International Journal of Statistics and Probability; Vol. 4, No. 1; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Robust Approach to Identifying Differential Circulating miRNAs in Breast Cancer

Sanjoy K. Sinha[1] & Abdus Sattar[2]

[1] School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6, Canada

[2] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Correspondence: Sanjoy K. Sinha, School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6, Canada. E-mail: sinha@math.carleton.ca

Received: December 5, 2014 Accepted: December 22, 2014 Online Published: January 27, 2015

doi:10.5539/ijsp.v4n1p155 URL: http://dx.doi.org/10.5539/ijsp.v4n1p155

Abstract

This article proposes and explores a robust approach to identifying differential circulating miRNAs in the plasma of patients with breast cancer. The proposed approach, developed in the framework of the M-estimation, is used to provide protection against potential outliers in miRNA expression data. As the study involves multiple comparisons with a large number of circulating miRNAs, robust multiple tests are adopted at a given level of false discovery rate (FDR). Also, due to the uncertainties in the underlying distributions of the miRNA expression data sets, the p-values of the multiple tests are approximated using a permutation method. The empirical properties of the proposed robust tests are studied in simulations. An application is provided using miRNA expression data from a breast cancer study.

Keywords: breast cancer, circulating miRNA, false discovery rate, outlier, permutation test, robust estimation

1. Introduction

MicroRNAs (miRNAs) are short non-coding segments of RNA that are thought to regulate gene expression through sequence-specific base-pairing with target mRNAs (Lee and Ambros, 2001). The miRNA platform is different from the traditional mRNA gene expression array platform in that the mRNA arrays measure gene expression from specific genes while the miRNA array measures expression of specific miRNAs which represent signaling from many genes. Most miRNAs are not specific to any single gene, but rather a group of genes. Thousands of miRNAs have been identified in many different organisms to date using genetics, molecular cloning and predictions from bioinformatics (Ambros, 2003). The molecular classification of human tumors based on mRNA microarray profiling is an area of intense genetic research (e.g., Blenkiron et al., 2007; Iorio and Croce, 2009; van Schooneveld et al., 2012; Volinia and Croce, 2013). A number of classifiers have been developed for human breast tumors in recent years, including the use of miRNA expression data as prognostic tools.

In cancer studies, miRNAs have been found to be deregulated in tissue specific patterns which uniquely classify each type of tumor studied to date (Calin and Croce, 2006). A group of miRNAs are known to be deregulated in breast cancer (Calin and Croce, 2006; Iorio et al., 2005), with specific miRNAs correlated to breast cancer subtype, prognosis, and treatment resistance (Qian et al., 2009). Further, functional studies have shown the mechanisms through which these miRNAs are closely involved in tumor biology of the breast (Kong et al., 2010). In the circulation, miRNAs have been detected at unexpectedly high levels and found to be the most stable nucleic acid in peripheral blood. This important discovery has prompted researchers to investigate circulating miRNAs as a novel biomarker for minimally invasive early cancer detection (Iorio and Croce, 2009).

This research was motivated by a recent breast cancer study of patients with breast cancer and healthy mammography - screened controls at the University Hospitals Case Medical Center (UHCMC) (Leidner et al., 2013). Details of the study design are given Section 5. Breast cancer is one of the leading causes of cancer deaths among women. High false positive rates from mammography often lead to unnecessary biopsies each year, which in turn increases health care costs as well as anxieties associated with screening processes. Expression profiling of circulating miRNAs in blood of breast cancer patients is currently being investigated for the development of a test for breast

cancer screening. Such tools would be useful for developing blood-based alternative tests for cancer screening and/or diagnosis. In this case-control study, a genome-wide miRNA dataset collected during 2009–2010 contained expression levels of miRNAs in the circulation of 20 breast cancer patients and 20 healthy controls using an Illumina miRNA microarray with the expression of 1145 miRNAs. The goal of the study was to identify a unique set of deregulated (differentially expressed) miRNAs that would be associated with having breast cancer.

Statistical analysis of the aforementioned miRNA expression data involves multiple testing with a large number of miRNAs. The problem is to develop valid statistical tests, which can control false discovery rates (Benjamini and Hochberg, 1995; Storey, 2002), and are also potentially more powerful than other naive tests performed under strong distributional assumptions. The use of false discovery rates has received increased popularity in recent years due to multiple hypothesis testing in high-dimensional genomics data analysis (Storey and Tibshirani, 2003). For example, in the the miRNA experiment, microarrays measure the expression levels of thousands of genes from a single biological sample. Microarrays can be applied to samples collected from two biological conditions, such as patient versus healthy control. A goal of the study is to identify miRNAs that are differentially expressed between two biological conditions, which involves performing a hypothesis test on each miRNA. False discovery rates are widely used to deal with false positives that wrongly identify differentially expressed miRNAs.

Multiple hypothesis tests based on the classical least squares estimators of location parameters are generally sensitive to potential outliers in the data. The goal of this paper is to develop a robust approach which can bound the influence of outliers in the data when estimating the model parameters. In particular, the miRNA data exhibit a special feature of outliers in the expression levels. The standard approach to analyzing the data based on the least squares estimators has been found to be influenced by the outliers. In this note, we investigate a robust approach to analyzing the data. This robust approach is developed in the framework of the Huber's M-estimation (Huber, 1964, 1981) of location and scale parameters.

Finally, to approximate the p-values of the tests associated with the multiple comparisons, the permutation method is commonly used (e.g., Fitzmaurice, Lipsitz, and Ibrahim, 2007). Permutation tests are useful when there is insufficient information about the distribution of the outcome variable, or if the distribution of the test statistic cannot be easily computed. We adopt the permutation method to approximate the p-values of the robust multiple tests as considered here.

The paper is organized as follows. Section 2 introduces the model and notation for the robust estimation. Section 3 describes the computation of permutation p-values and false discovery rates used in multiple testing. Section 4 studies empirical properties of the proposed tests based on simulations. Section 5 presents an application of the proposed method using the miRNA expression data introduced earlier. Section 6 concludes the paper with some discussion.

