

METHODOLOGY ARTICLE

Open Access

# Age-adjusted nonparametric detection of differential DNA methylation with case-control designs

Hanwen Huang<sup>1</sup>, Zhongxue Chen<sup>2\*</sup> and Xudong Huang<sup>3</sup>

## Abstract

**Background:** DNA methylation profiles differ among disease types and, therefore, can be used in disease diagnosis. In addition, large-scale whole genome DNA methylation data offer tremendous potential in understanding the role of DNA methylation in normal development and function. However, due to the unique feature of the methylation data, powerful and robust statistical methods are very limited in this area.

**Results:** In this paper, we proposed and examined a new statistical method to detect differentially methylated loci for case control designs that is fully nonparametric and does not depend on any assumption for the underlying distribution of the data. Moreover, the proposed method adjusts for the age effect that has been shown to be highly correlated with DNA methylation profiles. Using simulation studies and a real data application, we have demonstrated the advantages of our method over existing commonly used methods.

**Conclusions:** Compared to existing methods, our method improved the detection power for differentially methylated loci for case control designs and controlled the type I error well. Its applications are not limited to methylation data; it can be extended to many other case-control studies.

**Keywords:** Nonparametric method, One-sided test, Combining p-value

## Background

Essential for normal development and an influence on a variety of processes related to DNA integrity and function, DNA methylation plays an important role for epigenetic gene regulation in both development and disease [1]. It is associated with processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and carcinogenesis [2-4]. Aberrant DNA methylation, such as hypomethylation of oncogenes and hypermethylation of tumor suppressor genes, leads to genomic instability and tumorigenesis [5-10].

With the development of high-throughput platforms, genome-wide analysis of large-scale DNA methylation patterns and their impacts on gene regulation have received a significant amount of attention. We are proposing an age-adjusted nonparametric method to detect

differentially methylated loci that can account for age effects that has advantages over existing methods the limitations of which we explain next. Typically, methylation levels in Illumina methylation assays are quantified in terms of the  $\beta$ -value calculated from the intensity ratio of methylated (M) to unmethylated (U) alleles. Specifically,  $\beta = \frac{\max(M,0)}{\max(M,0)+\max(U,0)+100}$ , where M and U are the intensities of red and green dyes, respectively, for the GoldenGate and VeraCode Methylation assays, or the signals A and B, respectively, for the Illumina assay. The striking feature of the  $\beta$ -values is that they are continuous and range from 0 (totally unmethylated) to 1 (fully methylated).

With more and more DNA methylation data generated from the high-throughput DNA methylation platforms, powerful and efficient statistical methods to handle these complex data are becoming highly demanded. One of the important research topics in this area is to detect differentially methylated loci in case and control studies. The commonly used methods for this purpose are the

\* Correspondence: [zc3@indiana.edu](mailto:zc3@indiana.edu)

<sup>2</sup>Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, 1025 E. 7th Street, Bloomington, IN 47405, USA  
Full list of author information is available at the end of the article

Student's *t*-test and linear regression. Recently, a number of model-based approaches have been proposed in the literature. Siegmund [11] introduced a Bernoulli-lognormal mixture model to perform clustering analysis on methylation data generated using MethyLight. Houseman [12] proposed a  $\beta$ -mixture model to classify different tissue types using a recursive-partitioning algorithm for high-dimensional data from Illumina arrays. Wang [13] developed a likelihood based uniform-normal-mixture model to select differentially methylated loci between case and control groups using the Illumina array. The basic idea of Wang [13] is to describe the data using a three-component mixture model and treat completed methylated, unmethylated, and partially methylated loci differently. Wang [13] showed that, compared to the Student's *t*-test under some situations, their method increases detection power [13]. However, the aforementioned methods assume that the methylation profiles follow some special distributions that are known in terms of a finite number of parameters. Obviously if the underlying true distribution is far from the proposed ones, such assumptions will lead to biased results in practice.

Another complexity of the methylation study comes from the presence of other potential confounders such as age effects. As shown in [14-17], age is by far the strongest demographic risk factor for cancer, and there is substantial evidence that aging affects DNA methylation of specific loci, including cancer-related genes. Based on these observations, it is necessary to adjust age effects in the analysis of detecting differentially methylated loci. If the relationship between the methylation and age is more complex than a linear one, a traditional linear regression leads to inaccurate results. To solve this problem, Chen [16] proposed a method by first dividing the samples into several age groups and then combined the p-values obtained from each individual group together to form a new test. This method has been shown to increase the power in contrast to the traditional *t*-test without age adjustment or regression model with age adjusted linearly. However, within each group, a p-value is obtained from a simple *t*-test that is less robust and thus leaves room for improvement.

