $p_{si} \leq p_a$, $p_b < p_{si} < p_c$, and $p_{si} \geq p_d$ and based on where the maximum falls, we assign $s_{si} = 0$, $s_{si} = \frac{1}{2}$ or $s_{si} = 1$ respectively. Since the error rates are expected to be low at around 10%, we can reasonably assume $p_a = 0.1, p_b = 0.4, p_c = 0.6$ and $p_d = 0.9$.

$$P(p_{si}|y_{si}, n_{si}) = \frac{P(y_{si}|p_{si}, n_{si})P(p_{si}|n_{si})}{P(y_{si}|n_{si})}$$

$$= \frac{\binom{n_{si}}{y_{si}} {p_{si}}^{y_{si}}(1-p_{si})^{n_{si}-y_{si}}P(p_{si})}{P(y_{si}|n_{si})}.$$

Since the distribution function $P(y_{si}|n_{si})$ and binomial coefficient $\binom{n_{si}}{y_{si}}$ remain the same, $\mathrm{P}(p_{si}|y_{si}, n_{si}) \propto {p_{si}}^{y_{si}}(1 - p_{si})^{n_{si}-y_{si}}P(p_{si})$. Here we only need to maximize $f(p_{si}) = {p_{si}}^{y_{si}}(1 - p_{si})^{n_{si}-y_{si}}P(p_{si})$. However, we don't know $P(p_{si})$, which is the overall distribution of methylation status of all CpG sites. We propose first to use one-third rule and estimate $P(p_{si})$ with this result, then use this estimated $\hat{P}(p_{si})$ to maximize $f(p_{si})$ under three restrictions.

We denote the observed methylation level $\hat{p}_{si} = y_{si}/n_{si}$, and let $\hat{P}(s_{si} = 0) = \hat{\pi}_s^{(0)}$, $\hat{P}(s_{si} = 1/2) = \hat{\pi}_s^{(1/2)}$ and $\hat{P}(s_{si} = 1) = \hat{\pi}_s^{(1)}$. Since $f(p_{si}) = {p_{si}}^{y_{si}}(1 - p_{si})^{n_{si}-y_{si}}P(p_{si})$ is convex function on p_{si}, we can find in which restricted region the maximum of this likelihood function falls based on $\hat{p}_{si}$ as follows:

(1) If $\hat{p}_{si} \le p_a$, $f(\hat{p}_{si})$ is the maximum for restricted region $p_{si} \le p_a$; $f(p_b)$ is the maximum for region $p_b < p_{si} < p_c$; $f(p_d)$ is the maximum for region $p_{si} \ge p_d$. We assign s_{si} according to the region of p_{si} which produced the maximized $f(p_{si})$.

(2) If $p_b \le \hat{p}_{si} \le p_c$, $f(p_a)$ is the maximum for restricted region $p_{si} \le p_a$; $f(\hat{p}_{si})$is the maximum for region $p_b < p_{si} < p_c$; $f(p_d)$is the maximum for region $p_{si} \ge p_d$. We assign s_{si} according to the region of p_{si} which produced the maximized $f(p_{si})$.

(3) If $\hat{p}_{si} \ge p_d$, $f(p_a)$ is the maximum for restricted region $p_{si} \le p_a$; $f(p_c)$ is the maximum for region $p_b < p_{si} < p_c$; $f(\hat{p}_{si})$is the maximum for region $p_{si} \ge p_d$. We assign s_{si} according to the region of p_{si} which produced the maximized $f(p_{si})$.

(4) If $p_a < \hat{p}_{si} < p_b$, $f(p_a)$ is the maximum for restricted region $p_{si} \le p_a$; $f(p_b)$ is the maximum for region $p_b < p_{si} < p_c$; $f(p_d)$is the maximum for region $p_{si} \ge p_d$. We assign s_{si} according to the region of p_{si} which produced the maximized $f(p_{si})$.

