BMC Bioinformatics

Methodology article

Ranking analysis of *F*-statistics for microarray data

Yuan-De Tan[1,2], Myriam Fornage[2] and Hongyan Xu*[3]

Address: [1]College of Life Sciences, Hunan Normal University, Changsha, 410081, China, [2]Institute of Molecular Medicine, the University of Texas-Houston, Houston, Texas 77030, USA and [3]Department of Biostatistics, Medical College of Georgia, Augusta, Georgia, 30912, USA

Email: Yuan-De Tan - tanyuande@hotmail.com; Myriam Fornage - myriam.fornage@uth.tmc.edu; Hongyan Xu* - hxu@mcg.edu

* Corresponding author

Published: 6 March 2008

BMC Bioinformatics 2008, **9**:142 doi:10.1186/1471-2105-9-142

Received: 23 November 2007
Accepted: 6 March 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/142

Abstract

Background: Microarray technology provides an efficient means for globally exploring physiological processes governed by the coordinated expression of multiple genes. However, identification of genes differentially expressed in microarray experiments is challenging because of their potentially high type I error rate. Methods for large-scale statistical analyses have been developed but most of them are applicable to two-sample or two-condition data.

Results: We developed a large-scale multiple-group *F*-test based method, named ranking analysis of *F*-statistics (RAF), which is an extension of ranking analysis of microarray data (RAM) for two-sample t-test. In this method, we proposed a novel random splitting approach to generate the null distribution instead of using permutation, which may not be appropriate for microarray data. We also implemented a two-simulation strategy to estimate the false discovery rate. Simulation results suggested that it has higher efficiency in finding differentially expressed genes among multiple classes at a lower false discovery rate than some commonly used methods. By applying our method to the experimental data, we found 107 genes having significantly differential expressions among 4 treatments at <0.7% FDR, of which 31 belong to the expressed sequence tags (ESTs), 76 are unique genes who have known functions in the brain or central nervous system and belong to six major functional groups.

Conclusion: Our method is suitable to identify differentially expressed genes among multiple groups, in particular, when sample size is small.

Background

Microarray gene expression technology, which profiles the expression of multiple genes in parallel [1,2], affords the means for globally exploring physiological and pathological processes [3] regulated by the coordinated expression of thousands of genes [4]. However, identification of genes that are differentially expressed in large-scale gene expression experiments requires global statistical methods rather than traditional statistical methods based on single hypothesis testing. A variety of multiple-testing procedures, such as the Bonferroni procedure, Holm procedure [5], Hochberg procedure [6], Benjamini-Hochberg (BH) procedure [7], and Benjamini-Liu (BL) procedures [8] have already been developed for testing a large family of null hypotheses. The first three methods bound the family-wise-error rate (FWER) that is the probability of at least one false positive over all tests and hence remain too stringent and have lower power for finding genes from the real data sets. The last two methods have an upper bound for the false discovery rate (FDR) with both strong and weak controls [9] and require a large sample size for valid p-values. Tusher et al. [9] has proposed a ranking statistic approach based on permutation for resampling. However, permutation is not a desirable approach to estimating null distribution in microarray data [10-12] because in general a microarray dataset has a large number of genes but small sample sizes [13] due to cost. Permutation fails to remove treatment effect and due to small sample sizes the difference of treatment effects between permutated groups may become a main component in differences between group means so that the estimated null distribution is not well approximate to the true null distribution ([13] and also see Appendix in Tan et al. [14]). For example, Xie et al. [12] found that the estimated null F-distribution based on permutation has a larger variance and a heavier tail compared to the true null F-distribution, which leads to a potential loss of power. Similar phenomenon was also observed in comparison of the estimated null t-distribution to the true null t-distribution [14]. To remove the group or treatment effects on the estimated null distribution, Tan et al. [14] developed a random splitting (RS) approach. Since treatment effects are completely eliminated, the estimated null distribution obtained by the RS method is smooth, stable and approximate true null distribution well.

