Chen et al. BMC Bioinformatics 2014, 15:142
http://www.biomedcentral.com/1471-2105/15/142

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data

Zhongxue Chen[1*], Hanwen Huang[2] and Qingzhong Liu[3]

Abstract

Background: Because of its important effects, as an epigenetic factor, on gene expression and disease development, DNA methylation has drawn much attention from researchers. Detecting differentially methylated loci is an important but challenging step in studying the regulatory roles of DNA methylation in a broad range of biological processes and diseases. Several statistical approaches have been proposed to detect significant methylated loci; however, most of them were designed specifically for case-control studies.

Results: Noticing that the age is associated with methylation level and the methylation data are not normally distributed, in this paper, we propose a nonparametric method to detect differentially methylated loci under multiple conditions with trend for Illumina Array Methylation data. The nonparametric method, Cuzick test is used to detect the differences among treatment groups with trend for each age group; then an overall p-value is calculated based on the method of combining those independent p-values each from one age group.

Conclusions: We compare the new approach with other methods using simulated and real data. Our study shows that the proposed method outperforms other methods considered in this paper in term of power: it detected more biological meaningful differentially methylated loci than others.

Keywords: Cuzick test, Nonparametric test, Trend test

Background

DNA methylation is an epigenetic mark that has important effects on transcriptional regulation, chromosomal stability, genomic imprinting, and X-inactivation, [1,2]. In addition, it is associated with many human diseases, including various types of cancer [3-10].

Due to the recent advances of BeadArray technology, high-throughput genome-wide methylation data can be routinely generated by Infinium Methylation Assays. This provides good opportunities for researchers to simultaneously study hundreds of thousands of DNA methylation loci. However, it also requires sophisticated and advanced statistical methods to analyze this kind of data.

The raw data generated from BeadArray are fluorescent intensities for each locus; they need appropriate preprocesses, such as background correction and normalization. Then a summarized β-value is generated from about 30 replicates in the same array as follows: $\beta = \frac{\max(M,0)}{\max(U,0)+100}$, where M is the average signal from a methylated allele while U is that from an unmethylated allele. Obviously, the β-values are continuous numbers between 0 and 1, with 0 stands for totally unmethylated and 1 for completely methylated.

Due to the non-normality of the β-value [11-13], those commonly used statistical methods, such as t-test for case control designs, ANOVA for multiple conditions, or linear regression with age as a predictor, usually have low power to detect differentially methylated loci [13,14]. Some statistical approaches with or without adjusting the age-effect, which has been found highly associated with DNA methylation [15,16], have been proposed to detect differentially methylated loci for case-control designs [11-13]. However, very little work has been done for the situation where there are three or more groups (e.g., conditions, or treatments).

* Correspondence: zc3@indiana.edu
[1]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, 1025 E. 7th street, PH C104, Bloomington, IN 47405, USA
Full list of author information is available at the end of the article

BioMed Central

© 2014 Chen et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

In a previous study, we compared some statistical tests with age-effect adjustment for DNA methylation data with three treatments, and found that the method based on the nonparametric Kruskal-Wallis (KW) test is usually more powerful than other methods, such as ANOVA and regression method [14]. However, if there is a trend associated with treatments or conditions, KW based test is no longer the optimal method since it ignores this information. In this case, a more powerful statistical approach is desirable.

In this paper, we propose a new statistical approach to detecting differentially methylated loci for methylation data with multiple conditions with trend. In this method, we also adjust the age-effect in a similar way that we used before. More specifically, we first group subjects into several categories based on their age; we then apply a nonparametric trend test and get a one-sided p-value for each age category. An overall p-value is then obtained through combining those individual p-values. The performance of the new approach is assessed through comparing it with other methods using simulated data and a real methylation data set with three treatments. The R code for the new method is provided (please see the Additional file 1: R code).

Methods

Existing methods

In a recent paper, we have proposed several methods based on combining independent p-values to adjust the effect of age for genome-wide methylation data with multiple conditions [14]. Since those commonly used methods, such as regression models with age as continuous or categorical covariate, have poor performances [12], we compare the proposed approach with the following ones, which have the best performances among current methods based on our previous study [14].

