Hindawi
International Journal of Genomics
Volume 2018, Article ID 6591634, 9 pages
https://doi.org/10.1155/2018/6591634

Research Article

Detecting Differentially Variable MicroRNAs via Model-Based Clustering

Xuan Li,[1] Yuejiao Fu ,[1] Xiaogang Wang ,[1] Dawn L. DeMeo,[2] Kelan Tantisira,[2] Scott T. Weiss,[2] and Weiliang Qiu [2]

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

Correspondence should be addressed to Yuejiao Fu; yuejiao@mathstat.yorku.ca

Received 31 December 2017; Revised 12 June 2018; Accepted 26 June 2018; Published 12 July 2018

Academic Editor: Wilfred van IJcken

Identifying differentially variable (DV) genomic probes is becoming a new approach to detect novel genomic risk factors for complex human diseases. The F test is the standard equal-variance test in statistics. For high-throughput genomic data, the probe-wise F test has been successfully used to detect biologically relevant DNA methylation marks that have different variances between two groups of subjects (e.g., cases versus controls). In addition to DNA methylation, microRNA (miRNA) is another important mechanism of epigenetics. However, to the best of our knowledge, no studies have identified DV miRNAs. In this article, we proposed a novel model-based clustering method to improve the power of the probe-wise F test to detect DV miRNAs. We imposed special structures on covariance matrices for each cluster of miRNAs based on the prior information about the relationship between variances in cases and controls and about the independence among them. Simulation studies showed that the proposed method seems promising in detecting DV probes. Based on two real datasets about human hepatocellular carcinoma (HCC), we identified 7 DV-only miRNAs (hsa-miR-1826, hsa-miR-191, hsa-miR-194-star, hsa-miR-222, hsa-miR-502-3p, hsa-miR-93, and hsa-miR-99b) using the proposed method, one (hsa-miR-1826) of which has not yet been reported to be related to HCC in the literature.

1. Introduction

Investigating the relationship between genomics and complex human diseases has greatly improved our understanding of the molecular mechanisms of, and the interplay of environmental factors and genomic factors to, the complex human diseases. High-throughput data from cutting-edge technologies have substantially facilitated the unbiased discovery of the genetic risk factors for many diseases. The standard approach to identify disease-associated genomic probes is to test if the mean level (e.g., DNA methylation) between cases and controls is significantly different. In addition to the mean, the variance is another important summary statistic. The larger the variance is, the more information the data could provide. However, the information about variance has not been directly used to detect disease-associated genomic probes until recent years.

Several groups of researchers have recently identified DNA methylation marks that have different variances between cases and controls [1–3]. They observed that (1) for differentially variable (DV) DNA methylation marks the variability in cases is usually higher than that in controls and (2) DV DNA methylation marks are biologically relevant. DNA methylation is an example of an epigenetic modification. Such modification leads to heritable changes via regulation of gene expression, without changing the genetic code. DNA methylation inhibits gene expression by adding a methyl group to the cytosine or adenine DNA nucleotides. Another example of an epigenetic modification is microRNAs (miRNAs) that are short noncoding 18–25-nucleotide-long RNA and negatively regulate mRNA translation [4, 5]. However, to the best of our knowledge, no studies have investigated differential variability for miRNAs. The main objective of this article is to develop statistical methods to detect DV miRNAs between cases and controls.

The *F* test is the classical method to test for equal variance between two groups of subjects, which evaluates whether the ratio of sample variances between two groups is significantly different from one. For high-throughput genomic data, such as DNA methylation data, the probe-wise *F* test could be used. That is, we first perform the *F* test for each probe to test for equal variances between cases and controls. We then calculate FDR-adjusted *p* value to control for multiple testing, where FDR stands for false discovery rate. If the FDR-adjusted p value < 0.05 for a DNA methylation mark, we then claim that this DNA methylation mark is differentially variable between cases and controls. The advantages of this probe-wise approach include flexibility (one model per probe) and easy implementation. However, DV probes might be governed by the same underlying mechanism. Statistically speaking, DV probes might follow the same distribution. Similarly, non-DV probes might also follow the same distribution. We hypothesize that these underlying distributions of variances could help us improve the power of the *F* test to detect DV probes.

A few methods have been proposed in the literature to borrow information across probes to detect differentially variable genomic probes. For example, Bar et al. [6] proposed a mixture-model approach for parallel detection of differential variances in genomic data analysis, by assuming that the ratio of sample variances between two groups for a given probe is drawn from a three-component mixture. Bar et al. [7] introduced a bivariate model (N3) to account for both differential expression and differential variation in high-throughput data analysis, by assuming that both means and variances follow three-component mixture distributions. Bar and Schifano [8] proposed a unified three-component mixture model, the L_2N model, that can be used to detect either differential expression (mean) or differential variation, by modeling the differences of means and variances (dispersions) between two groups of samples. In the L_2N model, one log-normal component is used to fit underexpressed (dispersed) probes, one log-normal component is used to fit overexpressed (dispersed) probes, and one normal component is used to fit nondifferentially expressed (dispersed) probes. These models characterize the distributions of the summary statistics (e.g., mean, variance, or difference of means), instead of the observed expression levels.

In this article, we propose a mixture of three-component multivariate normal distributions to fit the expression levels of miRNAs to identify DV miRNAs between cases and controls.

2. Method

2.1. Model. We assume that miRNAs belong to one and only one of the following three clusters: (1) miRNAs having higher variances in cases than in controls (denoted as the OV cluster), (2) miRNAs having equal variances between cases and controls (denoted as the EV cluster), and (3) miRNAs having smaller variances in cases (denoted as the UV cluster). We followed Qiu et al. [9] to directly model the marginal distributions of miRNAs in the 3 clusters. In this article, we modified Qiu et al.'s marginal model [9] to allow the detection of DV probes. We assume that (1) data have been normalized to remove the effects of confounding factors, such as chip effect and batch effect, and (2) data have been transformed so that the distributions of miRNA expressions are close to normal distributions.

