Insights in Genetics and Genomics

Research Article Open Access

New Score Tests for Equality of Variances in the Application of DNA Methylation Data Analysis [Version 1, 1 Approved, 1 Approved with Reservations]

Weiliang Qiu[1]**, Xuan Li[2]*, Jarret Morrow[1], Dawn L DeMeo[1], Scott T Weiss[1], Xiaogang Wang[2] and Yuejiao Fu[2]

[1]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, USA
[2]Department of Mathematics and Statistics, York University, Canada

**Corresponding author: Weiliang Qiu, Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, 181 Longwood Avenue, Boston, 02115, USA, Tel: 1-617-525-0841; Email: stwxq@channing.harvard.edu

*: Contribute Equally

Original Submission

Received: December 10, 2016
Accepted: December 20, 2016
Published: December 24, 2016
Last Updated: February 22, 2017

Open Peer Review Status: 1 Approved, 1 Approved with Reservations

How to cite this article: Weiliang Qiu, Xuan Li, Jarret Morrow, Dawn DeMeo, Scott Weiss, Xiaogang Wang, Yuejiao Fu. New Score Tests for Equality of Variances in the Application of DNA Methylation Data Analysis [Version 1, 1 Approved, 1 Approved with Reservations]. Insights Genet Genomics. (2016) 1: 3.1

Abstract

Recently, DNA methylation marks with differential variability have been reported to have biological meanings. Ahn and Wang (2013) proposed a joint test for testing if two distributions have the same mean and the same variance for DNA methylation data analysis. Their joint test statistic is a quadratic form of a vector of two test statistics. The first test statistic aims to test for equality of means, while the second test (denoted as AWvar) statistic aims to test for equality of variances. However, the performance of the AWvar test has not been studied yet. In this study, we evaluated the performance of the AWvar test and proposed three improved AWvar tests (denoted as iAWvar.Levene, iAWvar.BF, and iAWvar.Ltrim) based on Levene test, Brown-Forsythe test and trimmed-mean-based Levene test, respectively. Systematic simulation studies and a real DNA methylation data analysis showed that (1) AWvar test worked well under normality assumption; (2) the three improved AWvar tests had much higher testing power than AWvar if normality assumption is violated; and (3) iAWvar.Levene, iAWvar.BF and iAWvar.Ltrim had slightly better performance than their corresponding counterparts: Levene test, Brown-Forsythe test, and trimmed-mean-based Levene test.

Keywords

Levene Test; Brown-Forsythe Test; Logistic Regression; Systematic Simulation; Methylation

Introduction

Recently a series of papers in DNA methylation data analysis [1-7] proposed to identify DNA methylation marks that have different variances between two groups of subjects (e.g., cases and controls). DNA methylation is a biochemical process that can regulate gene expression without changing genetic code by adding a methyl group to the cytosine DNA nucleotides. Variable DNA methylation has been associated with cancers and many complex diseases.

It is a well-studied research topic to test for equality of variance in the science of Statistics. The classical test for the equality of variances is the F-test, which is based on the ratio of two sample variances. Under the null hypothesis that the two distributions are normal distributions with same variances, the F-test statistic follows the F distribution. The main limitation of the F-test is that it is sensitive (i.e., type I error rate would be much higher than the nominal value) to the violation of the normality assumption. This is relevant to the investigation of epigenetics, as high-throughput DNA methylation data might contain outliers (e.g., caused by technical failure) and DNA methylation levels might not be normally distributed [8].

More than 50 robust equal-variance tests have been proposed since the main limitation of the F-test was reported [9]. Compared 56 equal-variance testing procedures using simulation studies and found that Brown-Forsythe test [10] (denoted as BF test) was one of the best tests in terms of keeping nominal type I error rate, while having adequate power.

We have recently performed a comparative study of tests for homogeneity of variances with application to DNA methylation data [8]. We found that the trimmed-mean-based Levene test (denoted as Ltrim test) and BF test outperformed other tests for most scenarios.

Recently [11] proposed a *joint* test aiming to test if two distributions have the same mean and the same variance. The joint test statistic is a quadratic form of a vector of two test statistics. The first test statistic aims to test for equality of means. The second test statistic aims to test for the equality of variance. We denote the second test as the *AWvar* test [11]. Evaluated the performance of their joint test. However, they did not evaluate the performance of the *AWvar* test.

In this article, we evaluated the performance of the *AWvar* test and proposed three improved AWvar tests (denoted as *iAWvar.Levene*, *iAWvar.BF*, and *iAWvar.Ltrim*), based on Levene, BF, and Ltrim, respectively. We did systematic simulation studies and a real DNA methylation data analysis to compare the performances of AWvar, Levene, BF, Ltrim, iAWvar.Levene, iAWvar.BF, and iAWvar.Ltrim.

