A POWERFUL STATISTICAL METHOD FOR IDENTIFYING DIFFERENTIALLY METHYLATED MARKERS IN COMPLEX DISEASES

SURIN AHN

Email: surin.ahn@gmail.com

TAO WANG†

Department of Epidemiology and Population Health, Albert Einstein College of Medicine of Yeshiva, 1300 Morris Park Ave, Bronx, NY 10461
Email: tao.wang@einstein.yu.edu

DNA methylation is an important epigenetic modification that regulates transcriptional expression and plays an important role in complex diseases, such as cancer. Genome-wide methylation patterns have unique features and hence require the development of new analytic approaches. One important feature is that methylation levels in disease tissues often differ from those in normal tissues with respect to both average and variability. In this paper, we propose a new score test to identify methylation markers of disease. This approach simultaneously utilizes information from the first and second moments of methylation distribution to improve statistical efficiency. Because the proposed score test is derived from a generalized regression model, it can be used for analyzing both categorical and continuous disease phenotypes, and for adjusting for covariates. We evaluate the performance of the proposed method and compare it to other tests including the most commonly-used t-test through simulations. The simulation results show that the validity of the proposed method is robust to departures from the normal assumption of methylation levels and can be substantially more powerful than the t-test in the presence of heterogeneity of methylation variability between disease and normal tissues. We demonstrate our approach by analyzing the methylation dataset of an ovarian cancer study and identify novel methylation loci not identified by the t-test.

---

† This work is supported in part by the CTSA Grant UL1 RR025750 and KL2 RR025749 and TL1 RR025748 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH roadmap for Medical Research, R21HG006150 from National Human Genome Research Institute (NHGRI).

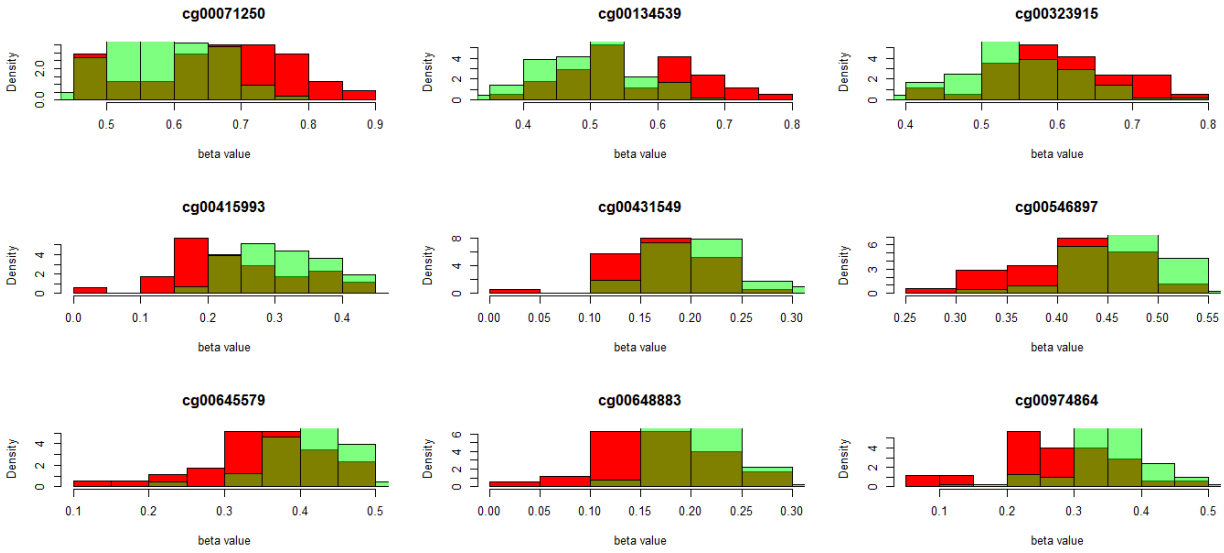
1. Introduction

DNA methylation is an important epigenetic modification that regulates transcriptional expression and plays an important role in complex diseases including cancer. [1] Recently, tremendous amounts of DNA methylation data have been generated from high-throughput DNA methylation platforms for many complex diseases. Compared to patterns of other molecular profiling, e.g. gene expression, DNA methylation has unique features. One example is that not only the mean but also the standard deviation of methylation levels can vary across age groups.[2,3] New statistical approaches designed to incorporate these features are desirable because they could be more robust and efficient than conventional methods. As such, Chen et al. proposed a test to evaluate the overall statistical significance of association by combining p-values from different age groups and showed it was more robust and usually more powerful than existing tests. [4]