For the multi-class microarray data, the analysis of variance (ANOVA) is useful to identify differentially expressed genes [4]. In ANOVA, the *F*-test is a generalization of the *t*-test that allows for comparison of more than two samples [15]. However, due to small sample sizes, the classical F-test is also subject to the same problems as the *t*-test: bias and unstable estimates of gene-specific variances. To tackle this issue, many authors [15-19] proposed modified *F*-statistics. However, like the classical *F*-test, these modified *F*-tests still suffer from high false-positive rates because (i) the sample size is often so limited that the asymptotic F distribution is not accurate enough to obtain a valid *p*-value and (ii) they appeal to multiple-testing procedures such as the Bonferroni procedure or the BH-procedure. As mentioned above, these multiple-testing procedures have a basic requirement that sample sizes are large enough for valid p-values. In microarray data, the requirement is not realistic. Based on consideration of these problems, we here propose a novel statistical method for the analysis of multi-class gene-expression data called Ranking Analysis of *F*-statistics (RAF). RAF is a natural extension of our previous work, i.e., the ranking analysis of microarrray (RAM) for two-class *t*-tests [14]. It works on finding genes that are differentially expressed among multiple treatment groups by comparing the ordered real *F*-statistics with the ordered estimated null *F*-statistics and implementing a two-simulation strategy to estimate the false discovery rate (FDR).

Methods

Animal model and design

Studies were performed on male stroke-resistant SHR/N (CRiv) (SHRSR) and stroke-prone SHR/A3 (Heid) (SHRSP) rats from a breeding colony maintained by the investigators as previously described [20]. Age-matched male rats from each strain (N = 12 SHRSP and 12 SHRSR) were fed a standard rat chow and water ad libitum until age 8 weeks. Subsequently, animals from each strain were randomized to one of 2 dietary regimens (N = 6 in each strain × diet group): a "stroke-permissive diet" high in sodium (HS) (0.63% potassium, 0.37% sodium) and 1% NaCl drinking solution; a "stroke-protective diet" low in sodium and high in potassium (LS) (1.3% potassium, 0.35% sodium) and regular drinking water. All animals were housed at $23°C$ on a 12-hour light-dark cycle. Rats were sacrificed at 12 weeks of age, and brain tissue was collected for RNA extraction and subsequent microarray analysis. The study protocols were approved by the Animal Care Committee of the University of Texas – Houston. Since strain and dietary factor each have only two levels, we here treated them as one-way in statistical analysis instead of two-way, that is, we are neither interested in strain effects alone nor in dietary effects alone but focus on their combined contributions to gene expression. Thus, HS-SHRSPs, LS-SHRSPs, HS-SHRSRs, and LS-SHRSRs are viewed as four treatment groups for the purpose of the analyses.

Microarray experiment

Microarray analysis was performed as described by Lockhart et al. [21]. Briefly, 10 μg total RNA extracted from each of the 24 rats was used to synthesize cDNA, which was then used as a template to generate biotinylated cRNA. cRNA was fragmented and hybridized to a Test2 chip to verify quality and quantity of the samples. Each sample was then hybridized to a RGU34A array (Affymetrix, Santa Clara, CA). After hybridization, each array was washed and scanned, and fluorescence values were measured and normalized using the Affymetrix Microarray Suite v. 5.0 software.

Ranking **F-*Test***

Let x_{gij} be the expression value for replicate j of gene g in group i where $g = 1,..., N$ (number of genes), $j = 1,..., r_{gi}$ (number of replicate observed values of gene g in group i) and $i = 1,..., n$ (number of groups). The traditional *F*-statistic in one-way ANOVA may be expressed as

$$F_g = \frac{\sigma^2(G_g)}{\sigma^2(e_g)} \qquad (1)$$

where σ^2 (G_g) and σ^2 (e_g) are inter- and intra-group variances of the expression values of gene g, respectively. In the conventional *F*-tests, for example, significance of p = 0.01 in the context of the standard *F* distribution is for a single hypothesis to be tested; therefore, it is unsuitable to microarray data because in a microarray experiment for 10,000 genes we would expect to identify 100 genes by chance [9]. To address this problem, an alternative approach is to rank genes by magnitude of their *F* values so that F_1 is the largest value, F_2 is the second largest value, etc., and F_{g*} is the $g*$th largest value where $g*$ is a rank position of gene g. Thus, we have a ranking *F*-test where

$$F_{g*} - f_{g*} > \Delta \tag{2}$$

indicates that the expression variation of gene g among multiple groups (or multiple conditions) is significant. In Eq. (2), f_{g*} is the expectation of F_{g*} under the null hypothesis and Δ is a given threshold.