For a given miRNA, we denote X_i as the preprocessed expression for the *i*th subject, $i = 1, \dots, m$, where $m = m_c + m_n$, m_c is the number of cases, and m_n is the number of controls. For the *k*th cluster ($k = 1$, 2, or 3), we assume that the expressions of the m_c cases are identically distributed with mean μ_{kc} and variance σ_{kc}^2. We assume that the expressions of the m_n controls are identically distributed with mean μ_{kn} and variance σ_{kn}^2. According to Qiu et al. [9], X_i's are marginally correlated with correlation ρ_{kc} for cases and ρ_{kn} for controls. We also assume that (1) cases and controls are independent, and (2) the $m \times 1$ random vector $(X_1, \dots, X_m)^T$ follows a multivariate normal distribution. For the OV cluster, we require that $\sigma_{1c}^2 > \sigma_{1n}^2$. For the UV cluster, we require that $\sigma_{3c}^2 < \sigma_{3n}^2$. For the EV cluster, we require that $\sigma_{2c}^2 = \sigma_{2n}^2$. We allow the means and correlations to be different between cases and controls in the EV cluster.

We used the EM algorithm [10] to estimate the model parameters μ_{kc}, σ_{kc}^2, μ_{kn}, and σ_{kn}^2. The posterior probability $p_{gk} = Pr(g\text{th miRNA in } k\text{th cluster} \mid x) = \pi_1 f_1(x)/[\pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x)]$ is used to assign the *g*th miRNA to one of the 3 clusters, where $f_k(x)$ is the density function of the multivariate normal distribution for the *k*th cluster. If p_{g1} is the largest posterior probability among p_{g1}, p_{g2}, and p_{g3}, then the *g*th miRNA will be assigned to the 1st cluster (i.e., OV cluster). The supplementary document gives the details about the model and the corresponding parameter estimation procedure.

2.2. Real Datasets. We downloaded two miRNA datasets from NIH's Gene Expression Omnibus (GEO) [11]: GSE67138 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67138) and GSE67139 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67139). Both datasets are from the same project that aims at detecting miRNAs differentially expressed between human hepatocellular carcinoma (HCC) tumor tissues with and without vascular invasion. GSE67138 is the first batch containing 57 samples (34 invasive tumor tissues and 23 noninvasive tumor tissues), while GSE67139 is the second batch containing 120 samples (60 invasive tumor tissues and 60 noninvasive tumor tissues). The expression levels of miRNAs in both GEO datasets were measured by using Affymetrix Multispecies miRNA-1 Array (GPL8786). Both datasets contain 847 miRNAs.

We checked the data quality by visualizing the plot (Figure A1) of percentiles across arrays and the scatterplot (Figure A2) of the first two principal components. Both plots indicate that the two datasets have been cleaned and have good quality (i.e., no apparent outlying miRNAs, outlying arrays, or technical batch effects). Hence, we directly used the two datasets in the further analyses. Since GSE67139 has a larger sample size than GSE61738, we regarded GSE67139 as the discovery set and GSE67138 as the validation set.

2.3. Simulation. We conducted 4 sets of simulation studies. In the first set (denoted as SimI), we generated miRNA data from the proposed marginal mixture model, where estimated model parameters for GSE67139 (i.e., the discovery set) are used as the true values of the model parameters ($\pi_1 = 0.31$, $\pi_2 = 0.58$, $\pi_3 = 0.11$, $\mu_{1c} = -0.14$, $\sigma^2_{1c} = 1.49$, $\rho_{1c} = 0.08$, $\mu_{1n} = 0.14$, $\sigma^2_{1n} = 0.45$, $\rho_{1n} = 0.32$, $\mu_{2c} = 0.03$, $\sigma^2_{2c} = 1.01$, $\rho_{2c} = 0.04$, $\mu_{2n} = -0.03$, $\sigma^2_{2n} = 1.01$, $\rho_{2n} = 0.11$, $\mu_{3c} = 0.13$, $\sigma^2_{3c} = 0.28$, $\rho_{3c} = 0.04$, $\mu_{3n} = -0.13$, $\sigma^2_{3n} = 1.69$, $\rho_{1n} = -0.01$). We generated 100 datasets, each of which has 1000 miRNAs for 50 cases and 50 controls. Thirty one percent (310) of the 1000 miRNAs are in the OV cluster. Eleven percent (110) of the miRNAs are in the UV cluster. The remaining 58% (580) miRNAs are in the EV cluster.

In the second set (denoted as SimII), we generated miRNA data from a mixture of 3 multivariate t distribution with the same mean vectors and covariance matrices as those in SimI and with 3 degrees of freedom. SimII is used to evaluate the performance of the proposed method when the normality assumption for any one of the three clusters (OV, EV, and UV) is violated.

In the third set (denoted as SimIII) of the simulation studies, we generated miRNA data from the same model as that in SimI, except that the marginal correlation within-subject groups were set to zero ($\rho_{kc} = 0$ and $\rho_{kn} = 0$). SimIII is used to evaluate the performance of the proposed method when there are no marginal correlations.

In the fourth set (denoted as SimIV) of the simulation studies, we generated miRNA data from the same model as that in SimII, except that the marginal correlations within-subject groups were set to zero ($\rho_{kc} = 0$ and $\rho_{kn} = 0$). SimIV is used to evaluate the performance of the proposed method when there are no marginal correlations and when the normality assumption for any one of the three clusters (OV, EV, and UV) is violated.

