BMC Bioinformatics

Research article

A close examination of double filtering with fold change and t test in microarray analysis

Song Zhang*[1] and Jing Cao[2]

Address: [1]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA and [2]Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA

Email: Song Zhang* - song.zhang@utsouthwestern.edu; Jing Cao - jcao@smu.edu

* Corresponding author

Published: 8 December 2009

BMC Bioinformatics 2009, **10**:402 doi:10.1186/1471-2105-10-402

Received: 8 September 2009
Accepted: 8 December 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/402

Abstract

Background: Many researchers use the double filtering procedure with fold change and *t* test to identify differentially expressed genes, in the hope that the double filtering will provide extra confidence in the results. Due to its simplicity, the double filtering procedure has been popular with applied researchers despite the development of more sophisticated methods.

Results: This paper, for the first time to our knowledge, provides theoretical insight on the drawback of the double filtering procedure. We show that fold change assumes all genes to have a common variance while *t* statistic assumes gene-specific variances. The two statistics are based on contradicting assumptions. Under the assumption that gene variances arise from a mixture of a common variance and gene-specific variances, we develop the theoretically most powerful likelihood ratio test statistic. We further demonstrate that the posterior inference based on a Bayesian mixture model and the widely used significance analysis of microarrays (SAM) statistic are better approximations to the likelihood ratio test than the double filtering procedure.

Conclusion: We demonstrate through hypothesis testing theory, simulation studies and real data examples, that well constructed shrinkage testing methods, which can be united under the mixture gene variance assumption, can considerably outperform the double filtering procedure.

Background

With the development of microarray technologies, researchers now can measure the relative expressions of tens of thousands of genes simultaneously. However, the number of replicates per gene is usually small, far less than the number of genes. Many statistical methods have been developed to identify differentially expressed (DE) genes. The use of fold change is among the first practice. It can be inefficient and erroneous because of the additional uncertainty induced by dividing two intensity values. There are variants of Student's *t* test procedure that conduct a test on each individual gene and then correct for multiple comparisons. The problem is, with a large number of tests and a small number of replicates per gene, the statistics are very unstable. For example, a large *t* statistic might arise because of an extremely small variance, even with a minor difference in the expression.

The disadvantage of fold-change approach and *t* test has been pointed out by a number of authors [1,2]. There are approaches proposed to improve estimation of gene variances by borrowing strength across genes [1,3,4]. Despite the flaw, fold change and *t* test are the most intuitive approaches and they have been applied widely in practice.

To control the error rate, many researchers use fold change and *t* test together, hoping that the double filtering will provide extra confidence in the test results. Specifically, a gene is flagged as DE only if the *p*-value from *t* test is smaller than a certain threshold and the fold change is greater than a cutoff value. For example, in [5], 90 genes were found to be DE with two cutoff values (*p*-value < 0.01 and fold change > 1.5). There are numerous examples in the literature that implement the double filtering procedure with fold change and *t* statistic [6-9]. We argue, however, that the double filtering procedure provides higher confidence mainly because it produces a shorter list of selected genes. Given the same number of genes selected, a well constructed shrinkage test can significantly outperform the double filtering method.

Fold change takes the ratio of a gene's average expression levels under two conditions. It is usually calculated as the difference on the log_2 scale. Let x_{ij} be the log-transformed expression measurement of the ith gene on the jth array under the control ($i = 1,\cup, n$ and $j = 1,\cup, m_0$), and γ_{ik} be the log-transformed expression measurement of the ith gene on the kth array under the treatment ($k = 1,\cup m_1$). We define $\bar{x}_i = \frac{1}{m_0}\sum_{j=1}^{m_0} x_{ij}$ and $\bar{\gamma}_i = \frac{1}{m_1}\sum_{k=1}^{m_1} \gamma_{ik}$.

Fold change is computed by

$$fc_i \equiv \bar{\gamma}_i - \bar{x}_i. \quad (1)$$

As for the traditional *t* test, it is usually calculated on the log_2 scale to adjust for the skewness in the original gene expression measurements. The *t* statistic is then computed by

$$t_i = \frac{\bar{\gamma}_i - \bar{x}_i}{\sqrt{s_i^2\left(\frac{1}{m_0}+\frac{1}{m_1}\right)}}, \quad (2)$$

where s_i^2 is the pooled variance of x_{ij} and γ_{ik}. Comparing (1) and (2), it is obvious that fold change and *t* statistic are based on two contradicting assumptions. The underlying assumption of fold change is that all genes share a common variance (on the log_2 scale), which is implied by the omission of the variance component in (1). On the other hand, the inclusion of s_i^2 in (2) suggests that *t* test assumes gene-specific variances. In order for a gene to be flagged as DE, the double filtering procedure would require the gene to have extreme test scores under the common variance assumption as well as under the gene-specific variance assumption. It is analogous to using the intersection of the rejection regions defined by fold change and *t* statistic.