2. Robust Estimation

Let y_{is} represent the expression level on the miRNA s obtained from subject i ($s = 1, \ldots, S; i = 1, \ldots, N$) and x_i represent a binary covariate that is defined to be 1 if the subject is diagnosed with cancer, and 0, if not. To study the effect of this group indicator on the expression level, consider a simple linear model in the form

$$y_{is} = \beta_{0s} + \beta_{1s} x_i + \epsilon_{is}, \ \ s = 1, \ldots, S, \tag{1}$$

where the random error term ϵ_{is} is assumed to have mean 0 and variance σ_s^2. To determine whether miRNA s is differentially expressed in the patient and control groups, we can set the null hypothesis as $H_0 : \beta_{1s} = 0$. Under standard assumptions including normality, one can consider an ordinary t-test, $t = \hat{\beta}^*_{1s}/s.e.(\hat{\beta}^*_{1s})$, based on the least squares estimator $\hat{\beta}^*_{1s}$ and its standard error, $s.e.(\hat{\beta}^*_{1s})$, where an asymptotic p-value of the test can be used to estimate the false discovery rate. But as indicated earlier, the ordinary least squares estimator is generally sensitive to potential outliers in the the data. To bound the influence of such outliers, here we adopt a robust M-estimation technique based on Huber's "Proposal 2" (Huber, 1981) for simultaneous estimation of location and scale parameters.

To describe the robust method, model (1) can be reexpressed (suppressing suffix s for the miRNA) in the form

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i, \tag{2}$$

where $\mathbf{x}_i = (1, x_i)^t$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$. The M-estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ may be obtained by solving the equations

$$\sum_{i=1}^{N} \psi_c \left(\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \mathbf{x}_i = \mathbf{0}, \tag{3}$$

where ψ_c is the Huber's psi function, $\psi_c(r) = \max\{-c, \min(r, c)\}$, with a tuning constant c. A common choice of c is 1.345, which ensures a certain level efficiency of the M-estimator for the "true" underlying distribution.

The scale estimator $\hat{\sigma}$ of σ is obtained by solving the equation under Proposal 2:

$$\sum_{i=1}^{N} \left\{ \psi_c^2 \left(\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) - k \right\} = 0, \tag{4}$$

where k is a tuning constant chosen as $k = E\left(\psi_c^2\right)$, which ensures an unbiased estimating equation for $\hat{\sigma}$. Equations (3) and (4) are solved simultaneously using an iterative method for the M-estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$.

The variance of the robust estimator $\hat{\boldsymbol{\beta}}$ may be approximated by a sandwich-type variance-covariance matrix,

$$\mathbf{V} = \mathbf{M}^{-1}\mathbf{Q}\mathbf{M}^{-1}, \tag{5}$$

where

$$\mathbf{M} = \sum_{i=1}^{N} \psi_c' \left(\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \mathbf{x}_i \mathbf{x}_i^t,$$

and

$$\mathbf{Q} = \sum_{i=1}^{N} \psi_c^2 \left(\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \mathbf{x}_i \mathbf{x}_i^t,$$

with $\psi_c'(r)$ being the derivative of $\psi_c(r)$ with respect to r.

Note that as an alternative to the scale estimator $\hat{\sigma}$ under Proposal 2, we may also consider a computationally simpler robust estimator, $\tilde{\sigma} = \text{median}_i|y_i - \mathbf{x}_i^t\tilde{\boldsymbol{\beta}}|/0.6745$, where $\tilde{\boldsymbol{\beta}}$ is obtained by solving (3) with $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ being replaced by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}$, respectively. Between the two scale estimators, $\tilde{\sigma}$, also referred to as the adjusted median absolute deviation (MAD), is generally less efficient than $\hat{\sigma}$, although $\tilde{\sigma}$ is considered to be more robust with a higher breakdown point. We study both scale estimators and investigate their relative performance in multiple testing.

3. Multiple Testing

3.1 False Discovery Rate

As for the miRNA experiment, we are to conduct multiple hypothesis tests simultaneously for many miRNAs. In particular, we are to test the null hypotheses, $H_0 : \beta_{1s} = 0$, simultaneously for all miRNAs, $s = 1, \ldots, S$. A common problem with such multiple testing is that the chances of obtaining a positive result can be high even if all the null hypotheses are true. A popular approach to the multiple testing problem is to control the *false discovery rate* (FDR) (Benjamani and Hochberg, 1995). The FDR is defined by the expected proportion of false positives among all rejected hypotheses. Let V denote the number of false positive results (type I errors) and R denote the total number of rejected hypotheses. Then the FDR is defined by

$$\text{FDR} = E\left[\frac{V}{R \vee 1}\right], \tag{6}$$

where $R \vee 1 = \max(R, 1)$.

Let p_s denote the p-value of the hypothesis test, $H_0 : \beta_{1s} = 0$, for miRNA s ($s = 1, \ldots, S$). Also, let FDR(u) denote the FDR when rejecting the null hypotheses with $p_s \le u$ for $u \in [0, 1]$. Then in terms of empirical values of V and R, the FDR can be defined as

$$\text{FDR}(u) = E\left[\frac{V(u)}{R(u) \vee 1}\right], \tag{7}$$

where $V(u)$ is the number of null hypotheses with $p_s \le u$ (type I errors) and $R(u)$ is the total number of hypotheses with $p_s \le u$, $s = 1, \ldots, S$.

An estimator of FDR(u), proposed by Storey (2002), is defined by

$$\widehat{\text{FDR}}_\delta(u) = \frac{\hat{\pi}_0(\delta)u}{\{R(u) \vee 1\}/S}, \tag{8}$$

where $\hat{\pi}_0(\delta)$ is an estimator of the proportion of true null hypotheses, $\pi_0 \equiv S_0/S$, with S_0 being the number of true null hypothesis. The estimator $\hat{\pi}_0(\delta)$ is defined by

$$\hat{\pi}_0(\delta) = \frac{W(\delta)}{(1-\delta)S}, \tag{9}$$

where $W(\delta) = S - R(\delta)$ and $\delta \in (0, 1)$. A common choice of δ is 0.5. Details about the computation of the empirical FDRs are given later.

3.2 Permutation Test

Permutation tests are commonly used in genomics. These are useful when there is insufficient information about the distribution of the outcome variable, or if the distribution of the test statistic cannot be easily computed. In the case of the miRNA expression data, to compute the p-value of the hypothesis test, $H_0 : \beta_{1s} = 0$ ($s = 1, \ldots, S$), it is sensible to relax the normality assumption, as this may not be a valid assumption for all the miRNA data sets considered in the study. Therefore, instead of finding the p-values of the tests analytically under the normality assumption, we approximate the p-values by using the permutation method which does not require any distributional assumption. The algorithm for finding approximate permutation p-values is described below.

1. Consider the expression data from miRNA s. Set $s = 1$.

2. Compute the observed value of the test statistic, $T_s = \hat{\beta}_{1s}/s.e.(\hat{\beta}_{1s})$, where $\hat{\beta}_{1s}$ is an M-estimator of β_{1s}, and $s.e.(\hat{\beta}_{1s})$ is an estimate of the standard error of $\hat{\beta}_{1s}$ obtained from the sandwich-type variance-covariance matrix (5).