Estimation of $\mathbf{f_{g*}}$

In the ANOVA framework, we have

$$\begin{aligned}\sigma^2(G_g) &= \frac{\sum_{i=1}^n r_{gi}(\overline{x}_{gi}-\overline{x}_g)^2}{n-1} = \frac{\sum_{i=1}^n r_{gi}[(\tau_{gi}+\overline{e}_{gi})-(\mu_g+\overline{e}_g)]^2}{n-1} \\ &= \frac{\sum_{i=1}^n r_{gi}(\tau_{gi}-\mu_g)+\sum_{i=1}^n r_{gi}(\overline{e}_{gi}-\overline{e}_g)^2}{n-1} \\ &= \sigma^2(\tau_g)+\sigma^2(\overline{e}_g)\end{aligned} \tag{3}$$

where $\overline{x}_{gi} = \tau_{gi} + \overline{e}_{gi}$ and $\overline{x}_g = \mu_{gi} + \overline{e}_g$; τ_{gi} and $\overline{e}_{gi}$ are treatment effect and average random error in group i, respectively; $\overline{x}_g$ and $\overline{e}_g$ are the overall observed mean and the overall average error for gene g, respectively, and μ_g is the overall expected mean for gene g and r_{gi} is the number of replicate observed values of gene g in group i. Therefore, the inter-group variance σ^2 (G_g) consists of two parts: variance of treatment effects on expression of gene g, σ^2 (τ_g), and average error variance, σ^2 $(\overline{e}_g)$. Thus, *F*-statistic can be rewritten as

$$F_g = \frac{\sigma^2(G_g)}{\sigma^2(e_g)} = \frac{\sigma^2(\tau_g)+\sigma^2(\overline{e}_g)}{\sigma^2(e_g)} = \frac{\sigma^2(\tau_g)}{\sigma^2(e_g)} + \frac{\sigma^2(\overline{e}_g)}{\sigma^2(e_g)} = \frac{\sigma^2(\tau_g)}{\sigma^2(e_g)} + f_g. \tag{4}$$

Therefore, the null hypothesis is equivalent to $F_g = f_g$ because σ^2 $(\tau_g) = 0$ under the null hypothesis. Note that σ^2 $(\tau_g) = 0$ means the treatment effects $\tau_{gi} = \ldots = \tau_{gn} = \mu_g$. In order to do a ranking *F*-test, it is necessary to obtain a good estimate of f_{g*}. In the two-group scenario, Tusher et al. [9] employed a permutation approach to estimate the expected *t*-statistics. The permutation process cannot completely clear the treatment effect in the ranked *d*-statistics so that the estimated ranked *d*-statistics distribution is biased against its null distribution and unstable, in particular, when sample sizes are small (see Appendix A in Tan et al., [14]). Tan et al. [14] developed a "Randomly Splitting" (RS) approach to estimate the null distribution of *t*-statistics. In this study, we extended the RS approach to estimating the null distribution of *F*-statistics.

In the RS approach, one sample consisting of r_{gi} replicates is drawn from group i. Since only one sample is drawn from a group, sample i represents group i. Within a sample all the observed expression values of gene g come from the same group. These values have the same overall expected mean μ_g and the same treatment effect τ_{gi} on expression of gene g except for expression noises. A sample of r_{gi} replicate values for gene g is randomly split into two sub-samples denoted by s = 1 and s = 2. If let $\overline{x}_{gis}^J$ be the mean of sub-sample s of sample i for gene g at split J(J = 1,..., M), then $\overline{x}_{gis}^J$ can be expressed as

$$\overline{X}_{gis}^J = \mu_g + \tau_{gi} + \overline{e}_{gis}^J \tag{5}$$

where $\overline{e}_{gis}^J = \sum_{j=1}^{m_{gis}^J} e_{gisj}^J / m_{gis}^J$ and m_{gis}^J and e_{gisj}^J are replicate number and noise in the observed expression value j in sub-sample s in group i for gene g at split J, respectively. $\overline{e}_{gi}$ is estimated by the difference between two sub-sample means in sample i for gene g at split J,

$$\begin{aligned}\overline{e}_{gi}^J &= \frac{1}{2}\left(\overline{x}_{gi1}^J - \overline{x}_{gi2}^J\right) \\ &= \frac{1}{2}\left[\left(\mu_g + \tau_g + \overline{e}_{gi1}^J\right) - \left(\mu_g + \tau_g + \overline{e}_{gi2}^J\right)\right]. \\ &= \frac{1}{2}\left(\overline{e}_{gi1}^J - \overline{e}_{gi2}^J\right)\end{aligned} \tag{6}$$