In this paper, we propose and examine a novel means to detecting differentially methylated loci and, that is, an age-adjusted nonparametric method that can account for age effects, given that substantial evidence exists that aging affects DNA methylation of specific loci, including cancer-related genes. Our method stems from the rank-based nonparametric methods that focus on a comparison of two entire empirical distribution functions rather than only two means. More specifically, we first divide the subjects into several age groups; then for each group, a nonparametric test is performed on each locus, and an individual p-value is reported. An overall p-value for each locus is estimated

through combining all the individual p-values computed previously for that locus in different age groups. Our method does not depend on any distribution assumption but rather adjusts for age effects in an efficient way. We demonstrate the powerfulness of our method using both simulated and real datasets.

## Methods

Assume all the subjects are from  $K$  different age groups. In the  $k^{\text{th}}$  group,  $k = 1, \dots, K$ , there are  $n_{1k}$  control subjects and  $n_{2k}$  case subjects. For each DNA methylation marker, let  $y_{ijk}$ ,  $i = 1, \dots, n_{jk}$ ,  $j = 1, 2$ ,  $k = 1, \dots, K$ , denote the observation of  $\beta$ -value for the  $i^{\text{th}}$  subject in  $j^{\text{th}}$  treatment (1 for control and 2 for case) and  $k^{\text{th}}$  age group. To adjust the age influence on the methylation level, the linear model takes the form

$$y_{ijk} = a_j * treatment_j + b_k * age_k + \varepsilon_{ijk},$$

$$for i = 1, \dots, n_{jk}, j = 1, 2, k = 1, \dots, K,$$

Where  $a_j$  and  $b_k$  are regression coefficients and  $\varepsilon_{ijk}$  follows a i.i.d normal distribution. The normality assumption is clearly invalid for the  $\beta$ -values which have limited range between 0 and 1. Moreover, the relationship between the  $\beta$ -value and age is likely to be more complicated than a linear one. Therefore more powerful and robust nonparametric methods are desirable.

Here we propose a new nonparametric method which does not depend on any distribution assumption and meanwhile allows for the adjustment of covariates. We process as follows. For each age group  $k$  ( $k = 1, \dots, K$ ), we test the difference between the control group  $y_{i1k}$  ( $i = 1, \dots, n_{1k}$ ) and the case group  $y_{i2k}$  ( $i = 1, \dots, n_{2k}$ ). Our goal is to test whether or not the two methylation groups follow the same distribution. Toward this end, the Wilcoxon rank sum test is a useful tool when there are reasons to believe that the outcome variables of interest may fail certain distributional assumptions required for parametric methods. However, as discussed in Baumgartner [18], Wilcoxon rank sum test is not suitable for situations where the expected values of the two populations are close to each other. To overcome this problem, they proposed a more powerful nonparametric test to handle the general two-sided two-sample problem [18]. Neuhaeuser further extended the two-sided two-sample test to a one-sided test that can detect if one population is stochastically larger than the other [19]. Our p-value calculation for each age group is based on Neuhaeuser's one-sided test whose statistics can be explicitly formulated as

$$B_k = \frac{1}{2}(B_{1k} - B_{2k}),$$

where

$$B_{1k} = \frac{1}{n_{1k}} \sum_{i=1}^{n_{1k}} \left( G_i - \frac{n_{1k} + n_{2k}}{n_{1k}} i \right) \left| G_i - \frac{n_{1k} + n_{2k}}{n_{1k}} i \right|$$

$$B_{2k} = \frac{1}{n_{2k}} \sum_{j=1}^{n_{2k}} \left( H_j - \frac{n_{1k} + n_{2k}}{n_{2k}} j \right) \left| H_j - \frac{n_{1k} + n_{2k}}{n_{2k}} j \right|$$

Here  $G_i, i = 1, 2, \dots, n_{1k}$  and  $H_j, j = 1, 2, \dots, n_{2k}$  are the ranks of the samples from the  $k^{\text{th}}$  case and control groups, respectively. Due to the intractable asymptotic distribution for the test statistics  $B$ , it is hard to find a close form for the relationship between p-value and  $B$ . We will use numerical fit to approximate this relationship. We first obtain the empirical distribution of  $B$  based on  $10^7$  permutations and then fit the distribution function piecewise exponentially to obtain the empirical relationship. The p-value for the  $k^{\text{th}}$  age group can be calculated according to this empirical formula.

