Research Article

Data-Driven-Based Approach to Identifying Differentially Methylated Regions Using Modified 1D Ising Model

Yuanyuan Zhang [ID],[1] **Shudong Wang,**[2] **and Xinzeng Wang** [ID][3]

[1]*School of Information and Control Engineering, Qingdao University of Technology, Qingdao, Shandong, China*
[2]*College of Computer and Communication Engineering, China University of Petroleum (East China), Qingdao, Shandong, China*
[3]*College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, Shandong, China*

Correspondence should be addressed to Yuanyuan Zhang; yyzhang1217@163.com and Xinzeng Wang; wangelxz@126.com

Background. DNA methylation is essential for regulating gene expression, and the changes of DNA methylation status are commonly discovered in disease. Therefore, identification of differentially methylation patterns, especially differentially methylated regions (DMRs), in two different groups is important for understanding the mechanism of complex diseases. Few tools exist for DMR identification through considering features of methylation data, but there is no comprehensive integration of the characteristics of DNA methylation data in current methods. *Results.* Accounting for the characteristics of methylation data, such as the correlation characteristics of neighboring CpG sites and the high heterogeneity of DNA methylation data, we propose a data-driven approach for DMR identification through evaluating the energy of single site using modified 1D Ising model. Applied to both simulated and publicly available datasets, our approach is compared with other popular methods in terms of performance. Simulated results show that our method is more sensitive than competing methods. Applied to the real data, our method can identify more common DMRs than DMRcate, ProbeLasso, and Wang's methods with a high overlapping ratio. Also, the necessity of integrating the heterogeneity and correlation characteristics in identifying DMR is shown through comparing results with only considering mean or variance signals and without considering relationship of neighboring CpG sites, respectively. Through analyzing the number of DMRs identified in real data located in different genomic regions, we find that about 90% DMRs are located in CGI which always regulates the expression of genes. It may help us understand the functional effect of DNA methylation on disease.

1. Introduction

DNA methylation is an important epigenetic modification which plays an essential role in gene expression [1, 2] and cancers [3–5]. Aberrant methylation status, such as hypermethylation in promoter, often leads to gene silencing. It is an important mechanism of antioncogene inactivation [6]. Global hypomethylation always leads to the emergence of cancers through affecting the stability of chromatin [7]. There are pieces of evidence showing that abnormal methylation patterns are related to many cancers and other diseases [8–13]. Also, some genomic regions have been found instable in methylation, which increases methylation variability in cancer and then causes cancer heterogeneity [14–16]. Therefore, identification of aberrant methylation patterns is important to understand the pathogenesis of diseases.

With the development of high-throughput technologies, there are two main technologies to quantify genome-wide DNA methylation, bisulfite microarray, and sequencing which provide great opportunities for revealing the epigenetic mechanisms of diseases. Array technologies, Illumina Infinium HumanMethylation 27 K and 450 K, are often used to study complex diseases owing to their low cost and high-resolution ratio popularly. There are two designs of data form, β-values and M-values, used in identifying aberrant methylated patterns. β-value measures the proportion of methylated intensities out of total intensities, and M-value is calculated as the log2 ratio of the intensities of methylated

probe versus unmethylated probe. The relationship between β-value and M-value is shown as the following equation:

$$M = \log_2\left(\frac{\beta}{1-\beta}\right). \tag{1}$$

It is shown that M-value may have more statistically valid for differential analysis of methylation [17].

Nowadays, various approaches have been proposed on the basis of microarray data to extract DNA methylation patterns, including differentially methylated loci [18–26] and differentially methylated region (DMR) [27–36]. Existing DMR detecting methods always consider some data characteristics to develop different assumptions. Bump hunter [27], DMRcate [28], and ProbeLasso [29] were developed through hypothesizing that the mean difference in methylation level of different groups is a primary cause in DMR identification; therefore, DMRs are identified through considering difference of mean signal between normal and cancer samples. Considering the heterogeneity of cancer samples [37], Wang et al. [30] developed an approach based on integrating mean and variance signal to identify DMRs. It is noteworthy that DMR methods integrating more information, such as mean and variance signals, always have better performance than those integrating less information, such as considering only mean signal [30]. Therefore, considering that additional characteristic, highly correlated neighboring CpG sites [38] in methylation data are rarely integrated in existing DMR methods; a data-driven method is developed based on integrating more information of data.