Method

In this section, we first review the definition of the AWvar test and then propose three improved AWvar tests.

AWvar Test

For a given DNA methylation mark (i.e., CpG site), [11] defined the following statistic for testing equality of variance between two distributions:

$$U_2 = \sum_{i=1}^{n_1+n_0} z_i (y_i - \overline{y}),$$

where n_1 is the number of cases, n_0 is the number of controls, y_i is a binary variable indicating case-control status (i.e., y_i=1 indicates that the *i* -th subject is a case and y_i=0 indicates a control), $\overline{y}$ is the average of y_i, *i*=1,...,n_1+n_0 (i.e., $\overline{y}$ is equal to the proportion of cases), z_i is the squared within-group deviation. That is,

$$z_i = \begin{cases} (x_i - \overline{x}_0)^2 & \text{if subject } i \text{ is a control,} \\ (x_i - \overline{x}_1)^2 & \text{if subject } i \text{ is a case,} \end{cases}$$

where x_i is the DNA methylation level at a given DNA methylation mark for the *i*-th subject, $\overline{x_0} = \sum_{i=1}^{n_1+n_0} x_i(1-y_i)/n_0$ (i.e., the sample mean DNA methylation level for controls) and $\overline{x_1} = \sum_{i=1}^{n_1+n_0} x_i y_i / n_1$ (i.e., the sample mean DNA methylation level for cases).

For the logistic regression model $\log it(p_i|z_i) = \beta_0 + \beta_1 z_i$,

where $p_i = pr(y_i = 1|z_i)$ the score test statistic T_2 is asymptotically chi squared distributed with degree of freedom 1 under the null hypothesis that $\beta_1 = 0$:

$$T_2 = \frac{U_2}{\widehat{Var}(U_2)} \xrightarrow{\beta_1=0} x_1^2,$$

Where $\widehat{Var}(U_2) = \overline{y}(1-\overline{y}) \sum_{i=1}^{n_1+n_0} (z_i - \overline{z})^2$, and

$$\overline{z} = \sum_{i=1}^{n_1+n_0} z_i / (n_1 + n_0)$$

(i.e., the sample mean of z_i).

Three Improved AWvar Tests

Since z_i is sensitive to outliers, we borrow the ideas of three Levene tests (Levene test, Brown-Forsythe test, and trimmed-mean based Levene test) to propose three improved AWvar tests by implementing robust versions of z_i so that the improved AWvar tests are less sensitive to outliers.

The first improved AWvar test (denoted as iAWvar.Levene) is based on the Levene test (c.f. **Supplementary Document Section A**). The idea is to replace the squared deviation of x_i by absolute deviation:

$$z_i^* = \begin{cases} \left| x_i - \overline{x_0} \right|, & \text{if subject } i \text{ is a control,} \\ \left| x_i - \overline{x_1} \right|, & \text{if subject } i \text{ is a case.} \end{cases}$$

For the logistic regression $logit(p_i^* \mid z_i^*) = \beta_0^* + \beta_1^* z_i^*$, where $p_i^* = pr(y_i = 1 \mid z_i^*)$, the score test statistic T_2^* is asymptotically chi squared distributed with degree of freedom 1 under the null hypothesis that $\beta_1^* = 0$:

$$T_2^* = \frac{U_2^*}{\widehat{Var}(U_2^*)} \xrightarrow{\beta_1^*=0} x_1^2,$$

where $$U_2^* = \sum_{i-1}^{n_1+n_0} z_i^*(y_i - \bar{y}),$$

$$\widehat{Var}(U_2^*) = \bar{y}(1-\bar{y}) \sum_{i=1}^{n_1+n_0} (z_i^* - \bar{z}^*)^2,$$

and $$\bar{z}^* = \sum_{i=1}^{n_1+n_0} z_i^* / (n_1 + n_0).$$

The second improved AWvar test (denoted as iAWvar.BF) is based on the Brown-Forsythe test (c.f. **Supplementary Document Section B**). The idea is to replace the within-group mean by within-group median in iAWvar.Levene:

$$z_i^{**} = \begin{cases} |x_i - \tilde{x}_0|, & \text{if subject } i \text{ is a control,} \\ |x_i - \tilde{x}_1|, & \text{if subject } i \text{ is a case,} \end{cases}$$

where $\tilde{x}_0$ is the median for controls and $\tilde{x}_1$ is the median for cases. For the logistic regression $\log it(p_i^{**} \mid z_i^{**}) = \beta_0^{**} + \beta_1^{**} z_i^{**}$, where $p_i^{**} = pr(y_i = 1 \mid z_i^{**})$, the score test statistic T_2^{**} is asymptotically chi squared distributed with degree of freedom 1 under the null hypothesis that $\beta_1^{**} = 0$:

$$T_2^{**} = \frac{U_2^{**}}{\widehat{Var}(U_2^{**})} \xrightarrow{\beta_1^{**}=0} x_1^2$$

where

$$U_2^{**} \equiv \sum_{i=1}^{n_1+n_0} (z_i^{**} = \bar{z}^{**}),$$

$$\widehat{Var}(U_2^{**}) = \bar{y}(1-\bar{y}) \sum_{i=1}^{n_1+n_0} (z_i^{**} - \bar{z}^{**})^2,$$

and $$\bar{z}^{**} = \sum_{i=1}^{n_1+n_0} z_i^{**} / (n_1 + n_0).$$

The third improved AWvar test (denoted as iAWvar.Ltrim) is based on the trimmed-mean based Levene test (c.f. **Supplementary Document Section C**). The idea is to use trimmed within-group means. Denote $\breve{x}_0$ and $\breve{x}_1$ as the 5% trimmed means for cases and controls, respectively. The 5% trimmed mean for a sample is the sample mean after trimmed 5% lowest values and 5% highest values. Let

$$z_i^{\dagger} = \begin{cases} |x_i - \breve{x}_0|, & \text{if subject } i \text{ is a control,} \\ |x_i - \breve{x}_1|, & \text{if subject } i \text{ is a case,} \end{cases}$$

For the logistic regression $\log it(p_i^{\dagger} \mid z_i^{\dagger}) = \beta_0^{\dagger} + \beta_1^{\dagger} z_i^{\dagger}$, where $p_i^{\dagger} = pr(y_i = 1 \mid z_i^{\dagger\dagger})$, the score test statistic $T_2^{\dagger}$ is asymptotically chi squared distributed with degree of freedom 1 under the null hypothesis that $\beta_1^{\dagger} = 0$:

$$T_2^{\dagger} = \frac{U_2^{\dagger}}{\widehat{Var}(U_2^{\dagger})} \xrightarrow{\beta_1^{\dagger}=0} x_1^2,$$

where $$U_2^{\dagger} = \sum_{i=1}^{n_1+n_0} z_i^{\dagger}(y_i - \bar{y}),$$

$$\widehat{Var}(U_2^{\dagger}) = \bar{y}(1-\bar{y}) \sum_{i=1}^{n_1+n_0} (z_i^{\dagger} - \bar{z}^{\dagger})^2,$$

and $$\bar{z}^{\dagger} = \sum_{i=1}^{n_1+n_0} z_i^{\dagger} / (n_1 + n_0).$$

Simulation Studies

We used the same systematic simulation design as the one used in [8] to compare the 7 equal-variance tests: AWvar, Levene, BF, Ltrim, iAWvar.Levene, iAWvar.BF, and iAWvar.Ltrim. Briefly, we performed 2 sets of simulation studies. The first set of simulation studies (denoted as Simulation I) is based on the simulation studies in [11]. Specifically, for a given CpG site, DNA methylation levels for cases and controls were generated from a mixture of two normal, two Student's t, or two chi-squared distributions. DNA methylation levels of all CpG sites were independently generated. The second set of simulation studies (denoted as Simulation II) is based on the simulation studies in [7]. DNA methylation levels were generated from a mixture of Bayesian hierarchical models. Specifically, for a CpG site, given its variance the DNA methylation levels were generated from normal distributions. The variances themselves are random variables from a scaled inverse chi squared distribution. Each pair of CpG sites was marginally correlated with each other. In each simulation scenario, we generated 100 data sets so that we can evaluate the variations of estimated type I error rates or estimated power. In each simulated data set, 1000 CpG sites were generated. We set the case group and control group to have the same number of subjects. Please refer to [8] for details.

The simulation studies evaluated the effects of (1) sample size, (2) the presence of heterogeneity of means, (3) violation of the normality assumption, and (4) outliers on the performances of the 7 equal-variance tests. We considered three sample sizes (number of subjects per-group=20, 50, or 200) to evaluate the effect of sample size on the performance of the 7 tests of equality of variance. To evaluate the effect of inequality of means, we considered 4 scenarios: 2 distributions have the same mean and same variance; 2 distributions have the same mean, but different variances; 2 distributions have different means, but the same variances; and 2 distributions have

different means and different variances. To evaluate the effect of non-normal distribution, we considered t distributions and chi squared distributions, in addition to normal distributions. When evaluating the effect of outliers, we followed [7] and replaced the DNA methylation level of one randomly picked case subject by the maximum DNA methylation level across all CpG sites and all subjects.

