**J. Biomedical Science and Engineering, 2010, 3, 317-321** **JBiSE**
doi:10.4236/jbise.2010.33043 Published Online March 2010 (http://www.SciRP.org/journal/jbise/).

# A mixture model based approach for estimating the FDR in replicated microarray data

**Shuo Jiao, Shun-Pu Zhang**

Department of Statistics University of Nebraska Lincoln, NE, USA.
Email: szhang3@unl.edu

Received 5 November 2009; revised 27 November 2009; accepted 3 December 2009.

## ABSTRACT

**One of the mostly used methods for estimating the false discovery rate (FDR) is the permutation based method. The permutation based method has the well-known granularity problem due to the discrete nature of the permuted null scores. The granularity problem may produce very unstable FDR estimates. Such instability may cause scientists to over- or under-estimate the number of false positives among the genes declared as significant, and hence result in inaccurate interpretation of biological data. In this paper, we propose a new model based method as an improvement of the permutation based FDR estimation method of SAM [1] The new method uses the *t*-mixture model which can model the microarray data better than the currently used normal mixture model. We will show that our proposed method provides more accurate FDR estimates than the permutation based method and is free of the problems of the permutation based FDR estimators. Finally, the proposed method is evaluated using extensive simulation and real microarray data.**

**Keywords:** FDR; T-Mixture Model; Microarray; Genes

## 1. INTRODUCTION

Genome-wide expression data generated from the microarray experiments are widely used to uncover the functional roles of different genes, and how these genes interact with each other. A key step to achieve this is to identify the differentially expressed (DE) genes under different experimental conditions. Such information can be used to identify disease biomarkers that may be important in the diagnoses of different types of diseases. Earlier statistical approaches for detecting DE genes focused mostly on parametric methods which are easily subject to model misspecification problems. Some of the well-known parametric methods for detecting DE genes include the two sample *t*-test [2], the analysis of variance approach [3], a regression approach [4], the regularized t-statistic method (Bayes-t test) [5,6]), the semi-parametric hierarchical mixture method [7], and the parametric EB method [8]. Recently, the availability of replicated microarrays has made it possible to use the nonparametric methods to detect the DE genes. The nonparametric methods require much less stringent dis-tributional assumptions, and thus can provide more robust results than the parametric methods. Some of the well-known nonparametric methods for analyzing microarrays include the Significance Analysis of Microar-ray (SAM) of [1], the nonparametric EB method [9,10], the non-parametric t-test with adjusted p-value [11], the Wilcoxon Rank Sum test [12], samroc [13] and the normal mixture model method (MMM) of [14].

In this paper, we will focus our attention on SAM, one of the most popular methods in microarray data analysis. SAM indentifies DE genes by computing a modified *t*-statistic as the test score of a gene and finding the genes with test scores exceeding an adjustable threshold. The false discovery rate (FDR) was then estimated by a permutation based method. More specifically, the number of false positive (FP) genes among the significant genes is estimated as the median of the numbers of scores exceeding the cutoffs in each permuted set of null scores.

Since the permutation based approach estimates the FDR by counting the number of FP genes exceeding some cutoffs, we will call it the empirical method in this paper. Due to its nature, there are two drawbacks with the empirical method: 1) the granularity problem – the FDR estimates based on the counted number of FP genes tend to be unstable when the actual number of FP genes is small; 2) the zero FDR problem – the estimated FDR may be zero when the range of the permuted null scores is smaller than that of test scores and when the cutoffs are more extreme than the endpoints of permuted null scores. These two drawbacks are illustrated in the **Figures 1, 2** and **3**.

In this paper, we will propose a *t*-mixture model based approach as an improvement of the empirical FDR estimation method of SAM. Our method aims to solve the two aforementioned drawbacks of the current empirical

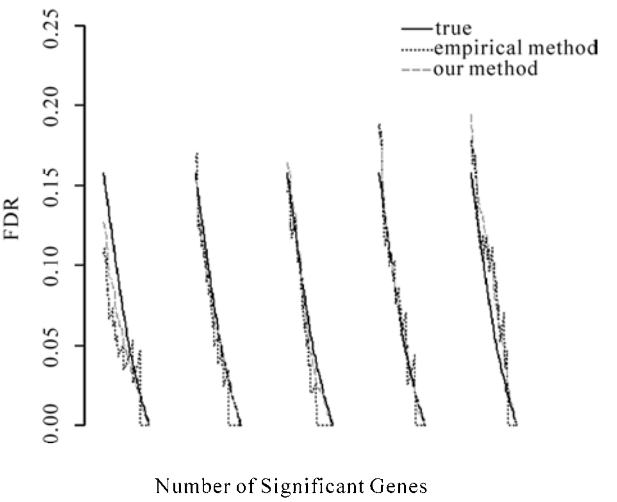

**Figure 1.** Comparison of true FDR, the empirical FDR estimator $\widehat{FDR}$ and the model based FDR estimator $\widehat{FDR}1$ for two sample microarray data. 5 replicates are listed. Total number of significant genes is decreasing from 100 to 1 (left to right) for each replicate.