Motivated by Ising model which describes matter phase transition considering the strong interaction among neighbor molecular, we consider DNA methylation in genome by 1D (one-dimension) Ising model which can integrate the effect of neighbor sites. For each site, the status depends on its differential significance (p-value) and that of their neighbors with correlation characteristic. Generally, if the status of the site is significant, the more the accordance of the neighbor sites with the site, the lower the energy. If the status of the site is nonsignificant and those of its neighbors are significant, there are strong correlation between the site and its neighbors; we think that the site may also have low energy by integrating all information. The reason is that methylation level of a site is affected by multiple factors expect for disease; the information of neighbor sites can amend the bias of the site caused by other confounders. DMRs are identified as regions with low energy.

2. Material and Methods

We develop a data-driven approach to detecting DMRs (see Figure 1) which considers the data characteristics, correlation of neighboring CpG sites, and the high heterogeneity.

Step 1 (calculate site-level energy). Motivated by the principle of 1D Ising model, we define the site-level energy as follows:

$$e_i^f = -\sum_{j\neq i}\left|J_{ij}\right|\cdot\left(s_i^f + s_j^f\right),$$

$$s_i^f = \begin{cases} -\log\left(pvalue_i^f\right), & pvalue_i^f \leq 0.05 \\ 0, & pvalue_i^f \geq 0.05, \end{cases} \tag{2}$$

where e_i^f represents the energy of site i in feature f; J_{ij} represents the correlation of sites i and j in normal samples; and s_i^f describes the signal value of i which represents the difference between tumor and normal samples in feature f. $pvalue_i^f$ in function (2) is used to describe whether the site i is significantly different between two groups under the feature f. Here, if p-value is less than 0.05, we believe that this site can provide energy to distinguish the two types of samples under this feature. Otherwise, no energy is provided. The smaller the p-value is, the more the energy it provides. Therefore, we use the negative log of $pvalue_i^f$ for energy when p-value is less than 0.05; otherwise, zero is used.

$pvalue_i^f$ is calculated using two paired t-tests and one-sided *Pitman-Morgan* test to describe the mean and variance signals, respectively, which are often used in other works [30]. Considering high heterogeneity of methylation in cancers, we integrate the mean and variance signal to define site-level energy by parameter λ; that is, $f \in \{mean, variance\}$. For mean and variance signals, the statuses are denoted by s_i^{mean} and s_i^{var}, respectively. Therefore, the site-level energy is denoted as follows:

$$e_i = \lambda \cdot e_i^{mean} + (1-\lambda)\cdot e_i^{var}, \tag{3}$$

where

$$e_i^{mean} = -\sum_{j\neq i}\left|J_{ij}\right|\cdot\left(s_i^{mean} + s_j^{mean}\right)$$

$$e_i^{var} = -\sum_{j\neq i}\left|J_{ij}\right|\cdot\left(s_i^{var} + s_j^{var}\right) \tag{4}$$

$$\lambda = \frac{e_i^{mean}}{e_i^{mean} + e_i^{var}}.$$

λ represents the weight of mean signal to total signals. For each site, the lower the energy is, the more possible the difference site between case and control is.

Step 2 (identify candidate DMRs). To identify candidate DMRs, we define the total energy E of region R_k as follows:

$$E_{R_k} = \sum_{i\in R_k} e_i. \tag{5}$$

We use a greedy algorithm with the following steps to identify candidate DMRs. Considering the question of what conditions of a site are required to add candidate regions, we use a permuted method. First, we permute the sample labels n times and calculate the permuted energy of each site using (3). Second, permuted energy of all sites is sequenced in ascending order and the value at 5% is selected as the threshold τ. Therefore, the greedy algorithm is executed in
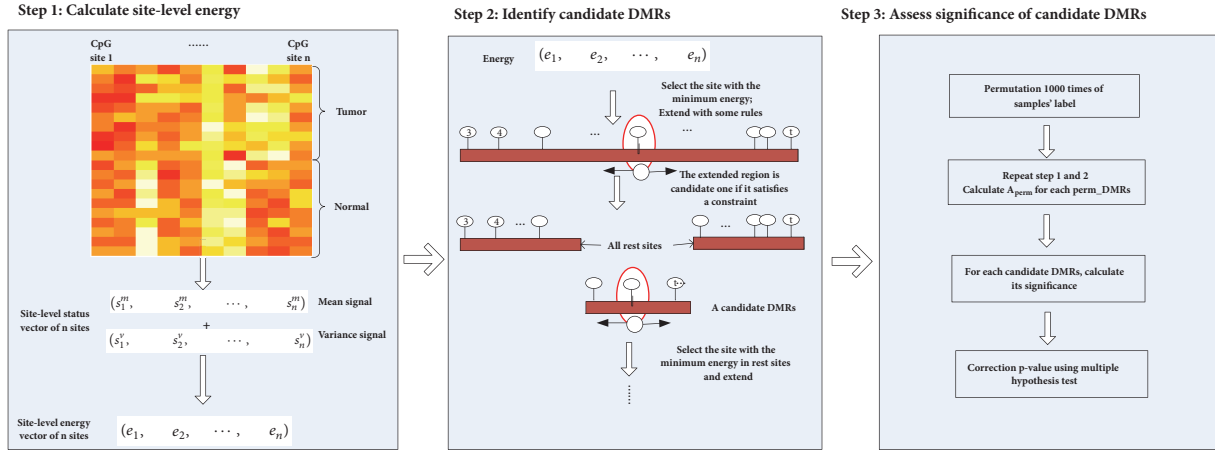
FIGURE 1: Flowchart of the proposed approach. Step 1, single site-level energy is calculated based on modified 1D Ising model. Step 2, candidate DMRs are identified using a greedy algorithm. Step 3, for each candidate DMR, the significance is assessed through permuting the sample labels.