As a consequence, we have  $K$  p-values from the left-sided test, denoted by  $p_{lk} (k = 1, \dots, K)$ , and  $K$  p-values from the right-sided test, denoted by  $p_{rk} = 1 - p_{lk} (k = 1, \dots, K)$ . Then combining  $K$  left-sided p-values together gives a statistic

$$T_l = -2 \sum_{k=1}^K \log(p_{lk})$$

Similarly, combining  $K$  right-sided p-values together gives a statistic

$$T_r = -2 \sum_{k=1}^K \log(p_{rk})$$

Under the null hypothesis of no difference between the two treatment groups,  $p_{lk}$  and  $p_{rk}$  are uniform  $[0,1]$  random variables for  $k = 1, \dots, K$ . Therefore, according to Fisher [20], both  $T_l$  and  $T_r$  will follow a chi-square distribution with degree of freedom  $2K$ . We define a new variable:

$$T = \max\{T_l, T_r\},$$

then we have [21]

$$2\alpha - \alpha^2 \leq \Pr(T > x) \leq 2\alpha, \text{ where } \alpha = 1 - F_{\chi_{2K}^2}(x), \text{ and } F_{\chi_{2K}^2} \text{ is the CDF of } \chi_{2K}^2.$$

Thus, for small  $\alpha$ , we can approximate the p-value of  $T$  by its upper bound  $2\alpha$  as

$$\Pr(T > x) \approx 2\alpha$$

More details can be found in [21-23]. For large  $\alpha$ , we will fit a formula empirically through permutation. We

call the above proposed method “combined Baumgartner-Weiß-Schindler (BWS) test”.

## Results and discussion

### Empirical fit of the p-value formula for one-sided BWS test

The asymptotic distribution function of  $B$  is complex and in practice permutation method is often used. The permutation results depend on the sample size. But as mentioned in Baumgartner [18], for a two-sided test, the asymptotic distribution can be approximated by the permutation method quite well even with a small sample size (as small as 10). We first derive the empirical formula to fit the asymptotic distribution using the permutation method. Toward this end, we sample two subpopulations from the same distribution (e.g. standard normal), each of which has a sample size of 30. Then, the whole populations are permuted  $10^7$  times, and a one-sided BWS test statistic  $B$  is calculated for each permuted sample. Then we fit the empirical cumulative distribution of  $B$  using a piecewise exponential function to arrive at the following empirical formula

$$P(B) = \begin{cases} e^{-0.699 - 1.255^*B} & (0 \leq B < 1.5) \\ e^{-0.895 - 1.153^*B + 0.0173^*B^2} & (1.5 \leq B < 9), \text{ and} \\ e^{-2.895 - 0.786^*B} & (B \geq 9) \end{cases}$$

$$P(-B) = 1 - P(B).$$

The node points we used here are 1.5 and 9. We find that the choice of node points has very little influence on the final analysis results for both simulated and real data. Note that the sample size we used for deriving this formula is 30. We have also tried some other choices and found that the results are quite similar. Thus, we recommend the above formula to be used in practice for problems with a sample size larger than 10.

### Simulation results

The first simulation settings are for the evaluation of the type I error rate for the proposed method. For the purposes of comparison, we also include the results from the combined  $t$ -test proposed in [16], linear regression and combined Wilcoxon methods for all simulated and real datasets. We assume that there are 6 age groups, and each group includes 100 subjects, 50 controls and 50 cases. For each age group, we also assume the  $\beta$ -values follow a three-component mixture distribution as in Wang [13]. Let  $\tau_1$  and  $\tau_2$  be the two threshold values. The first and the second components are uniform distributions  $U_{[0, \tau_1]}$  and  $U_{[\tau_2, 1]}$ . The third component is a truncated normal distribution  $N_{[\tau_1, \tau_2]}(\mu, \sigma^2)$ . The probabilities for a measurement to fall into these three components are  $\pi_1, \pi_2, \pi_3$  respectively. Under the null hypothesis, the two treatment groups are sampled from the same distribution. The mean of the truncated

**Table 1 Estimated Type I error rates at significant level 0.05 based on the four methods under different parameter settings of the three-component mixture distributions with  $\tau_1 = 0.1$ ,  $\tau_2 = 0.9$  and  $\delta_\mu = 0.05$**

Parameter Setting	$\pi_1$	$\pi_2$	$\pi_3$	$\mu$	$\sigma$	t-test	Linear regression	Wilcox	BWS
1	0.3	0.5	0.2	0.3	0.1	0.0513	0.0521	0.0514	0.0458
2	0.4	0.5	0.1	0.2	0.1	0.0495	0.0494	0.0519	0.0492
3	0.4	0.5	0.1	0.3	0.2	0.0511	0.0519	0.0503	0.0495
4	0.5	0.1	0.4	0.3	0.2	0.0528	0.0521	0.0544	0.0511
5	0.4	0.2	0.4	0.2	0.1	0.0509	0.0510	0.0472	0.0464

normal distribution is taken to be  $\mu + k\delta_\mu$  for the  $k^{\text{th}}$  age group. The simulation is repeated 10000 times. Table 1 lists the proportion of times that null is rejected using the four different methods under different parameter settings. Table 1 shows that the nominal type I error rate of 0.05 is well controlled by all methods.