Another phenomenon that has recently received attention is the increased methylation variability at relevant loci of cancer. [5-7] It has been found that differential variability between normal and cancer tissues can be very useful for identifying methylation markers of cancer [6-8] However, commonly-used statistical methods, such as the t-test and linear regression, which do not directly detect differences in variability, are statistically inefficient in the presence of heterogeneity of methylation variability. In the statistical literature, various tests, e.g. the Barlett's test[9] and the Levene's test[10], have been proposed for testing homogeneity of variance between two groups. In general, the Levene's test is less sensitive than the Bartlett's test to departures from normality. Figure 1 shows methylation distributions of several representative loci in cancer and normal tissues from an ovarian cancer study.[3] One important feature of these loci is both the mean and variability of methylation levels are different between cancer and normal tissues. For these loci, it may be useful to combine information from both the first and second moments of methylation distribution to improve power to identify methylation markers. One approach to combine the results for testing mean and variability is Fisher's method of combining p-values. However, it requires that the mean and variance are independent, which is often not true for methylation data. Another approach is to use tests, e.g. Kolmogorov-Smirnov test, to compare the empirical distribution of methylation data, which, however, is often not statistically efficient[11].

In this article, we propose a new statistical test that incorporates changes in both mean and variability to identify methylation markers of diseases, and demonstrate how jointly testing the mean and variability can identify methylation markers that are otherwise missed by testing the mean alone. More specifically, we first define two score tests for testing methylation differences in mean and variability, respectively, under a generalized regression model. Then, we develop a new joint test by combining these two statistics, while accounting for their correlation. As such, the new test may not require intensive sampling approaches to evaluate p-values. We evaluate the performance of the proposed approach and compare it to the conventional tests including the commonly-used two-sample t-test through simulations. We show that the validity of the proposed test is robust to departures of the normal distribution of methylation levels and can be substantially more powerful than the t-test in the presence of heterogeneity of variability between two groups. Finally, we apply our approach to the methylation data of an ovarian cancer study and identify cancer relevant loci that other tests could fail to identify.

Fig1: Histograms of DNA methylation values of pretreatment cancers and control groups at 9 selected methylation loci. Red bars represent cancers and green bars represent controls.



2. Methods

We consider detecting the association of individual methylation loci with disease based on a case-control study. For individual i $(i = 1, 2, ..., n)$, the trait value is denoted as Y_i, and the methylation value is denoted as X_i. To identify methylation loci that are relevant to disease, we consider the statistical hypothesis $H_0 : \mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$ versus $H_1 : \mu_0 \neq \mu_1$ and $\sigma_0^2 \neq \sigma_1^2$, in which μ_0 and μ_1 are means of methylation levels for controls and cases, respectively, and σ_0^2 and σ_1^2 are the corresponding variances.

To compare the average methylation levels of disease and normal tissues, we consider a generalized linear model,

$$logit[P(Y_i = 1)] = \alpha + \beta X_i,$$

in which α and β are regression coefficients. Under this model, a score statistic to test the difference of the average methylation levels of two groups is given by $U_1 = \sum_i (Y_i - \bar{Y}) X_i$. By treating X_i as the variable, the variance of the score statistics can be estimated by $\hat{\sigma}^2_{U_1} = \sum_i (Y_i - \bar{Y})^2 \hat{\sigma}^2_X$, where $\hat{\sigma}^2_X$ is the estimated variance of methylation levels. As such, the score test can be formed by

$$T_1 = \frac{U_1^2}{\hat{\sigma}^2_{U_1}}.$$

This test is closely related to the commonly-used t-test as they both test the difference of means between two groups and has a centered χ^2_1 under the null hypothesis for a large sample size.