Assuming a common variance for all the genes apparently is an oversimplification. The assumption of gene-specific variances, however, leads to unstable estimates due to limited replicates from each gene. A more realistic assumption might lie in between the two extremes, i.e., modeling gene variances by a mixture of two components, one being a point mass at the common variance, another being a continuous distribution for the gene-specific variances. Under this mixture variance assumption, a DE gene could have a large fold change or a large *t* statistic, but not necessarily both. Taking intersection of the rejection regions flagged by fold change and *t* statistic, as is adopted by the double filtering procedure, might not be the best strategy under the mixture variance assumption.

The goal of the paper is not to propose a new testing procedure in microarray analysis, but to provide insight on the drawback of the widely used double filtering procedure with fold change and *t* test. We present a theoretically most powerful likelihood ratio (LR) test under the mixture variance assumption. We further demonstrate that two shrinkage test statistics, one from the Bayesian model [10] and the other from the significance analysis of microarrays (SAM) test [1], can be united as approximations to the LR test. This association explains why those shrinkage methods can considerably outperform the double filtering procedure. A simulation study and real microarray data analyses are then presented to compare the shrinkage tests and the double filtering procedure.

Methods

A Likelihood Ratio Test

For gene i, we use $f_i = p_v f_{i1} + (1 - p_v) f_{i2}$, a mixture of two components f_{i1} and f_{i2}, to denote the density under the null hypothesis that the gene is not DE under two experiment conditions. Density f_{i1} is defined under the gene-specific variance assumption, f_{i2} is defined under the common variance assumption, and p_v is the mixing probability. Similarly, we use $g_i = p_v g_{i1} + (1 - p_v) g_{i2}$ to denote the density under the alternative hypothesis, with g_{i1} and g_{i2} defined in a similar fashion as f_{i1} and f_{i2}. For example, in the context of testing DE genes, we can assume $f_{i1} = N(\mu_i, \sigma_i^2)$, $f_{i2} = N(\mu_i, \sigma_0^2)$, $g_{i1} = N(\mu_i + \Delta_i, \sigma_i^2)$, and $g_{i2} = N(\mu_i + \Delta_i, \sigma_0^2)$, where σ_0^2 is the assumed common variance, σ_i^2 is the gene-specific variance, μ_i is the mean expression level under the control, and Δ_i is the difference in the expression levels between two conditions. Under the null

hypothesis $H_0 : \Delta_i = 0$, the likelihood ratio test statistic, which is the most powerful among all test statistics, is

$$\begin{aligned} R_i &= \frac{p_v g_{i1} + (1-p_v) g_{i2}}{p_v f_{i1} + (1-p_v) f_{i2}} \\ &= \frac{g_{i1}}{f_{i1}} \frac{p_v f_{i1}}{p_v f_{i1} + (1-p_v) f_{i2}} + \frac{g_{i2}}{f_{i2}} \frac{(1-p_v) f_{i2}}{p_v f_{i1} + (1-p_v) f_{i2}} \\ &= \frac{g_{i1}}{f_{i1}} w_i + \frac{g_{i2}}{f_{i2}} (1 - w_i). \end{aligned} \tag{3}$$

The R_i statistic is a weighted sum of two ratios g_{i1}/f_{i1} and g_{i2}/f_{i2}, with weight $w_i = p_v f_{i1}/[p_v f_{i1} + (1 - p_v) f_{i2}]$. Under the normality assumption, it is easy to show that $R_i = w_i h_1(|t_i|) + (1 - w_i) h_2(|fc_i|)$, where fc_i and t_i are fold change and t statistic, as defined in (1) and (2). Both $h_1(\cdot)$ and $h_2(\cdot)$ are monotonic increasing functions.

The rejection region of the LR test is defined by $R_i > \lambda_R$, where λ_R is the threshold to attain a certain test size. In order to reject H_0, it requires that either $|fc_i|$ is large, or $|t_i|$ is large, or both. In this sense, the LR test rejection region is more like a union of the rejection regions defined by fold change and t statistic. On the other hand, the double filtering procedure with fold change and t statistic would require both $|fc_i|$ and $|t_i|$ to be large. This practice is analogous to using the intersection of the two rejections regions determined by $|fc_i|$ and $|t_i|$. Compared with the LR test, the double filtering procedure will lose power. The "loss of power" has two meanings. First, for a given false discovery rate (FDR), the double filtering procedure produces a shorter list of identified genes for further investigation. Second, for a given number of identified genes, the list produced by the double filtering procedure has a higher FDR. The double filtering procedure offers a false sense of confidence by producing a shorter list.

The LR test statistic R_i requires one to know the true values of parameters p, μ_i, σ_i^2, σ_0^2, and Δ_i, which are usually unknown in reality. One strategy is to estimate R_i by $\hat{R}_i$, where the maximum likelihood estimates (MLE) of the unknown parameters are plugged into (3). Unfortunately, with a small number of replicates from each gene, the MLE would be extremely unstable and lead to unsatisfactory testing results.