The second simulation settings are for assessing the power of the proposed method under alternative hypothesis, i.e. the case and control subjects are sampled from different distributions. We still assume that the  $\beta$ -values follow the similar three-component mixture distributions. Two scenarios are considered here. In the first scenario, we use different means for different treatment groups. More specifically, we let the truncated normal mean for the control sample be constant  $\mu$ , but for the case sample vary as  $\mu + k\delta_\mu$  for the  $k^{\text{th}}$  age group. We replicate the simulation 10000 times for each scenario. The power is defined as the proportion of times that the p-value is less than 0.05. The first two rows in Table 2 list the results for this scenario. In the second scenario, we let the two threshold values vary as  $\tau_1 + k\delta_\tau$  and  $\tau_2 - k\delta_\tau$  for the  $k^{\text{th}}$  case group but keep them constants as  $\tau_1$  and  $\tau_2$  for all control groups. The results are listed in the last two rows of Table 2. For the first scenario, the mean values are different for the case and control groups. As expected, all four methods have increasing powers as the signal increases. For the second scenario, the expected values are the same, but the variances are different between the two treatment groups. In this situation, our proposed method is more powerful in detecting the difference than the other three methods. Therefore, the proposed method can detect not only the location difference but also the scale difference between the two distributions.

The third settings assume that the  $\beta$ -values for both the case and control subjects follow the beta-distributions. Let

$s_1 = s_2 = 4$ . For the case group, the  $\beta$ -values are sampled from a beta-distribution  $Beta(s_1 + \delta, s_2 - \delta)$ . For the control group, the  $\beta$ -values are sampled from a beta-distribution  $Beta(5s_1 + \delta, 5s_2 - \delta)$ . Here  $\delta$  takes six different values,  $-3\lambda$ ,  $-2\lambda$ ,  $-\lambda$ ,  $0.5\lambda$ ,  $\lambda$ , and  $1.5\lambda$ , one for each age group.

Based on the above setting, it can be easily shown that the mean of the distribution for the case group is  $\frac{s_1 + \delta}{s_1 + s_2}$  while the mean of the distribution for the control group is  $\frac{5s_1 + \delta}{5s_1 + 5s_2}$ . The mean difference between the two treatment groups is  $\frac{4\delta/5}{s_1 + s_2}$  which will increase with  $\delta$ . Table 3 lists the empirical powers of four methods for different  $\lambda$  values. In all situations, the most powerful method is combined DWS. Since the variance of two distributions are different, even in situation where  $\delta = 0$ , the power of the proposed method can still reach 0.67, whereas the other three methods have no power at all. As  $\lambda$  decreases, the mean differences become smaller and make it more difficult to distinguish between the two treatment groups. As expected, the powers decrease for all methods as  $\lambda$  decreases. For small  $\lambda$ , the power from the combined DWS method is much bigger than those from the other three methods. The performances of the combined t-test and combined Wilcoxon methods are quite similar, and both are much better than the linear regression method.

### Real data results

We also applied our proposed method to the United Kingdom Ovarian Cancer Population Study (UKOPS) [15] to select differentially methylated loci between ovarian cancer cases and healthy controls. The data were created using Illumina Infinium Human Methylation27 Beadchip and downloaded from the NCBI Gene

**Table 2 Estimated powers of the four methods at significant level 0.05 under different parameter settings for the three-component mixture distributions with  $\tau_1 = 0.1$ ,  $\tau_2 = 0.9$**

Parameter Setting	$\pi_1$	$\pi_2$	$\pi_3$	$\mu$	$\sigma$	t-test	Linear regression	Wilcox	BWS
$\delta_\mu = 0.03$	0.3	0.5	0.2	0.3	0.1	0.475	0.479	0.749	0.836
$\delta_\mu = 0.05$	0.3	0.5	0.2	0.3	0.1	0.889	0.892	0.951	0.988
$\delta_\tau = 0.03$	0.45	0.1	0.45	0.5	0.05	0.048	0.047	0.078	0.727
$\delta_\tau = 0.06$	0.45	0.1	0.45	0.5	0.05	0.048	0.047	0.092	0.877