To test the difference in methylation variability between disease and normal tissues, we first define a variability score for each sample by $Z_i = (X_i - \bar{X})^2$, in which $\bar{X}$ is the sample mean of methylation levels. With the variability score, a similar logistic regression can be constructed with the variability score as the independent variable. Then, the score statistic is given by $U_2 = \sum_i (Y_i - \bar{Y}) Z_i$. It can be easily seen that this score statistic is proportional to the difference of estimated variances between disease and normal tissues, i.e. $U_2 \propto \hat{\sigma}_1^2 - \hat{\sigma}_0^2$. The variance of U_2 can be estimated by $\hat{\sigma}_{U2}^2 = \sum_i (Y_i - \bar{Y})^2 \hat{\sigma}_Z^2$, in which $\hat{\sigma}_Z^2$ is the estimated variance of the variability score. As such, the score test based on the variability score is

$$T_2 = \frac{U_2^2}{\hat{\sigma}_{U_2}^2}.$$

Similarly, T_2 also has χ_1^2 under the null hypothesis for a large sample size.

A joint test statistic for both mean and variability of methylation levels may be simply formed as $T_1 + T_2$ that has a χ_2^2 under the null hypothesis, or by Fisher's method for combing p-values when T_1 and T_2 are independent. However, T_1 and T_2 are generally not independent. To take into account the correlation between T_1 and T_2, it is necessary to estimate the covariance of U_1 and U_2. To do this, we denote the joint score statistic as $U_{joint} = (U_1, U_2)$ and its variance-covariance matrix can conveniently be estimated by $\hat{\Sigma}_{Ujoint}^2 = \sum_i (Y_i - \bar{Y})^2 \hat{\Sigma}_S^2$, in which $\hat{\Sigma}_S^2$ is the estimated variance-covariance matrix of X and Z. Then, the joint test is defined by

$$T_{joint} = U_{joint} \Sigma_{U_{joint}}^{-1} U_{joint}^T.$$

For a large sample size, T_{joint} has a centered χ_2^2 under the null hypothesis. When sample size is small, we could use a fast permutation procedure by randomly shuffling the order of the trait values of Y_is. Of note, the inverse of $\hat{\Sigma}_S^2$ does not require to be calculated at each replicate.

3. *Results*

3.1 Simulation study

We evaluated the performance of the proposed joint test through simulations. To evaluate the type I error rate, we first considered a case-control study with various sample sizes (n=20, 30, 50 and 100) for each group and sampled methylation values of cases and controls from various distributions (the standard normal, t distribution with 10 degrees of freedom or χ^2 with 2 degrees of freedom). For each scenario, we used 10,000 replicates to evaluate type I error rate. It is also of interest to examine the false positive rate at a more stringent threshold as a large number of loci are now routinely examined in methylation studies. We simulated 10 million replicates to evaluate type I error rate for a sample size of 100 cases and 100 controls. With this large number of simulations, we estimate the false positive rate with reasonable accuracy for a threshold of 10^{-5}. Finally, we examined the type I error rate of the proposed test after adjusting for batch effects. We assumed different proportions of cases and controls were assayed in two batches (30% in batch 1 for cases and 70% in batch 1 for controls), yielding difference methylation variability between cases and controls due to batch effects.

Table 1: The empirical type I error rate at the statistical significance level of 0.05

Distribution	Sample size	t-test	Levene	KS	T_{joint}	$T_{permutation}$
N(0,1)	20	0.050	0.040	0.037	0.039	0.053
	30	0.047	0.043	0.033	0.039	0.049
	50	0.050	0.042	0.037	0.045	0.050
	100	0.050	0.050	0.036	0.048	0.050
t_{10}	20	0.048	0.049	0.041	0.034	0.051
	30	0.050	0.043	0.037	0.036	0.050
	50	0.049	0.048	0.042	0.042	0.050
	100	0.052	0.047	0.037	0.046	0.051
χ_2^2	20	0.050	0.039	0.033	0.034	0.049
	30	0.051	0.044	0.034	0.038	0.050
	50	0.050	0.046	0.041	0.036	0.049
	100	0.054	0.040	0.048	0.043	0.049

We further compared power of the proposed joint test with the t-test, Levene's test and Kolmogorov-Smirnov test (KS) at the statistical significance level of 0.05. First, we simulated the methylation values of controls from a standard normal distribution, and cases from a normal distribution with various means and standard deviations (sd). The sample size was set at 100 for each group. Second, we simulated situations when different levels of heterogeneity exist in cancer tissues by sampling cases from a mixed normal distribution,

$$\pi_0 N(0,1) + (1-\pi_0) N(\mu^2, \sigma^2).$$

In this simulation, we set π_0 at 0.5 and varied μs and σ^2s to simulate different changes in the mean and variances between cancer and normal tissues. The sample size was set at 200 for each group. For each scenario we used 1,000 replicates to evaluate power. P-values of the proposed method were assessed using both the asymptotic distribution and the empirical null distribution obtained by the permutation procedure. The number of permutation was set at 1,000.