(5) If $p_b < \hat{p}_{si} < p_c$, $f(p_a)$ is the maximum for restricted region $p_{si} \le p_a$; $f(p_c)$ is the maximum for region $p_b < p_{si} < p_c$; $f(\hat{p}_{si})$ is the maximum for region $p_{si} \ge p_d$. We assign s_{si} according to the region of p_{si} which produced the maximized $f(p_{si})$.

2.2 EVALUATION CRITERIA

In the simulation section, we will evaluate the accuracy, precision and bias of several methods by comparison. Accuracy and precision are compared between the Bayes' method and the 1/3 rule; bias is compared among the Bayes' method, the 1/3 rule, the observed methylation status, and the proportion of methylated reads from all observed reads without considering the single cell information as seen in bulk cell DNA methylation analysis. Accuracy is obtained from the proportions of correct predictions for the three true methylation status groups, and for overall methylation levels of a region. Precision is calculated as the standard deviations of predicted methylation regions. To evaluate bias, we compute the average of differences between mean of predicted data and the real global methylation level for the region: $\pi_s^{(1/2)}/2 + \pi_s^{(1)}$.

3.0 SIMULATION

3.1 SIMULATION SETTING

We simulate the data for K number of CpG sites using the following steps:

1. Assign the overall methylation levels by selecting the proportions of unmethylated, 50% methylated and methylated CpG sites as ($\pi_s^{(0)}$, $\pi_s^{(1/2)}$ and $\pi_s^{(1)}$). In this simulation study, we consider three scenarios: (1) $\pi_s^{(0)}$ = 0.4, $\pi_s^{(1/2)}$ = 0.1, and $\pi_s^{(1)}$ = 0.5; (2) $\pi_s^{(0)}$ = 0.3, $\pi_s^{(1/2)}$ = 0.2, and (3) $\pi_s^{(1)}$ = 0.5; $\pi_s^{(0)}$ = 0.2, $\pi_s^{(1/2)}$ = 0.1, and $\pi_s^{(1)}$ = 0.7. Three scenarios are related to 55%, 60% and 75% overall methylation levels respectively.
2. For each CpG site, we simulated total number of reads n uniformly 2 to 31.
3. Then randomly assign the methylation status for each CpG sites with proportion ($\pi_s^{(0)}$, $\pi_s^{(1/2)}$ and $\pi_s^{(1)}$). For CpG sites, assigned p_{si} as the true proportion of methylated reads by considering the error rates: $p_{si} \sim Unif(0, p_a)$ if $s_{si} = 0$; $p_{si} \sim Unif(p_b, p_c)$ if $s_{si} = 1/2$; $p_{si} \sim Unif(p_d, 1)$ if $s_{si} = 1$.
4. Since CG-contents may affect the overall reads, we simulated total number of reads n uniformly 2 to 41 if $s_{si} = \frac{1}{2}$ or 1, while sampled from 2 to 31 for CpG sites with $s_{si} = 0$.

5. Simulate the observed methylated reads y_{si} at each CpG site from a binomial distribution, i.e. $y_{si}|p_{si} \sim Bin(n_{si}, p_{si})$.
6. Apply the 1/3 rule on $\hat{p}_{si} = y_{si}/n_{si}$, and obtain the hypothesized methylation level for later calculation. Similarly, our Bayes' Rule method is applied on $\hat{p}_{si}$ to get the hypothesized methylation statuses.
7. Repeat the settings for 100 times and took the averaged results.

3.2 SIMULATION RESULTS

3.2.1 DNA METHYLATION DETECTION ACCURACY

Table 1 shows the comparison between the Bayes' method and the 1/3 rule about the accuracy of DNA methylation level prediction under sample size of $K = 100{,}000$. The comparison is conducted under 2 scenarios of the overall distribution methylation status on all CpG sites ($\pi_s^{(0)}$, $\pi_s^{(1/2)}$ and $\pi_s^{(1)}$): Scenario I is $\pi_s^{(0)}$ = 0.4, $\pi_s^{(1/2)}$ = 0.1, and $\pi_s^{(1)}$ = 0.5; Scenario II is $\pi_s^{(0)}$ = 0.3, $\pi_s^{(1/2)}$ = 0.2, and $\pi_s^{(1)}$ =0.5; Scenario III is $\pi_s^{(0)}$ = 0.2, $\pi_s^{(1/2)}$ = 0.1, and $\pi_s^{(1)}$ =0.7.