Combined ANOVA test

We assume there are K conditions (treatments) and G age groups. For each age group g $(g = 1,2,\ldots,G)$, we apply an ANOVA test and obtain a p-value $p_g^A NOVA$. We know that under the null hypothesis that this locus is not differentially methylated among K conditions in any age group, $-2\log\left(\prod_{g=1}^{G} p_g^{ANOVA}\right)$ has a chi-square distribution with 2G degrees of freedom (df) since these G p-values are independent. Therefore, the overall p-value can be obtained through combining p-values by Fisher test [17]:

$$p_{ANOVA} = \chi^2_{df=2G}\left(\chi^2 > -2\sum_{g=1}^{G}\log\left(p_g^{ANOVA}\right)\right) \qquad (1)$$

Combined KW test

Similarly, we replace ANOVA by the nonparametric Kruskal-Wallis test for each age group and obtain an overall p-value with p_g^{ANOVA} being replaced by the p-value p_g^{KW} from KW test:

$$p_{KW} = \chi^2_{df=2G}\left(\chi^2 > -2\sum_{g=1}^{G}\log\left(p_g^{KW}\right)\right) \qquad (2.)$$

Methods for combining p-values

Besides the Fisher method mentioned above, we can also use Z-test to combine p-values from independent tests [18-20]. We calculated the weighted Z statistic using individual p-values from each age group: $Z = \sum_{g=1}^{G} n_g \Phi^{-1}\left(1-p_g\right) / \sum_{g=1}^{G} n_g^2$, where n_g is the total sample size in age group g and Φ is the cumulative distribution function (CDF) of the standard normal distribution. It can be shown that under the null hypothesis this statistic has the standard normal distribution. The overall p-value is then calculated as 1- Φ(Z) by the one-sided z-test.

The proposed method

If there is a monotonic trend of the outcome (i.e., β-value here) over the treatments, we can use the more powerful one-sided Cuzick test [21] to obtain a p-value for each age group g (g = 1,2,…,G). The Cuzick test statistic for age group g is calculated as:

$$C_g = \frac{\sum_{i=1}^{N_g} r_{gi} s_{gi} - N_g\left(N_g+1\right)\sum_{i=1}^{K} s_{gi} p_{gi}/2}{\sqrt{\frac{1}{12} N_g{}^2\left(N_g+1\right)\left(\sum_{i=1}^{K} s_{gi}^2 p_{gi} - \left(\sum_{i=1}^{K} s_{gi} p_{gi}\right)^2\right)}} \qquad (3)$$

where N_g is the total number of subjects in age group g, r_{gi} is the rank of the ith of the N_g subjects, s_{gk} is the score of the kth $(k = 1, 2, \ldots, K)$ treatment, K is the number of treatments, $p_{gk} = \frac{n_{gk}}{N_g}$, and n_{gk} is the number of subjects in the kth treatment within the gth age group. For the kth treatment, we assign a score s_{gk} to each of the n_{gk} subjects. In this paper, we set $s_{gk} = k$ $(k = 1, 2, \ldots, K)$, that is, we use scores 1,2,…,K.

It can be shown that under the null hypothesis, the statistic C_g (g = 1,2,…,G) in (3) has an asymptotic standard normal distribution [21]. If there is an increasing trend over the K treatments, we should use the one-sided p-value, $p_{r,g} = Prob(Z > c_g) = 1 - \Phi(c_g)$. On the other hand if there is a decreasing trend over the K treatments, we should use the other one-sided p-value, $p_{l,g} = Prob(Z < c_g) = \Phi(c_g)$. The

statistic $W_1 = -2\log\left(\prod_{g=1}^{G} p_{l,g}\right)$ has an asymptotic chi-square distribution with 2G df under the null hypothesis according to Fisher [17]. Similarly, under the null hypothesis the statistic $W_2 = -2\log\left(\prod_{g=1}^{G}\left(1-p_{l,g}\right)\right)$ also has an asymptotic chi-square distribution with 2G df.

If we know the direction of the trend (increasing or decreasing), we should use either W_1 or W_2 to calculate the overall p-value. However, if the trend direction is unknown, which is usually the case in practice, we will use the maximum of the two statistics:

$$W = \max\left(W_1, W_2\right) \tag{4}$$

Since W_1 and W_2 are correlated, the null distribution of W is not easy to obtain. However, we have the following nice result [22-27].

Theorem 1 Under the null hypothesis, the survival function of W in (4) is asymptotically bounded by

$$2\gamma-\gamma^2 \le \Pr(W > w) \le 2\gamma, \tag{5}$$

where $\gamma = 1-\chi^2_{2G}(w)$, and $\chi^2_{2G}(.)$ is the cumulative distribution function of the chi-square distribution with 2G df.