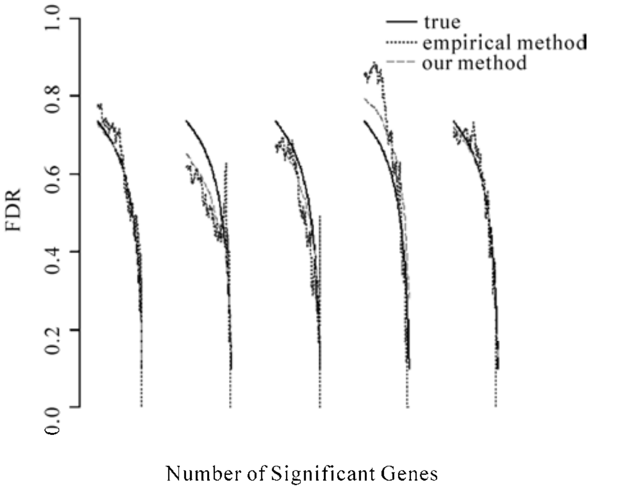

**Figure 2.** Comparison of true FDR, the empirical FDR estimator $\widehat{FDR}$ and the model based FDR estimator $\widehat{FDR}1$ for two sample microarray data. 5 replicates are listed. Total number of significant genes is decreasing from 150 to 1 (left to right) for each replicate.

FDR estimation method: The granularity and the zero FDR problems. The performance of our method is assessed by applying them to simulated and real microarray data.

## 2. METHODS

### 2.1. SAM

#### 2.1.1. SAM algorithm

Let $Y_{ij}$ be the expression levels of genes *i* under array *j*

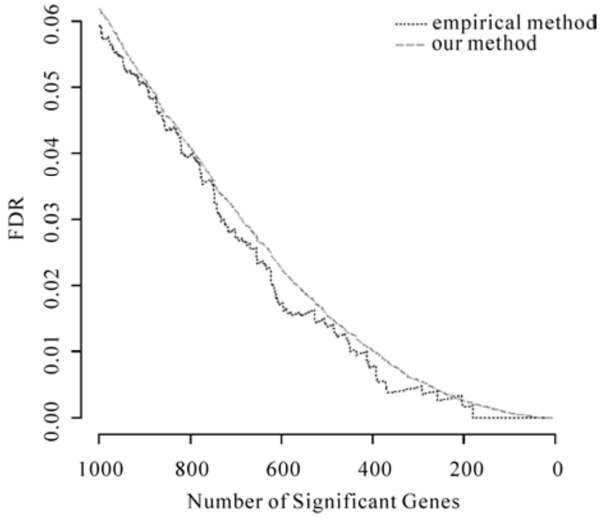

**Figure 3.** Comparison of the empirical FDR estimator $\widehat{FDR}$ and the model based FDR estimator $\widehat{FDR}1$ for Leukemia microarray data.

($i$=1,…,$n$; $j$=1,… $j_1$, $j_1$ +*1*,…, $j_1 + j_2 = J$ ), and the first $j_1$ and last $j_2$ arrays are obtained under two conditions. We need to test if gene *i* has differential expressions under the two conditions.

In SAM, the test statistic is defined as:

$$Z_i = \frac{Y_{i(1)} - Y_{i(2)}}{\sqrt{(1/j_1 + 1/j_2)s_i^2 + s_0}},$$

where $Y_{i(1)}$, $Y_{i(2)}$ are the sample means under two conditions; $s_i^2$ is the pooled sample variance; $s_0$ is the fudge factor. The null score $z_i^b$ is then computed by applying the test statistic to the *b-th* set of permuted data.