As we can see from Table 1, both methods capture correct methylation predictions well. Generally, Bayes' method outperforms 1/3 rule in terms of higher correct proportions at the global level. Bayes' method shows higher accuracy at underlying methylation levels of 0.5 and 1, whereas 1/3 rule performs better at s_{si} = 0.

Table 1 Correct call proportions of the 2 methods with sample size 100,000 under 3 scenarios

	Scenario I	Scenario II	Scenario III
Bayes' method			
s_{si} = 0	0.993	0.986	0.989
s_{si} = 0.5	0.741	0.809	0.748
s_{si} = 1	0.997	0.995	0.998
Global	0.970	0.955	0.971
One-third rule			
s_{si} = 0	0.994	0.994	0.994
s_{si} = 0.5	0.728	0.727	0.727
s_{si} = 1	0.991	0.991	0.991
Global	0.966	0.939	0.965

3.2.2 DNA METHYLATION DETECTION PRECISION

The standard errors of these 2 methods under different sample regions R and different scenarios are shown in Table 2. The 2 method show similar results with Bayes' method acting slightly better under Scenario II and Scenario III.

Table 2 Standard deviations of the 2 methods with different sample regions $ under 2 scenarios

		R=50	R=100	R=200
Scenario I	**Bayes' method**	0.334	0.105	0.034
	One-third rule	0.334	0.105	0.034
Scenario II	**Bayes' method**	0.311	0.098	0.032
	One-third rule	0.316	0.100	0.032
Scenario III	**Bayes' method**	0.287	0.092	0.028
	One-third rule	0.289	0.093	0.029

3.2.3 DNA METHYLATION DETECTION BIAS

The comparison of bias among all the 4 methods under different sample sizes K and scenarios is shown in Table 3. In Scenario I and Scenario II, the Bayes' method outperformed others; Bayes' method and 1/3 rule produced less bias than the other two methods. In Scenario III, the method of $\sum \boldsymbol{y}_{si} / \sum \boldsymbol{n}_{si}$ produced the smallest bias with the second smallest obtained from the Bayes' method. It might be due to a high global methylation level in this scenario.

Table 3 Biases produced by the 4 methods with different sample sizes K

	K=100	K=1000	K=10,000	K=100,000
Scenario I				
Bayes' method	-0.017	-0.016	-0.015	-0.016
One-third rule	-0.022	-0.022	-0.021	-0.022
$\hat{\boldsymbol{p}}_{si}$	-0.023	-0.025	-0.023	-0.024
$\sum \boldsymbol{y}_{si} / \sum \boldsymbol{n}_{si}$	0.030	0.029	0.030	0.030
Scenario I				
Bayes' method	-0.008	-0.011	-0.010	-0.010
One-third rule	-0.019	-0.022	-0.021	-0.021
$\hat{\boldsymbol{p}}_{si}$	-0.023	-0.026	-0.025	-0.025
$\sum \boldsymbol{y}_{si} / \sum \boldsymbol{n}_{si}$	0.023	0.022	0.023	0.022
Scenario III				
Bayes' method	-0.007	-0.009	-0.009	-0.009
One-third rule	-0.017	-0.019	-0.019	-0.019
$\hat{\boldsymbol{p}}_{si}$	-0.038	-0.039	-0.039	-0.039
$\sum \boldsymbol{y}_{si} / \sum \boldsymbol{n}_{si}$	0.001	0.0001	0.001	0.001

3.3 CONCLUSTIONS OF SIMULATIONS

From the simulation results from different situations, we can see that the biggest advantage of using the Bayes' method is a more accurate methylation level prediction. Producing a less biased global mean is also a strength of this method compared to the one-third rule. Its benefit in the terms of precision is nuanced. Thus, overall, the Bayes' method has a better performance than the one-third rule.