The above theorem can be proved using the concept of associated variables due to Esary, Proschan and Walkup [28] and Theorem 2 of Owen [29]. From theorem 1, the overall p-value can be estimated by the upper bound 2γ. It is easily seen that when the true p-value of W is small, the difference between the true and the estimated p-values is negligible.

We can also estimate the overall p-value by the weighted Z-test:

$$p_Z = 1-\Phi\left(\left|\sum_{g=1}^{G} w_g c_g\right|\right) \tag{6}$$

where the weight $w_g = \frac{n_g}{N}$, and n_g (g = 1,2,...,G) is the number of total subjects of the K treatments within the gth age group. The validity of (6) is easily seen: under the null hypothesis c_g and therefore $w_g c_g$ has an asymptotic standard normal distribution; a two-sided p-value then can be obtained through (6).

Simulation settings

To assess the performance of the proposed method, we use simulated data to compare the proposed test with current methods in terms of controlling type I error rate and power. We assume there are three different treatments (i..e, K = 3) and six age groups (i.e., G = 6). For each treatment we assume the β-value has the same following distributions over the six age groups: (i) uniform U(a,b) where $0 \le a < b \le 1$, (ii) truncated normal TN $(\mu, \sigma^2, 0, 1)$ (or simply TN (μ, σ^2), and (iii) Beta distribution Beta (c,d) with various parameters. We consider relatively small sample sizes in our simulation study. To reflect practical situations, we either choose 20 samples for each of the three treatments (sample size = (20, 20, 20)), or set the sample sizes as 15, 20, and 25 (sample size = (15, 20, 25)), respectively, for the three treatments. Since the proposed test is designed to detect differentially methylated loci when there is a monotonic trend over the treatments, we simulate β-values with increasing or decreasing mean values over the three treatments for the alternative hypotheses. For example, in simulation, we first generate 20 β-values (sample size = (20, 20, 20)) from three uniform distributions (denoted by a = (0,0,0.25), b = (1,1,1)), U(0,1), U(0,1), and U(0.25, 1) for each of the three treatment groups. The significance level is set to be 0.05 in simulation study. The type I error rate and power are estimated by the proportions of rejection with 10^4 replicates.

A real data set

The real methylation data set of the United Kingdom Ovarian Cancer Population Study (UKOPS) [16], which is one of the largest available Illumina methylation data sets, will be used for real data application. This data set originally includes 274 healthy controls, 131 pre-treatment cases, and 135 post treatment cases. If the DNA methylation of a locus is positively associated with the disease, we would expect that the methylation rates are increasing from control to post-treatment then to pre-treatment. On the other hand, if the association is negative, there would be a decreasing trend over the three conditions: control, post-treatment, and pre-treatment. In either of the two situations, we can use the proposed test.

The above mentioned methylation data were generated by the Illumina Infinium Human Methylation27 BeadChip and can be downloaded from the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) with the accession number GSE19711. For this data set, there are 27578 loci. After a data quality control process, we removed 60 subjects with BS values less than 4000 or the coverage rates less than 95%. All of the subjects are separated into 6 age groups (50-55, 55-60, 60-65, 65-70, 70-75, and 75 and over). Table 1 lists the resulting numbers of subjects in each age by treatment group. For each locus, we apply the proposed test and other methods.

Results

Simulation results

For the new method and the combined ANOVA and KW tests, we only report the results using Fisher method to combine independent p-values, as the results using Z-test are very similar. Table 2 reports the empirical type I error rates for the proposed method, the combined ANOVA test and the combined KW test,

Table 1 Number of samples in age by treatment group used in the paper after data quality control step

Age group	Control	Pre-treat	Post-treat	Total
50_55	14	15	16	45
55_60	61	17	25	103
60_65	64	17	22	103
65_70	35	17	21	73
70_75	63	24	22	109
75_over	20	18	9	47
Total	257	108	115	480

from the simulation study. It is clearly shown that even if the sample size is relatively small and the underlying distribution is not normal, all the methods, including the ANOVA based test, control type I error rate quite well.