2.4. Statistical Analysis. We compared the proposed method (denoted as gs) with sixteen existing differential-variance detecting methods by using both the real datasets and the simulated datasets. The ten equal variance tests are (1) the F test (denoted as F), (2) Ahn and Wang's score test [12] (denoted as AW), (3) Phipson and Oshlack's AD test [13] (denoted as PO.AD), (4) Phipson and Oshlack's SQ test [13] (denoted as PO.SQ), (5) Levene's test [14] (denoted as L), (6) Brown and Forsythe's test [15] (denoted as BF), (7) trimmed-mean-based Levene's test [15] (denoted as Ltrim), (8) improved AW test based on Levene's test [16] (denoted as iL), (9) improved AW test based on the BF test [16] (denoted as iBF), and (10) improved AW test based on the trimmed-mean-based Levene's test [16] (denoted as iTrim). The remaining six methods are based on Bar et al.'s [7] N3 model and Bar and Schifano's [8] L_2N model. Both N3 and L_2N models have been implemented in the R package DVX [8]. For both N3 and L_2N, DVX outputs raw p values, q values, and posterior probabilities p_{gk} that the probe g belongs to cluster k given its expression profile and estimated model parameters, $k = 1, 2, 3$. Hence, for both N3 and L_2N, we used three methods to assign probes to two clusters: DV probes and non-DV probes. The first method is based on the q value. If a miRNA has a q value < 0.05, we claim it is differentially variable; otherwise, we claim it is nondifferentially variable. The second method is based on the false discovery rate- (FDR-) adjusted p value. If a miRNA has an FDR-adjusted p value < 0.05, we claim it is differentially variable; otherwise, we claim it is nondifferentially variable. The third method is based on the posterior probabilities. We assign a miRNA to cluster $k*$ if the posterior probability p_{gk*} is the largest among the 3 posterior probabilities, p_{g1}, p_{g2}, and p_{g3}. We denote the 3-miRNA assignment methods as N3.q (L_2N.q), N3.f (L_2N.f), and N3 (L_2N), respectively.

In real data analysis, we followed Qiu et al.'s [9] data preprocessing steps. That is, we first performed the same Box-Cox transformation for each expression level, and then for each miRNA, we performed mean centering and scaling operations so that the mean expression level is 0 and the variance is 1. We then applied the 17 methods (the gs method and the 16 existing methods) to the discovery set (GSE67139) to detect DV miRNAs between invasive tumors and noninvasive tumors. For the 10 probe-wise tests (F, AW, PO.AD, PO.SQ, L, BF, Ltrim, iL, iBF, and iTrim), we obtained FDR-adjusted p values. If a miRNA has an FDR-adjusted p value < 0.05, we claim that this miRNA has significantly different variances between invasive tumors and noninvasive tumors. We then applied the same procedure to the validation set (GSE67138). We claim that a miRNA is a validated DV miRNA (1) if the miRNA is DV in both discovery and validation sets and (2) if the sign of the difference ($s_c^2 - s_n^2$) is the same in both datasets, where s_c^2 and s_n^2 are sample variances for cases and controls, respectively. We next calculated the proportion of the validated DV miRNAs (i.e., validation rate) pValid $= n_{12}/n_1$, where n_1 is the number of DV miRNAs in the discovery set (GSE67139) and n_{12} is the number of significant DV miRNAs sharing the same difference direction of variances in both data sets. To estimate the variation of the validation rate pValid, we obtained the 100 bootstrap validation rates based on 100 bootstrap discovery and validation sets. We then test if the median bootstrap validation rate of the gs method is the same as that of each of the other 16 methods by two-sided Wilcoxon signed rank tests.

For the validated DV miRNAs detected by the gs method, we also checked if they are validated differentially expressed (DE) miRNAs by using R Bioconductor package *limma* [17]. A miRNA is a validated DE miRNA if the FDR-adjusted p value for testing equal mean expression between cases and controls is <0.05 in both the discovery and validation sets and if the sign of the mean difference $\bar{x}_c - \bar{x}_n$ is the same in both discovery and validation sets, where $\bar{x}_c$ and $\bar{x}_n$ are the sample means of the cases and controls, respectively. Denote S_{DVonly} as the set of miRNAs that are validated DV, but not validated DE. Denote S_{DEonly} as the set of miRNAs

Table 1: Information about the validated DV miRNAs.

Method	nSig	n.OV	n.UV	nValid	nValid.OV	nValid.UV	pValid
gs	358	262	96	67	66	1	0.19
F	472	349	123	99	96	3	0.21
AW	141	136	0	33	33	0	0.23
PO.AD	202	186	0	68	68	0	0.34
PO.SQ	141	136	5	32	32	0	0.23
L	201	185	16	72	70	2	0.36
BF	175	164	11	54	53	1	0.31
Ltrim	181	168	0	58	56	2	0.32
iL	199	183	16	70	68	2	0.35
iBF	174	163	11	53	52	1	0.30
iTrim	181	168	13	57	55	2	0.31
L2N	225	121	104	30	29	1	0.13
L2N.q	173	69	104	17	16	1	0.10
L2N.f	157	60	97	16	15	1	0.10
N3	247	141	106	34	33	1	0.14
N3.q	202	96	106	25	24	1	0.12
N3.f	178	74	104	18	17	1	0.10

nSig: the number of the DV miRNAs detected in the discovery set (GSE67139); n.OV: the number of OV miRNAs detected in the discovery set; n.UV: the number of UV miRNAs detected in the discovery set; nValid: the number of validated DV miRNAs; nValid.OV: the number of validated OV miRNAs; nValid.UV: the number of validated UV miRNAs; pValid = nValid/nSig.

that are validated DE, but not validated DV. Denote S_{both} as the set of miRNAs that are both validated DE and validated DV.

We applied the miRSystem [18] to predict the target genes of miRNAs in each of the 3 sets: S_{DVonly}, S_{DEonly}, and S_{both}. The miRSystem also provides the enriched KEGG pathways for the predicted target genes.

For simulated datasets, we calculated the magnitude of agreement between the true cluster memberships of miRNAs and the detected cluster memberships by each of the 17 methods by using the Jaccard index [9, 19]. The maximum value of the Jaccard index is one, indicating perfect agreement. The minimum value of the Jaccard index is zero, indicating that the agreement is by chance. We also evaluate the performances using false positive rate (FPR) (i.e., the proportion of detected DV probes among the true non-DV probes) and false negative rate (FNR) (i.e., the proportion of detected non-DV probes among the true DV probes). The smaller the FPR (FNR) is, the better the performance is.