In the SAM manual [15], the following algorithm is given to detect DE genes. First, all genes are ranked by the magnitude of their test scores $Z_i$ so that $Z_{(1)}$ is the largest test score and $Z_{(i)}$ is the *i-th* largest test score. For the *b-th* set of null scores, the same procedure is applied so that $z_{(i)}^b$ is the *i-th* largest null score in the *b-th* set of null scores. The expected relative difference is then defined as $z_{(i)}^E = \sum_{b=1}^{B} z_{(i)}^b / B$. After that, a scatter plot of $Z_{(i)}$ vs. $z_{(i)}^E$ is plotted. In the scatter plot, some points are displaced from the $Z_{(i)} = z_{(i)}^E$ line with a distance greater than $\Delta$, a pre-specified threshold. In [16], the author pointed out that the estimated total number of significant (TS) genes and FP genes obtained using the SAM algorithm can be written as:

$$\widehat{TS} = \#\{i; Z_{(i)} > \delta_U \text{ or } Z_{(i)} < \delta_L\}, \text{ and} \quad (1)$$

$$\widehat{FP} = \sum_{b=1}^{B} \#\{i; z_{(i)}^{b} > \delta_U \text{ or } z_{(i)}^{b} < \delta_L\} / B, \qquad (2)$$

where $\delta_U$ and $\delta_L$ are the upper and lower cutoffs decided by the pre-specified threshold $\Delta$. For simplicity, we only consider symmetric cutoffs ($|\delta_U| = |\delta_L|$) in this paper though extensions to asymmetric cutoffs are straightforward. Under symmetric cutoffs, (1) and (2) can be written as:

$$\widehat{TS}(\delta) = \#\{i; |Z_{(i)}| > \delta\} \qquad (3)$$

$$\widehat{FP}(\delta) = \sum_{b=1}^{B} \#\{i; |z_{(i)}^{b}| > \delta\} / B \qquad (4)$$

### 2.1.2. Empirical FDR Estimator of SAM

Given a gene-specific significance level $\alpha \in (0, 1]$ and assume that we have obtained the *p*-values for all the genes under consideration, the FDR of [17] is defined as:

$$FDR = E[\frac{N(\alpha)}{TS(\alpha)}], \qquad (5)$$

where $N(\alpha)$ is the number of genes among the EE genes whose *p*-values are less than or equal to $\alpha$, and $TS(\alpha)$ is the number of genes among all the genes whose *p*-values are less than or equal to $\alpha$ (or it is the total number of significant genes). Instead of controlling gene-specific significance level $\alpha$, SAM usually controls the total number of significant genes by setting a corresponding cutoff $\delta$, hence (5) can be re-written as:

$$FDR = E[\frac{N(\delta)}{TS(\delta)}], \qquad (6)$$

where $N(\delta)$ is the number of EE genes with absolute value of $Z_i$ greater than $\delta$, and $TS(\delta)$ is the total number of genes with absolute value of $Z_i$ greater than $\delta$.

It was shown in [18] that the FDR can be approximated by

$$FDR \approx \frac{E[N(\delta)]}{E[TS(\delta)]}. \qquad (7)$$

Since $N(\delta)$ is the number of false positive among the EE genes, denote the proportion of EE genes by $\pi_0$, (7) becomes

$$FDR \approx \frac{\pi_0 E[FP(\delta)]}{E[TS(\delta)]}, \qquad (8)$$

where $FP(\delta)$ is the number of FP if all the genes are EE. $FP(\delta)$ and $TS(\delta)$ can be estimated by $\widehat{FP}(\delta)$ and $\widehat{TS}(\delta)$ in (3) and (4), respectively. As a result, the empirical FDR estimator of SAM is

$$\widehat{FDR} = \frac{\hat{\pi}_0 \widehat{FP}(\delta)}{\widehat{TS}(\delta)}, \qquad (9)$$

As mentioned before, this empirical FDR estimator of SAM has the granularity problem and the zero FDR problem. In the following sections, we solve these problems by proposing a model based FDR estimation method.