Table 3 lists the empirical power values for the three methods under various situations. As expected, the proposed test always has higher power values than those of the combined ANOVA and KW tests. This demonstrates that the proposed test which uses the trend information can improve the detecting power. It should point out that in the simulation study, we assign scores 1, 2, and 3 to the three treatments. However, the effect sizes between treatments 1 and 2 and that between treatments 2 and 3 are not set to be 1 to 2, respectively, which makes the scores (1,2,3) optimal; therefore, the proposed test have the best power. In words, we don't use the optimal scores for the Cuzick test to reflect the real situations when the optimal scores are unknown. This can be seen from the powers of the new test with different scores (e.g., (1,1,2), and (1,3,2)) in the last two columns of Table 3. For many situations considered in Table 3, the scores (1,1,2) are closer to the optimal scores, which are determined by the effect sizes of treatments 2 vs. 1, and treatments 3 vs. 1, than the default ones, (1,2,3); therefore, it is not surprising that the new test with scores (1,1,2) has larger power values than those from the one with scores (1,2,3). However, for most of the situations, the scores (1,3,2) do not use the trend correctly and hence has lower power compared with the other two.

Results from the real data application

The proposed test and the combined ANOVA and KW tests are applied to the real data mentioned above. Due to the multiple comparison issue and the correlation among loci, it is desirable but difficult to obtain a meaningful cut-off p-value to determine differentially methylated loci. We

Table 2 Empirical size for each method at significance level 0.05 with 10^4 replicates from the simulation study

Simulation setting (3 treatments)			Combined ANOVA	Combined K-W	New
Distribution	Sample size	Parameters			
Uniform U(a,b)	(20,20,20)	a = (0,0,0), b = (1,1,1)	0.051	0.045	0.047
		a = (0,0,0), b = (0.5,0.5,0.5)	0.051	0.045	0.045
		a = (0.5,0.5,0.5), b = (1,1,1)	0.055	0.046	0.047
	(15,20,25)	a = (0,0,0), b = (1,1,1)	0.052	0.043	0.046
		a = (0,0,0), b = (0.5,0.5,0.5)	0.052	0.044	0.050
		a = (0.5,0.5,0.5), b = (1,1,1)	0.049	0.040	0.043
Truncated Normal TN (μ ,σ^2)	(20,20,20)	μ = (0.5,0.5,0.5), σ = (1,1,1)/5	0.050	0.043	0.048
		μ = (0.5,0.5,0.5), σ = (1,2,3)/5	0.058	0.050	0.045
		μ = (0.2,0.2,0.2), σ = (1,1,1)/5	0.049	0.050	0.043
		μ = (0.8, 0.8, 0.8), σ = (1,1,1)/5	0.046	0.043	0.048
	(15,20,25)	μ = (0.5,0.5,0.5), σ = (1,1,1)/5	0.050	0.046	0.045
		μ = (0.5,0.5,0.5), σ = (1,1.2,1.3)/5	0.053	0.041	0.033
		μ = (0.2,0.2,0.2), σ = (1,1,1)/5	0.050	0.046	0.051
		μ = (0.8, 0.8, 0.8), σ = (1,1,1)/5	0.049	0.044	0.048
Beta (c,d)	(20,20,20)	c = (1,1,1), d = (1,1,1)	0.050	0.044	0.049
		c = (1,1,1), d = (5,5,5)	0.046	0.045	0.043
		c = (5,5,5), d = (1,1,1)	0.048	0.044	0.045
		c = (5,5,5), d = (5,5,5)	0.049	0.041	0.044
	(15,20,25)	c = (1,1,1), d = (1,1,1)	0.049	0.044	0.046
		c = (1,1,1), d = (5,5,5)	0.045	0.042	0.047
		c = (5,5,5), d = (1,1,1)	0.049	0.049	0.048
		c = (5,5,5), d = (5,5,5)	0.052	0.044	0.052

Table 3 Empirical power for each method at significance level 0.05 with 10^4 replicates from the simulation study