3. Result

For the real data analyses, the numbers of the DV miRNAs in the discovery set (GSE67139), and the numbers and proportions of the validated DV miRNAs are shown in Table 1. The gs method detected 358 DV probes based on the discovery set (GSE67139), 67 of which are validated in the validation set (GSE67138). Among the 67 validated DV miRNAs, 66 miRNAs are OV and only one miRNA is UV. The proportion of the validated DV miRNAs is 0.19 for the gs method, which is higher than those of the N3 and L_2N methods but lower than those of the 10 probe-wise tests. However, the gs method had the highest median bootstrap validation rate among all 17 methods (Figure 1). For all the 17 methods, the number (nValid.OV) of the validated OV miRNAs is much larger than the number (nValid.UV) of the validated UV miRNAs. This observation is consistent with that observed by other researchers using DNA methylation data [3].

We got 392 DE miRNAs based on the discovery set (GSE67139), among which 217 DE miRNAs were validated. There are only 7 miRNAs in S_{DVonly} (hsa-miR-1826, hsa-miR-191, hsa-miR-194-star, hsa-miR-222, hsa-miR-502-3p, hsa-miR-93, and hsa-miR-99b), the parallel boxplots of which are shown in Figure A3. S_{DEonly} contains 157 miRNAs (Table A1), the parallel boxplots of which are shown in Figure A4. S_{both} contains 60 miRNAs (Table A2), the parallel boxplots of which are shown in Figure A5.

Based on the miRSystem analysis, there are 1639 genes (Table A3) targeted by the 7 miRNAs in S_{DVonly}, 8141 targeted genes (Table A4) for the 157 miRNAs in S_{DEonly}, and 6893 targeted genes (Table A5) for the 60 miRNAs in S_{both}. The 1639 genes targeted by the 7 miRNAs in S_{DVonly} are significantly enriched (raw p value < 0.05) in 6 KEGG pathways (calcium signaling pathway, salivary secretion, amyotrophic lateral sclerosis (ALS), MAPK signaling pathway, PPAR signaling pathway, and Alzheimer's disease) (Table A6). The 8141 genes targeted by the 157 miRNAs in S_{DEonly} are significantly enriched in only one KEGG pathway (antigen processing and presentation) with raw p value $= 2.70E-2$ (Table A7). The 6893 genes targeted by the 60 miRNAs in S_{both} are enriched in two KEGG pathways (O-glycan biosynthesis and glycine serine and threonine metabolism) (Table A8).

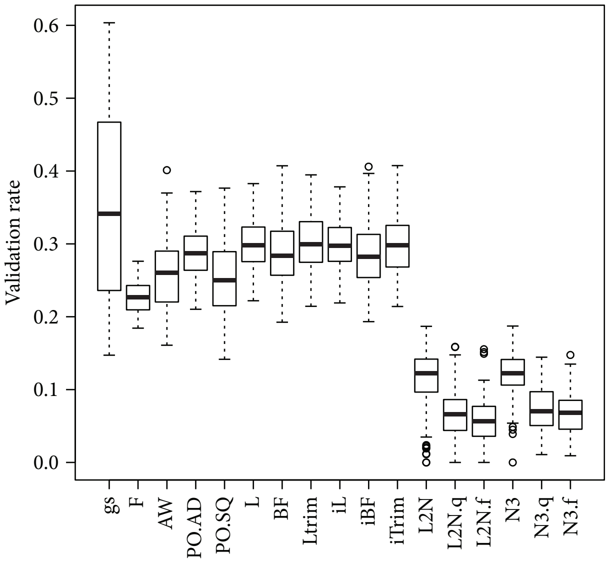

Figure 1: Boxplots of validation rates based on 100 bootstrap samples of the discovery set (GSE67139) and the validation set (GSE67138).

For the first and the third simulation studies (SimI and SimIII) where data were generated from a mixture of multivariate normal distributions, the values of the Jaccard index obtained by the gs method are close to one (the perfect agreement). Figures 2 and 3 showed that the boxplots of the Jaccard index obtained by the gs method are higher than boxplots obtained by the other 16 methods. The small p values in Table A9 showed that the gs method had significantly higher values of the Jaccard index than the 16 existing equal-variance tests. Figures 2 and 3 and Tables A10 and A11 also showed that the gs method had significantly smaller FPR and FNR than the other 16 methods, except that the gs method had significantly higher median FNR in SimI (Figure 2).

Figures 4 and 5 and Table A9 showed that for SimII and SimIV where data was generated from a mixture of multivariate t distributions, the N3 and L_2N methods had the highest Jaccard index values (median Jaccard index > 0.50), while the remaining 11 methods had a median Jaccard index < 0.4. The gs method still had much higher values of the Jaccard index than the 10 probe-wise tests (F, AW, PO.AD, PO.SQ, L, BF, Ltrim, iL, iBF, and iTrim). Figures 4 and 5 and Tables A10 and A11 also showed that the gs method had a lower FPR than the 10 probe-wise tests, but had higher FNR than these 10 probe-wise tests. Although the N3 and L2N methods had very low FPR, they had FNR close to one.

Tables A9–A12 showed the p values of two-sided Wilcoxon signed rank tests to evaluate if the median measures (Jaccard index, FPR, FNR, and bootstrap validation rate) are the same as those of each of the other 16 methods, respectively. All p values are smaller than 0.05, indicating that all the differences are statistically significant.

4. Discussion

In this article, we proposed a novel model-based clustering method (the gs method) to detect miRNAs having different variances between cases and controls. The proposed method is different from probe-wised equal-variance tests in that it does not involve hypothesis testing. The real data analysis showed that the gs method had a larger median bootstrap validation rate than the 16 existing equal-variance detecting methods. The four simulation studies showed that the gs method outperformed the 16 existing equal-variance detection methods if the miRNA data follow a mixture of multivariate normal distributions. If the data were generated from other distributions, such as a mixture of multivariate t distributions, the gs method had a lower FPR and a higher FNR than the 10 probe-wise tests. Since controlling FPR is more important than controlling FNR, the gs method is promising in the genomic data analysis.