A Bayesian model [10] was constructed under the mixture variance assumption to detect DE genes. The inference is made based on the marginal posterior probability of a gene being DE, denoted by $z_i = P(\Delta_i \neq 0 \mid X, Y)$. Here $X = \{x_{ij}\}$ and $Y = \{y_{ik}\}$ are the collection of gene expression data under the two conditions. We will show that, like $\hat{R}_i$, z_i is also an approximation to R_i. The difference between $\hat{R}_i$ and z_i is that the former plugs in the point estimates (MLE) of unknown parameters, while the latter marginalizes the unknown parameters with respect to their posterior distribution. In the Bayesian inference, the uncertainty from various sources are accounted for in a probabilistic fashion.

Similar to the Bayesian mixture model, some existing methods also try to strike a balance between the two extreme assumptions of a common variance and gene-specific variances. The SAM statistic slightly modifies the t-statistic by adding a constant to the estimated gene-specific standard deviation in the denominator. We will present it as being motivated by a mixture model on the variances (standard deviations). Furthermore, the SAM statistic can be directly written as a weighted sum of t statistic and fold change. Thus both the Bayesian method and the SAM method are approximations to the LR test under the mixture variance assumption, and they can achieve better performance than the double filtering procedure.

The Bayesian Mixture Model

Cao *et al.* [10] proposed a Bayesian mixture model to identify DE genes, which has shown comparable performance to frequentist shrinkage methods [1,11]. With parameters $(\mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v)$ defined similarly as in the LR test, gene expression measurements x_{ij} and y_{ij} are modeled by normal distributions with a mixture structure on the variances,

$$\begin{aligned} x_{ij} \mid \mu_i, \sigma_i^2, \sigma_0^2, p_v &\sim p_v N(\mu_i, \sigma_i^2) + (1 - p_v) N(\mu_i, \sigma_0^2), \\ y_{ik} \mid \mu_i, \Delta_i, \sigma_i^2, \sigma_0^2, p_v &\sim p_v N(\mu_i + \Delta_i, \sigma_i^2) + (1 - p_v) N(\mu_i + \Delta_i, \sigma_0^2). \end{aligned} \tag{4}$$

A latent variable r_i is used to model the expression status of the ith gene,,

$$\begin{cases} \Delta_i = 0, & \text{if } r_i = 0, \\ \Delta_i \sim N(0, s_\Delta^2), & \text{if } r_i = 1. \end{cases}$$

where $r_i = 0/1$ indicates that gene i is non-DE/DE and it is modeled by a Bernoulli distribution: $r_i \mid p_r \sim Bernoulli(p_r)$. For σ_i^2 and σ_0^2, it is assumed that $\sigma_i^2 \overset{iid}{\sim} IG(a_\sigma, b_\sigma)$ and $\sigma_0^2 \sim IG(a_0, b_0)$. Here $IG(a, b)$ denotes an inverse gamma distribution with mean $b/(a - 1)$. The other hyper-priors

include, $\mu_i \sim N(0, s_\mu^2)$, $p_r \sim U(0, 1)$, and $p_v \sim U(0, 1)$. More details can be found in [10].

We make inference based on $z_i = P(r_i = 1 \mid X, Y) = P(\Delta_i \neq 0 \mid X, Y)$, the marginal posterior probability that gene i is DE. A gene is flagged as DE if $z_i > \lambda_z$, where λ_z is a certain cutoff. We argue that the Bayesian rejection region defined by $z_i > \lambda_z$ is an approximation to the LR test rejection region defined by $R_i > \lambda_R$. First we have

$$z_i = \int P(r_i = 1 \mid \mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v, p_r, X, Y) dP(\mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v, p_r \mid X, Y). \quad (5)$$

Here $P(\mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v, p_r \mid X, Y)$ is the joint posterior distribution of $(\mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v, p_r)$, marginalized with respect to other random parameters (e.g., μ_j and $\sigma_j^2, j \neq i$).

It is easy to show that

$$\begin{aligned} & P(r_i = 1 \mid \mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v, p_r, X, Y) \\ &= \frac{p_r[p_v g_{i1} + (1-p_v) g_{i2}]}{p_r[p_v g_{i1} + (1-p_v) g_{i2}] + (1-p_r)[p_v f_{i1} + (1-p_v) f_{i2}]} \\ &= \frac{1}{1 + \frac{1-p_r}{p_r} \cdot \frac{p_v f_{i1} + (1-p_v) f_{i2}}{p_v g_{i1} + (1-p_v) g_{i2}}} \\ &= \frac{1}{1 + \frac{1-p_r}{p_r} \cdot \frac{1}{R_i}}. \end{aligned} \quad (6)$$