the following steps: (1) Select the CpG site with the lowest energy as a seed site if its energy is smaller than τ, which is considered as the starting regions R_k. (2) Select the neighbor site of the current region with the lowest energy and add the site to R_k if energy of the site is smaller than τ. (3) Continue with the neighbors of R_k and keep adding the site to R_k until energy of the neighbor sites of R_k is greater than or equal to τ. (4) Choose another site with a seed one in the remaining sites except for R_k and repeat the above steps which can obtain another candidate region until the energy of any site of the rest is greater than τ.

Step 3 (assess significance of candidate DMRs). To assess significance of candidate DMRs, we need to calculate p-value for each candidate region. We make the null hypothesis that a candidate region is not a DMR. If p-value is less than a significance level, the null hypothesis is rejected. The identification of DMR is equivalent to determining whether a candidate region is associated with the sample label. Therefore, p-value of a candidate region is calculated through permuting sample label. We complete the process of Step 2 for each permutation. For the t-th permutation, we can obtain n_t permuted DMRs. The energy of permuted DMR R_{perm_i} is denoted by $E_{perm_t,i}$, $i = 1, \ldots, n_t$. For each candidate DMR R_k, the significance is measured by the p-value which is calculated as

$$p = \frac{\sum_{t=1}^{1000} \sum_{i=1}^{n_t} I\left(E_{perm_t,i}/size\left(R_{perm_i}\right) < E_{R_k}/size\left(R_k\right)\right)}{\sum_{t=1}^{1000} n_t}, \quad (6)$$

where $I(x)$ is an indicator function and $size(R)$ represents the numbers of CpG sites in region R. To consider the multiple testing, we use a function p.adjust in R and compare the results of different parameters; the results obtained with Bonferroni were the most conservative. Therefore, the p-values are multiplied by the number of comparisons using Bonferroni correction. The candidate DMR is considered

significant if the p-value corrected by Bonferroni is smaller than 0.05.

3. Results and Discussion

In calculating the energy of each site, we hypothesized that a locus was associated with only two sites adjacent to its left and right for simplifying the correlation characteristic. To show the performance of the proposed method, we compare the method with bump hunting [27], DMRcate [28], ProbeLasso [29], and Wang's method [30] in simulation data. Applying to the real data, we compare the DMRs identified by our method and Wang's method [30] based on β-value and M-value, respectively. Also, based on M-value, we compare our method with DMRcate [28] and ProbeLasso [29]. Finally, the necessity of integrating characteristics of methylation data is analyzed.

3.1. Simulation Study. We generate simulation data referencing Wang's method [30] which consider the real characteristics of methylation data based on M-value. For case-control design, like Wang's method, we consider methylation measure X following a conditional scaled normal distribution:

$$\begin{aligned} X \mid Y &= 1, \\ Z = z &\sim \sqrt{z} \cdot N\left(\mu, \Delta^T \Sigma \Delta\right), \\ X \mid Y &= 0, \\ Z = z &\sim \sqrt{z} \cdot N\left(0, \Sigma\right), \end{aligned} \quad (7)$$

where $z \sim Beta(1,1)$, $Y = 1$ and $Y = 0$ represent tumor and matched-control samples, respectively; the vector $\mu = (\mu_1, \mu_2, \ldots, \mu_h)^T$ and $\Delta = (v_1, v_2, \ldots, v_h)^T$ represent mean and variance signals, respectively; an element Σ_{ij} in matrix $\Sigma_{h \times h}$ is $\sigma \times \rho^{|i-j|}$ which describes the correlation characteristic of neighboring sites i and j; and h is the number of consecutive sites in a defined cluster. In each simulation, we generate 100

TABLE 1: Confusion matrix.

| | | Identified results of a method | |
		Number of differential methylated sites within DMRs	Number of non-differential methylated sites without in DMRs
Real DMRs	Number of differential methylated sites within DMRs	d True positive	c False negative
	Number of non-differential methylated sites without in DMRs	b False positive	a True negative

TABLE 2: Different parameter settings.