Simulation setting (3 treatments)			Combined ANOVA	Combined K-W	New[1]	New[2]	New[3]
Distribution	Sample size	Parameters					
Uniform U(a,b)	(20,20,20)	a = (0,0,0.25), b = (1,1,1)	0.699	0.607	0.877	0.962	0.069
		a = (0,0.1,0.1), b = (0.5,0.5,0.5)	0.450	0.339	0.724	0.830	0.726
		a = (0.6,0.6,0.5), b = (1,1,1)	0.460	0.338	0.695	0.821	0.027
	(15,20,25)	a = (0,0,0.25), b = (1,1,1)	0.809	0.692	0.926	0.980	0.957
		a = (0,0.1,0.1), b = (0.5,0.5,0.5)	0.433	0.319	0.618	0.758	0.218
		a = (0.6,0.6,0.5), b = (1,1,1)	0.482	0.380	0.754	0.854	0.860
Truncated Normal TN (μ, σ^2)	(20,20,20)	μ = (0.1,0.1,0.2), σ = (1,1,1)/5	0.451	0.394	0.743	0.862	0.052
		μ = (0.1,0.1,0.2), σ = (1,1.2,1.3)/5	0.773	0.642	0.962	0.954	0.200
		μ = (0.5,0.5,0.4), σ = (1,1,1)/5	0.691	0.656	0.918	0.976	0.054
		μ = (0.5,0.5,0.4), σ = (1,1.2,1.3)/5	0.402	0.374	0.696	0.820	0.032
	(15,20,25)	μ = (0.1,0.1,0.2), σ = (1,1,1)/5	0.464	0.428	0.786	0.886	0.948
		μ = (0.1,0.1,0.2), σ = (1,1.2,1.3)/5	0.735	0.643	0.959	0.952	0.713
		μ = (0.5,0.5,0.4), σ = (1,1,1)/5	0.775	0.738	0.949	0.981	0.827
		μ = (0.5,0.5,0.4), σ = (1,1.2,1.3)/5	0.382	0.364	0.756	0.838	0.852
Beta (c,d)	(20,20,20)	c = (1,1,1), d = (30,40,50)	0.596	0.442	0.889	0.723	0.432
		c = (1,1.2,1.5), d = (40,40,40)	0.490	0.609	0.962	0.920	0.329
		c = (30,40,50), d = (1,1,1)	0.578	0.450	0.899	0.745	0.420
		c = (40,40,40), d = (1,1.2,1.5)	0.488	0.620	0.972	0.924	0.369
	(15,20,25)	c = (1,1,1), d = (30,40,50)	0.608	0.405	0.861	0.727	0.998
		c = (1,1.2,1.5), d = (40,40,40)	0.426	0.602	0.952	0.912	0.559
		c = (30,40,50), d = (1,1,1)	0.618	0.409	0.888	0.752	0.458
		c = (40,40,40), d = (1,1.2,1.5)	0.450	0.606	0.958	0.919	0.995

Note: [1]the prosed test with scores (1,2,3), [2]the prosed test with scores (1,1,2), [3]the prosed test with scores (1,3,2).

therefore report the numbers of loci with p-values less than a given cutoff value from each method. We choose different cutoff values: 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, and 10^{-8}. The results are reported in Table 4. For each of the given cutoff p-values, the proposed test always detects more loci than the other methods. In addition, most of the loci detected by the combined ANOVA and KW tests were also detected by the proposed test. For example, when the cutoff p-value is 10^{-5}, the combined ANOVA test , the combined KW test, and the proposed test detected 479, 551, and 1283 loci, respectively, when Fisher method was used to combine p-values. Out of the 479 loci detected by the combined ANOVA test, 471 were also detected by the new test; out of the 551 loci detected by the combined KW test, only 7 were not detected by the proposed test.

This indicates that the proposed test is more powerful than other methods that are compared in this study. It is noticeable that the methods of combining independent p-values (i.e., Fisher test and Z-test) have

Table 4 Number of significant differentially methylated loci detected by each method for each given cutoff p-value

Method	1e-3		1e-4		1e-5		1e-6		1e-7		1e-8	
	F	Z	F	Z	F	Z	F	Z	F	Z	F	Z
T1 (Combined ANOVA)	981	1079	655	690	479	499	350	375	257	275	189	208
T2 (Combined KW)	1359	1340	823	859	551	590	381	401	261	277	172	185
T3 (New)	2915	3117	1855	1951	1283	1310	905	929	674	686	513	521
T1 and T2	926	980	615	656	442	474	306	338	221	235	152	167
T1 and T3	931	1018	639	670	471	491	346	367	252	269	187	206
T2 and T3	1294	1279	806	832	544	577	377	396	259	276	170	184
T1, T2, and T3	895	954	605	642	437	468	303	336	220	234	151	166

similar performance here, although the Z-test usually gives a few more significant loci.