3. Permute the miRNA expression data randomly and assign them to the patient-control groups, so that it destroys any association between the two biological conditions and the miRNA expression, as defined by the null hypothesis. Compute the test statistic based on the permutation sample. Produce a series of test statistics, $(T_s^1, \ldots, T_s^R)$, for R permutation samples by repeating this step a large number of times, R.

4. Obtain an approximate p-value, p_s, of the test as the proportion of permutation samples with $T_s^r \geq T_s$.

5. Repeat Steps 1–4 to obtain the next p-value for $s = 2$, and so on for $s = 3, \ldots, S$.

Using Steps 1–5 above, we obtain a series of p-values, $p_1, \ldots, p_S$, for the S miRNAs considered. This set of p-values is then used to obtain empirical false discovery rates as defined in the previous section.

4. Simulation Study

We ran a simulation study to explore the performance of the proposed robust tests for selecting the miRNAs that are differentially expressed in two biological conditions, patient versus control. We considered comparing 1500 miRNAs based on expression data from 20 patients and 20 controls. The data were generated using the simple linear model, $y_{is} = \beta_{0s} + \beta_{1s}x_i + \epsilon_{is}$ ($s = 1, \ldots, S$), with $E(\epsilon_{is}) = 0$, $\text{Var}(\epsilon_{is}) = \sigma_s^2$, and with x_i being the group indicator. The intercept parameters β_{0s} were chosen uniformly from the set of values $\{5, 6, 7, 8, 9\}$. Also, for the slope parameters β_{1s}, 10% values were chosen uniformly from the set of non-zero values $\{4, 4.5, 5, 5.5, 6\}$ (i.e., 150 miRNAs were considered differentially expressed) and the remaining 90% β_{1s}'s were chosen as 0 (not differentially expressed) under the null hypothesis. The random errors ϵ_{is} were generated from normal distributions with mean 0 and variances σ_s^2 being chosen uniformly from the values $\{1, 2, 3, 4, 5\}$. In addition, to generate outliers in the expression data, 10% of the original ϵ_{is}'s were randomly replaced by values generated from a normal distribution with mean 0, but with a much larger variance $(10 + \sigma_s)^2$.

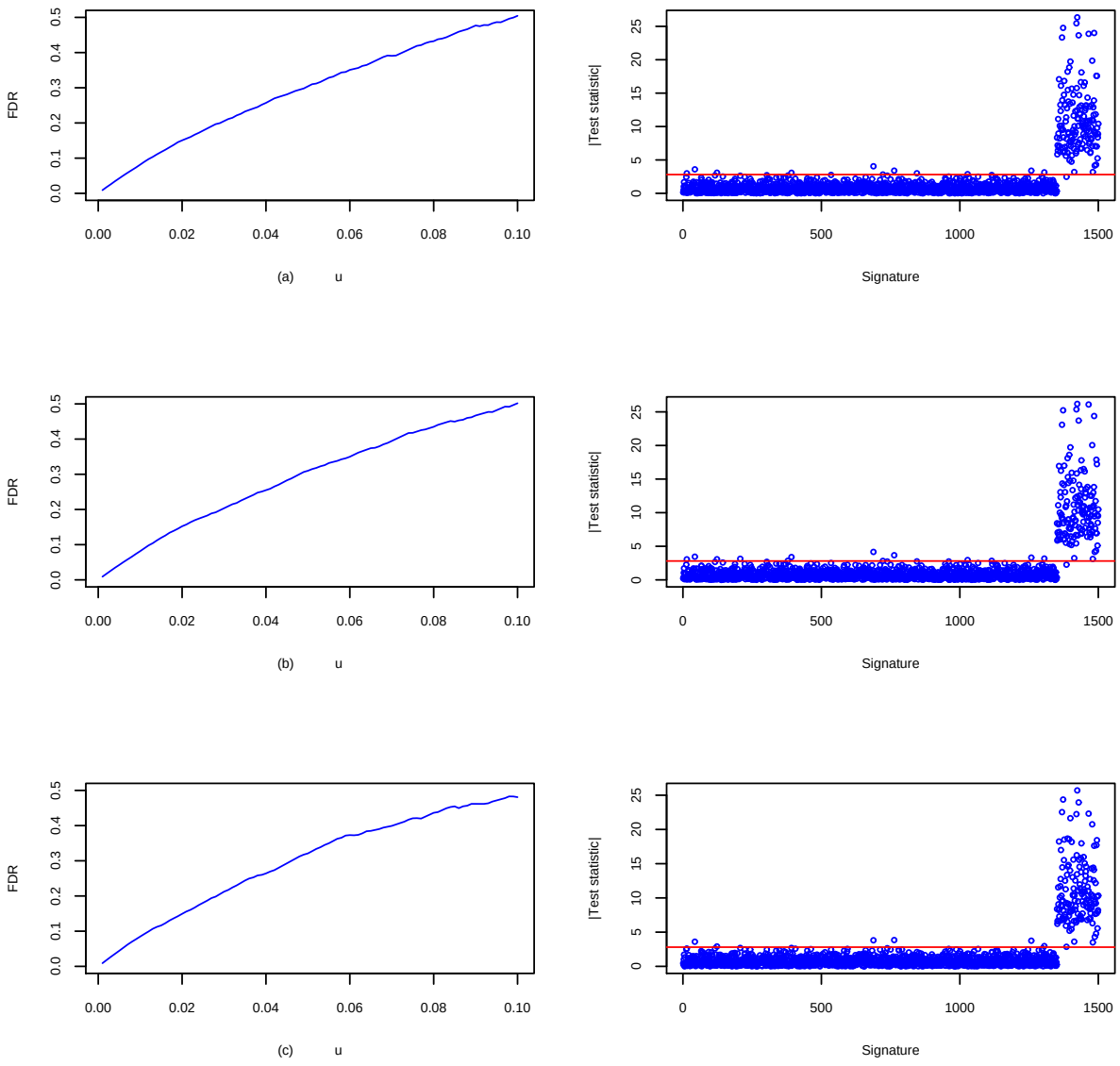

Figure 1. Empirical FDRs and test statistics for expression data with no outliers. Multiple tests are performed at the FDR level 0.05. Top panels (a) – Proposal 2; middle (b) – MAD; bottom (c) – LS.

To estimate the intercept parameters β_{0s}, slope parameters β_{1s}, and scale parameters σ_s, we considered the following three methods:

i) M-estimation with σ estimated by Huber's Proposal 2,
ii) M-estimation with σ estimated by MAD, and
iii) Least squares (LS) estimation.

Note that under the assumption of normality of the response variable y, the LS estimators of the regression parameters β_{0s} and β_{1s} are also the maximum likelihood estimators.