There were 36 pairs of different scenarios in Simulation I and 24 pairs of different scenarios in Simulation II. Within a pair, cases and controls have the same variance in one scenario and have different variances in another scenario. Hence, there were 48 pairs of scenarios in total. Forty-eight scenarios, in which cases and controls have the same variance of DNA methylation levels, were used to evaluate whether the empirical type I error rates are equal to or less than the nominal value 0.05. Specifically, for each simulated data set, we test for equality of variance for the 1000 CpG sites, separately. A test is positive if the p-value of the test is < 0.05. The proportion of positive tests among the 1000 tests is an estimate of the type I error rate. For the 100 simulated data sets of a given scenario, we will have 100 estimated type I error rates. We then tested the null hypothesis H_0 that the mean type I error rate is ≤ 0.05 by using one-sided one-sample t-test. The number (n_{reject}) of scenarios that rejected H_0 could be used to evaluate if a test for equality of variance is good or not. The smaller n_{reject} is, the better the equal-variance test is.

The other 48 scenarios, in which cases and controls have different variances of DNA methylation levels, were used to evaluate the power of each of the 7 equal-variance tests. Specifically, for each simulated data set, we test for equality of variance for the 1000 CpG sites, separately. A test is positive if the p-value of the test is < 0.05. The proportion of positive tests among the 1000 tests is an estimate of the power. For each scenario, we then obtained the median of the 100 estimated powers for each equal-variance test. We then ranked the median powers only for the equal-variance tests that did not reject the null hypothesis that the mean type I error rates is ≤ 0.05. If the median power of a test is the highest, then the rank of the test is 1. For ties, average ranks were used. If a test that rejected the null hypothesis that mean type I error rate ≤ 0.05, then we set its rank as missing value "-". We next obtained the median rank (denoted as m) for each of the 7 equal-variance tests across the scenarios. The smaller m is, the better the equal-variance test is. We drew the plot of n_{reject} versus m to visualize the relative performance of the 7 equal-variance tests.

Real Data Analyses

We used two publicly available DNA methylation data sets (GSE20080[12] and GSE37020[6]), which were downloaded from Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo), to compare the performances of the 7 equal variance tests. For both data sets, we are interested in detecting CpG sites differentially variable between samples with normal histology and samples with cervical intraepithelial neoplasia of grade 2 or higher (CIN2+). GSE20080 contains 30 normal samples and 18 CIN2+ samples, while GSE37020 contains 24 normal samples and 24 CIN2+ samples. DNA methylation levels in both GSE20080 and GSE37020 were measured by IlluminaHumanMethylation27 platform, which measures the methylation levels for 27578 CpG sites. We checked data quality for each of the 2 data sets and excluded CpG sites residing on SNPs or with missing values. After data QC and preprocessing (for details please refer to [8]), there were 22859 CpG sites appearing in both cleaned data sets.

We used GSE20080 as the discovery set and GSE37020 as the validation set. For a given CpG site in a given data set, we applied each of the 7 equal-variance tests to test for equality of variance. For a given equal-variance test, we claimed a CpG site in the analysis of GSE20080 as significantly differentially variable if the false discovery rate (FDR)[13] adjusted p-value for the CpG site is less than 0.05. For a significantly differentially variable (DV) CpG site in the analysis of GSE20080, if the corresponding un-adjusted p-value in the analysis of GSE37020 is less than 0.05, then we claimed that the significance in the analysis of GSE20080 is validated in the analysis of GSE37020.

Results

Results of Simulation Studies

Table 1 summarizes the median ranks of median estimated powers of the 7 equal-variance tests for the 48 scenarios, in which cases and controls have different variances of DNA methylation levels. The lower the median rank is, the better a method is. We can see that (1) AWvar worked well (median ranks =1) for data generated from normal distributions without outliers, no matter whether cases have the same mean methylation levels as controls or not; (2) however, for data generated from non-normal distributions (t distributions or chi square distributions), or for data containing outliers, AWvar did not perform well (median ranks are among the highest); (3) both Levene (n_{reject} =34) and iAWvar.Levene (n_{reject} =37) had very high values of n_{reject}. That is, they tend to have inflated type I error rates in most of the 48 scenarios; (4) Other 4 equal-variance tests (BF, Ltrim, iAWvar.BF, and iAWvar.Ltrim) tend to have higher type I error rate for data generated from the chi square distribution; (5) compared to the BF and Ltrim, iAWvar.BF and iAWvar.Ltrim had smaller median ranks and same n_{reject}; (6) For data that have large sample size (nPerGroup=200) and were generated from normal distributions, all 7 equal-variance tests tend to have same median ranks; (7) Although iAWvar.Levene had highest value of n_{reject}, it had smallest median ranks for scenarios in which iAWvar.Levene kept nominal type I error rate; and (8) iAWvar.Ltrim had the smallest median rank of the median powers (m=1.5); (9) In terms of keeping nominal type I error rate, BF, iAWvar.BF and AWvar are better than the other 4 equal variance tests (Levene, Ltrim, iAWvar.Levene, and iAWvar.Ltrim).