Discussion

We proposed a new statistical approach based on combing p-values and the Cuzick test, which is a nonparametric one-sided test. Through simulation study and real data application, we show that if there exists a monotonic (not necessarily linear) trend over the treatments, the proposed test is more powerful than other methods. Figure 1 plots the mean β-value of each of the three treatments over the six age groups for loci with p-values less than 10^{-3} from the proposed test. From Figure 1, we can see there is a decreasing trend among the three treatments (i.e., for the β-value, pre-treatment < post-treatment < control) for all of the six age groups; while from for those loci with large p-values, such trend does not exist for any of the six age groups (see Additional file 2: Figure S1).

Although many methods can detect those loci which are strongly differentially methylated among different treatments, it is important to detect loci having small effects as they are biological meaningful and provide useful information for set-based analyses, such as gene, gene-set, and pathway analyses which use those detected differentially methylated loci as input [30].

To use the Cuzick test, we need to assign a score for each of the treatment. Here we assign 1, 2, and 3 to the control, post-treatment, and pre-treatment, respectively. In practice, if we have the information of the effects for each treatment, we can use this information to assign scores. For example, for the K-1 treatments 2, 3, …, K, if the effect sizes are m_2, …, m_K compared to treatment 1, we can assign scores 0, m_2, …, m_K to those treatments for the proposed test. However, if we only know that there is a monotonic trend, we can choose 1, 2, …, K (equivalent to 0, 1, …, K-1) as the scores. Although, the performance of the proposed test can be improved by assigning optimal scores, which are determined by the true effects, to the treatments; in general, it is impractical to obtain the optimal scores. In addition, the optimal scores for each locus may not be the same across age groups (see Figure 1).

Like other large scale data, such as microarray data and genome-wide association study data, the multiple comparison is an important but challenging issue. Although some procedures have been proposed to control either family-wise error rate or false discovery rate, it remains an open topic in this area. One possible direction is to use the so-called "effective number" estimated from correlations among the loci [31].

Conclusions

We propose a new statistical approach to detecting methylated loci for high-throughput methylation data with multiple groups. This approach is based on the nonparametric Cuzick test, which is robust and powerful if there exists a trend over groups. Through simulated and

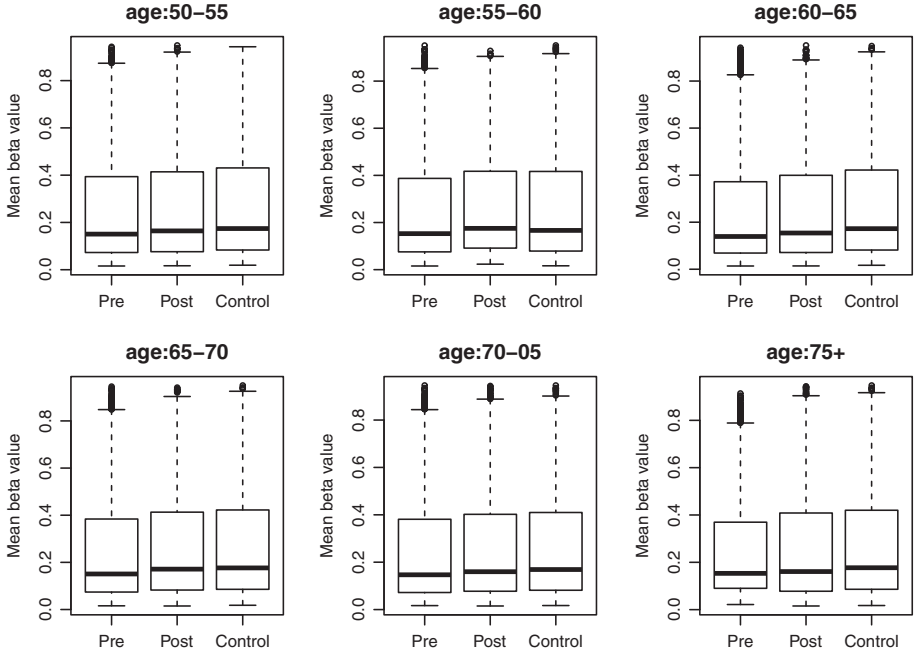

Figure 1 The mean β-value of loci with p-value less than 10^{-3} from the proposed test over the three treatment groups by the age group. For each age group, there is a trend among the three treatments: pre-treatment has smaller β-value than the post-treatment group, which in turn has smaller β-value than the control group.

real data, we show that the proposed test outperforms existing methods.

Additional files

Additional file 1: R code.