After finding the estimates of the model parameters and associated p-values of the multiple tests by each of the above three methods, we calculate the corresponding empirical FDRs to identify miRNAs that are differentially expressed in the two biological conditions. Figures 1 and 2 exhibit the FDR(u) against u, and the absolute values of

the test statistics against all 1500 miRNAs obtained under the three methods. The null hypotheses that are rejected at the empirical FDR level 0.05 correspond to the values of the test statistics above the horizontal line. Figure 1 shows the plots for data with no outliers, and Figure 2 repeats them for data with outliers.

It is clear from Figure 1 that when data are not contaminated with outliers, all three methods perform equally well, as almost all of the differentially expressed miRNAs (true alternatives) are identified correctly by the three methods. In particular, Proposal 2 identified 160 miRNAs that are differentially expressed, MAD identified 162 miRNAs, and LS method identified 156 miRNAs. On the other hand, when data are contaminated with outliers, the two robust methods appear to perform much better than the ordinary LS method. Proposal 2 and MAD identified 161 and 160 differentially expressed miRNAs, respectively, whereas the LS method identified only 130 of the miRNAs that are differentially expressed, as shown in Figure 2. Clearly, unlike the robust methods, the LS method fails to correctly identify many of the miRNAs that are differentially expressed.

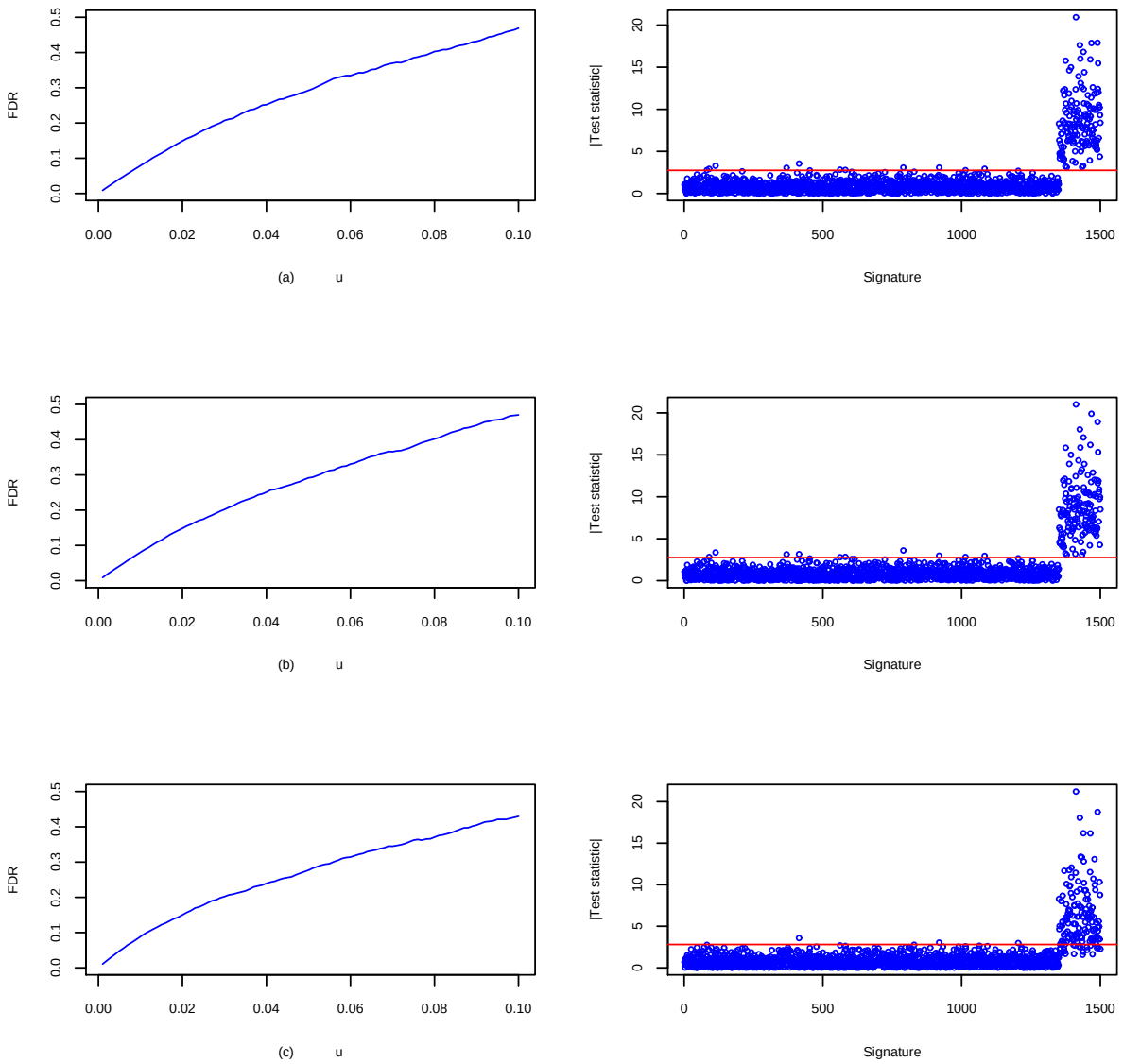

Figure 2. Empirical FDRs and test statistics for expression data with outliers. Multiple tests are performed at the FDR level 0.05. Top panels (a) – Proposal 2; middle (b) – MAD; bottom (c) – LS.

5. Application: miRNA Expression Data

Here we present an analysis of the miRNA expression data from the breast cancer study introduced earlier in Section 1. The published raw data can be found in the National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (GEO) with the accession number GSE41526. Details of the study design, clinical specimen collection, sample handling, RNA isolation, and miRNA expression profiling can be found in Leidner et al. (2013). The experiment involved a case-control study design where blood samples were collected from 20 newly diagnosed breast cancer patients and 20 controls recruited from individuals undergoing screening mammography at UHCMC during 2009–2010. To minimize any technical errors and experimental bias due to labs, technicians, and elapsed times, all blood samples were processed on the same day in the same lab. Plasma samples were de-identified and lab personnel were blinded to subset status to avoid any potential bias and/or batch effects. The Illumina Human v2 Microarray (MI-101-1124, Illumina) was utilized to profile circulating levels of 1145 miRNAs.

5.1 Data Normalization

We first consider normalizing the expression datasets to remove any machine artifacts. Normalization is routinely performed in microarray analysis, as observed expression levels include variation during the preparation of samples, manufacture of the arrays, and the processing of the arrays (see Hartemink et al., 2001 and Parmigiani et al., 2003 for more details).