Table 1 shows the empirical type I error rate at the statistical significance of 0.05 for the proposed joint test (T_{joint}), the joint test based on permutation ($T_{permuation}$), the Levene's test Kolmogorov-Smirnov test (KS), and the t-test. We can see all tests maintained a good control of type I error rate under simulated scenarios. However, T_{joint} tended to be slightly conservative when sample size is small (n<50) and the distribution is highly skewed (χ_2^2 distribution). For a more stringent threshold of 10^{-5}, we found a similar pattern of the type I error rate for T_{joint}, which tended to be slightly conservative with the type I error rate at around 0.4×10^{-5}.

Table 2 shows the type I error rate of different tests when there is a difference in methylation variability between cases and controls due to batch effects. We can see the proposed test maintained a good control of type I error rate by incorporating a batch variable for adjustment for batch effects, while the Levene's test tend to have an inflated type I error rate.

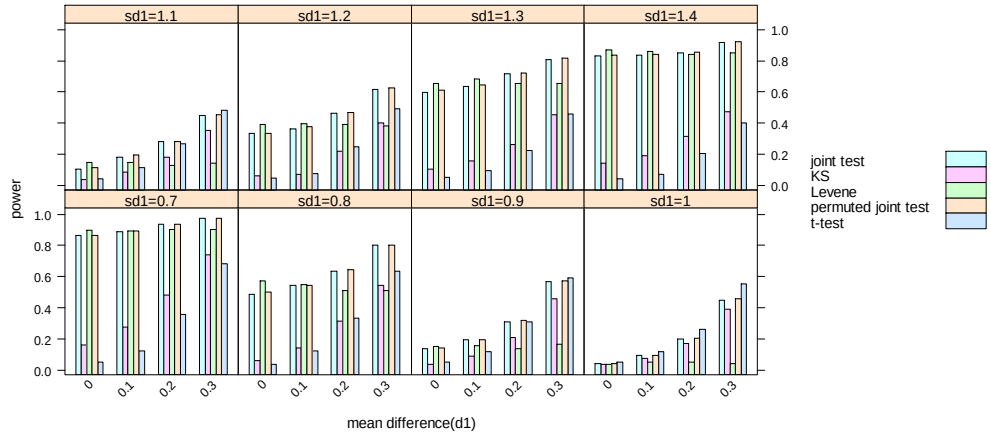
Table 2: The empirical type I error rate at the statistical significance level of 0.05 in the presence of heterogeneity in variability between cases and controls due to batch effects (n=100)

SD1*	SD2	t-test	KS	Levene	T_{joint}	$T_{permutation}$
1	1.1	0.053	0.036	0.061	0.059	0.058
1	1.2	0.046	0.038	0.066	0.042	0.044
1	1.3	0.038	0.033	0.068	0.046	0.047
1	1.4	0.046	0.039	0.091	0.042	0.042
1	1.5	0.045	0.031	0.090	0.042	0.042

*SD1 and SD2 are standard deviations of methylation values in batch 1 and 2, respectively. 70% cases are assumed to be assayed in batch 1 and 30% controls are assayed in batch 2.

Fig 2: The empirical power of the proposed test and the two-sample t-test at significance level of 0.05 to detect methylation loci associated with disease. (a) Controls are simulated from a standard normal distribution and cases are simulated with varied means and standard deviations (sds). The x-axes indicate varied means of cases and different panels represent varied sds. The sample size is 100 for each group. (b) Controls are simulated from a standard normal distribution and cases are simulated from a mixture normal distribution, i.e. $0.5N_0(0,1)+0.5N_1(d,sd)$. The x-axes indicate the means of $N_1(d,sd)$ and panels represent sds of $N_1(d,sd)$. The sample size is 200 for each group.