It can be seen from Eq. (6) that μ_g and τ_{gi} are cleared in difference between two sub-sample means, which is unrelated to sample size. Thus, the average random error variance σ^2 $(\overline{e}_g)$ in Eq.s (3) and (4) can be estimated by

$$\sigma^2(\overline{e}_g^J) = \frac{\sum_{i=1}^n r_{gi}(\overline{e}_{gi}^J-\overline{e}_g^J)^2}{n-1} \tag{7}$$

where $\overline{e}_g^J = \sum_{i=1}^n \overline{e}_{gi}^J / n$ is estimate of mean $(\overline{e}_g)$ of expression noise of gene g among groups at split J. Variance $\sigma^2(\overline{e}_g)$ is an estimate of expectation (σ^2(graphic file="1471-2105-9-142-i17.gif"/>)) of inter-group variance (σ^2 (G_g)) under the null hypothesis at split J. We therefore have

$$f_g^J = \frac{\sigma^2(\bar{e}_g^J)}{\sigma^2(e_g)}. \qquad (8)$$

Note that since treatment effect is completely removed from the difference between two sub-sample means, the difference is pure noise. We rank f_g^J across all g and let f_{g*}^J denote the value in ordered position $g*$ at split J. After running M splits, we have M values of f_{g*}^J for position $g*$. Thus f_{g*} in Eq. (2) can be estimated by the average of f_{g*}^J over all M splits, i.e., $\bar{f}_{g*} = \sum_{J=1}^{M} f_{g*}^J / M$.

Estimation of FDR

To identify genes whose expression is significantly changed among multiple conditions, it is necessary to estimate the FDR for a given threshold [7,22]. Here we propose a two-simulation approach for FDR estimation [14]. Consider a series of threshold values Δ_k(k = 1,...L) and let N_k be the number of genes that are claimed as significant by RAF at threshold Δ_k. N_k comprises two parts: the number $N_k(t)$ of the true positives and the number $N_k(f)$ of the false positives, i.e., $N_k = N_k(t) + N_k(f)$. Thus, given a threshold Δ_k, FDR is defined as $\lambda_k = N_k(f)/N_k$. $N_k(f)$ is unknown, hence λ_k must be estimated. Many approaches such as BH procedure [7,22], BL procedure [8], Storey's procedure [23,24], and Pounds and Cheng's procedure [25] have been proposed to estimate the FDR. These approaches, however, are based on the assumption that the tests are independent. As mentioned previously, this assumption may not be met in practice. Therefore, these methods may not be suitable to our ranking test. Based on the fact that sampling distribution fluctuates around the expected distribution via permutation, Tusher et al. [9] developed a permutation-based estimator to estimate FDR in the ranking tests. It has been proved, however, in theory and in simulation that when the sample sizes are small, the number of permutations is very limited so that the treatment effects cannot be removed in the permutated data [14]. As a result, the estimator is biased for a given threshold. Here we extend the interval approach by Tan et al. [14] to the ranking analysis of F-statistics. In this approach, we first construct an estimated interval of the true FDR, and then we find a reasonable estimate of FDR. This interval is based on the complete and partial null distributions given by two simulations.

In simulation 1, for each gene, n samples (groups) each having r replicates are generated from normal distributions with a set of sample means ($\bar{y}_{g1},...,\bar{y}_{gn}$) and a set of sample error variances [s^2 (e_{g1}),..., s^2 (e_{gn})]. Here we set $\bar{y}_{g1} = ... = \bar{y}_{gn} = \bar{x}_{gi}$ and i is a randomly chosen group from the observed data, for each of a half of the genes with the null effect that the group variance is zero, i.e., σ^2 (G_g) = 0 and $\bar{y}_{gi} = \bar{x}_{gi}$ for each of the other half with unknown effect that the group variance is larger than or equal to zero, i.e., σ^2 (G_g) ≥ 0. s^2 (e_{gi}) is set to be equal to σ^2 (e_{gi}) where $\bar{x}_{gi}$ and σ^2 (e_{gi}) are the observed values from the real microarray data set.