Several model-based clustering algorithms have been proposed to detect DV genomic probes in the literature, such as Bar et al.'s [7] N3 methods, and Bar and Schifano's [8] L_2N methods. The N3 methods and L2N methods do not seem to work as well as the gs method under the simulation scenarios in this article. This is probably partly due to the gs method that directly models the observed expression levels, while the N3 and L_2N methods model the summary statistics (e.g., mean, variance, or difference of means). Using summary statistics might cause the loss of information. Moreover, the N3 and L_2N methods applied a couple of approximations to derive the marginal densities, while approximations might cause deviations from true marginal densities.

In the simulation studies, the proposed method outperformed the 10 probe-wised tests, including the classic F test that has been reported to outperform other equal-variance tests when the normality assumption is held [20, 21]. The reason why the gs method performed better than the F test in SimI and SimIII, where the normality assumption for any one of the three clusters (OV, EV, and UV) is held, is that the gs method could borrow information across miRNAs

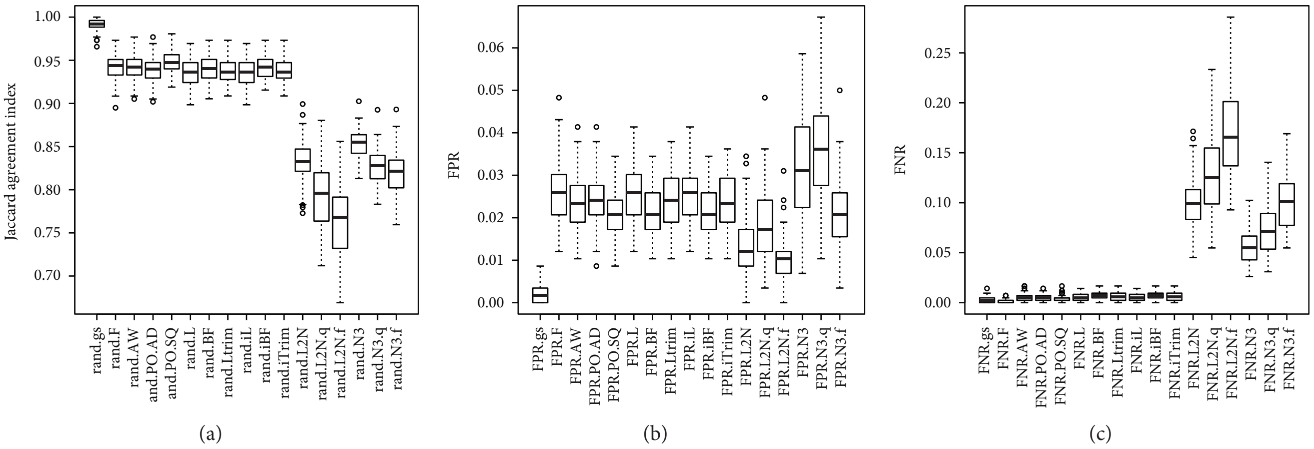

Figure 2: The boxplots of the 100 estimated Jaccard indices, FPR, and FNR based on the 100 simulated datasets in SimI (generating data from a mixture of multivariate normal distributions with nonzero marginal correlations). The closer the Jaccard index is to one, the better the performance of the method is. The closer the FPR (FNR) is to zero, the better the performance of the method is.

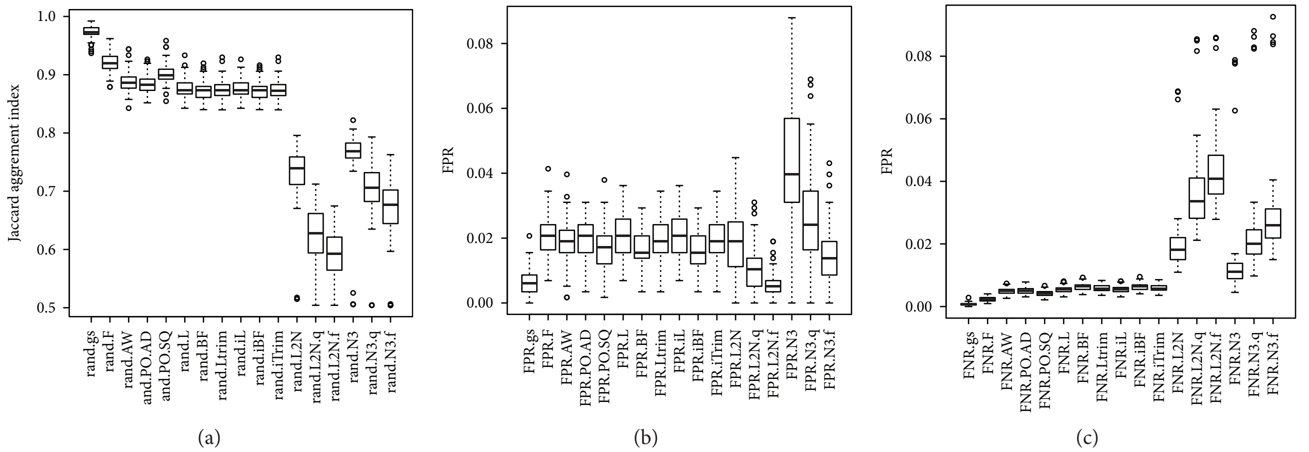

Figure 3: The boxplots of the 100 estimated Jaccard indices based on the 100 simulated datasets in SimIII (generating data from a mixture of multivariate normal distributions with zero marginal correlations). The closer the Jaccard index is to one, the better the performance of the method is. The closer the FPR (FNR) is to zero, the better the performance of the method is.