Given parameters $(\mu_i, \Delta_i, \sigma_0^2, \sigma_i^2, p_v, p_r)$, $P(r_i = 1 \mid \mu_i, \Delta_i, \sigma_0^2, \sigma_{i^2}, p_v, p_r, X, Y)$ is an increasing function of R_i. Rejecting H_0 for $R_i > \lambda_R$ is equivalent to rejecting for $P(r_i = 1 \mid \lambda_i, \Delta_i, \sigma_0^2, \sigma_{i^2}, p_v, p_r, X, Y) > \lambda_z$, with $\lambda_z = \lambda_R / [\lambda_R + (1 - p_r)/p_r]$. Thus the two test statistics, $P(r_i = 1 \mid \mu_i, \Delta_i, \sigma_0^2, \sigma_{i^2}, p_v, p_r, X, Y)$ and R_i, are equivalent. Expression (5) demonstrates that z_i is obtained from $P(r_i = 1 \mid \mu_i, \Delta_i, \sigma_0^2, \sigma_{i^2}, p_v, p_r, X, Y)$ by integrating with respect to the unknown parameters under the joint posterior distribution. If the integral does not have a closed form, we can conduct numerical integration to calculate z_i through the Gibbs sampling algorithm [12,13]. The uncertainty from those unknown parameters are accounted for in a probabilistic fashion. It is in this sense that we consider z_i a good approximation to the LR test statistic R_i.

The SAM Test

The SAM statistic [1] is defined as

$$d_i = \frac{\bar{x}_i - \bar{y}_i}{s_i + s_0},$$

where s_i is the gene-specific standard deviation, and s_0 is a constant that minimizes the coefficient of variation. Although it might not be the original intention of the authors [1], a test statistic like d_i can be motivated by a model with a mixture structure on gene standard deviations. We begin with a simple case where $x_{ij} \sim N(\mu_i, \delta_i^2)$ and $y_{ik} \sim N(\mu_i + \Delta_i, \delta_i^2)$, and the null hypothesis is $H_0 : \Delta_i = 0$. Given δ_i, the LR test statistic is

$$R_i^* \equiv \frac{\bar{x}_i - \bar{y}_i}{\delta_i}.$$

We assume a mixture structure on gene standard deviations, where $\delta_i = \sigma_i$ with probability p_v and $\delta_i = \sigma_0$ with probability $1 - p_v$. We can then approximate R_i^* by

$$R_i^* \approx \frac{\bar{x}_i - \bar{y}_i}{E(\delta_i)} = \frac{\bar{x}_i - \bar{y}_i}{p\sigma_i + (1-p)\sigma_0} = \frac{1}{p} \cdot \frac{\bar{x} - \bar{y}_i}{\sigma_i + \frac{1-p}{p}\sigma_0}.$$

Replacing σ_i with s_i and $\frac{1-p}{p}\sigma_0$ with s_0, we can see that d_i and R_i^* only differ by a factor of $1/p$, which does not change the ordering of test statistics. The above derivation suggests that the SAM statistic can also be considered an approximation to the LR test statistic under the mixture variance (standard deviation) assumption. We can also write d_i as a weighted sum of t_i and fc_i:

$$\begin{aligned} d_i &= \frac{s_i}{2(s_i + s_0)} \cdot \frac{\bar{x}_i - \bar{y}_i}{s_i} + \frac{s_0}{2(s_i + s_0)} \cdot \frac{\bar{x}_i - \bar{y}_i}{s_0} \\ &= \frac{s_i}{2(s_i + s_0)} t_i + \frac{1}{2(s_i + s_0)} fc_i. \end{aligned} \quad (7)$$

Recall that under the mixture variance assumption, the LR test statistic is $R_i = w_i h_1(|fc_i|) + (1 - w_i) h_2(|t_i|)$, where $h_1(\cdot)$ and $h_2(\cdot)$ are both monotonic increasing functions. Both d_i and R_i define rejection regions that are analogous to the union of the rejection regions defined by t test and fold change. In other words, the SAM procedure rejects H_0 for large $|t_i|$, or large $|fc_i|$, or both. The SAM statistic is a better approximation to the LR test statistic than the double filtering procedure.

As a side note, Cui *et al*. [11] proposed a shrunken *t* test procedure based on a variance estimator that borrow information across genes using the James-Stein-Lindley shrinkage concept. This variance estimator shrinks individual variances toward a common value, which conceptually serves the same purpose as the mixture variance model. From this perspective, we also consider the shrunken *t* statistic an approximation to the LR test statistic.

Results and Discussion

Simulation Study

We conducted a simulation study to compare the double filtering procedure to the shrinkage methods. The simulation truth is specified as follows. We tested 1000 genes with 100 genes being truly DE. Without loss of generality, we set $\mu_i = 0$. We further assumed

$$\begin{cases} \Delta_i = 0, & \text{if } r_i = 0, \\ \Delta_i \sim \mathrm{N}(0, 2), & \text{if } r_i = 1, \end{cases}$$

and

$$\begin{cases} \sigma_i^2 = 0.25, & \text{if } v_i = 0, \\ \sigma_i^2 \sim \mathrm{IG}(4, 1), & \text{if } v_i = 1, \end{cases}$$

Three scenarios were considered. Scenario 1: 90% of the genes with gene-specific variances and 10% of the genes with a common variance, and 3 replicates per gene under each condition. Scenario 2: same as Scenario 1, but with 6 replicates per gene under each condition. Scenario 3: all the genes having a gene-specific variance, and 3 replicates per gene under each condition. For each scenario we repeated the simulation 1000 times.