B sets of simulation data are obtained from this procedure. Each is subject to the ranking analysis described in the previous section. For simulation data set b, every ranked position has thus its corresponding F value that is denoted by F_{g*1}^L (b = 1,..., B). Here those that are called significant by comparing F_{g*1}^b to $\bar{f}_{g*}$ at a given threshold Δ_k are counted as N_{1k}^b across all ranking positions. Let $N_{1k} = \max_{J=1}^{B} N_{1k}^b$ Given the fact that a small part of genes have unequal means in the samples, the simulation data set produces a partially null F-distribution. In other words, it may produce $N_{1k}^b > N_k(f)$, possibly leading to $\lambda_k = N_{1k}/N_k > 1$. To avoid this situation, suppose N_{1k}(k = 1,..., L) takes the maximum value $N(m)$ at $\Delta_{k=m}$, we define

$$\lambda_{1k} = \frac{2N_{1k}}{N(m)+N_{1k}} \qquad (9)$$

as a function of threshold Δ_k for estimating FDR where we set $N_{1k} = N(m)$ for all $\Delta_k < \Delta_m$(k = 1,..., L). Obviously, λ_{1k} is a decreasing function of the threshold and bounded between 0 and 1. For example, λ_{1k} = 1 when $N_{1k} = N(m)$, and λ_{1k} = 2/3 when $N_{1k} = N(m)/2$. Results from the simulation study in Figure 1 indicate that $\lambda_{1k} \geq \lambda_k$ (true value of FDR at threshold Δ_k) when threshold Δ_k is smaller than a value Δ* but $\lambda_{1k} \leq \lambda_k$ when $\Delta_k > \Delta$* (see Figure 1).

The second simulation for estimating FDR is carried out in the following fashion. n samples (groups) each having r replicates for each gene are generated from normal distributions with a set of sample means, $\bar{y}_{g1} = ... = \bar{y}_{gn} = \bar{x}_{gi}$ and a set of sample variances s^2 (e_{g1}) = σ^2 (e_{g1}),..., s^2 (e_{gn}) = σ^2 (e_{gn}).

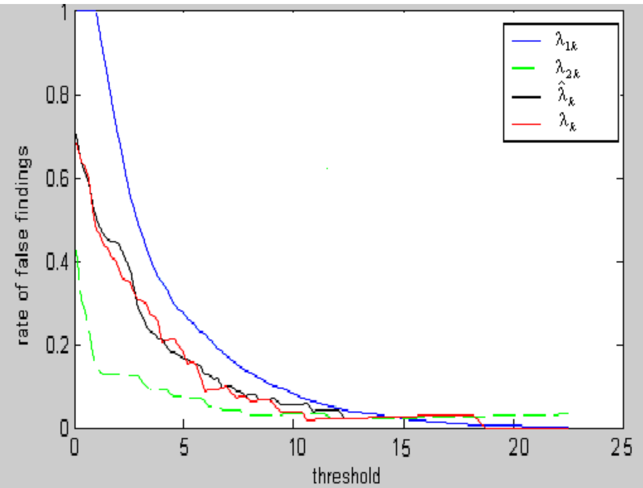

Figure 1
Profile of estimates of FDRs for a series of thresholds. λ_{1k} and λ_{2k} are two threshold functions from simulations 1 and 2 and were used to construct an estimation interval for estimate of FDR at threshold Δ_k. λ_k and $\hat{\lambda}_k$ are true and estimated FDRs at threshold Δ_k, respectively, where k = 1, 2,...,L.

We also produce B sets of data from simulation 2. As in simulation 1, for each simulation data set, every ranked position also has its corresponding F-value denoted as F^b_{g*2} (b = 1,..., B). Let $F_{g*2} = \min_{b=1}^{B} F^b_{g*2}$. The positives found by comparing F^b_{g*2} to F_{g*2} at a given threshold Δ_k are counted as N^b_{2k} across all ranking positions. Here let $N_{2k} = \sum_{b=1}^{B} N^b_{2k} / B$. Unlike the first simulation, here the simulation data sets produce B null F-distributions, so N_{2k} should be approximate to the true number of false positives $N_k(f)$. However, when threshold Δ_k is large, it is possible to have $N_k = 0$ so that N_{2k}/N_k is undefined. To avoid this situation, we define

$$\lambda_{2k} = \frac{N2k}{Nk+N2k} \tag{10}$$

as the second function of threshold (see Figure 1). In particular, we let $\lambda_{2k} = 1$ if $N_k = N_{2k} = 0$ because $\lambda_{2k} = 1$ when $N_k = 0$ and $N_{2k} > 0$.