(i.e., the estimation of the model parameters uses the information provided by all the miRNAs).

The gs method had lower FPR than the 10 probe-wised equal-variance tests in SimII and SimIV, where the data were generated from a mixture of multivariate *t* distribution (i.e., the normality assumption for any one of the three clusters (OV, EV, and UV) is violated) (Figure A6). However, the gs method had higher FNR than the 10 probe-wised tests in SimII and SimIV. The results of SimII and SimIV also suggest that the large variation of the validation rate in the real data analyses (Figure 1) may be due to the violation of the underlying assumption that the expression levels are from the mixture of multivariate normal distributions.

Figure A7 showed that the distributions of the original real datasets are quite different from a mixture of normal distributions. Note that in genomic data analysis, majority probes are supposed to be nondifferentially expressed. Hence, we can use a histogram and QQ plot to roughly check if data are from a mixture of normal distributions. We followed Qiu et al.'s [9]data preprocessing steps. That is, we first performed the same Box-Cox transformation to each expression level, and then for each miRNA, we performed mean-centering and scaling operations so that the mean expression level is 0 and the variance is 1. Figure A7 showed that even after the Box-Cox transformation and scaling, the data distributions are still quite different from normal distributions. Further investigation is warranted to develop more robust model-based clustering methods.

In the real data analysis, the gs method detected 67 validated DV miRNAs (66 OV and 1 UV), seven of which are DV only. The 7 DV-only miRNAs (hsa-miR-1826, hsa-miR-191, hsa-miR-194-star, hsa-miR-222, hsa-miR-502-3p, hsa-miR-93, and hsa-miR-99b) were targeted to 1639 genes based on the miRSystem analysis. Except for hsa-miR-1826, all DV-only miRNAs have been associated with HCC. Elyakim et al. [22] showed that miR-191 is a candidate oncogene target for hepatocellular carcinoma therapy. Law and Wong [23] reported the association of miR-194 with

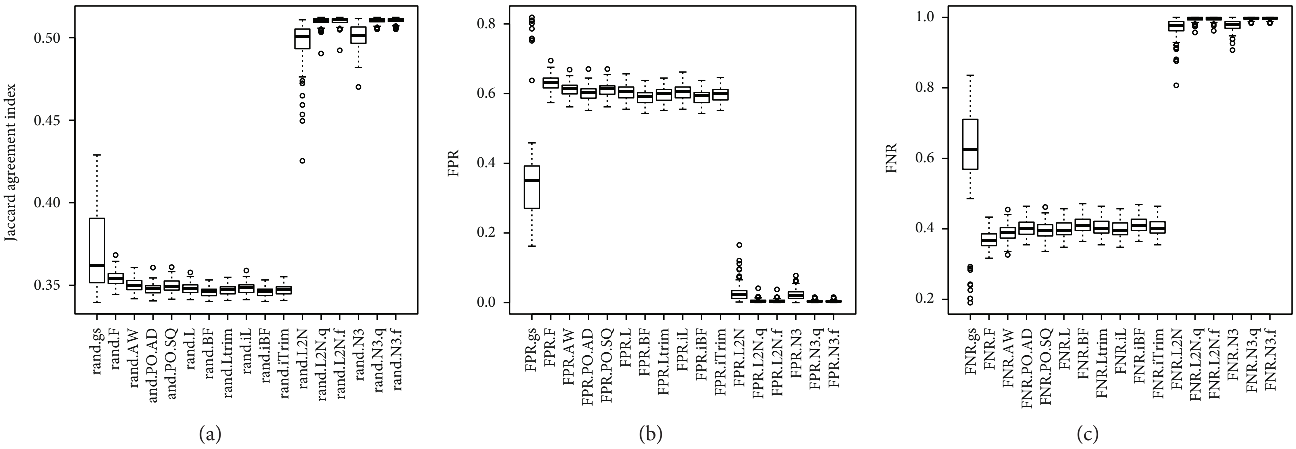

Figure 4: The boxplots of the 100 estimated Jaccard indices based on the 100 simulated datasets in SimII (generating data from a mixture of multivariate t distributions with nonzero marginal correlations). The closer the Jaccard index is to one, the better the performance of the method is. The closer the FPR (FNR) is to zero, the better the performance of the method is.

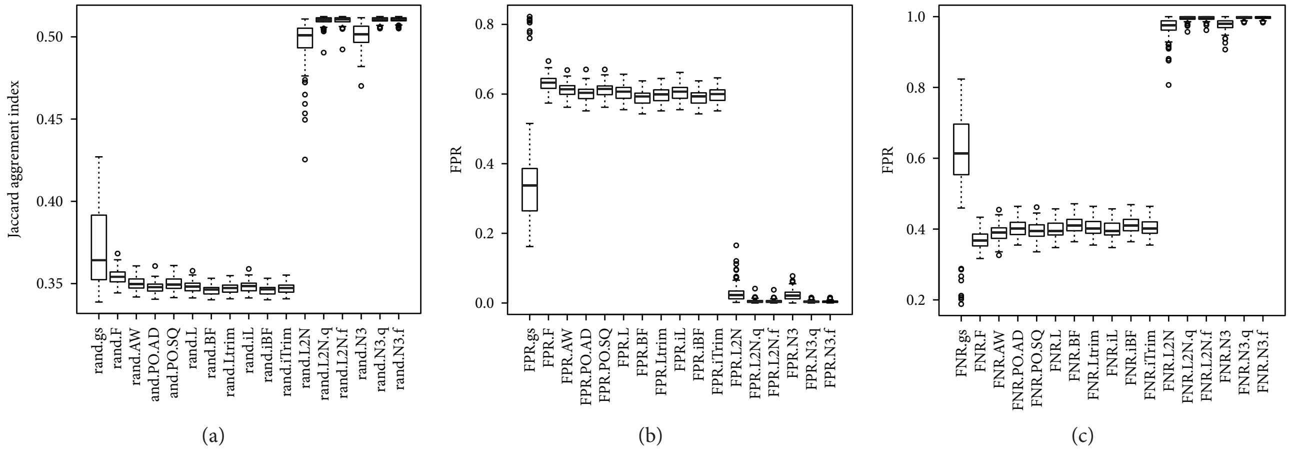

Figure 5: The boxplots of the 100 estimated Jaccard indices based on the 100 simulated datasets in SimIV (generating data from a mixture of multivariate t distributions with zero marginal correlations). The closer the Jaccard index is to one, the better the performance of the method is. The closer the FPR (FNR) is to zero, the better the performance of the method is.