Table 1: Rank of the median powers of the 100 simulated data sets for each of the 48 simulation scenarios.

nPerGroup	out	eqM	Distr	Levene	BF	Ltrim	AWvar	iAWvar. Levene	iAWvar. BF	iAWvar. Ltrim
20	no	no	c.chisq	-	2	-	3	-	1	-
20	no	no	chisq	-	-	-	-	-	-	-
20	no	no	c.N	-	5	3	1	-	4	2
20	no	no	N	-	5	3	1	-	4	2
20	no	no	t	-	5	2	3	-	4	1
20	no	yes	chisq	-	2	-	-	-	1	-
20	no	yes	N	-	5	3	1	-	4	2
20	no	yes	t	-	4	2	5	-	3	1
20	yes	no	c.chisq	-	4	2	5	-	3	1
20	yes	no	chisq	-	4	2	5	-	3	1
20	yes	no	c.N	2	6	4	7	1	5	3
20	yes	no	N	-	4	2	5	-	3	1
20	yes	no	t	-	4	2	5	-	3	1
20	yes	yes	chisq	-	4	2	5	-	3	1
20	yes	yes	N	-	4	2	5	-	3	1
20	yes	yes	t	2	5	5	5	1	5	5
50	no	no	c.chisq	-	2	-	3	-	1	-
50	no	no	chisq	-	-	-	-	-	-	-
50	no	no	c.N	2	6	4	1	-	5	3
50	no	no	N	-	3.5	1.5	-	-	3.5	1.5
50	no	no	t	-	4	2	5	-	3	1
50	no	yes	chisq	-	3	-	1	-	2	-
50	no	yes	N	-	3.5	1.5	-	-	3.5	1.5
50	no	yes	t	-	4	2	5	-	3	1
50	yes	no	c.chisq	-	4	2	5	-	3	1
50	yes	no	chisq	-	4	2	5	-	3	1
50	yes	no	c.N	2	6	4	7	1	5	3
50	yes	no	N	-	2.5	2.5	5	-	2.5	2.5
50	yes	no	t	-	4	2	5	-	3	1
50	yes	yes	chisq	-	4	2	5	-	3	1
50	yes	yes	N	1	3.5	3.5	6	-	3.5	3.5
50	yes	yes	t	2	6	3.5	7	1	5	3.5
200	no	no	c.chisq	-	1.5	-	3	-	1.5	-
200	no	no	chisq	-	-	-	1	-	-	-
200	no	no	c.N	4	4	4	4	4	4	4
200	no	no	N	4	4	4	4	4	4	4
200	no	no	t	-	1.5	-	3	-	1.5	-
200	no	yes	chisq	-	2	-	3	-	1	-
200	no	yes	N	4	4	4	4	4	4	4
200	no	yes	t	2	6	3.5	1	-	5	3.5
200	yes	no	c.chisq	-	3.5	1.5	5	-	3.5	1.5
200	yes	no	chisq	-	3.5	1.5	5	-	3.5	1.5
200	yes	no	c.N	3.5	3.5	3.5	7	3.5	3.5	3.5
200	yes	no	N	4	4	4	4	4	4	4
200	yes	no	t	-	-	-	1	-	-	-
200	yes	yes	chisq	-	4	1.5	5	-	3	1.5
200	yes	yes	N	4	4	4	4	4	4	4
200	yes	yes	t	2	6	5	7	1	4	3
	n_{reject}			34	4	11	5	37	4	11
	m			2	4	2	5	3.5	3.25	1.5

"out=yes" means scenarios in which data contain outliers; "eqM=yes" means scenarios in which mean methylation levels of cases is equal to that of controls; "-" in the cells indicates the null hypothesis H0 that the type I error rate of the equal-variance test is ≤ 0.05 was rejected based on 100 simulated data sets for the scenario; nreject = number of scenarios where an equal-variance test rejected H0; m= median rank of the median powers among the 7 equal variance tests (for ranks with ties, average ranks were used); c.N and c.chisq indicate conditional normal and chi squared distributions, respectively.