Thus, an interval for FDR estimation at threshold Δ_k can be constructed between λ_{1k} and λ_{2k}. The third function of threshold for FDR estimation is given as

$$\lambda_{3k} = \alpha_k\lambda_{1k} + \beta_k\lambda_{2k}$$

where $\alpha_k = \min(\lambda_{1k}, \lambda_{2k})/(\lambda_{1k} + \lambda_{2k})$ and $\beta_k = 1 - a_k$. λ_{3k} plays the role of weight in balancing λ_{1k} and λ_{2k}. Therefore, at threshold Δ_k, a putative probability that a false discovery is found in the genes called significant by RAF is

$$\hat{\lambda}_k = \frac{1}{3}(\lambda_{1k} + \lambda_{2k} + \lambda_{3k}). \tag{12}$$

Note that as shown in the simulation result section, λ_{2k} is an underestimate of λ_k and λ_{1k} is an overestimate of FDR when the threshold $\Delta_k < \Delta^*$. However, the situation is reversed when threshold $\Delta_k > \Delta^*$. This is because N_{1k} becomes very small when $\Delta_k > \Delta^*$ so that λ_{1k} becomes very small whereas, from Eq. (10), λ_{2k} slowly decreases if $N_k > N_{2k}$ or increases if $N_k < N_{2k}$ as threshold increases. In addition, when the microarray data have no treatment effects for all the genes detected, then $\lambda_{1k} = \lambda_{2k} = \lambda_{3k} = 1$, leading to $\hat{\lambda}_k = 1$

In order to smooth $\hat{\lambda}_k$ between thresholds Δ_k and Δ_{k+1}, we define a recursive formula modifying the probability $\hat{\lambda}_k$ as

$$\hat{\lambda}_k = \lambda_k p_k + \lambda_{k+1} q_k \tag{13}$$

where $p_k = (N_k - N_{k+1})/(1 + N_k - N_{k+1})$ and $q_k = 1 - p_k$. Eq. (13) suggests that $\lambda_{k+1} = \lambda_k$ if $N_k = N_{k+1}$. The number of false discoveries among those found to be significant at threshold Δ_k in the observed data is estimated by $\hat{N}_k(f) = \hat{\lambda}_k N_k$. Figure 1 shows that the curve of $\hat{\lambda}_k$ agrees well with that of λ_k.

Results

Estimation of the null distribution of F*-statistics*

To examine if the empirical distributions obtained by the RS approach are appropriate for the analysis of the expression data, we simulated a microarray data set consisting of 3770 genes and four groups each having 6 replicates using one group mean and error variance for each gene. Thus, the simulation without treatment effect generated a set of pure noise data.

A set of 3770 F_k values was computed from the simulated data set. We applied our RS approach to this simulated data set to generate $\bar{f}_k$ over 50 splits. This set of 3770 F_k values formed null distribution of F-statistics, which is called f-distribution. To display the profile that our estimate of *f*-distribution is approximate to the null *f*-distribution, we plotted the ranked F_k versus ranked $\bar{f}_k$. The result displays in Figure 2 where all ranked F - $\bar{f}$ dots roughly fall on a diagonal line as expected by two sets of the same ranked distributions. These results suggest that the $\bar{f}$ -distribution is indeed an approximate estimate of *f*-distribution.

Estimation of FDR

Since it is in general unknown if a given gene expresses differently among multiple conditions, it is not necessarily best to use real data of gene expression to evaluate an FDR estimator. But simulation is a useful approach to doing such a task. Therefore, we also conducted a computer simulation for comparing expression status (significant or not) of a gene identified by a method with its real status. This simulation was also based on our real data set of 3770 genes. Treatment effect τ on expression variation was set for 30 % of the genes and assigned in 4 groups. The mean expression value of gene g was set $\bar{y}_{g1} = \bar{x}_g + 2\tau, \bar{y}_{g2} = \bar{x}_g + \tau, \bar{y}_{g3} = \bar{x}_g - \tau$ and $\bar{y}_{g4} = \bar{x}_g - 2\tau$ for the 4 groups where $\tau = 100U$, $0 < U \le 1$, $\bar{x}_g$ is overall observed average for gene g, and each group has 6 replicates. Obviously, treatment effect τ on expression changes randomly with genes in our simulation, which would make it more difficult to identify differentially expressed genes than the simulations with a fixed treatment effect. Figure 1 displays a comparison between RAF estimated and true FDRs. One can see that the RAF estimate curve is very close to the true FDR curve given a series of thresholds.