## 2.2. The *T*-mixture Model (TMM) Based FDR Eestimation Approach

Let $f$ be the probability density of the test score $Z_i$ and $f_0$ be the density of null score $z_i^b$. In TMM, it is assumed that the data are from several components with distinguished *t*-distributions. In other words, both $f$ and $f_0$ are considered to be a mixture of the *t*-distributions with probability density function:

$$h(z; \psi_g) = \sum_{i=1}^{g} \pi_i \phi(z; \mu_i, \Sigma_i, \nu_i), \qquad (10)$$

where $\phi(z; \mu_i, \Sigma_i, \nu_i)$ denotes the *t*-distribution density function with mean $\mu_i$, variance $\Sigma_i$, and degrees of freedom $\nu_i$. The coefficients $\pi_i$ are the mixing proportions and *g* is the number of components, which can be selected adaptively. $\psi_g$ denotes all the unknown parameters ($\pi_i$, $\mu_i$, $\Sigma_i$, $\nu_i$) $| i = 1,...g$ in (5). The mixture model is fitted by maximum likelihood using an expectation conditional maximization (ECM) algorithm [19]. The final model is selected based on Bayesian Information Criterion (BIC). More details on how to fit the TMM to microarray data can be found in [20]. It was reported in their paper that not only does the TMM approach provide more accurate estimates of the densities, but also it enjoys computational efficiency since it was demonstrated in [20] that one only needs to use one set of permuted null scores to fit the *t*-mixture model. More specifically, instead of using all $z_i^b$ *'s (size=n*B)* to fit the *t*-mixture model, a random sample with size *n* can be drawn from $\bigcup_{b=1}^{B}\bigcup_{i=1}^{n} z_i^b$ and used as the null statistics.

Since the test statistic $Z_i$ and the null statistic $z_i$ (because only one set of null score is used now, we will denote the null statistic as $z_i$ instead of $z_i^b$) have the densities $f$ and $f_0$, respectively, it is easy to see from (8) that

$$FDR \approx \pi_0 \frac{E[FP(\delta)/n]}{E[TS(\delta)/n]}$$

$$
\begin{aligned}
&=\pi_0 \frac{E\sum_{i=1}^{n} I(|z_i| \geq \delta)/n}{E\sum_{i=1}^{n} I(|Z_i| \geq \delta)/n} \\
&=\pi_0 \frac{P(|z| \geq \delta)}{P(|Z| \geq \delta)} \\
&=\pi_0 \frac{\int_{|z|\geq\delta} f_0(z)dz}{\int_{|z|\geq\delta} f(z)dz} \qquad (11)
\end{aligned}
$$

where $\delta$ is chosen such that a given number of significant genes is detected. Equation (11) can be viewed as the model based formula of FDR.

Assume that we have available the estimators $\hat{f}$ and $\hat{f}_0$ of $f$ and $f_0$ from the TMM, respectively, then the corresponding model based FDR estimator for (11) is

$$\widehat{\mathrm{FDR}}_1 = \hat{\pi}_0 \frac{\int_{|z|\geq\delta} \hat{f}_0(z)dz}{\int_{|z|\geq\delta} \hat{f}(z)dz}, \qquad (12)$$

The model based FDR estimator (12) has the following advantages compared to the empirical FDR estimator of SAM:

1) It does not have the granularity problem of the empirical FDR estimator (9);

2) It provides non-zero FDR estimate for any $\delta$, while (9) only provides non-zero FDR when cutoffs are within the two endpoints of the range of the permuted null scores;

3) Unlike (9), the numerator and the denominator of (12) are not subject to the sampling variability.

## 3. RESULTS

### 3.1. Simulated Data

In the simulation, $j_1 = j_2 = 4$ replicates and *n* = 5,000 genes are generated while 200 of them are assumed to be differentially expressed. For the DE genes, the data under condition 1 are generated from *N*(2,1) and the data under condition 2 are generated from *N*(0,1). The EE genes are generated from *N*(0,1) regardless of the conditions. For the generated data, we calculate the true FDR and estimated FDR for a grid of total number of significant genes ranging from 100 to 1 (in decreasing order). This procedure is repeated for five times. **Figure 1** shows comparisons of true FDR, empirical FDR estimator $\widehat{FDR}$ defined by (9), and the model based FDR estimator $\widehat{FDR}_1$ defined by (12).

As we can see, the instability of empirical $\widehat{FDR}$ increases significantly as it decreases to 0, which shows its granularity problem. Another fact worth noticing is that $\widehat{FDR}$ tends go to zero faster than the true FDR, which is the zero FDR problem. It can be seen that the true FDR strictly decreases as the total number of significant genes decreases. However, the empirical $\widehat{FDR}$ does not show this characteristic. In contrast, $\widehat{FDR}_1$ captures the decreasing trend very well and does not have the erratic jumps of $\widehat{FDR}$. To check how well these two FDR estimators approximate the true FDR, we calculate the mean squared error for both of them. MSE for $\widehat{FDR}$ is 0.00045 and MSE for $\widehat{FDR}_1$ is 0.00021, which shows that our method outperforms the empirical method.