For the Bayesian mixture model, we specified noninformative priors so that the posterior inference is dominated by information from data. We let $s_\mu^2 = s_\Delta^2 = 5.0$ where 5.0 is sufficiently large for expression levels on the logarithm scale. To specify the hyper-parameters for the inverse gamma priors, first we set $a_\sigma = a_0 = 2.0$ so that the inverse gamma priors have an infinite variance. Then we set the prior means, $\frac{b_\sigma}{a_\sigma - 1}$ and $\frac{b_0}{a_0 - 1}$, equal to the average of the sample variances to solve for b_σ and b_0. Finally, we chose $a_r = b_r = a_v = b_v = 1$, which corresponds to uniform priors for p_r and p_v.

Five test statistics were compared: the marginal posterior probability (z_i) of a gene being DE based on the Bayesian mixture model, the SAM statistic, the shrunken *t* statistic, the *t* statistic, and the double filtering with *t* statistic and fold change greater than 2. The first three graphs in Figure 1 plot the FDR versus the total number of selected genes under the above three scenarios. The shrinkage methods (the Bayesian model, the SAM test, and the shrunken *t* test) have comparable performance. The double filtering procedure performs better than the traditional *t* statistic, but it is obviously outperformed by the three shrinkage methods. We have tried different fold change cutoff values for the double filtering procedure (e.g., setting the cutoff at 1.5) and the results did not change materially. Given the same number of selected genes, the shrinkage methods can identify more truly DE genes than the double filtering procedure. Note that under the gene-specific variance assumption (Scenario 3), the *t* test, which theoretically is the most powerful likelihood ratio test, still performs the poorest. This result indicates the usefulness of shrinkage in microarray studies, where only a small number of replicates are available for each gene. In short, the simulation study shows that for a given number of selected genes, well constructed shrinkage methods can outperform the double filtering procedure.

In Scenario 1 and 2 of the simulation study, the true variance distribution is specified as the mixture of a point mass and an inverse gamma distribution, which might lead to a result that is biased in favor of a shrinkage method. Here we conduct another simulation study with a "real" variance distribution, denoted as Scenario 4. Specifically, let x_{ij} (j = 1, ... , m_{0i}) and y_{ik} (k = 1, ... , m_{1i}) be the observed expression levels from a real microarray study.

Define the residual vector $\mathbf{e_i} = (e_{i1}, \ldots, e_{i,m_{0i}+m_{1i}})'$ by

$$e_{il} = \begin{cases} x_{il} - \bar{x}_i, & \text{for } l = 1, \cdots, m_{0i}, \\ y_{i,(l-m_{0i})} - \bar{y}_i, & \text{for } l = m_{0i}+1, \cdots, m_{0i}+m_{1i}. \end{cases}$$

Then $\mathbf{e_i}$ can be considered a set of random errors sampled based on the true variance distribution. We simulate 1000 data sets according to the following steps. For iteration s (s = 1, ∪, 1000) and gene i (i = 1, ... , n),

1. obtain a random permutation of $(e_{i1}, \ldots, e_{i,(m_{0i}+m_{1i})})$, denoted by $\mathbf{e_i^{(s)}}$;

2. generate $\Delta_i^{(s)}$ as described in the previous simulation scenarios;

3. for j = 1, ... , m_{0i}, compute $x_{ij}^{(s)} = e_{ij}^{(s)}$, and for k = 1, ... , m_{1i}, compute $y_{ik}^{(s)} = \Delta_i^{(s)} + e_{i,(m_{0i}+k)}^{(s)}$, where $e_{ij}^{(s)}$ is the jth element of $\mathbf{e_i^{(s)}}$.