Parameters	μ_1	μ_2	ν	σ	ρ
1	-2	2	1.5	0.3	0.7
2	-2	2	2.5	0.3	0.7
3	-3	3	1.5	0.3	0.7
4	-3	3	2.5	0.3	0.7
5	-2	2	1.5	0.3	0.4

tumor and control samples with 10000 methylation sites. The genomic positions of the 10000 sites are simulated by that of the first 10000 chromosome 1 on 450k array [30]. The clusters are obtained based on the genomic position of sites using R package "bump hunter".

To show the performance of different methods, we select ten clusters randomly and set them as real DMRs as follows: for tumor samples, the first three are simulated mean signal only through setting μ_i in vector μ which follows the uniform distribution $unif(\mu_a, \mu_b)$; the next three are simulated variance signal only through setting $v_i = \alpha + \varepsilon$ for tumor samples, where α is a basic value and ε is a random value following $unif(0, 0.5)$; the last four are both simulated mean and variance signals through adding mean and variance signal in tumor samples by μ_i and v_i. The sensitivity (SE) and specificity (SP) are defined according to the confusion matrix shown in Table 1:

$$
\begin{aligned}
SE &= \frac{d}{c + d}, \\
SP &= \frac{a}{a + b}.
\end{aligned}
\tag{8}
$$

We compare the performance of our method with bump hunting [27] and Wang's method [30] in simulated data based on different values of different parameters (see Table 2). To improve the significance, we implement 10 times for each set of parameters and calculate the mean values of specificity and sensitivity. It is shown that our method has higher sensitivity than other methods (see Table 3).

Furthermore, to avoid the deviation of the numbers of identified DMRs, the numbers of true positive (TP) and false positive (FP) DMRs are compared through calculating the overlap between identified DMRs and embedded true DMRs (see Table 3).

In this study, an identified DMR is known as a true positive one if the intersection of the DMR and some true

DMR contains more than L CpG sites. L is defined by multiplying the length of the true DMR by θ. The greater the θ is, the higher the degree of overlap between the true positive DMR and the real area is. It is also shown that our method has higher matching degree with true DMRs when either $\theta = 0.2$ or $\theta = 0.5$. More simulations studies, changing simulation parameters, are described in supplementary material (see Tables S1-S3 for comprehensive analysis). It is shown that our method had better performance than other methods in identifying DMRs when there are high heterogeneity and correlation characteristics of methylation data.

3.2. Real Data Application. The real data we used is breast cancer (BRCA) which is available at TCGA. Preprocessing of DNA methylation data is implemented with reference to [35]. We permute sample labels 500 times and calculate the permuted energy for each site for each time. The value in descending order of 5% is obtained for each time and τ is obtained averaging these 500 values.

To compare the performance of our method, we apply the method and Wang's method to DNA methylation based on β-value and M-value, respectively, in our experiment. The thresholds τ are set as -2 and -5 for β-value and M-value data, respectively. The results show that our method can identify more DMRs based on either β-value or M-value data with a high overlapping ratio (see Table 4), especially that based on M-values. Therefore, we implement the next analysis based on the results of M-values data.

Based on M-value, we also compare our method with DMRcate [28] and ProbeLasso [29] and calculate the overlap of different methods (see Table 5). It is shown that our method has more common DMRs than any of the other methods.

Through analyzing the location in genome of these DMRs, we find that about 90% DMRs (7081 in 7871) are located in CGI which are CpG enrichment regions. This is consistent with the fact that aberrant methylation in CGI always influences the expression of genes. To understand the possible functions of DMRs, we evaluate the enrichment of these DMRs according to the location relative to genes (see Figure 2; numbers 1 to 6 represent different regions TSS1500, TSS200, 5'UTR, 1stExon, gene body, and 3'UTR, respectively). It is shown that most DMRs are enriched in gene body which is consistent with the recent studies reporting that aberrant methylation in gene body has essential role in cancer occurrence and development [39]. One example which is identified based on M-value but not for β-value is

TABLE 3: Comparison of different methods in different parameters.