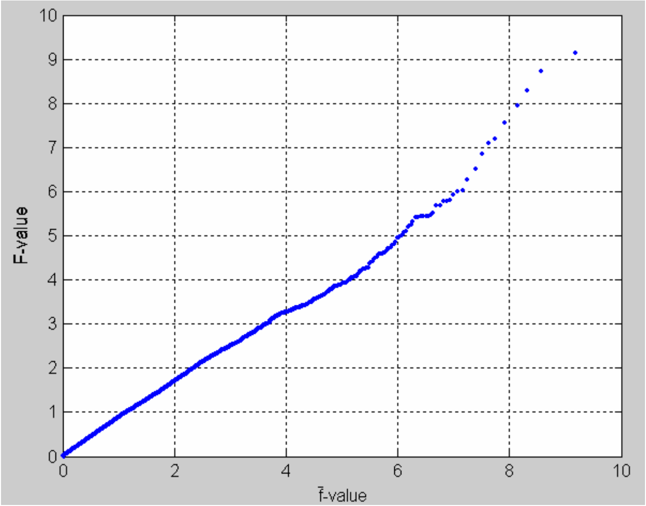

Figure 2 $\bar{f}$

The dot-plot of *F*-values versus $\bar{f}$ -values. *F*-values and $\bar{f}$ -values obtained from the simulated microarray data of 3770 genes were ranked where $\bar{f}$ -values were yielded by the random splitting approach. All ranked *F* - $\bar{f}$ dots roughly fall on a diagonal line as expected by two sets of the same ranked distributions.

Efficiencies of different methods in finding genes differentially expressed among multiple groups

To evaluate different methods, we generated 30 simulation data sets of 3770 genes with the same simulation procedure described above where treatment effect τ was randomly assigned to 10% of the genes in 4 groups, each with 6 replicates. We compared four typical methods with these simulation datasets, of which the Bonferroni (B) procedure and Benjamini-Hochberg (BH) procedure are conventional multiple-testing procedures based on a series of *p*-values obtained from the classical *F*-test; SAM is a ranking method using the Fisher linear discrimination [9]. Our method also is a ranking method but based on the classical *F*-test. Although the *F*-test based on the hierarchical error model (HEM) proposed by Cho and Lee [26] also is suitable to multiple-sample data, the HEM method has consistent performance with the SAM and has no estimate of FDR. Therefore we did not take the HEM method into account of our comparisons among methods. Table 1 summarizes the results obtained by applying these four methods to the 30 simulation datasets where efficiency of a method in finding genes differentially expressed among multiple groups was comprehensively evaluated by averaging number of called significances (NCS), estimated number of false positives (ENFP), true number of false positives (TNFP), and differences (d values) between ENFPs and TNFPs within a given range of FDR over these 30 simulation data sets. Here we measure the conservativeness of a method by the conservative degree $C(d \ge 0)$, defined as the proportion of simulations with $d \ge 0$. In Table 1, as expected, the B procedure gave the most conservative findings and the lowest power among these methods. Similarly, the BH procedure also yielded a very conservative result in which 96.7 percent of ENFPs were larger than TNFPs, in other words, conservation degree reached 96.7%. For the two ranking methods, Table 1 displays the results in the cases of FDR at 6 levels $0.04 < \lambda \le 0.05$, $0.03 < \lambda \le 0.04$, $0.02 < \lambda \le 0.03$, $0.01 < \lambda \le 0.02$, $10^{-4} < \lambda \le 0.01$, and $\lambda < 10^{-4}$. It is clear that RAF has slightly larger means of NCS at all these 6 FDR levels than SAM. In RAF, the means of ENFP are all higher than the

Table 1: Efficiencies of different methods in identifying genes differentially expressed among four groups each with 6 replicates in 30 simulated datasets