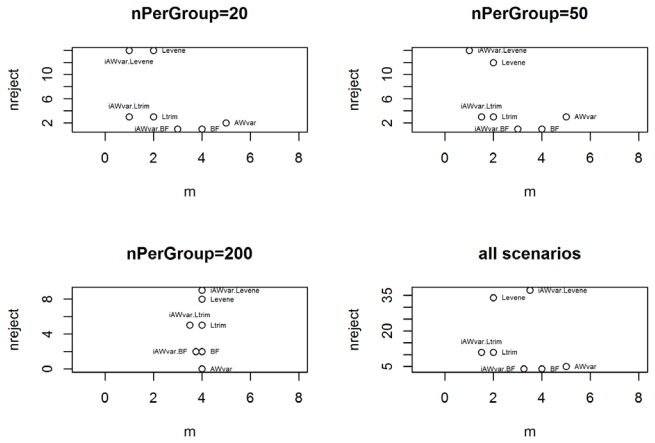

Figure 1: Plots of n_{reject} versus m, where n_{reject} is the number of scenarios where an equal-variance test rejected the null hypothesis H_0 that mean type I error rates is ≤ 0.05 and m is the median rank of the median powers. The upper-left, upper-right, bottom-left panels are the plots where n_{reject} and m were obtained based on scenarios with sample size 20, 50, or 200 subjects per group, respectively. The bottom-right panel is the plot where n_{reject} and m were obtained based on all scenarios.

Figure 1 showed the plots of n_{reject} versus m for the 7 equal-variance tests for scenarios with nPerGroup = 20, 50, or 200, or for all 48 scenarios, separately. If an equal-variance test is good, it should have both small n_{reject} and small m. Hence a good equal-variance test should appear in the bottom left corner of the plots in Figure 1. Figure 1 showed that (1) AWvar has low n_{reject} in all 4 plots, indicating that AWvar is good at keeping nominal type I error rate, compared to other 6 tests; 2) AWvar has the largest m in all 4 plots, indicating that AWvar tends to be less powerful than other 6 tests; (3) iAWvar.Levene, iAWvar.BF, and iAWvar. Ltrim have smaller m than AWvar, indicating that the 3 improved AWvar tests tend to be more powerful than AWvar; (4) compared to AWvar, iAWvar. BF has smaller or similar n_{reject} in all 4 plots, which indicating iAWvar. BF tends to perform better than AWvar in terms of both type I error rate and power; (5) compared to the BF and Ltrim, iAWvar.BF and iAWvar.Ltrim had smaller m and same n_{reject}; and (6) iAWvar.Levene's n_{reject} values are the highest among the 7 equal-variance tests in all 4 plots, although its m are the smallest for scenarios with nPerGroup=20 and nPerGroup=50.

Online Supplementary Figures S1 to S16 showed the parallel boxplots of the 100 estimated type I error rates and estimated power for each pair of the 96 scenarios that were designed to evaluate the type I error rate and the power of the 7 equal-variance tests. From these boxplots, we observed that (1) AWvar performed best when data were generated from the normal distribution without outliers (**Online Supplementary Figures S1 and S2**); (2) AWvar performed badly when data were generated from a distribution with outliers and without equal mean (e.g. **Online Supplementary Figures S6, S14, and S16**); (3) sample size had a large effect on the performance of the 7 equal-variance tests. For example, the power of the

7 tests were less than 0.5 when nPerGroup=20, while power was greater than 0.99 when nPerGroup=200 for the scenarios where data were generated from normal distributions without outliers (e.g. **Online Supplementary Figure S5**); and (4) departure from normality had effects on all 7 tests (e.g., comparing **Online Supplementary Figures S1, S7 and S11**).

Results of the Real Data Analysis

For the real data set GSE20080, the numbers of DV CpG sites (i.e., CpG sites with FDR-adjusted p-value <0.05) obtained by the 7 equal-variance tests are 0 (AWvar), 448 (Levene), 13 (BF), 39 (Ltrim), 330 (iAWvar.Levene), 0 (iAWvar.BF), and 14 (iAWvar.Ltrim), respectively. The cross table of overlapping DV CpG sites are shown in Table 2. We can see that the 448 DV CpG sites detected by Levene test contain the 330 DV CpG sites detected by iAWvar.Levene. The 330 DV CpG sites detected by iAWvar.Levene in turn contain the 39 DV CpG sites detected by Ltrim. The 39 DV CpG sites detected by Ltrim in turn contain the 14 DV CpG sites detected by iAWvar.Ltrim and the 13 DV CpG sites detected by BF. And 11 of the 13 DV CpG sites detected by BF were also detected by iAWvar.Ltrim.

Table 2: Cross table of overlapping DV CpG sites among the 7 equal-variance tests in the analysis of GSE20080.