(a)



(b)

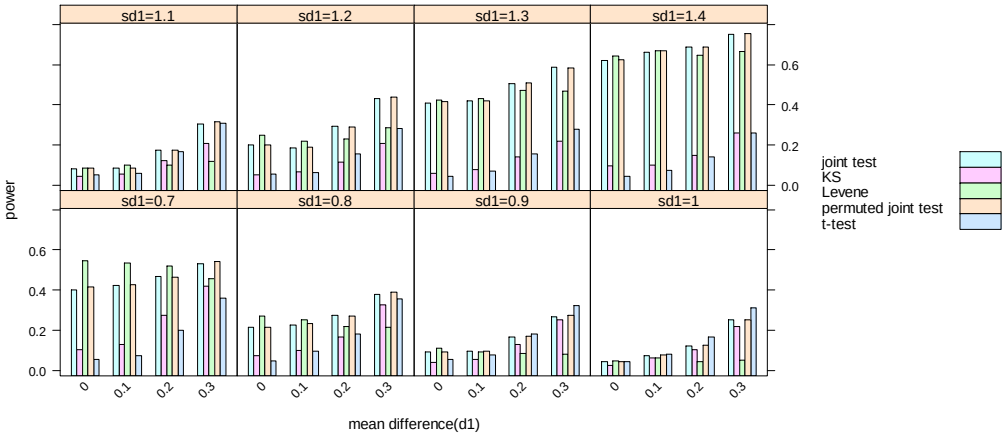


Figure 2 compares the empirical power of different tests at the significance level of 0.05 to detect methylation loci associated with disease under various situations. Based on our simulations, we have the following observations. First, T_{joint} was slightly less powerful than $T_{permutation}$. In situations when cases were sampled from an admixture distribution, the gain of power for $T_{permutation}$ appeared more obvious, which might reflect the conservative nature of the asymptotic test when the normal assumption does not hold. Second, the proposed tests were substantially more powerful than the t-test in the presence of heterogeneity of methylation variability between cases and controls. Third, T_{joint} and $T_{Permutation}$ were only slightly less powerful than the t-test when there was no heterogeneity of variability between cases and controls.

3.2 Application to an ovarian cancer study

To demonstrate the utility of the proposed test, we applied the proposed method to the data of United Kingdom Ovarian Cancer Population Study (UKOPS)[3]. This dataset is available at the NCBI Gene Expression Omnibus (http:///www.ncbi.nlm.nih.gov/geo) with accessing number GSE19711. The data includes 266 cases with 131 treatment and 135 post-treatment patients, and 274 age-matched healthy controls. To avoid the heterogeneity between age groups, we chose to analyze the 50-60 year group with 35 pretreatment patients and 82 controls. The data with 27,578 GpG loci were generated by Infinium assay with the HumanMethylation27 DNA Analysis beadchip. After background correction and normalization for the raw fluorescent intensities, a summarized value, i.e. β value, is calculated based on about 30 replicates in the same array by max(M,0)/[max(M,0)+max(U,0)+100], where M is the average signal from a methylated allele and U is from an unmethylated allele. Hence, the range of the β value is between 0 (unmethylated) and 1 (fully methylated). Because of the small sample size, we calculated both T_{joint} and $T_{permutation}$, and compared them to other tests. For computational reasons, the number of permutations for each locus was determined adaptively. Initially, 10^3 simulations were performed. If the resulting empirical p value was less than 0.01, 10^4 simulations were performed. If the p value from 10^4 simulations was less than 0.001, 10^5 simulations were performed.

Table 3: Number of loci with p-values smaller than the given cutoff from different tests

P-value	t-test	Levene	KS	T_{joint}	$T_{permutation}$	t-test and T_{joint} ($T_{permutation}$)
<0.01	750	157	1044	1047	1318	549(750)
<0.001	267	18	353	463	582	214(267)
<0.0001	62	4	85	169	250	51(62)

Fig 3: The correlation of –log10 p-values between the t-test, T_{joint} and $T_{permutation}$ for comparing pre-treatment cases and controls in the age group of 50-60 years. (a) the t-test and T_{joint} (b) the t-test and $T_{permutation}$ and (c) T_{joint} and $T_{permutation}$.