Parameters	Bump hunting				DMRcate				ProbeLasso				Wang's method				Our method			
	SP	SE	No. TP (FP)[a]	No. TP (FP)[b]	SP	SE	No. TP (FP)[a]	No. TP (FP)[b]	SP	SE	No. TP (FP)[a]	No. TP (FP)[b]	SP	SE	No. TP (FP)[a]	No. TP (FP)[b]	SP	SE	No. TP (FP)[a]	No. TP (FP)[b]
1	0.99	0.12	0(1)	0(1)	1.00	0.25	2(0)	2(0)	1.00	0.16	3(1)	1(3)	0.99	0.53	6(1)	3(4)	**0.99**	**0.81**	**9(0)**	**7(2)**
2	1.00	0.23	2(0)	0(2)	1.00	0.25	2(0)	2(0)	1.00	0.12	2(1)	1(2)	0.99	0.38	5(0)	3(2)	**0.99**	**0.90**	**10(0)**	**10(0)**
3	1.00	0.15	0(2)	0(2)	1.00	0.25	2(0)	2(0)	1.00	0.16	3(1)	1(3)	1.00	0.59	9(0)	4(5)	**0.99**	**0.82**	**8(0)**	**6(2)**
4	1.00	0.26	1(1)	2(0)	1.00	0.25	2(0)	2(0)	1.00	0.16	3(1)	1(3)	1.00	0.52	5(0)	3(2)	**0.99**	**0.93**	**10(0)**	**10(0)**
5	1.00	0.18	3(0)	1(2)	1.00	0.25	2(0)	2(0)	1.00	0.12	2(1)	1(2)	1.00	0.45	8(1)	3(6)	**0.99**	**0.79**	**9(1)**	**7(3)**

[a] $\theta = 0.2$; [b] $\theta = 0.5$

TABLE 4: Comparison results of our method with Wang's method based on β-value and M-value.

	Our method	Wang's method	Overlapping ratio
β-value	2127	1618	78.9% (1276/1618)
M-value	7871	3070	93.0% (2856/3070)
Overlapping ratio	88.3% (1879/2127)	98.6% (1595/1618)	- - -

TABLE 5: Overlap results of different methods in identifying DMRs.

	Our	Wang's	DMRcate	ProbeLasso
ProbeLasso	**6720**	1566	2932	*9077*
DMRcate	**6097**	2579	*7577*	2932
Wang's	**2856**	*3070*	2579	1566
Our	*7871*	**2856**	**6097**	**6720**

∗Italic numbers indicate the numbers of DMRs identified by different methods. Black ones are overlap numbers of two methods.

Distribution of DMRs relative to genes
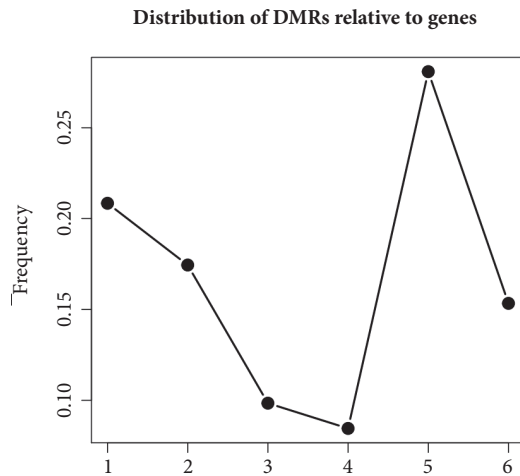
FIGURE 2: Distribution of DMRs relative to genes.

shown in Figure 3. It has obvious significant difference in variance signal but not in mean signal.

3.3. Validation of Identified DMRs. To illuminate the necessity of integrating the characteristics of DNA methylation data, we compare the results when only considering one signal (mean or variance) and without considering relationship of neighboring sites, respectively. Firstly, we calculate the energy of each site in (3) when $\lambda = 1$ (only mean signal) and when $\lambda = 0$ (only variance signal), respectively, and identify DMRs. Secondly, we do not consider the correlation of neighboring sites; the energy of each site is calculated when $e_i^{mean} = s_i^{mean}$ and $e_i^{var} = s_i^{var}$ in (3).

Take chromosome 1 as an example, the identified DMRs are shown in Table 6. It is shown that there is prominence in identifying DMRs when integrating mean, variance, and correlation characteristics more than when only considering variance signal ($\lambda=0$) and not considering correlation characteristic. Although there are more DMRs when $\lambda=1$ than those of integrated method, most of these DMRs contain fewer number of sites. Therefore, we think that the mean signal may be more effective in identifying differentially methylated loci than DMRs.