	Levene	BF	Ltrim	AWvar	iAWvar.Levene	iAWvar.BF	iAWvar.Ltrim
Levene	448	13	39	0	330	0	14
BF	13	13	13	0	13	0	11
Ltrim	39	13	39	0	39	0	14
AWvar	0	0	0	0	0	0	0
iAWvar.Levene	330	13	39	0	330	0	14
iAWvar.BF	0	0	0	0	0	0	0
iAWvar.Ltrim	14	11	14	0	14	0	14

Table 3: Number of DV CpG sites for GSE20080 and number/proportion of DV CpG sites validated by GSE37020. nValidated/pValidated is the number/proportion of DV CpG sites that were validated in GSE37020.

Test	nSig	nValidated	pValidated (%)
Levene	448	276	61.6
BF	13	10	76.9
Ltrim	39	28	71.8
AWvar	0	NA	NA
iAWvar.Levene	330	226	68.5
iAWvar.BF	0	NA	NA
iAWvar.Ltrim	14	11	78.6

The numbers/proportions of validated DV CpG sites are shown in Table 3. We can see that iAWvar.Ltrim had the highest validation ratio (78.6\%), followed by BF (76.9\%), Ltrim (71.8\%), iAWvar.Levene (68.5\%), and Levene (61.6\%).

Discussion

In this article, we evaluated the performance of the AWvar score test and proposed three improved AW score tests for equality of variance. The simulation studies showed that the AWvar test is good at keeping nominal type I error rate for all 48 scenarios and had highest power for the scenarios where data were generated from normal distributions without outliers. For other scenarios, the other 6 tests performed better than the AWvar test in terms of power. Note that AWvar score test statistic can be rewritten as the difference of the two sample variances[11] and that F test, which is the ratio of two sample variances, is sensitive to non-normality and outliers. Hence, we expect that AWvar test is sensitive to non-normality and outliers too.

Levene, BF, and Ltrim tests are robust versions of F test, which are robust to outliers and departures of normality. In this article, we proposed three improved AWvar tests (iAWvar.Levene, iAWvar.BF, and iAWvar.Ltrim) based on Levene, BF, and Ltrim, respectively. The three improved AW score tests (iAWvar.Levene, iAWvar.BF, and iAWvar.Ltrim) had slightly larger power than, and had similar type I error rate to, their counterparts (Levene, BF, and Ltrim). For larger sample size nPerGroup=200, all 7 equal-variance tests had similar power, however, Levene and iAWvar.Levene tend to have high type I error rates.

The results of the real data analysis showed that Levene and iAWvar.Levene are much powerful than other 4 tests since Levene and iAWvar.Levene detected more than 8 times DV CpG sites than the other 5 equal-variance tests. The results also showed that the DV CpG sites detected by Levene and iAWvar.Levene contain the DV CpG sites detected by other 5 equal-variance tests. iAWvar.Ltrim had the highest proportion of validated DV CpG sites and iAWvar.Levene had higher proportion of validated DV CpG sites than Levene.

The results also showed that AWvar test is not as much as powerful than the other 6 equal-variance tests. The fact that iAWvar.BF did not find any DV CpG sites for GSE20080 indicates that real data are more complicated than simulated data sets and more investigations of the 3 improved AWvar tests are warranted.

Supplementary Table S1 listed the 14 DV CpG sites obtained by iAWvar.Ltrim. The test statistics and pvalues for both GSE20080 and GSE37020 are listed. The functional annotation clustering obtained by DAVID functional annotation tool[14] (**Supplementary Table S2**) showed that the 13 genes corresponding to the 14 DV CpG sites are related to biological functions of cell membrane, plasma membrane, and cell adhesion.

The parallel boxplots of methylation levels of the 14 DV CpG sites versus disease status (CIN2+ samples versus normal samples) are shown in **Supplementary Figure S17** (for GSE20080) and **Supplementary Figure S18** (for GSE37020). The variabilities of these 14 DV CpG sites are much larger in CIN2+ samples than in normal samples, which is consistent with what observed by [6].

The top DV CpG site cg00027083 (p-value=6.24 x 10^{-7}, FDR adjusted p-value=0.0053 in the analysis of GSE20080) detected by iAWvar.Ltrim is near the gene *EPB41L3* on chromosome 18. The full name of *EPB41L3* is erythrocyte membrane protein

band 4.1 like 3, which is a protein-coding gene. According to GeneCards[15], *EPB41L3* is a "*tumor suppressor that inhibits cell proliferation and promotes apoptosis*".

By searching PubMed using the keywords "*EPB41L3 and cervical*", we found that it is known in the literature that DNA methylation in the *EPB41L3* gene is associated with CIN2+[16-26]. For example, it has been reported that *EPB41L3* is a potential biomarker in cervical cancer and is often silenced by cancer-specific promoter methylation[22]. *EPB41L3* has also been used to construct classifier to triage women to reduce adverse events and costs associated with unnecessary colposcopy[16,18-21,24,25].