Additional file 2: Figure S1. The mean β-value of loci with p-value greater than 10^{-3} from the proposed test over the three treatment groups by the age group. For each age group, there is no obvious trend over the three treatments for the β-value.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ZC, HH, and QL were jointly responsible for the development of the algorithm and the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The authors would like to thank the Editor and five reviewers for their constructive comments, which resulted in an improved presentation of this paper. The first author also acknowledges the support from the faculty research funds awarded by the School of Public Health, Indiana University Bloomington.

Author details
[1]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, 1025 E. 7th street, PH C104, Bloomington, IN 47405, USA. [2]Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602, USA. [3]Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA.

Received: 10 December 2013 Accepted: 6 May 2014
Published: 15 May 2014

References
1. Rakyan VK, 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.
2. Bock C: **Analysing and interpreting DNA methylation data.** *Nat Rev Genet* 2012, **13**(10):705–719.
3. Baylin SB, Ohm JE: **Epigenetic gene silencing in cancer–a mechanism for early oncogenic pathway addiction?** *Nat Rev Cancer* 2006, **6**(2):107–116.
4. Feinberg AP, Tycko B: **The history of cancer epigenetics.** *Nat Rev Cancer* 2004, **4**(2):143–153.
5. Jabbari K, Bernardi G: **Cytosine methylation and CpG, TpG (CpA) and TpA frequencies.** *Gene* 2004, **333**:143–149.
6. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer.** *Nat Revi Genet* 2002, **3**(6):415–428.
7. Kulis M, Esteller M: **DNA methylation and cancer.** *Adv Genet* 2010, **70**:27–56.
8. Laird PW: **Principles and challenges of genome-wide DNA methylation analysis.** *Nat Rev Genet* 2010, **11**(3):191–203.
9. Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Péquignot E: **Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene.** *Nature* 1999, **402**(6758):187–191.
10. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D: **Increased methylation variation in epigenetic domains across cancer types.** *Nat Genet* 2011, **43**(8):768–775.
11. Wang S: **Method to detect differentially methylated loci with case-control designs using Illumina arrays.** *Genet Epidemiol* 2011, **35**(7):686–694.
12. Chen Z, Liu Q, Nadarajah S: **A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data.** *Bioinform* 2012, **28**(8):1109–1113.
13. Huang H, Chen Z, Huang X: **Age-adjusted nonparametric detection of differential DNA methylation with case-control designs.** *BMC Bioinform* 2013, **14**(1):86.
14. 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.
15. 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.
16. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr 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.
17. Fisher RA: *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd; 1932.
18. 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.
19. Chen Z: **Is the weighted z-test the best method for combining probabilities from independent tests?** *J Evol Biol* 2011, **24**(4):926–930.
20. Chen Z, Nadarajah S: **On the optimally weighted z-test for combining probabilities from independent studies.** *Comput Stat Data Anal* 2014, **70**:387–394.
21. Cuzick J: **A wilcoxon type test for trend.** *Stat Med* 1985, **4**(4):543–547.
22. Chen Z, Huang H, Ng HKT: **Testing for Association in Case–control Genome-wide Association Studies with Shared Controls.** In *Statistical Methods in Medical Research.* 2013. Published online before print February 1, 2013, doi: 101177/0962280212474061.
23. Chen Z: **Association tests through combining p-values for case control genome-wide association studies.** *Stat Probabil Lett* 2013, **83**(8):1854–1862.
24. Chen Z, Ng HKT: **A Robust Method for Testing Association in Genome-Wide Association Studies.** *Hum Hered* 2012, **73**(1):26–34.
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: **A new association test based on Chi-square partition for case-control GWA studies.** *Genet Epidemiol* 2011, **35**(7):658–663.
27. Chen Z, Huang H, Ng HKT: **An Improved Robust Association Test for GWAS with Multiple Diseases.** *Stat Probabil Lett* 2014, **91**:153–161.
28. Esary JD, Proschan F, Walkup DW: **Association of random variables, with applications.** *Ann Math Stat* 1967, **38**:1466–1474.
29. Owen AB: **Karl Pearson's meta-analysis revisited.** *Ann Statist* 2009, **37**(6B):3867–3892.
30. Sun H, Wang S: **Penalized logistic regression for high-dimensional DNA methylation data with case-control studies.** *Bioinform* 2012, **28**(10):1368–1375.
31. 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.

doi:10.1186/1471-2105-15-142
Cite this article as: Chen *et al.*: Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data. *BMC Bioinformatics* 2014 **15**:142.

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

BioMed Central