**Table 3 Change of the power with  $\lambda$  for four different methods when the distributions are  $Beta(s_1 + \delta, s_2 - \delta)$  and  $Beta(5s_1 - \delta, 5s_2 + \delta)$  for the case and control groups;  $\delta$  takes the values of  $-3\lambda, -2\lambda, -\lambda, 0.5\lambda, \lambda,$  and  $1.5\lambda$  for the six age groups and  $s_1 = s_2 = 4$**

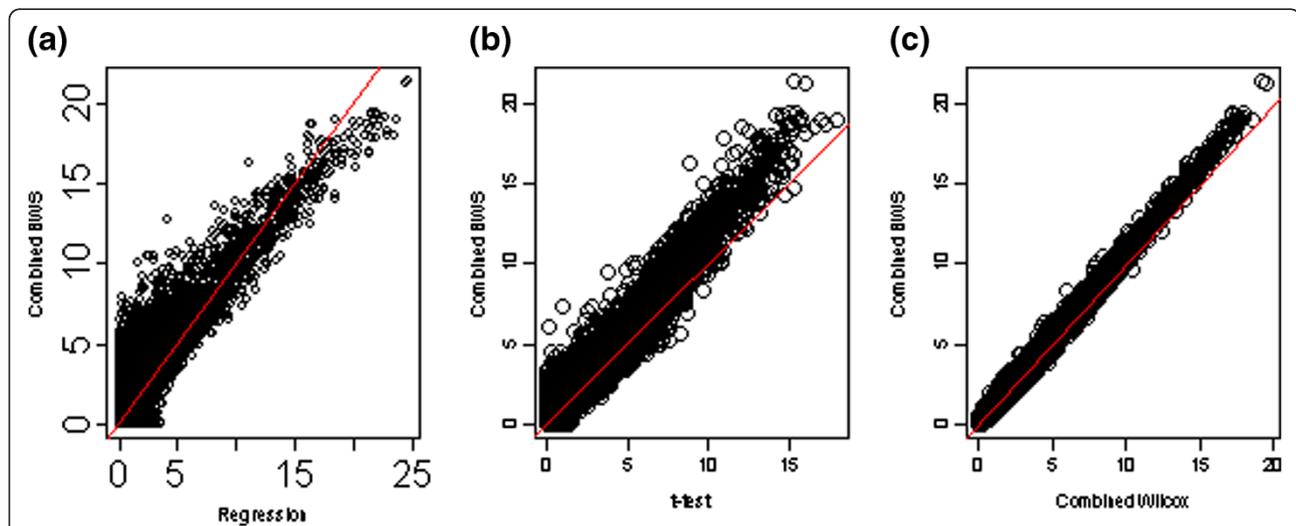
$\lambda$	Combined <i>t</i> -test	Linear regression	Combined Wilcoxon	Combined BWS
0	0.056	0.053	0.073	0.669
0.05	0.075	0.057	0.094	0.703
0.1	0.182	0.087	0.200	0.832
0.15	0.402	0.141	0.412	0.939
0.2	0.701	0.212	0.687	0.988
0.25	0.911	0.298	0.893	0.999

Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under the accession number GSE19711. There were 274 control samples and 131 pre-treatment case samples, and the number of loci was 27578. For the data quality control, we removed 29 patients (15 controls and 14 treatments) with BS conversion efficiency value < 4000 or coverage rate < 95% [15]. For each treatment group, the samples were divided into 6 age groups (50–55, 55–60, 60–65, 65–70, 70–75, and 75 and over). We further removed 12 patients in the pre-treatment group whose ages exceeded 78 since there were no such patients in the control group. The final sample size for each individual group is the same as the one listed in Table 2 of Chen [16] except that the “75 and over” group has 13 pre-treatment samples instead of 25. This dataset was analyzed by both Wang [13] and Chen [16] in their papers. Wang [13] did not consider the age effect, and only 96 cases and 136 controls were included in their analysis

after further screening; while Chen [16] included the 12 patients with ages exceeding 78; thus their results are different from ours even though the same method was used.