These results of literature searches show that the significant results obtained by iAWvar.Ltrim have biological meanings. Therefore, the improved AWvar tests could help identify DNA methylation marks that could potentially uncover the molecular differences between diseased samples and normal samples. It would be an interesting future work to find what additional information that differential variable DNA methylation marks could bring.

Acknowledgement

This work was sponsored by NIH P01 HL 132825.

References

1. Feinberg AP, RA Irizarry. 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. 2010; 107: 1757-1764.
2. Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med. 2010; 2: 49ra67.
3. Issa JP. Epigenetic variation and cellular Darwinism. Nat Genet. 2011; 43: 724-726.
4. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011; 43: 768-775.
5. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012; 13: 166-178.
6. Teschendorff AE, M Widschwendter. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. Bioinformatics. 2012; 28: 1487-1494.
7. Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. Genome Biol. 2014; 15: 465.
8. Li X, Qiu W, Morrow J, DeMeo DL, Weiss ST, et al. A Comparative Study of Tests for Homogeneity of Variances with Application to DNA Methylation Data. PLoS One. 2015; 10: e0145295.
9. Conover WJ, ME Johnson, MM Johnson. A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. Technometrics. 1981; 23: 351-361.
10. Brown MB, AB Forsythe. Robust tests for equality of variances. Journal of the American Statistical Association. 1974; 69: 364-367.
11. Ahn S, T Wang. A powerful statistical method for identifying differentially methylated markers in complex diseases. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2013; 69-79.
12. Teschendorff AE, Allison Jones, Heidi Fiegl, Alexandra Sargent, Joanna J Zhuang, et al. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome medicine. 2012; 4: 24.
13. Benjamini Y, Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995; 57: 289-300.
14. Huang da W, BT Sherman, RA Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4: 44-57.
15. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997; 13: 163.
16. Eijsink JJ, Lendvai Á, Deregowski V, Klip HG, Verpooten G, et al. A four-gene methylation marker panel as triage test in high-risk human papillomavirus positive patients. Int J Cancer. 2012; 130: 1861-1869.
17. Guerrero-Setas D, Pérez-Janices N, Blanco-Fernandez L, Ojer A, Cambra K, et al. RASSF2 hypermethylation is present and related to shorter survival in squamous cervical cancer. Mod Pathol. 2013; 26: 1111-1122.
18. Vasiljević N, Scibior-Bentkowska D, Brentnall AR, Cuzick J, Lorincz AT. Credentialing of DNA methylation assays for human genes as diagnostic biomarkers of cervical intraepithelial neoplasia in high-risk HPV positive women. Gynecol Oncol. 2014; 132: 709-714.
19. Brentnall AR, Vasiljević N, Scibior-Bentkowska D, Cadman L, Austin J, et al. A DNA methylation classifier of cervical precancer based on human papillomavirus and human genes. Int J Cancer. 2014; 135: 1425-1432.

20. Boers A, Bosgraaf RP, van Leeuwen RW, Schuuring E, Heideman DA, et al. DNA methylation analysis in self-sampled brush material as a triage test in hrHPV-positive women. Br J Cancer. 2014; 111: 1095-1101.

21. Louvanto K, Franco EL, Ramanakumar AV, Vasiljević N, Scibior-Bentkowska D, et al. Methylation of viral and host genes and severity of cervical lesions associated with human papillomavirus type 16. Int J Cancer. 2015; 136: E638-645.

22. Huisman C, van der Wijst MG, Falahi F, Overkamp J, Karsten G, et al. Prolonged re-expression of the hypermethylated gene EPB41L3 using artificial transcription factors and epigenetic drugs. Epigenetics. 2015; 10: 384-396.

23. Blanco-Luquin I, Guarch R. Differential role of gene hypermethylation in adenocarcinomas, squamous cell carcinomas and cervical intraepithelial lesions of the uterine cervix. Pathol Int. 2015; 65: 476-485.

24. Brentnall AR, Vasiljevic N, Scibior-Bentkowska D, Cadman L, Austin J, et al. HPV33 DNA methylation measurement improves cervical pre-cancer risk estimation of an HPV16, HPV18, HPV31 and EPB41L3 methylation classifier. Cancer Biomark. 2015; 15: 669-675.

25. Lorincz AT, Brentnall AR, Scibior-Bentkowska D, Reuter C, Banwait R, et al. Validation of a DNA methylation HPV triage classifier in a screening sample. Int J Cancer. 2016; 138: 2745-2751.

26. Boers A, Wang R, van Leeuwen RW, Klip HG, de Bock GH, et al. Discovery of new methylation markers to improve screening for cervical intraepithelial neoplasia grade 2/3. Clin Epigenetics. 2016; 8: 29.