4. Conclusions

In this paper, we proposed a data-driven method to identify DMRs through integrating the characteristics of methylation data. Simulation study has shown that our method is more sensitive than the two alternative methods. Through application to real data, we compared the results of DMRs identified based on β-value and M-value, respectively, and our method showed further better performance than Wang's method. Based on M-value data, the necessity of integrating all characteristics of data is shown through comparing the DMRs identified by different measures. It is also shown that the integration of multiple information is effective in identifying DMR.

Currently, the available DMR identification methods are insufficient in integrating data characteristics. Most methods only consider mean signal, and high heterogeneity of methylation data is not considered. The recent work by Wang et al. is developed through accounting for high heterogeneity, and they obtained some meaningful results. Therefore, integrating more information to identify DMR is required.

Although our method integrates the characteristics of high heterogeneity and correlation of neighbor sites of methylation data, we only consider the correlation of the two neighbors of the site limited by the Ising model. Therefore, first, we hope to integrate more comprehensive information based on biological a priori knowledge to build an appropriate model in the further work; second, in view of the strong relationships of CpG sites, we hope to identify DNA methylation patterns based on building methylation network which has been widely used in identification of disease-related genes [40–43].

Data Availability

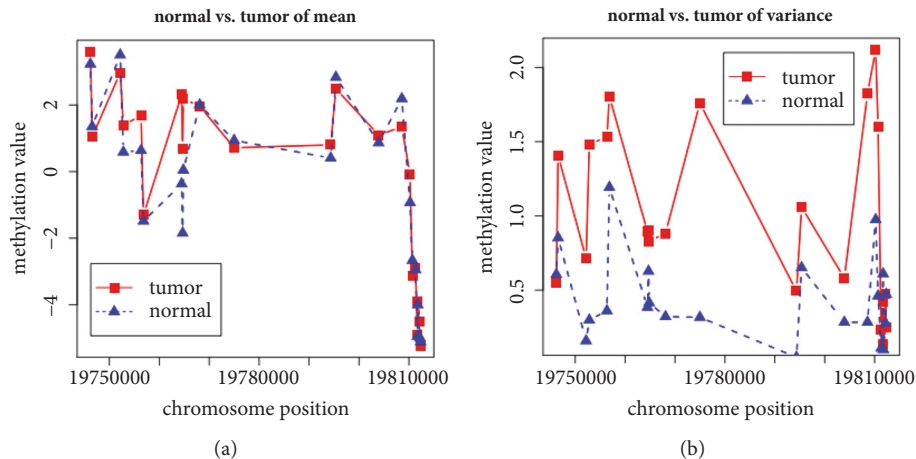The real data we used is breast cancer (BRCA) which is available at TCGA (https://cancergenome.nih.gov/).

FIGURE 3: An example of DMRs identified based on M-value but not for β-value. The difference of these two DMRs in mean signal (a) and variance signal (b) between normal and tumor samples.

TABLE 6: A Comparison of the results to show the necessity of integrating data characteristics.

	Integrated method	$\lambda = 1$	$\lambda = 0$	Without considering correlation
Numbers of DMRs	883	2374	370	34
Overlapping (a/b)*	100% (883/883)	64.5% (1532/2374)	100% (370/370)	94.1% (32/34)

∗a is the numbers of overlapping DMRs compared with integrated method (1532>883 means that more DMRs coincide with one DMR identified by integrated method.); b is the numbers of DMRs identified by the corresponding method.

Conflicts of Interest

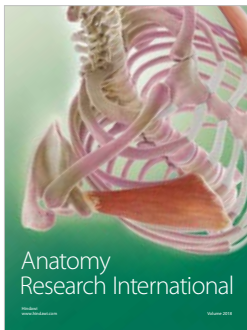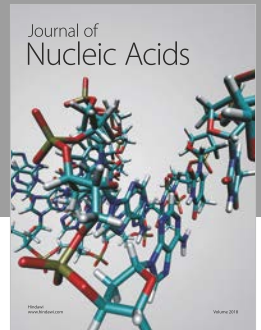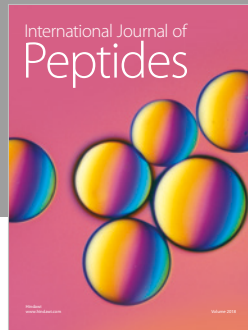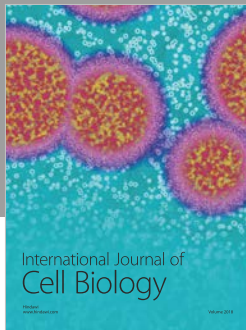There are no conflicts of interest to declare.