the metastatic behavior of HCC. Murakami et al [24] reported that miR-222 is increased in poorly versus moderately versus well-differentiated hepatomas. Jin et al. [25] reported that miR-502-3p suppressed cell proliferation, migration, and invasion in HCC by targeting SET. Li et al. [26] confirmed that the miR-106b-25 cluster, which miR-93 belongs to, is overexpressed in HCC. Morishita et al. [27] found that miR-99b is upregulated in HBV-infected HCC cells.

The 1639 genes, which are targeted by the 7 DV-only miRNAs, are enriched in 6 KEGG pathways (calcium signaling pathway, salivary secretion, amyotrophic lateral sclerosis (ALS), MAPK signaling pathway, PPAR signaling pathway, and Alzheimer's disease). All these 6 pathways have been linked to HCC in the literature. For example, Huang et al. [28] reported that increased mitochondrial fission induced cytosolic calcium signaling in HCC cells. Chen et al. [29] reported that in a mice study, DNA methylation marks that are differentially methylated between livers with HCC and livers without HCC are enriched in the salivary secretion pathway. Seol et al.'s [30] results suggest that Riluzole, an amyotrophic lateral sclerosis (ALS) drug, has an anticancer effect on HCC. Feng et al. [31] reported that cantharidic acid inhibits HCC cell proliferation by inducing cell apoptosis through the p38 MAPK signaling pathway. Nwosu et al. [32] reported that downregulated genes (HCC versus non-HCC) were enriched in the PPAR signaling pathway based on each of the 8 HCC datasets downloaded from the Gene Expression Omnibus (GEO). Jin et al. [33] reported that kynurenine 3-monooxygenase (KMO), an enzyme playing a critical role in Huntington's and Alzheimer's diseases, exhibits tumor-promoting effects towards HCC. Hence, DV-only miRNAs are biologically relevant to HCC.

There are no overlaps among the enriched pathways for the 3 sets of miRNAs in the real data analysis: S_{DVonly} (the set of miRNAs that are validated DV, but not validated DE), S_{DEonly} (the set of miRNAs that are validated DE, but not validated DV), and S_{both} (the set of miRNAs that are both validated DE and validated DV). This indicates that DV-only miRNAs might provide additional information about the molecular mechanisms of HCC than that provided by DE miRNAs.

In summary, the proposed gs method assumes expression levels from the mixture of multivariate normal distributions. The proposed gs method seems promising to detect differential variability based on our simulation studies. In the future, we will improve it into a robust version against the violation of the normality assumption on the component distributions.

Data Availability

The authors downloaded two microRNA datasets from NIH's Gene Expression Omnibus (GEO) [8]: GSE67138 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67138) and GSE67139 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67139).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research is partially supported by the NSERC Discovery Grants (RGPIN-2018-05846, RGPIN-2018-05981) and by the National Institute of Health of the United States (NIH/NHLBI 1 P01 HL 132825-01, NIH/NHLBI 5 R01 HL 125734).

Supplementary Materials

Details about the gs model and its parameter estimation via the EM algorithm. Supplemental Table A1: the list of the 157 DE-only miRNAs. Supplemental Table A2: the list of the 60 DV-and-DE miRNAs. Supplemental Table A3: the list of the 1639 genes targeted by the 7 miRNAs in S_{DVonly}. Supplemental Table A4: the list of the 8141 genes targeted by the 157 miRNAs in S_{DEonly}. Supplemental Table A5: the list of the 6893 genes targeted by the 60 miRNAs in S_{both}. Supplemental Table A6: the list of the 6 KEGG pathways enriched by the 1639 genes targeted by the 7 miRNAs in S_{DVonly}. Supplemental Table A7: the list of the 1 KEGG pathway enriched by the 8141 genes targeted by the 157 miRNAs in S_{DEonly}. Supplemental Table A8: the list of the 2 KEGG pathways enriched by the 6893 genes targeted by the 60 miRNAs in S_{both}. Supplemental Table A9: p values of two-sided Wilcoxon signed rank tests to test if the median Jaccard index obtained by the gs method is the same as that obtained by each of the other 16 methods. Supplemental Table A10: p values of two-sided Wilcoxon signed rank tests to test if the median FPR obtained by the gs method is the same as that obtained by each of the other 16 methods. Supplemental Table A11: p values of two-sided Wilcoxon signed rank tests to test if the median FNR obtained by the gs method is the same as that obtained by each of the other 16 methods. Supplemental Table A12: p values of two-sided Wilcoxon signed rank tests to test if the median proportion of validation obtained by the gs method is the same as that obtained by each of the other 16 methods based on 100 bootstrap samples. Figure A1: the plot of percentiles of log2 expression levels across arrays. Left panel: GSE67138. Right panel: GSE67139. Figure A2: the plot of the first and second principal components. Left panel: GSE67138. Right panel: GSE67139. Supplemental Figure A3: parallel boxplots of the 7 validated DV-only miRNAs. Left panel: GSE67138. Right panel: GSE67139. Supplemental Figure A4: parallel boxplots of the 157 validated DE-only miRNAs. Left panel: GSE67138. Right panel: GSE67139. Supplemental Figure A5: parallel boxplots of the 60 validated DV-and-DE miRNAs. Left panel: GSE67138. Right panel: GSE67139. Supplemental Figure A6: histograms and QQ plots for a simulated dataset in each of the 4 simulation scenarios. Supplemental Figure A7: histograms and QQ plots for the two real datasets. Note that in genomic data analysis, majority probes are supposed to be nondifferentially expressed. Hence, we can use a histogram and QQ plot to roughly check if data are from a mixture of normal distributions. *(Supplementary Materials)*

References

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

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

[3] A. E. Teschendorff and M. Widschwendter, "Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions," *Bioinformatics*, vol. 28, no. 11, pp. 1487–1494, 2012.