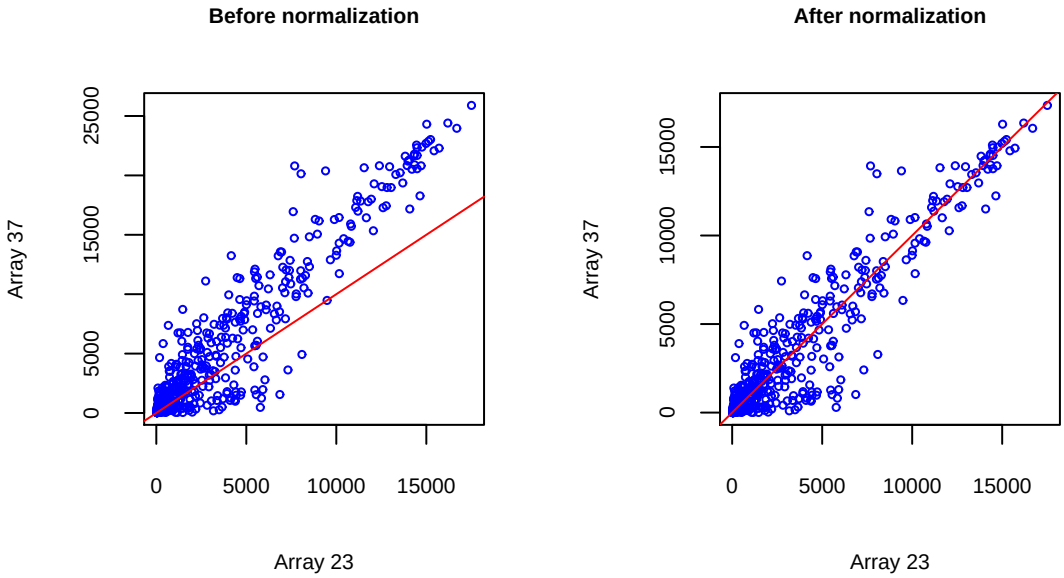

Figure 3. Intensities of two arrays plotted against each other before and after normalization.

To demonstrate the normalization procedure, Figure 3 presents intensities of two arrays, 23 and 37, plotted against each other before and after normalization. As we assume that the majority of miRNAs will not be differentially expressed, we would like the observations to scatter around the diagonal line, $y = x$. The scatter plot of the raw data (before normalization) deviates from the diagonal line, indicating the need for normalization. After normalization, the scatter plot (x-axis is the baseline array and y-axis is the normalized value of the array to be normalized) centers around the diagonal line. The array to be normalized is adjusted to have a similar overall brightness as the baseline array. The baseline array is chosen such that its median is closest to the overall median for all arrays.

For the normalization under the robust approach, we fit a linear model of the form $z = \alpha_0 + \alpha_1 z_0 + e$ for each pair of arrays, (z, z_0), by the M-estimation, where z_0 is the baseline array and z is an array to be normalized. The normalized values of array z are then obtained as $z^* = (z - \hat{\alpha}_0)/\hat{\alpha}_1$. Figure 4 presents the normalized expression profiles on the log2-transformed scale for 50 miRNAs as heat map, which is obtained such that the yellow (lighter) colour represents high expression levels and the orange (darker) colour represents low expression levels. Some

of the miRNAs appear to be highly expressed, although it is not clear which of them are differentially expressed between two biological conditions, case versus control. We adopt the proposed robust methodology to determine which of the 1145 miRNAs under study are differentially expressed between the two biological conditions.

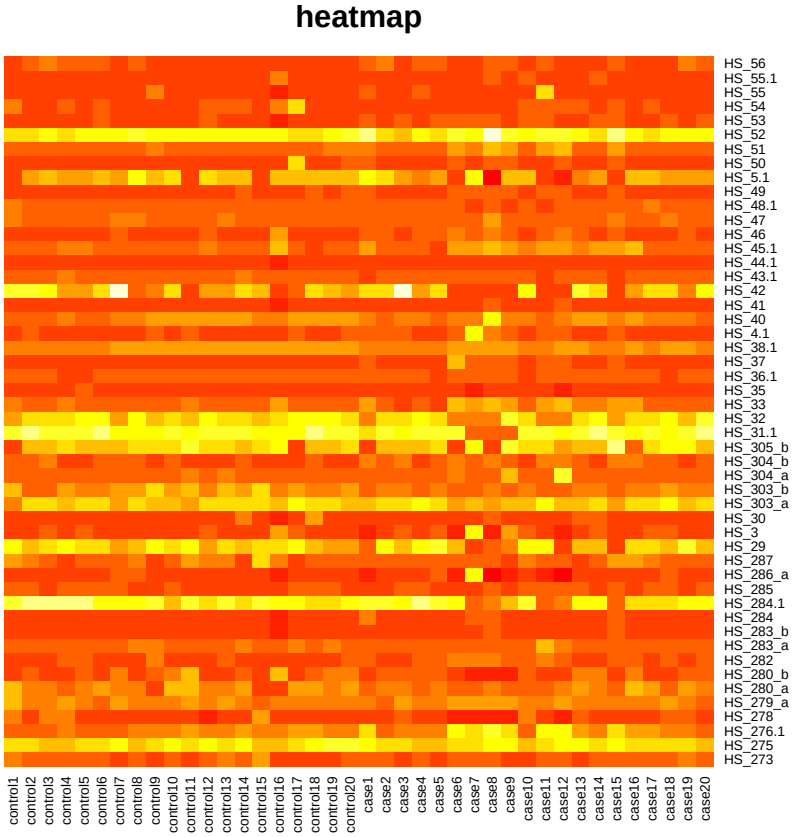

Figure 4. Expression profiles of 50 miRNAs as heat map.

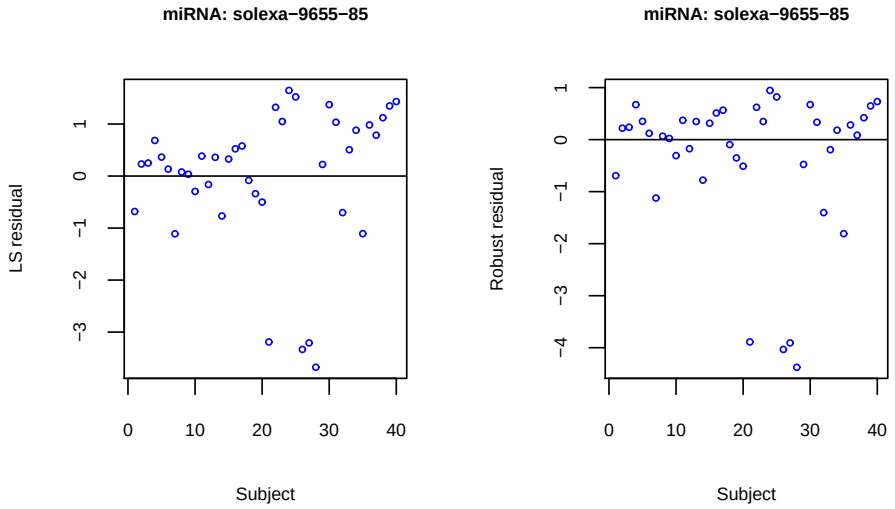

Figure 5. Non-robust (LS) and robust (Proposal 2) standardized residuals for a given miRNA. Both plots indicate four outliers in the residuals.