4.0 APPLICATION IN REAL DATA ANALYSIS

4.1 DATA

The data set is obtained from (Smallwood et al., 2014), which consists of 12 ovulated metaphase II oocytes (MII), 12 mouse embryonic stem cells (ESCs) cultured in two kinase inhibitors (2i), 20 ESCs cultured in serum conditions (Ser) and 7 negative control sc-BS-seq libraries. We selected the promoter regions among the dataset for analysis. Two kinase inhibitors have effect on reducing overall methylation levels.

We plan to study the methylation differences between different types of cells. Figure 2 shows an overall methylation level in promoter regions. The methylation levels from ESC Ser are higher compare to those from ESC 2i, showing the global hypomethylation for the cells in ESC 2i group (Figure 2 left panel). Another interesting finding is that methylation levels are very homogenous in MII group, which demonstrated that the methylation levels are changing heterogeneously with the development of cells. The control samples have very shallow sequencing depth compared to other groups with less than 1% of CpG sites covered.

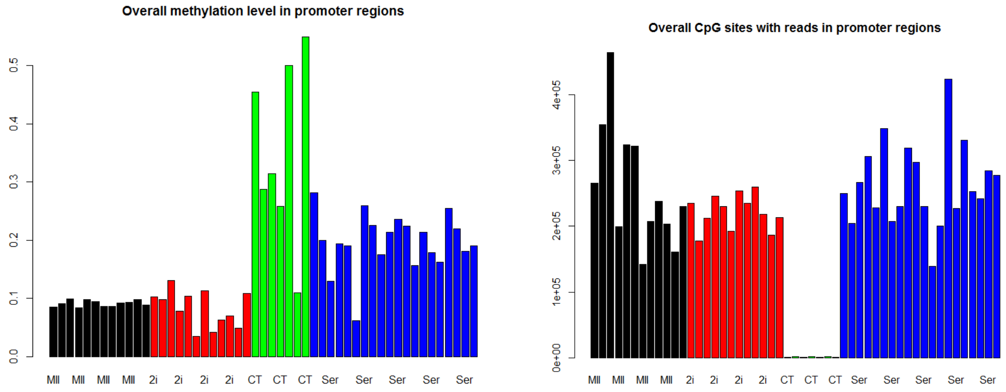

Figure 2 Overall methylation levels and total number of CpG sites in promoter regions

Methylation levels in ESC 2i samples are hypomethylated compared to those ESC Ser samples (left). The controls have very low coverage with less than 1% covered CpG sites (right).

4.2 ERROR RATES ESTIMATION

Ovulated metaphase II oocytes (MII) are haploid in which only one set of chomosome exists in these cells. Therefore, the methylation status for each of CpG sites can only be 0 or 100%. To estimate the error rates of methylation status, we study this haploid samples. We collect CpG sites with 2, 3, … and so on number of reads. CpG sites with most reads covered is 270. For each group of CpG sites, the status should be either T (unmethylated) or C (methylated) since all reads are from the same location. We assume the small number of Ts or Cs are due to errors and sum them together, then calculate the proportion of error reads among the total reads collected. Figure 3 shows estimated error rates vesus number of CpG sites covered by sequencing reads. For cells MII 1 – 12 with shallow sequencing depth, the error rate is increasing then decreasing over the number of reads covering the CpG sites. This is probably due to shallow sequencing depth, i.e., there are not enough CpG sites with large reads covered. However, when examing two cells with deeper sequencing depth, the error rates increase and then stay at around 10%, indicating the error rate is roughly at 10%. One panel (bottom left) is data for mixture of 120 bulk cells. Since methylation status in each cell is not homogeneous, it is expected to have larger estimated error rates because sequncing reads from different alleles may have different methylation status.