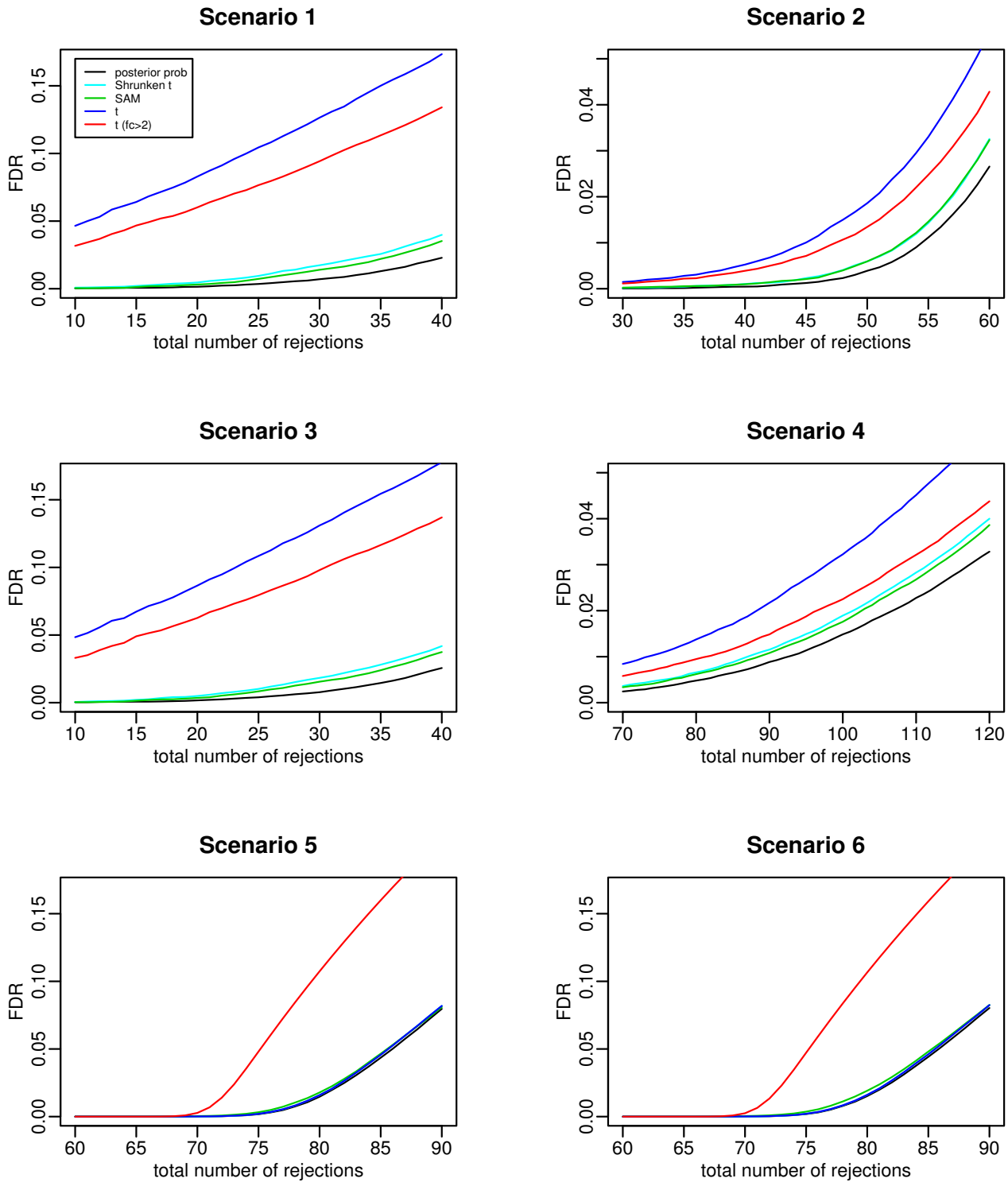

Figure 1

Comparison of the FDR given the total number of selected genes under Scenario 1-6 in the simulation study. The competing test statistics are the posterior probability based on the Bayesian model, the shrunken *t* statistic, the SAM statistic, the *t* statistic, and the double filtering procedure with *t* statistic and fold change.

The real data comes from a microarray study comparing the gene expressions of breast cancer tumors with *BRCA1* mutations, *BRCA2* mutations, and sporadic tumors [14]. The data set is available at http://research.nhgri.nih.gov/microarray/NEJM_Supplement. Here we only consider the *BRCA1* group and the *BRCA2* group. There are 3226 genes, with 7 arrays in the *BRCA1* group and 8 arrays in the *BRCA2* group. We analyzed the data on the log_2 scale. Following Storey and Tibshirani [15], we eliminated genes with aberrantly large expression values (>20), which left us with measurements on n = 3169 genes. The fourth graph in Figure 1 compares the different methods under Scenario 4, where the residual vector $\mathbf{e}_i$ was constructed based on the breast cancer data. We kept the same replicate number in the experiment, with 7 replicates per gene in one group and 8 replicates in the other group. The relative performance of the five methods remains unchanged as in the other scenarios.

In current microarray studies, the number of replicates per gene can be easily 30 or more due to the low cost of array and the easiness to collect patients. So we considered two scenarios with a relatively large number of replicates. Scenario 5: 90% of the genes with gene-specific variances and 10% of the genes with a common variance, and 30 replicates per gene under each condition. Scenario 6: all the genes having a gene-specific variance, and 30 replicates per gene under each condition. In each of the two scenarios, we assume there are 1000 genes with 100 genes being truly DE. The two graphs in the bottom panel of Figure 1 plot the FDR versus the total number of selected genes for the five test statistics under Scenario 5 and Scenario 6, respectively. The comparison demonstrates that when the replicate number is large, the performance of the traditional *t* test becomes comparable to the performance of the shrinkage methods, thanks to the more reliable estimate of gene variance component. More importantly, the drawback of the double filtering procedure becomes more obvious, which has substantially worse performance compared to the other methods, including the *t* test.