[4] J. Hayes, P. P. Peruzzi, and S. Lawler, "MicroRNAs in cancer: biomarkers, functions and therapy," *Trends in Molecular Medicine*, vol. 20, no. 8, pp. 460–469, 2014.

[5] B. Simonson and S. Das, "MicroRNA therapeutics: the next magic bullet?," *Mini Reviews in Medicinal Chemistry*, vol. 15, no. 6, pp. 467–474, 2015.

[6] H. Y. Bar, J. G. Booth, and M. T. Wells, "A mixture-model approach for parallel testing for unequal variances," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, pp. 1–21, 2012.

[7] H. Y. Bar, J. G. Booth, and M. T. Wells, "A bivariate model for simultaneous testing in bioinformatics data," *Journal of the American Statistical Association*, vol. 109, no. 506, pp. 537–547, 2014.

[8] H. Bar and E. D. Schifano, *Differential Variation and Expression Analysis*, BioRxiv, 2018.

[9] W. Qiu, W. He, X. Wang, and R. Lazarus, "A marginal mixture model for selecting differentially expressed genes across two types of tissue samples," *The International Journal of Biostatistics*, vol. 4, no. 1, 2008.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[11] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[12] S. Ahn and T. Wang, "A powerful statistical method for identifying differentially methylated markers in complex diseases," *Pacific Symposium on Biocomputing*, pp. 69–79, 2013.

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

[14] H. Levene, "Robust tests for equality of variances," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, Ed., pp. 278–292, Stanford University Press, Palo Alto, CA, USA, 1960.

[15] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

[16] W. Qiu, X. Li, J. Morrow et al., "New score tests for equality of variances in the application of DNA methylation data analysis," *Insights in Genetics and Genomics*, vol. 1, p. 3.1, 2017.

[17] M. E. Ritchie, B. Phipson, D. Wu et al., "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, article e47, 2015.

[18] T. P. Lu, C. Y. Lee, M. H. Tsai et al., "miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets," *PLoS One*, vol. 7, no. 8, article e42390, 2012.

[19] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[20] W. J. Conover, M. E. Johnson, and M. M. Johnson, "A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data," *Technometrics*, vol. 23, no. 4, pp. 351–361, 1981.

[21] X. Li, W. Qiu, J. Morrow et al., "A comparative study of tests for homogeneity of variances with application to DNA methylation data," *PLoS One*, vol. 10, no. 12, article e0145295, 2015.

[22] E. Elyakim, E. Sitbon, A. Faerman et al., "hsa-miR-191 is a candidate oncogene target for hepatocellular carcinoma therapy," *Cancer Research*, vol. 70, no. 20, pp. 8077–8087, 2010.

[23] P. T.-Y. Law and N. Wong, "Emerging roles of microRNA in the intracellular signaling networks of hepatocellular carcinoma," *Journal of Gastroenterology and Hepatology*, vol. 26, no. 3, pp. 437–449, 2011.

[24] Y. Murakami, T. Yasuda, K. Saigo et al., "Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues," *Oncogene*, vol. 25, no. 17, pp. 2537–2545, 2006.

[25] H. Jin, M. Yu, Y. Lin et al., "MiR-502-3P suppresses cell proliferation, migration, and invasion in hepatocellular carcinoma by targeting SET," *OncoTargets and Therapy*, vol. 9, pp. 3281–3289, 2016.

[26] Y. Li, W. Tan, T. W. L. Neo et al., "Role of the *miR-106b-25* microRNA cluster in hepatocellular carcinoma," *Cancer Science*, vol. 100, no. 7, pp. 1234–1242, 2009.

[27] A. Morishita, H. Iwama, S. Fujihara et al., "MicroRNA profiles in various hepatocellular carcinoma cell lines," *Oncology Letters*, vol. 12, no. 3, pp. 1687–1692, 2016.

[28] Q. Huang, H. Cao, L. Zhan et al., "Mitochondrial fission forms a positive feedback loop with cytosolic calcium signaling pathway to promote autophagy in hepatocellular carcinoma cells," *Cancer Letters*, vol. 403, pp. 108–118, 2017.

[29] H. Chen, W. Cai, E. S. H. Chu et al., "Hepatic cyclooxygenase-2 overexpression induced spontaneous hepatocellular carcinoma formation in mice," *Oncogene*, vol. 36, no. 31, pp. 4415–4426, 2017.

[30] H. S. Seol, S. E. Lee, J. S. Song et al., "Glutamate release inhibitor, Riluzole, inhibited proliferation of human hepatocellular carcinoma cells by elevated ROS production," *Cancer Letters*, vol. 382, no. 2, pp. 157–165, 2016.

[31] I. C. Feng, M. J. Hsieh, P. N. Chen et al., "Cantharidic acid induces apoptosis through the p38 MAPK signaling pathway in human hepatocellular carcinoma," *Environmental Toxicology*, vol. 33, no. 3, pp. 261–268, 2018.

[32] Z. C. Nwosu, D. A. Megger, S. Hammad et al., "Identification of the consistently altered metabolic targets in human hepatocellular carcinoma," *Cellular and Molecular Gastroenterology and Hepatology*, vol. 4, no. 2, pp. 303–323.e1, 2017.

[33] H. Jin, Y. Zhang, H. You et al., "Prognostic significance of kynurenine 3-monooxygenase and effects on proliferation, migration, and invasion of human hepatocellular carcinoma," *Scientific Reports*, vol. 5, no. 1, article 10466, 2015.