5.2 Robust Estimation of Model Parameters

We considered estimating the model parameters by the proposed robust method, as there were indication of outliers in the expression levels and the ordinary least squares estimators were found to be influenced by the outliers. Specifically, for a given miRNA s, we fit a linear model of the form $y_{is} = \beta_{0s} + \beta_{1s} x_i + \epsilon_{is}$, for $s = 1, \ldots, 1145$, where y_{is} is the log 2-transformed value of the normalized expression level for subject i on miRNA s, and the indicator variable x_i is defined to be 0 if subject i is in the healthy control group, and 1, if in the patient group. Figure 5 presents two standardized residual plots obtained from the non-robust LS and robust M-estimation methods for a given miRNA. Both plots indicate four large outliers, which were found to influence the ordinary least squares estimators.

5.3 Hypothesis Tests

In the next step, we use the proposed robust approach to identifying the miRNAs that are differentially expressed in the two medical conditions. Specifically, we test the null hypotheses, $H_0 : \beta_{1s} = 0$, for $s = 1, \ldots, S$, at a given FDR level. Since it may not be valid to assume that the expression levels follow a normal distribution for all miRNAs, instead of finding the asymptotic p-values, we consider approximating the p-values of the tests by using the permutation method based on 10,000 permutation samples, as described earlier. Also, for comparing the results with the asymptotic tests, we obtain p-values by naively assuming that the test statistic $T_s = \hat{\beta}_{1s}/s.e.(\hat{\beta}_{1s})$ follows a standard normal distribution. Figure 6 presents two sets of p-values, asymptotic versus permutation, plotted against each other for all the three methods considered. Clearly, the least squares method shows large discrepancies between the asymptotic and permutation p-values.

The two left panels of Figure 7 exhibit empirical FDRs at $\delta = 0.5$ calculated based on the asymptotic and permutation p-values. The absolute values of the test statistics and cutoff values at the FDR level 0.05 are shown (horizontal lines) under the asymptotic and permutation p-value methods in the two right panels of Figure 7. The cutoff points appear to be somewhat different under the two p-value methods.

Results from the hypothesis tests based on the permutation p-values are presented in Tables 1–3. Among the 1145 miRNAs under study, 49 are found to be differentially expressed by Proposal 2, 47 by MAD and 35 by the LS method. The miRNA "HS-303b" appears to be the most significant by both robust methods. On the other hand, the LS method fails to identify this as a differentially expressed miRNA between the two biological conditions. The LS method is heavily influenced by the outliers in the expression data, and consequently many of the miRNAs are not considered significant by this classical method.

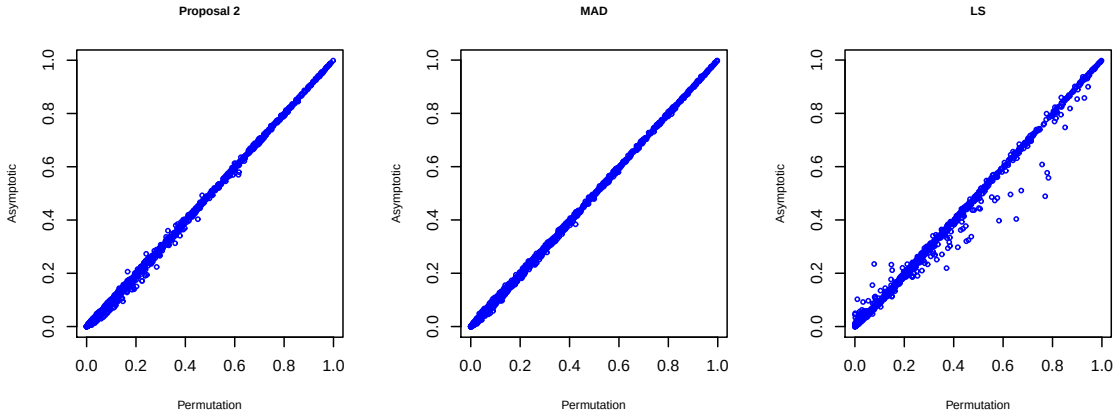

Figure 6. Plots of permutation p-values versus asymptotic p-values under three methods.