Experimental Datasets

In this section we compared the shrinkage methods with the double filtering procedure based on two microarray datasets. The first is the Golden Spike data [16] where the identities of truly DE genes are known. The Golden Spike dataset includes two conditions, with 3 replicates per condition. Each array has 14,010 probesets, among which 10,144 have non-spiked-in RNAs, 2,535 have equal concentrations of RNAs, and 1,331 are spiked-in at different fold-change levels, ranging from 1.2 to 4-fold. Compared with other spike datasets, the Golden Spike dataset has a larger number of probsets that are known to be DE, making it popular for comparing performance among different methods. Irizarry *et al.* [17] pointed out that "the feature intensities for genes spiked-in to be at 1:1 ratios behave very differently from the features from non-spiked-in genes". Following Opgen-Rhein and Strimmer [18], we removed the 2,535 probe sets for spike-ins with ratio 1:1 from the original data, leaving in total 11,475 genes and 1,331 known DE genes. Figure 2 plots the FDR under each testing procedures versus the total number of rejections. For the double filtering procedure, the fold change cutoff was set at 1.5 because only 248 genes have a fold change greater than 2.0. The figure indicates that the shrinkage methods (Bayesian, SAM, and shrunken *t*) have similar performance, and they outperform the double filtering procedure and *t* test.

The second is the breast cancer dataset [14] described in the simulation study. With the identities of truly DE genes unknown, we estimated the FDR for the SAM test, the shrunken *t* test, the *t* test, and the double filtering procedure, through the permutation approach described in [15]. For Bayesian methods, Newton *et al.* [19] proposed to compute the Bayesian FDR, which is the posterior proportion of false positives relative to the total number of rejections. However, the Bayesian FDR is incomparable to the permutation-based FDR estimate employed by frequentist methods [20]. Cao and Zhang [21] developed a generic approach to estimating the FDR for Bayesian methods under the permutation-based framework. A computationally efficient algorithm was developed to approximate the null distribution of the Bayesian test statistic, the posterior probability. The approach can provide

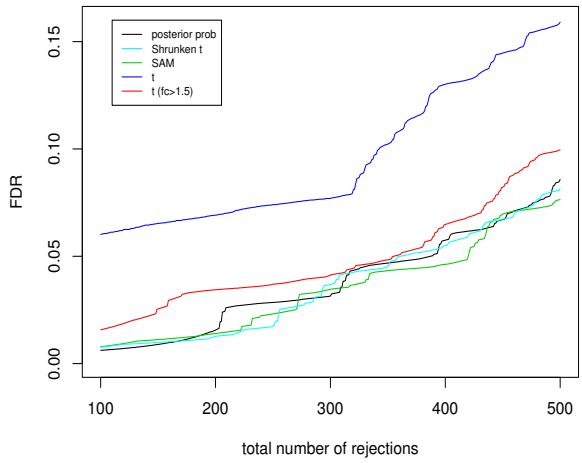

Figure 2
Comparison of the FDR given the total number of selected genes in the analysis of Golden Spike data. The test statistics include the posterior probability based on the Bayesian model, the shrunken *t* statistic, the SAM statistic, the *t* statistic, and the double filtering procedure with *t* statistic and fold change.

an unbiased estimate of the true FDR. Constructed under the same permutation-based framework, the resulting FDR estimate allows a fair comparison between full Bayesian methods with other testing procedures. We adopted the approach in [21] to estimate the FDR of the Bayesian mixture model (4). Figure 3 plots the permutation-based FDR estimates under each testing procedure versus the total number of rejections. It shows that the shrinkage methods can considerably outperform the double filtering procedure.

Conclusion

It has been a common practice in microarray analysis to use fold change and *t* statistic to double filter DE genes. In this paper, we provided a close examination on the drawback of the double filtering procedure, where fold change and *t* statistic are based on contradicting assumptions. We further demonstrated that several shrinkage methods (SAM, shrunken *t*, and a Bayesian mixture model) can be united under the mixture gene variance assumption. Based on the theoretical derivation, the simulation study, and the real data analysis, we showed compelling evidence that well constructed shrinkage methods can outperform the double filtering procedure in identifying DE genes. With publicly available softwares, these methods are as easy to implement as the double filtering procedure.

We acknowledge some researchers' argument that the double filtering procedure might work well because it filters out the genes that show relatively small differences between conditions, which are sometimes considered to be less biologically meaningful. This argument, however, is based on the criterion of so called "biological meaningfulness" instead of testing power. Although many biologists refer to fold change in terms of "biological meaningfulness", there is in fact no clear cut-off for it, and 2-fold is often invoked merely based on convenience. In addition, different normalization methods can differ quite drastically in terms of the fold changes they produce. So a particular cut-off in fold change could mean one thing using one method and quite another using a different method. Taken together, even if researchers decide to employ the double filtering procedure based on the rationale of "biological meaningfulness", it is still helpful to understand the potential loss in power.