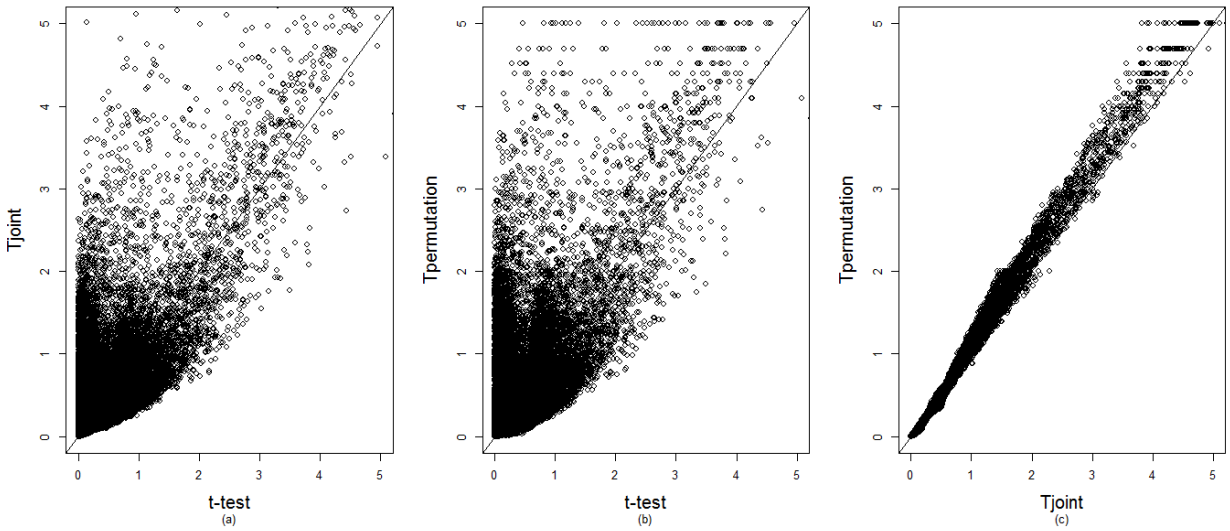


Table 3 shows the number of significant loci at different significance levels fordifferent tests. As expected, $T_{permutation}$ identified slightly more loci than T_{joint} because T_{joint} tends to be conservative for small sample sizes. However, T_{joint} and $T_{permutation}$ identified many more loci than the t-test. We further compared –log10 p-values of different methods for all loci (Figure 3). It can be seen that for many loci, T_{joint} and $T_{permutation}$ provided much lower p-values than the t-test, suggesting a large proportion of loci may have significant changes in the methylation variability between cases and controls. However, T_{joint} and $T_{permutation}$ had similar p-values, although $T_{permutation}$ tended to generate slightly smaller p-values. The analysis has also been performed on other age groups (60-70 years and >70 years) and yielded similar findings (data not shown).

4. Discussion

Although in recent cancer studies suggested the difference of methylation levels in both mean and variability was observed between cancer and normal tissue[5-7,12,13], so far most methods to identify differentially methylated loci examine the methylation mean and variability separately. To overcome this drawback, we propose a new statistical score test that achieves higher power than the t-test when there is heterogeneity in methylation variability between cases and controls. The traditional t-test gives less significant p-values in this case as it ignores information provided by the second moment of the methylation distribution. When there is no heterogeneity in methylation variability, the proposed method, although it is not optimal in terms of power,

generally has robust power. Additionally, because the proposed test is very simple and hence can be calculated in a fast fashion, it is computationally feasible to be applied to very large methylation datasets, e.g. Illumina 450K. Our simulations and application to an ovarian cancer demonstrated the utility of our new method for discovering new methylation markers of complex diseases.

Essentially, the proposed method is an attempt to combine tests for mean and variance of methylation levels between two groups. With the normal assumption of methylation levels, one may perform the t-test for comparing means and F-test for comparing variances; and the joint test statistic can be obtained by Fisher's method for combining p-values.[14] However, the normal assumption is in general not true for methylation data. Moreover, a normal transformation is often not feasible for a large number of genome-wide methylation loci, since each can have a unique distribution. One of the consequences due to departures of normal distribution is that test statistics for the mean and variance are no longer independent, resulting in an inflated type I error rate when Fisher's method of combining p-values is used. To obtain valid p-values, computationally extensive sampling procedures, e.g. permutation, may be necessary. However, for highly significant p values, sampling is not a trivial task as such a procedure can be very inefficient. To address the issue of correlation, we propose a score test in which the correlation between test statistics for the mean and variance can be naturally adjusted. Another consequence of non-normality, in particular when the distribution is skewed, is that the t-test may lead to loss in power. The underlying assumption of our test statistic is that there is a linear relationship between independent variables and risk of disease. Because the linear relationship does not hold when the distribution is skewed, the power of our method may also be sensitive to skewness of the methylation distribution, although the validity of our method is quite robust. Of note, the permutation procedure itself would not improve power in this case. Further research is necessary to develop or identify statistical tests that can maintain good power when the distribution is highly skewed.