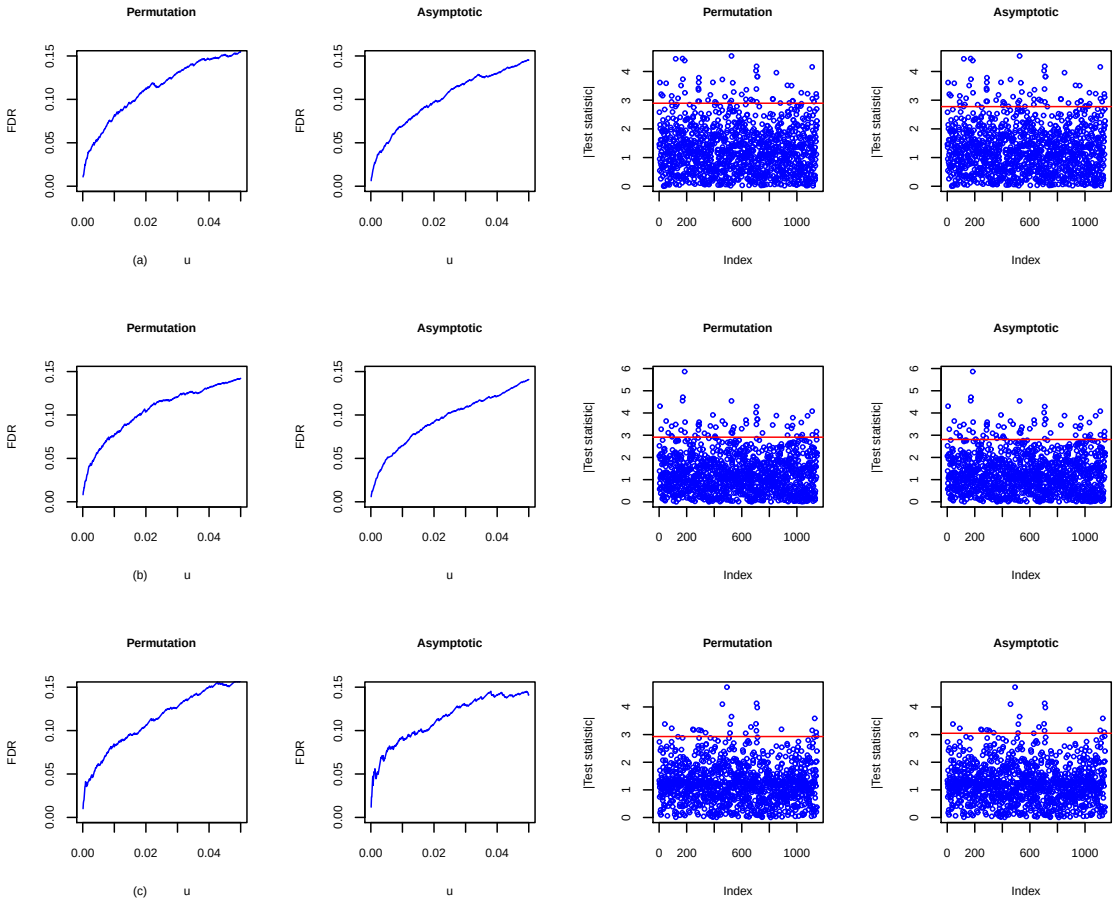

Figure 7. Empirical FDRs and test statistics under three methods. miRNAs with values of the test statistics above the horizontal line are considered significant at the FDR level 0.05. Top panels (a) – Proposal 2; middle (b) – MAD; bottom (c) – LS.

Table 1. p-values of 49 differentially expressed miRNAs determined by "Proposal 2" at 0.05 FDR level.

miRNA	p-value
HS-303b	0.0000
hsa-miR-199a-3p, hsa-miR-199b-3p	0.0000
HS-43.1	0.0001
hsa-miR-1184	0.0001
hsa-miR-379	0.0001
hsa-miR-525-3p	0.0001
HS-23	0.0002
hsa-miR-1183	0.0002
hsa-miR-151-5p	0.0002
hsa-miR-380	0.0002
HS-149	0.0003
hsa-miR-1182	0.0003
hsa-miR-376a	0.0003
hsa-miR-376c	0.0003
solexa-2952-306	0.0003
hsa-miR-612	0.0004
HS-105	0.0005
hsa-miR-1295	0.0006
HS-123	0.0008
HS-257	0.0008
hsa-miR-1181	0.0008
HS-304b	0.0009
hsa-miR-924	0.0011
hsa-miR-1281	0.0012
hsa-miR-585	0.0012
hsa-miR-200a*	0.0013
hsa-miR-202*:9.1	0.0014
hsa-miR-34c-5p	0.0014
hsa-miR-30c-2*	0.0018
hsa-miR-376b	0.0018
HS-193	0.0019
hsa-miR-1179	0.0020
hsa-miR-708*	0.0021
HS-283a	0.0022
hsa-miR-617	0.0022
hsa-miR-28-3p	0.0023
solexa-539-2056	0.0024
HS-45.1	0.0025
solexa-9081-91	0.0025
hsa-miR-30a*	0.0027
hsa-miR-33b	0.0029
hsa-miR-518f	0.0030
hsa-miR-518e*, hsa-miR-519a*,...	0.0032
hsa-miR-1197	0.0033
solexa-8926-93	0.0034
HS-192.1	0.0036
hsa-miR-130b	0.0036
hsa-miR-377	0.0036
hsa-miR-619	0.0036

Table 2. *p*-values of 47 differentially expressed miRNAs determined by "MAD" at 0.05 FDR level.

miRNA	*p*-value
HS-303b	0.0000
HS-304b	0.0000
HS-43.1	0.0000
hsa-miR-199a-3p, hsa-miR-199b-3p	0.0000
hsa-miR-376c	0.0000
HS-105	0.0001
hsa-miR-617	0.0001
hsa-miR-619	0.0001
hsa-miR-1197	0.0002
hsa-miR-379	0.0002
hsa-miR-924	0.0002
hsa-miR-1182	0.0003
hsa-miR-151-5p	0.0003
hsa-miR-376a	0.0003
hsa-miR-380	0.0003
hsa-miR-525-3p	0.0004
HS-149	0.0005
hsa-miR-1295	0.0005
hsa-miR-1322	0.0005
hsa-miR-612	0.0005
hsa-miR-202*:9.1	0.0008
hsa-miR-518e*, hsa-miR-519a*, ...	0.0008
hsa-miR-1183	0.0009
solexa-2952-306	0.0009
hsa-miR-1184	0.0010
hsa-miR-200a*	0.0010
hsa-miR-30c-2*	0.0010
hsa-miR-30b*	0.0011
solexa-9081-91	0.0014
HS-46	0.0015
hsa-miR-376b	0.0015
hsa-miR-28-3p	0.0020
hsa-miR-1281	0.0021
hsa-miR-1181	0.0022
hsa-miR-708*	0.0022
hsa-miR-645	0.0023
hsa-miR-1179	0.0025
HS-23	0.0026
hsa-miR-453	0.0026
HS-192.1	0.0027
HS-193	0.0027
hsa-miR-585	0.0027
hsa-miR-19b-2*	0.0028
HS-283a	0.0032
hsa-miR-199a*:9.1	0.0032
hsa-miR-130b	0.0034
hsa-miR-181a*	0.0034

Table 3. *p*-values of 35 differentially expressed miRNAs determined by "LS" at 0.05 FDR level.

miRNA	*p*-value
hsa-miR-185	0.0000
hsa-miR-151-5p	0.0001
hsa-miR-376a	0.0001
hsa-miR-486-3p	0.0001
hsa-miR-585	0.0001
hsa-miR-1184	0.0002
hsa-miR-376c	0.0003
solexa-603-1846	0.0003
hsa-miR-130b	0.0004
hsa-miR-379	0.0004
hsa-miR-33b	0.0008
hsa-miR-194*	0.0010
hsa-miR-199a-3p, hsa-miR-199b-3p	0.0010
hsa-miR-30a*	0.0011
HS-113	0.0012
hsa-miR-143	0.0012
solexa-539-2056	0.0012
solexa-8926-93	0.0013
HS-139	0.0015
hsa-miR-583	0.0015
hsa-let-7c	0.0017
hsa-miR-629	0.0018
HS-193	0.0019
hsa-miR-1224-5p	0.0021
hsa-miR-549	0.0022
HS-97	0.0023
solexa-2952-306	0.0023
solexa-9081-91	0.0024
hsa-miR-1257	0.0027
hsa-miR-130a	0.0029
hsa-miR-139-3p	0.0029
hsa-miR-654-5p	0.0031
solexa-3022-299	0.0031
solexa-9655-85	0.0031
HS-257	0.0033