Figure 1 shows the scatter plots for each locus based on the negative  $\log_{10}$  p-values derived from four different methods. Figure 1 (a) plots the results for the combined DWS test and the linear regression method, Figure 1 (b) plots the results for the combined DWS test and combined *t*-test, Figure 1 (c) plots the results for the combined DWS test and combined Wilcoxon-test. From Figure 1 (a), it can be seen that most of the loci with small p-value in the linear regression tend to have small p-value in our proposed method as well. However, there exist many loci whose p-values are large in the linear regression but small and significant in the combined DWS test, i.e., those points in Figure 1 (a) with x-values close to zero but y-values large than 3. This indicated that our proposed method is more powerful than the linear regression method in detecting the differentially methylated loci. Similar conclusions can be drawn from Figure 1 (b) and Figure 1 (c) for the comparison with the combined *t*-test and combined Wilcoxon test respectively.

Table 4 lists the number of the loci detected by each of the four tests based on four different significant levels,  $10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$ . Clearly at the same significant level, in terms of the number of significant loci detected, the most powerful method is combined DWS, the next one is combined Wilcoxon test, and then combined *t*-test and linear regression. Table 4 also reports the numbers of significant loci obtained by pairs of the proposed test and each of the other three methods for given significance levels. For example, when the cutoff p-value is  $10^{-3}$ , the combined BWS, the linear regression, the combined *t*-test,



**Figure 1 Scatter plots for negative  $\log_{10}$  p-values based on different methods.** The left panel is for the comparison between combined DWS and linear regression. The middle panel is for the comparison between combined DWS and combined *t*-test. The right panel is for the comparison between combined DWS and combined Wilcoxon test.

**Table 4 Number of loci with p-values less than the given cutoff significance levels from different methods**

Cutoff p-value	Linear regression (I)	Combined t-test (II)	Combined Wilcoxon test (III)	Combined BWS test (IV)	From both I and IV	From both II and IV	From both III and IV
$10^{-3}$	2038	2754	3143	3387	1884	2659	3081
$10^{-4}$	1438	1879	2152	2321	1352	1795	2117
$10^{-5}$	1120	1343	1495	1653	1059	1286	1479
$10^{-6}$	894	982	1109	1222	844	931	1099

and the combined Wilcoxon obtained 3387, 2038, 2754, and 3143 significant loci, respectively. However, among those 3387 loci that have p-values less than  $10^{-3}$  using the combined BWS method, there were 1884, 2659, and 3081 loci overlapping with the linear regression, the combined t-test, and the combined Wilcoxon methods, respectively. In other words, there were 1503, 728, and 306 significant loci were only obtained by the proposed test, but not by the linear regression, the combined t-test, and the combined Wilcoxon, respectively. In contrast, at the same significance level  $10^{-3}$ , there were only 154, 95, and 62 loci whose p-values from the new method are larger than  $10^{-3}$  but less than  $10^{-3}$  from the linear regression, the combined t-test, and the combined Wilcoxon, respectively. Therefore most of the loci detected by the other three methods are also detected by the proposed method. However, for the same cutoff level, there were many loci that were significant in the proposed method but not in the other three methods, a point that clearly demonstrated the advantages of our method over those three.