In the application to a real dataset from an ovarian cancer study, our method achieves higher statistical significance than the t-test at some loci. Indeed, a relatively large proportion of markers are only identified by the proposed test. The main reason for this might be that heterogeneity of methylation variability between cancer and normal tissue is a common phenomenon. In our study of both simulated and real datasets, the t-test performs better than our method when there is no difference in variance between cases and controls as an extra degree of freedom is used for testing variance in our method. However, Figure 2 (a) and (b) show that the relative power gain of the t-test is not very dramatic.

The proposed method can be generalized in different ways. In this paper we consider a case-control study. However, our score test is developed from a generalized linear regression model. As such, our method could be generalized for both continuous, e.g. age, and other categorical disease phenotypes. Another advantage of our method is that it can easily generalize to incorporate covariates. As such, our method can differentiate the true biological difference from the technical difference of variance between cases and controls, e.g. the batch effect, by incorporating an

addition batch variable as a covariate. As shown in our simulation result, our method maintained a good control of the type I error rate after adjustment for batch effects. When there is no obvious variable, the technical difference can also be corrected by using a "genomic control", in which the null distribution of the test statistic can be estimated from random methylation loci in the genome[15]. In addition, the application of our method can naturally extend beyond the analysis of a single methylation locus to the region-based (or gene-based) analysis under the framework of generalized linear regression. The advantage of the region-based analysis is it can make use of information of correlated loci in a spatial region. One challenge in applying the method for testing variances is the interpretation. Because the proposed test is an omnibus test that can simultaneously account for methylation mean and variability, it may be useful to further examine the independent effect of the change in methylation mean and variability when an association is identified. Various reasons could cause the change of methylation variability in disease tissues. One possibility is the heterogeneity of disease itself. However, it has also suggested that methylation variability may play an important biological role in the development of complex diseases[5]. Understanding the cause of heterogeneity of variance could have fundamental biological implications.

In summary, our results demonstrate that simultaneously testing differences in means and variances of methylation levels between cases and controls could identify disease related loci that are otherwise missed. Our method has the potential to be an efficient tool for screening potential methylation markers of diseases as our method does not require computationally intensive sampling to obtain valid p-values, and provides higher power than the t-test in the presence of differences in variability.

5. Acknowledgments

S.A is a high school student and worked as a summer intern in this project. T.W. is supported in part by the CTSA Grant UL1 RR025750 and KL2 RR025749 and TL1 RR025748 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH roadmap for Medical Research, R21HG006150 from National Human Genome Research Institute (NHGRI).

References

1. Laird, P.W. & Jaenisch, R. DNA methylation and cancer. *Hum Mol Genet* **3 Spec No**, 1487-95 (1994).
2. Christensen, B.C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* **5**, e1000602 (2009).
3. Teschendorff, A.E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* **20**, 440-6 (2010).
4. Chen, Z., Liu, Q. & Nadarajah, S. A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. *Bioinformatics* **28**, 1109-13 (2012).

5. Feinberg, A.P. & Irizarry, R.A. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A* **107 Suppl 1**, 1757-64 (2010).
6. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-75 (2011).
7. Jaffe, A.E., Feinberg, A.P., Irizarry, R.A. & Leek, J.T. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* **13**, 166-78 (2012).
8. Teschendorff, A.E. & Widschwendter, M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487-94 (2012).
9. Snedecor, G.W.a.C., William G. *Statistical Methods*, (Iowa State University Press, 1989).
10. Levene, H. *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, (Stanford University Press, 1960).
11. Chakravarti, L.a.R. Handbook of Methods of Applied Statistics. Vol. 1 392-394 (John Wiley and Sons, 1967).
12. Issa, J.P. Epigenetic variation and cellular Darwinism. *Nat Genet* **43**, 724-6 (2011).
13. Feinberg, A.P. *et al.* Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index (vol 3, 65er1, 2011). *Sci Transl Med* **2**(2010).
14. Perng, S.K.a.L., R.C. A test of equality of two normal population means and variances. *Journal of the American Statistical Association* **71**, 968-970 (1976).
15. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).