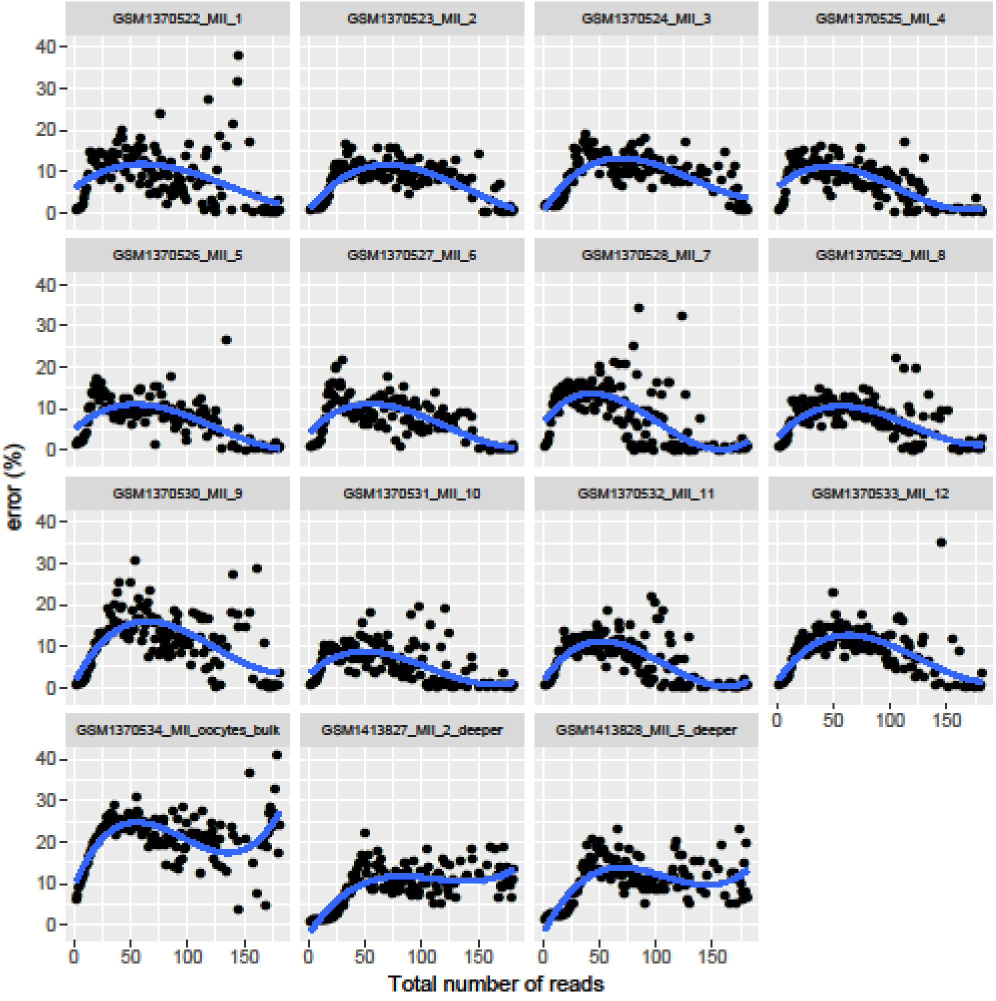

Figure 3 Estimated error rates of methylation status among MII cells

There are two cells (MII 2 and MII 5) with deeper sequencing depths. From these two deeper sequenced cells, we can roughly estimate that the error rate for methylation status call is about 10%.

4.3 APPLICATION IN DIFFERENTIALLY METHYLATION ANALYSIS

We then applied our method as data preprocessing in a differentially methylation analysis. After calling methylation status with our Bayes' method for CpG sites with multiple reads, we applied Beta-Binomial model in each promoter region to identify the differentially methylated ones. With clustering analysis, the differentially methylated promoter regions were clustered into groups of MII, ESC 2i and ESC Ser according to their similarity. Dendrogram was used to visualize the results (Figure 4). We can see that the cell types were segregated very well with only two from ESC Ser are misclassified as ESC 2i.