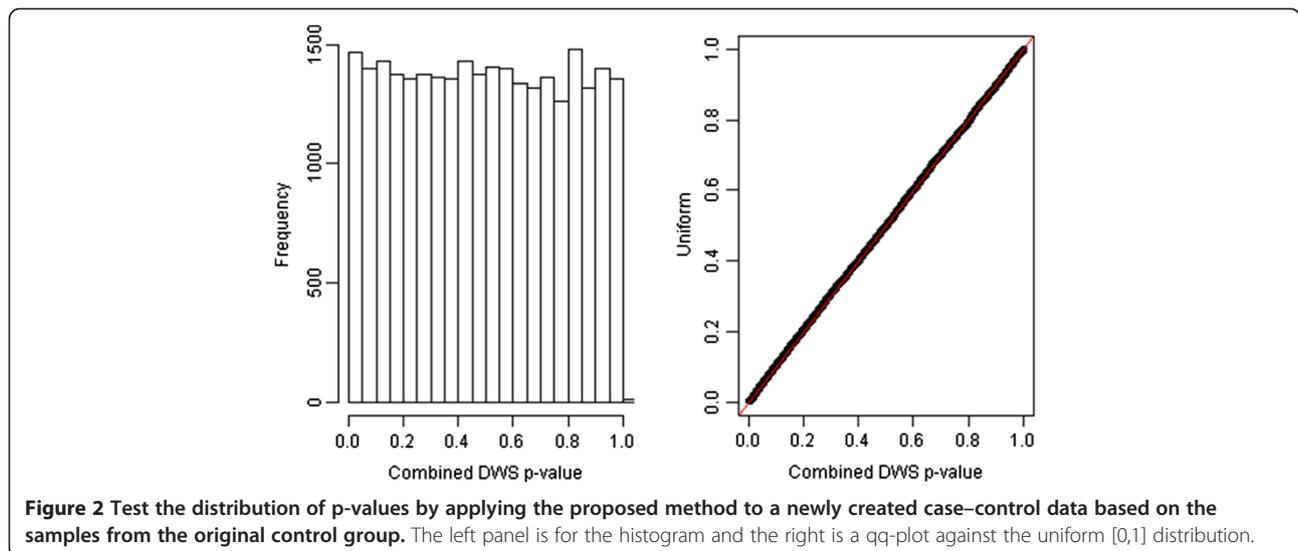
**Discussion**

To study whether or not the proposed method can control type I error rate as well, we created pseudo case and control samples. The way we did this was to first randomly divide the original control subjects into two parts

for each age group. Then we put one part into the new pseudo-control group and the other one into the new pseudo-case group. The distribution of p-values from applying the proposed method to this new case-control data set is shown in Figure 2 (a). It is very close to uniform distribution, and this finding is further confirmed by the qq-plot against the uniform [0,1] distribution as illustrated in Figure 2(b). Therefore, our method increased the detection power while it simultaneously controlled the type I error rate.

In this paper we chose different cutoff p-values to compare the performance of the proposed test with others. We did not consider the multiple testing issue, which is an important but difficult task for this area [4] and other areas where a large number of correlated variables are tested simultaneously [24-28]. The traditional correction methods for multiple comparisons, such as Bonferroni correction, are inappropriate here as they are too conservative due to the fact that many loci from the same gene are highly positively correlated. To account for the correlations among loci, methods based on the concept of “effective number” may be adopted [29].

There are many ways to combine p-values from independent tests [20,30-32]. In this paper, we chose to use Fisher test due to its robustness. It is possible, however, that other methods may be more powerful under some



**Figure 2 Test the distribution of p-values by applying the proposed method to a newly created case-control data based on the samples from the original control group.** The left panel is for the histogram and the right is a qq-plot against the uniform [0,1] distribution.

certain conditions. The combined p-value method used in this paper is based on the assumption that the test for every individual age group is independent from each other. However, if this assumption is questionable, the current combined p-value method needs to be modified such that it can handle the correlations among the individual tests as well. This is another research topic we will pursue in a future paper.

## Conclusions

In case-control methylation studies, the underlying distribution of the  $\beta$ -values is rarely known in advance, and clearly the normality assumption is not valid. Parametric models can be powerful if the assumptions for the models are valid, but they can also lead to biased results if the underlying true distribution is far deviated from the imposed ones. Thus, it is desirable to choose a powerful yet robust statistical test that does not depend on any underlying distribution assumption. In this article we proposed and examined a rank-based nonparametric method to detect differentially methylated loci. Through simulation, we showed that our proposed method is as powerful as the linear regression, *t*-test and Wilcoxon rank sum test methods if the mean differences between the two treatment groups are large. However, our method substantially outperformed the others in situations where the mean differences between the two groups were small while the variance differences were large.

Note that the age effects are strongly associated with methylation, and the ignoring age effect will cause a loss of power or a large number of false positives. Another advantage of the proposed method over many existed ones is that it combined the nonparametric test together with age adjustment. Our next goal was to generalize our method to adjust for more confounders other than the age such that it can have a much wider application in methylation research.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

HH devised the basic idea, analyzed the data and drafted the manuscript; ZC devised the basic idea and analyzed the data. XH participated in the study design, discussion and edit the manuscript. All authors read and approve the final manuscript.

## Acknowledgements

The authors would like to thank Ms. Kim Lawson for her professional editorial assistance, which highly improved the presentation of this paper. ZC also thanks the support from the faculty research funds awarded by the School of Public Health at the Indiana University Bloomington.

## Author details

<sup>1</sup>Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30605, USA. <sup>2</sup>Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, 1025 E. 7th Street, Bloomington, IN 47405, USA. <sup>3</sup>Neurochemistry Laboratory, Department of Psychiatry,

Massachusetts General Hospital and Harvard Medical School, Charlestown, MA 02129, USA.