Next, we compare the performances of the two methods when the two populations for the DE and EE genes are not so well separated. For this purpose, we conduct another simulation which tries to mimic the real data. The expression levels for the EE genes under the two conditions are generated from $N(\mu_{i1}, \sigma_i^2)$ and $N(\mu_{i1}, \sigma_i^2)$ with $\mu_{i1} = \mu_{i2} \sim N(0,2)$ $\sigma_i^2 \sim Gamma(4,2)$. The expression levels for the DE genes are generated similarly as the EE genes, except that $\mu_{i1}$ and $\mu_{i2}$ are generated from $N(0,2)$ separately. In this case, the grid of total number of significant genes ranges from 150 to 1 (in decreasing order). Comparison results are displayed in **Figure 2**.

It is seen from **Figure 2** that $\widehat{FDR}$ is very unstable and approximates true FDR poorly, which makes the estimates highly inaccurate. On the other hand, $\widehat{FDR}_1$ has a much smoother curve than $\widehat{FDR}$ and seems to be able to capture the decreasing trend of the true FDR very well. In addition, the fact that MSE for $\widehat{FDR}$ is 0.025 and for $\widehat{FDR}_1$ is 0.015 shows that our method gives a significantly better fit to the true FDR.

### 3.2. Real Data

The Leukemia data of [21] is one of the most studied gene expression data sets. This data set includes 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples for 7129 genes. In **Figure 3**, we estimate the FDRs for different number of significant genes using both our proposed model based FDR estimator and the empirical FDR estimator. As we expect, the model based FDR estimator gives a more stable estimate.

## 4. DISCUSSION

In this paper, we have proposed a *t*-mixture model based approach to improve the performance of SAM's empirical FDR estimator. We demonstrate that our method does not have the granularity and zero FDR problems as the empirical method. The results also show that our estimator provides more stable and accurate estimates of the FDR. The advantage of our method is more evident in the case when DE genes are not well separated with EE genes and the variances of expression levels for every gene are different. This is due to the fact that the permutation FDR estimator is more easily affected by the sampling variability.

## REFERENCES

[1] Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significant analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116-5121.

[2] Long, A. D., Mangalam, H. J., Chan, B. Y. P., Tolleri, L., Hatfield, W. G. and Baldi, P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical frame work, **276**, 19937-19944.

[3] Kerr, M.K., Martin, M. and Churchill, G. (2000) Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**, 819-837.

[4] Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L. P. (2001) An efficient and robust statistical modelling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, **11**, 1227-1236.

[5] Baldi, P. L. and Long, A. D. (2001) A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inference of gene changes. *Bioinformatics*, **17**, 509-519.

[6] Kendziorski, C. M., Newton, M. A., Lan, H. And Gould, M. N. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**, 819-837.

[7] Newton, M., Noueiry, A., Ahlquist, P., Sarkar, D. (2004) Detecting differential gene expression with a semi-parametric hierarchical mixture method. *Biostatistics*, **5(2)**, 155-176.

[8] Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3(1)**.

[9] Efron, B., Tibshirani R., Storey, J. D., Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment., *Journal of the American Statistical Association*, **96**, 1151-1160.

[10] Efron, B., Tibshirani, R., Gross, V. and Chu, G. (2000) Microarrays and their use in a comparative experiment, Technical report, Statistics Department, Standard University.

[11] Dudoit, S., Yang, H. Y., Callow, J. M. and Speed, P. T. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 111-139.

[12] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D. and Altman, R. B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-1461.

[13] Broberg, P. (2003) Ranking genes with respect to differential expression, *Genome Biology*, **4**, R41.

[14] Pan, W., Lin, J. and Le, C. (2003) A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & integrative genomics*, **3**, 117-124.

[15] Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V. SAM "significance analysis" of microarrays-users guide and technical document, http://www-stat.stanford.edu/~tibs/SAM/sam.pdf.

[16] Zhang, S. (2007) A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance, *BMC Bioinformatics*, **8**, 230.

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

[18] Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *PNAS*, **100**, 9440-9445.

[19] Liu, C. and Rubin, D. (1995) ML estimation of the t distribution using EM and its extensions ECM and ECME. *Statistica Sinica*, **5**, 19-39.

[20] Jiao, S. and Zhang, S. (2008) The t-mixture model approach for detecting differentially expressed genes in microarrays. *Functional & Integrative Genomics*, **8**, 181-186.

[21] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J.R. and Caligiuri, M. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **285**, 531-537.