Supplementary Materials

More simulated studies through changing the values of parameters are described in Tables S1–S3. *(Supplementary Materials)*

References

[1] J. Plumbridge, "The role of dam methylation in controlling gene expression," *Biochimie*, vol. 69, no. 5, pp. 439–443, 1987.

[2] R. J. Shaw, G. L. Hall, D. Lowe et al., "The role of pyrosequencing in head and neck cancer epigenetics: Correlation of quantitative methylation data with gene expression," *Archives of Otolaryngology—Head and Neck Surgery*, vol. 134, no. 3, pp. 251–256, 2008.

[3] P. M. Das and R. Singal, "DNA methylation and cancer," *Journal of Clinical Oncology*, vol. 22, no. 22, pp. 4632–4642, 2004.

[4] M. Widschwendter, K. D. Siegmund, H. M. Müller et al., "Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen," *Cancer Research*, vol. 64, no. 11, pp. 3807–3813, 2004.

[5] M. Ehrlich, "DNA methylation in cancer: too much, but also too little," *Oncogene*, vol. 21, no. 35, pp. 5400–5413, 2002.

[6] J. G. Herman and S. B. Baylin, "Gene silencing in cancer in association with promoter hypermethylation," *The New England Journal of Medicine*, vol. 349, no. 21, pp. 2042–2054, 2003.

[7] A. Eden, F. Gaudet, A. Waghmare, and R. Jaenisch, "Chromosomal instability and tumors promoted by DNA hypomethylation," *Science*, vol. 300, no. 5618, p. 455, 2003.

[8] T. Hinoue, D. J. Weisenberger, C. P. E. Lange et al., "Genome-scale analysis of aberrant DNA methylation in colorectal cancer," *Genome Research*, vol. 22, no. 2, pp. 271–282, 2012.

[9] L.-C. Li, S. T. Okino, and R. Dahiya, "DNA methylation in prostate cancer," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1704, no. 2, pp. 87–102, 2004.

[10] H. M. Müller, L. Ivarsson, H. Schröcksnadel et al., "DNA methylation changes in sera of women in early pregnancy are similar to those in advanced breast cancer patients," *Clinical Chemistry*, vol. 50, no. 6, pp. 1065–1068, 2004.

[11] J. L. Fleming, C. J. Phiel, and A. E. Toland, "The role for oxidative stress in aberrant DNA methylation in Alzheimer's disease," *Current Alzheimer Research*, vol. 9, no. 9, pp. 1077–1096, 2012.

[12] L. J. Martin and M. Wong, "Aberrant Regulation of DNA Methylation in Amyotrophic Lateral Sclerosis: A New Target of Disease Mechanisms," *Neurotherapeutics*, vol. 10, no. 4, pp. 722–733, 2013.

[13] J. Jelinek, V. Gharibyan, M. R. H. Estecio et al., "Aberrant DNA methylation is associated with disease progression, resistance to imatinib and shortened survival in chronic myelogenous leukemia," *PLoS ONE*, vol. 6, no. 7, 2011.

[14] B. Phipson and A. Oshlack, "DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging," *Genome Biology*, vol. 15, no. 9, p. 465, 2014.

[15] K. D. Hansen, W. Timp, H. C. Bravo et al., "Increased methylation variation in epigenetic domains across cancer types," *Nature Genetics*, vol. 43, no. 8, pp. 768–775, 2011.

[16] A. E. Jaffe, A. P. Feinberg, R. A. Irizarry, and J. T. Leek, "Significance analysis and statistical dissection of variably methylated regions," *Biostatistics*, vol. 13, no. 1, pp. 166–178, 2012.

[17] P. Du, X. Zhang, C.-C. Huang et al., "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, article 587, 2010.

[18] D. Wang, L. Yan, Q. Hu et al., "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data," *Bioinformatics*, vol. 28, no. 5, pp. 729–730, 2012.

[19] S. Wang, "Method to detect differentially methylated loci with case-control designs using Illumina arrays," *Genetic Epidemiology*, vol. 35, no. 7, pp. 686–694, 2011.

[20] Z. Chen, Q. Liu, and S. Nadarajah, "A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data," *Bioinformatics*, vol. 28, no. 8, Article ID bts093, pp. 1109–1113, 2012.

[21] H. Huang, Z. Chen, and X. Huang, "Age-adjusted nonparametric detection of differential DNA methylation with case-control designs," *BMC Bioinformatics*, vol. 14, article no. 86, 2013.

[22] Y. Zhang, J. Zhang, and J. Shang, "Quantitative identification of differentially methylated loci based on relative entropy for matched case-control data," *Epigenomics*, vol. 5, no. 6, pp. 631–643, 2013.