Received: 13 August 2012 Accepted: 20 February 2013

Published: 6 March 2013

## References

1. Baylin SB, Ohm JE: Epigenetic gene silencing in cancer—a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer* 2006, **6**(2):107–116.
2. Kuan PF, Wang S, Zhou X, Chu H: A statistical framework for Illumina DNA methylation arrays. *Bioinformatics* 2010, **26**(22):2849.
3. Rakan V, Down TA, Thorne NP, Flicek P, Kulesha E, Gräf S, Tomazou EM, Bäckdahl L, Johnson N, Herberth M: An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res* 2008, **18**(9):1518–1529.
4. Kuan PF, Chiang DY: Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. *Biometrics* 2012. doi:10.1111/j.1541-0420.2011.01730.x.
5. Feinberg AP, Tycko B: The history of cancer epigenetics. *Nat Rev Cancer* 2004, **4**(2):143–153.
6. Jabbari K, Bernardi G: Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 2004, **333**:143–149.
7. Jones PA, Baylin SB: The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 2002, **3**(6):415–428.
8. Kulis M, Esteller M: DNA methylation and cancer. *Adv Genet* 2010, **70**:27–56.
9. Laird PW: Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* 2010, **11**(3):191–203.
10. Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Péguignot E: Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* 1999, **402**(6758):187–191.
11. Siegmund KD, Laird PW, Laird-Offringa IA: A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics* 2004, **20**(12):1896–1904.
12. Houseman EA, Christensen BC, Yeh RF, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK: 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* 2008, **9**(1):365.
13. Wang S: Method to detect differentially methylated loci with case-control designs using Illumina arrays. *Genet Epidemiol* 2011, **35**(7):686–694.
14. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R: Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* 2009, **5**(8):e1000602.
15. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Nouchmeh H, Bell CG, Maxwell AP: Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* 2010, **20**(4):440–446.
16. Chen Z, Liu Q, Nadarajah S: A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. *Bioinformatics* 2012, **28**(8):1109–1113.
17. Chen Z, Huang H, Liu J, Ng HKT, Nadarajah S, Huang X, Deng Y: Detecting differentially methylated loci for Illumina Array methylation data based on human ovarian cancer data. *BMC Med Genomics* 2013, **6**(Suppl 1):S9.
18. Baumgartner W, Weiß P, Schindler H: A nonparametric test for the general two-sample problem. *Biometrics* 1998, **54**:1129–1135.
19. Neuhaus M: Exact tests for the analysis of case-control studies of genetic markers. *Hum Hered* 2002, **54**(3):151–156.
20. Fisher RA: *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd; 1932.
21. Owen AB: Karl Pearson's meta-analysis revisited. *Ann Statist* 2009, **37**(6B):3867–3892.
22. Chen Z, Ng HKT: A robust method for testing association in genome-wide association studies. *Hum Hered* 2012, **73**(1):26–34.
23. Chen Z: A new association test based on Chi-square partition for case-control GWA studies. *Genet Epidemiol* 2011, **35**(7):658–663.
24. Chen Z, Huang H, Ng HKT: Testing for association in case-control genome-wide association studies with shared controls. *Stat Methods Med Res* 2013, doi:10.1177/0962280212474061. First published on February 1, 2013.
25. Chen Z, Huang H, Ng HKT: Design and analysis of multiple diseases genome-wide association studies without controls. *Gene* 2012, **510**(1):87–92.

26. Chen Z, Liu J, Ng HKT, Nadarajah S, Kaufman HL, Yang JY, Deng Y: **Statistical methods on detecting differentially expressed genes for RNA-seq data.** *BMC Syst Biol* 2011, **5**(Suppl 3):S1.
27. Chen Z, McGee M, Liu Q, Kong YM, Huang X, Yang JY, Scheuermann RH: **Identifying differentially expressed genes based on probe level data for GeneChip arrays.** *Int J Comput Biol Drug Des* 2010, **3**(3):237–257.
28. Chen Z, Liu Q, McGee M, Kong M, Huang X, Deng Y, Scheuermann RH: **A gene selection method for GeneChip array data with small sample sizes.** *BMC Genomics* 2011, **12**(Suppl 5):S7.
29. Chen Z, Liu Q: **A New approach to account for the correlations among single nucleotide polymorphisms in genome-wide association studies.** *Hum Hered* 2011, **72**(1):1–9.
30. Chen Z: **Is the weighted z--test the best method for combining probabilities from independent tests?** *J Evol Biol* 2011, **24**(4):926–930.
31. Chen Z, Nadarajah S: **Comments on 'Choosing an optimal method to combine p--values' by Sungho Won, Nathan Morris, Qing Lu and Robert C. Elston, Statistics in Medicine 2009; 28: 1537–1553.** *Stat Med* 2011, **30**(24):2959–2961.
32. Lancaster H: **The combination of probabilities: an application of orthonormal functions.** *Austral J Statist* 1961, **3**:20–33.

doi:10.1186/1471-2105-14-86

**Cite this article as:** Huang et al.: Age-adjusted nonparametric detection of differential DNA methylation with case-control designs. *BMC Bioinformatics* 2013 **14**:86.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