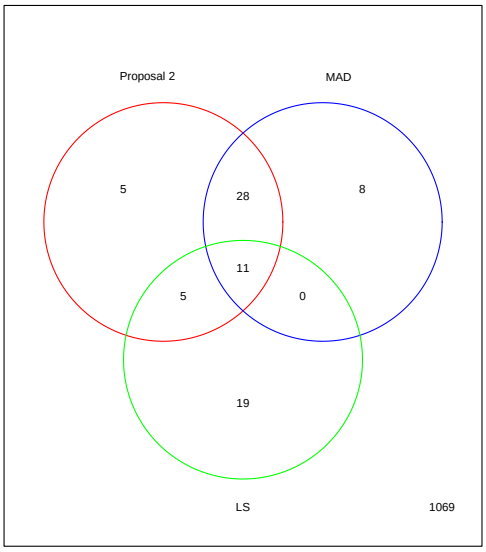

Figure 8. Number of significant miRNAs that overlap by three methods.

Figure 8 shows the number of significant miRNAs that overlap by the robust (Proposal 2 and MAD) and non-robust (LS) methods. Here the ordinary LS method appears to provide largely a different set of differentially expressed miRNAs, and is believed to be heavily influenced by the outliers in the expression levels. Among 35 miRNAs that are differentially expressed by the LS method, a large number of 19 miRNAs are not found to be significant (not differentially expressed) by any of the two robust methods.

6. Discussion

The purpose of this paper was to present a robust alternative to the classical least squares approach to identifying differentially expressed miRNAs between two biological conditions for breast cancer. The proposed approach developed in the framework of the M-estimation appears to be useful for bounding the influence of potential outliers in the expression levels when simultaneously conducting the multiple tests. The results from the simulation study were encouraging in that unlike the ordinary least squares approach, the proposed robust approach was able to correctly identify the differentially expressed miRNAs in the presence of outliers in expression levels. The application based on the miRNA expression data also demonstrates that the ordinary least squares method can be heavily influenced by outliers, and the proposed robust method is useful for bounding the influence of such outliers.

To approximate the p-values of the multiple tests, we consider using the permutation method, as the normality assumption may not be valid for all miRNA expression datasets. The proposed test is considered robust in that unlike the asymptotic test, it can provide valid p-values of the tests even for non-normal data. Under normality, the proposed robust permutation test would still provide competitive results.

We aimed at selecting differentially expressed miRNAs in the breast cancer study at a certain level of reliability. Conventional methods based on miRNA-specific p-values are often discouraged due to the multiplicity of the comparisons being performed. Several proposals suggest adjusted p-values to account for multiple comparisons (e.g., Dudoit et al., 2012). A second approach employs an empirical Bayes methodology and computes the posterior probability of differential expression between two biological conditions (Newton et al., 2001). However, perhaps the most popular approach is to report the false discovery rate (FDR) (Benjamani and Hochberg, 1995) for a group of biomarkers or for a given cutoff value of a test statistic of interest. When estimating the FDRs as considered in this paper, we suggest computing the test statistics and corresponding p-values based on the robust estimates of the model parameters.

We have considered only a single binary covariate in the regression model (1) to represent two biological conditions, as used for the multiple comparisons. This may be readily extended to the case of multiple covariates to account for any effects of demographic variables on the response of interest. In this case, the regression and scale parameters can still be estimated robustly using the M-estimation procedure.

Acknowledgements

Sanjoy Sinha is grateful for the support provided by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors thank Cheryl L. Thompson and Rom S. Leidner for providing the miRNA dataset, and helpful comments and suggestions that led to improvements in the presentation.

References

Ambros, V. (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell*, *113*, 673–676.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

Blenkiron, C., Goldstein, L. D., Thorne, N. P., Spiteri, I., Chin, S. F., Dunning, M. J., Barbosa-Morais, N. L., Teschendorff, A. E., Green, A. R., Ellis, I. O., Tavare, S., Caldas, C., & Miska, E. A. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biology*, *8*(10), R214.

Calin, G. A., & Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nature Reviews Cancer*, *6*(11), 857–866.

Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2012). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, *12*, 111–139.

Fitzmaurice, G. M., Lipsitz, S. R., & Ibrahim, J. G. (2007). A note on permutation tests for variance components

in multilevel generalized linear mixed models. *Biometrics*, *63*, 942–946.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., & Young, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. In *Proceedings SPIE 4266, Microarrays: Optical Technologies and Informatics*, *132*(4), http://dx.doi.org/10.1117/12.427981

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Iorio, M. V., & Croce, C. M. (2009). MicroRNAs in cancer: small molecules with a huge impact. *Journal of Clinical Oncology*, *27*(34), 5848–5856.

Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Menard, S., Palazzo, J. P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G. A., Querzoli, P., Negrini, M., & Croce, C. M. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Research*, *65*(16), 7065–7070.

Kong, W., He, L., Coppola, M., Guo, J., Esposito, N. N., Coppola, D., & Cheng, J. Q. (2010). MicroRNA-155 regulates cell survival, growth and chemosensitivity by targeting FOXO3a in breast cancer. *The Journal of Biological Chemistry*, *285*(23), 17869–17879.

Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, *294*(5543), 862–864.

Leidner, R. S., Li, L., & Thompson, C. L. (2013). Dampening enthusiasm for circulating microRNA in breast cancer. *PLoS ONE*, *8*(3), e57841. http://dx.doi.org/10.1371/journal.pone.0057841.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., & Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, *8*, 37–52.

Parmigiani, G., Garett, E. S., Irizarry, R. A., & Zeger, S. L. (2003). *The Analysis of Gene Expression Data – Methods and Software* (Editors). Springer–Verlag, New York.

Qian, B., Katsaros, D., Lu, L., Preti, M., Durando, A., Arisio, R., Mu, L., & Yu, H. (2009). High miR-21 expression in breast cancer associated with poor disease-free survival in early stage disease and high TGF-beta1. *Breast Cancer Research and Treatment*, *117*(1), 131–140.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, *64*, 479–498.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, *100*, 9440–9445.

van Schooneveld, E., Wouters, M. C. A., Van der Auwera, I., Peeters, D. J., Wildiers, H., Van Dam, P. A., Vergote, I., Vermeulen, P. B., Dirix, L. Y., & Van Laere, S. J. (2012). Expression profiling of cancerous and normal breast tissues identifies microRNAs that are differentially expressed in serum from patients with (metastatic) breast cancer and healthy volunteers. *Breast Cancer Research*, *14*, R34.

Volinia, S., & Croce, C. M. (2013). Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proceedings of the National Academy of Sciences*, *110*(18), 7413–7417.