[23] H. Sun and S. Wang, "Penalized logistic regression for high-dimensional DNA methylation data with case-control studies," *Bioinformatics*, vol. 28, no. 10, pp. 1368–1375, 2012.

[24] H. Sun, Y. Wang, Y. Chen, Y. Li, and S. Wang, "PETM: A penalized Exponential Tilt Model for analysis of correlated high-dimensional DNA methylation data," *Bioinformatics*, vol. 33, no. 12, pp. 1765–1772, 2017.

[25] Z. Chen, H. Huang, and Q. Liu, "Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data," *BMC Bioinformatics*, vol. 15, no. 1, article no. 142, 2014.

[26] Z. Chen, H. Huang, J. Liu et al., "Detecting differentially methylated loci for Illumina Array methylation data based on human ovarian cancer data," *BMC Medical Genomics*, vol. 6, supplement 1, no. 1, article no. S9, 2013.

[27] A. E. Jaffe, P. Murakami, H. Lee et al., "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International Journal of Epidemiology*, vol. 41, no. 1, Article ID dyr238, pp. 200–209, 2012.

[28] T. J. Peters, M. J. Buckley, A. L. Statham et al., "De novo identification of differentially methylated regions in the human genome," *Epigenetics & Chromatin*, vol. 8, no. 1, 2015.

[29] L. M. Butcher and S. Beck, "Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data," *Methods*, vol. 72, no. C, pp. 21–28, 2015.

[30] Y. Wang, A. E. Teschendorff, M. Widschwendter, and S. Wang, "Accounting for differential variability in detecting differentially methylated regions," *Briefings in Bioinformatics*, 2017.

[31] R. C. Slieker, S. D. Bos, J. J. Goeman et al., "Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array," *Epigenetics & Chromatin*, vol. 6, no. 1, 2013.

[32] T. Sofer, E. D. Schifano, J. A. Hoppin, L. Hou, and A. A. Baccarelli, "A-clustering: A novel method for the detection of co-regulated methylation regions, and regions associated with exposure," *Bioinformatics*, vol. 29, no. 22, pp. 2884–2891, 2013.

[33] M.-L. Ong and J. D. Holbrook, "Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways," *Aging Cell*, vol. 13, no. 1, pp. 142–155, 2014.

[34] C. Ruiz-Arenas and J. R. González, "Redundancy analysis allows improved detection of methylation changes in large genomic regions," *BMC Bioinformatics*, vol. 18, no. 1, 2017.

[35] Y. Zhang and J. Zhang, "Identification of functionally methylated regions based on discriminant analysis through integrating methylation and gene expression data," *Molecular BioSystems*, vol. 11, no. 7, pp. 1786–1793, 2015.

[36] R. Kolde, K. Märtens, K. Lokk, S. Laur, and J. Vilo, "Seqlm: An MDL based method for identifying differentially methylated regions in high density methylation array data," *Bioinformatics*, vol. 32, no. 17, pp. 2604–2610, 2016.

[37] P.-C. Tsai, T. D. Spector, and J. T. Bell, "Using epigenome-wide association scans of DNA methylation in age-related complex human traits," *Epigenomics*, vol. 4, no. 5, pp. 511–526, 2012.

[38] R. A. Irizarry, C. Ladd-Acosta, B. Carvalho et al., "Comprehensive high-throughput arrays for relative methylation (CHARM)," *Genome Research*, vol. 18, no. 5, pp. 780–790, 2008.

[39] P. A. Jones, "Functions of DNA methylation: islands, start sites, gene bodies and beyond," *Nature Reviews Genetics*, vol. 13, no. 7, pp. 484–492, 2012.

[40] Y. Zhang, M. Zhao, J. Su et al., "Novel Model for Cascading Failure Based on Degree Strength and Its Application in Directed Gene Logic Networks," *Computational & Mathematical Methods in Medicine*, vol. 2018, Article ID 8950794, 9 pages, 2018.

[41] D. K. Yip, L. L. Chan, I. K. Pang et al., "A network approach to exploring the functional basis of gene–gene epistatic interactions in disease susceptibility," *Bioinformatics*, vol. 34, no. 10, pp. 1741–1749, 2018.

[42] Y. Liu, M. Brossard, D. Roqueiro et al., "SigMod: An exact and efficient method to identify a strongly interconnected diseaseassociated module in a gene network," *Bioinformatics*, vol. 33, no. 10, pp. 1536–1544, 2017.

[43] A. Özgür, T. Vu, G. Erkan, and D. R. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277–i285, 2008.