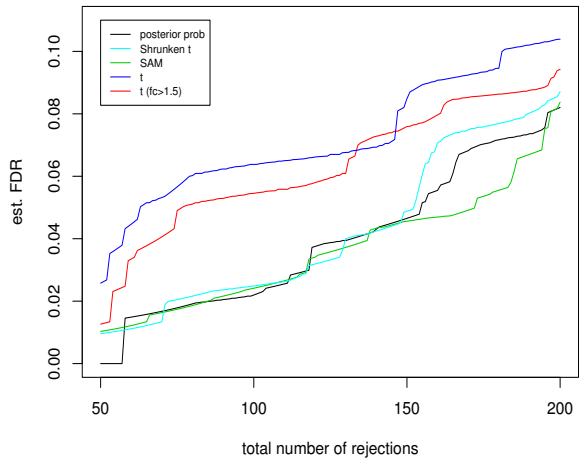

Figure 3
Comparison of the estimated FDR given the total number of selected genes in the analysis of the breast cancer data. The test statistics include the posterior probability based on the Bayesian model, the shrunken *t* statistic, the SAM statistic, the *t* statistic, and the double filtering procedure with *t* statistic and fold change.

Authors' contributions

SZ and JC conceived the study, conducted the examination on the double filtering procedure, analyzed the data, and drafted the paper. All authors read and approved the final manuscript.

Acknowledgements

This work has been supported in part by the U.S. National Institutes of Health UL1 RR024982. The authors thank the reviewers for their constructive comments and suggestions.

References

1. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to transcriptional responses to ionizing radiation.** *Proceedings of the National Academy of Sciences* 2001, **98:**5116-5121.
2. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, **19(15):**1945-1951.
3. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes.** *Bioinformatics* 2001, **17:**509-519.
4. Lonnstedt I, Speed T: **Replicated microarray data.** *Statistica Sinica* 2002, **12:**31-46.
5. Han T, Wang J, Tong W, Moore MM, Fuscoe JC, Chen T: **Microarray analysis distinguishes differential gene expression patterns from large and small colony Thymidine kinase mutants of L5178Y mouse lymphoma cells.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S9.
6. Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, Conte JV, Tomaselli G, Garcia JGN, Hare JM: **Gene expression in giant cell myocarditis: Altered expression of immune response genes.** *International Journal of Cardiology* 2005, **102(2):**333-340.
7. Li Y, Elashoff D, Oh M, Sinha U, St John MAR, Zhou X, Abemayor E, Wong DT: **Serum circulating human mRNA profiling and its utility for oral cancer detection.** *Journal of Clinical Oncology* 2006, **24(11):**1754-1760.
8. Quinn P, Bowers RM, Zhang X, Wahlund TM, Fanelli MA, Olszova D, Read BA: **cDNA microarrays as a tool for identification of biomineralization proteins in the coccolithophorid Emiliania hux-leyi (Haptophyta).** *Applied and Environmental Microbiology* 2006, **72(8):**5512-5526.
9. Sauer M, Jakob A, Nordheim A, Hochholdinger F: **Proteomic analysis of shoot-borne root initiation in maize (Zea mays L.).** *Proteomics* 2006, **6(8):**2530-2541.
10. Cao J, Xie X, Zhang S, Whitehurst A, White M: **Bayesian optimal discovery procedure for simultaneous significance testing.** *BMC Bioinformatics* 2009, **10:**5.

11. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates.** *Biostatistics* 2005, **6:**59-75.
12. Gelfand AE, Smith AFM: **Sampling-Based Approaches to Calculating Marginal Densities.** *Journal of the American Statistical Association* 1990, **85(410):**398-409.
13. Casella G, George EI: **Explaining the Gibbs sampler.** *The American Statistician* 1992, **46(3):**167-174.
14. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi O, Wilfond B, Borg A, Trent J: **Gene-expression profiles in hereditary breast cancer.** *New England Journal of Medicine* 2002, **344(8):**539-548.
15. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proceedings of the National Academy of Sciences* 2003, **100:**9440-9445.
16. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biology* 2005, **6(2):**R16.
17. Irizarry RA, Cope LM, Wu Z: **Feature-level exploration of a published Affymetrix GeneChip control dataset.** *Genome Biology* 2006, **7(8):**404.
18. Opgen-Rhein R, Strimmer K: **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach.** *Statistical Applications in Genetics and Molecular Biology* 2007, **6(1):**9.
19. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostatistics* 2004, **4:**155-176.
20. Storey JD, Dai JY, Leek JT: **The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments.** *Biostatistics* 2007, **8:**414-432.
21. Cao J, Zhang S: **Measuring statistical significance for full Bayesian methods in microarray analysis.** *Technical report* [http://smu.edu/statistics/TechReports/tech-rpts.asp].