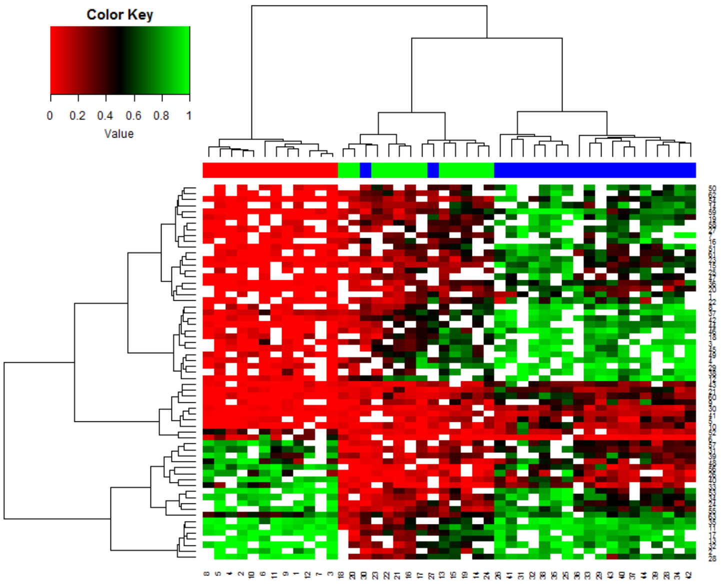

Figure 4 A two way hierarchical cluster analysis of the relative methylation of promoter regions

(Rows) measured on samples from 3 types of cells. Cell samples are identified on the top horizontal axis as red boxes (MII), green boxes (2i) and blue boxes (Ser). The methylation status of each region within each sample is presented in the image plot with values ranging from zero (red) to one (green; see color key).

5.0 DISCUSSION AND CONCLUSION

In this thesis, we proposed a method based on Bayes rule to improve the accuracy on estimation of methylation status for single cell Bisulfite sequencing data. Simulation result shows that this method has better accuracy of methylation status calling when compared to one-third rule. In terms of comparing the methylation level estimation for regions with different sizes (number of CpG sites), this proposed method has smaller bias and standard deviation compared to one-third rule method and the method to directly calculate from the observed methylation levels. The improvement in accuracy is the primary advantage of this method.

In real data application, this method segregated cell types accurately as a preprocessing method in differential methylation analysis. In the future, this method can be applied into the imputation of unobserved methylation status.

One limitation of our study is that the proposed method did not outperform other methods too much in reducing the bias obviously and improving efficiency. The benefit of accuracy in methylation status calling is consistent under various situations and discovering more other good features is in need for further study.

BIBLIOGRAPHY

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev, 16*(1), 6-21. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11782440. doi:10.1101/gad.947102

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schonegger, A., Klughammer, J., & Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep, 10*(8), 1386-1397. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25732828. doi:10.1016/j.celrep.2015.02.001

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet, 13*(7), 484-492. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22641018. doi:10.1038/nrg3230

Kriaucionis, S., & Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science, 324*(5929), 929-930. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19372393. doi:10.1126/science.1169786

Kulis, M., & Esteller, M. (2010). DNA methylation and cancer. *Adv Genet, 70*, 27-56. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20920744. doi:10.1016/B978-0-12-380866-0.60002-2

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., . . . Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature, 454*(7205), 766-770. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18600261. doi:10.1038/nature07107

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., . . . Reik, W. (2012). The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell, 48*(6), 849-862. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23219530. doi:10.1016/j.molcel.2012.11.001

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., . . . Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods, 11*(8), 817-820. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25042786. doi:10.1038/nmeth.3035

Szulwach, K. E., Li, X., Li, Y., Song, C. X., Han, J. W., Kim, S., . . . Jin, P. (2011). Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet, 7*(6), e1002154. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21731508. doi:10.1371/journal.pgen.1002154

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., . . . Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res, 20*(4), 440-446. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20219944. doi:10.1101/gr.103606.109