		NGCS			ENFP			TNFP			Difference between ENFP and TNFP		
Method	FDR	Mean (SD)	Min	Max	Mean (SD)	Min	Max	Mean (SD)	Min	Max	$\left\|\bar{d}\right\|$	*Var* (*d*)	*C*(*d* ≥ 0)
B procedure	λ = 0.05	59.6 (6.6)	46	73	3.0 (0.3)	2	4	0.0(0.0)	0	0	3.0		100%
BH Procedure	λ = 0.05	102.2 (9.9)	81	119	4.8 (1.0)	4	6	1.6 (1.4)	0	6	3.2		97%
SAM	0.04 <$\lambda \le$ 0.05	111.5(14.3)	89	129	5.1 (0.6)	5	6	5.6(2.8)	2	12	2.0	6.7	56.5%
	0.03 <$\lambda \le$ 0.04	106.8(13.2)	84	119	3.7 (0.6)	3	5	3.8(2.3)	0	8	1.5	4.0	66.7%
	0.02 <$\lambda \le$ 0.03	96.2(12.5)	80	119	2.3 (0.6)	1	3	3.1(1.7)	1	6	1.4	3.1	39.4%
	0.01 <$\lambda \le$ 0.02	91.0(12.7)	71	107	1.3 (0.47)	1	2	1.6(1.2)	0	4	0.9	1.1	67.5%
	0.00 <$\lambda \le$ 0.01	98.7(6.6)	94	108	0.9 (0.1)	1	1	1.5(1.1)	0	3	1.0	1.9	36.4%
	λ = 0.00	82.9(11.0)	66	108	0.0 (0.0)	0	0	1.0(0.6)	0	3	1.0	1.4	23.1%
RAF	0.04 <$\lambda \le$ 0.05	115.1 (9.2)	96	131	5.1 (0.4)	4	6	4.4(2.7)	1	9	2.2	7.3	75.0%
	0.03 <$\lambda \le$ 0.04	110.6(12.2)	85	128	3.9 (0.6)	3	5	3.2(2.1)	1	8	1.6	3.9	79.2%
	0.02 <$\lambda \le$ 0.03	103.6 (10.6)	86	120	2.7 (0.5)	2	3	2.1(1.5)	0	6	1.3	2.8	81.8%
	0.01 <$\lambda \le$ 0.02	100.7 (10.8)	81	118	1.7 (0.5)	1	2	1.1(0.9)	0	3	0.9	1.3	75.8%
	0.00 <$\lambda \le$ 0.01	100.8 (4.1)	96	112	1.1 (0.2)	1	2	0.7(1.0)	0	3	0.9	1.4	77.8%
	λ = 0.00	83.8 (7.1)	69	95	0.0 (0.0)	0	0	0.1(0.3)	0	1	0.1	0.1	86.2%

FDR, false discovery rate; NGCS, number of genes called significant; ENFP, estimated number of false positives; TNFP, true number of false positives.

$\left|\bar{d}\right| = \frac{1}{N_\lambda}\sum_{k=1}^{N_\lambda}\left|d_k\right|$ where $d_k = ENFP_K - TNFP_K$ and N_λ is number of $x < \lambda \le y$ in 30 simulations. $C(d \ge 0) = \sum_{k=1}^{N_\lambda} \mathrm{I}_k / N_\lambda$ where $\mathrm{I}_k = 1$ if $d_k \ge 0$, otherwise, $\mathrm{I}_k = 0$. $Var(d) = \frac{1}{N_\lambda - 1}\sum_{k=1}^{N_\lambda}(d_k - 0)^2 = \frac{1}{N_\lambda - 1}\sum_{k=1}^{N_\lambda} d_k^2$.

means of TNFP while in SAM the means of ENFP are all less than the means of TNFP. Table 1 also shows that RAF has 75~86.2% conservation degree in estimates of false positives under 5% of FDR whereas SAM has 23~66% of conservation degrees. These results suggest that RAF has the highest efficiency in finding genes differentially expressed among these four methods.

We also generated a simulated data set of 3770 genes where treatment effect τ was randomly assigned to 10% of genes but sample size for each group was changed from 6 replicates to 4. Table 2 displays the results obtained by SAM and RAF from this data set. It can be seen that SAM has very high FDRs while RAF still works well and detects 9 genes without false positives.

Array findings by RAF

We obtained a set of the observed data in which expression of 3770 genes was measured among four treatment groups HS-SHRSPs, LS-SHRSPs, HS-SHRSRs, and LS-SHRSRs. This set of microarray data is readily applicable to our ranking *F*-test analysis. Figure 3 shows a scattered-dot plot of *F*-values versus $\bar{f}$ -values obtained by the RS approach.

Figure 4 compares the observed plot of ranked F - $\bar{f}$ to the simulated one. One can see from Figure 4 that the observed F - $\bar{f}$ plot begins to deviate from the simulated F - $\bar{f}$ plot at about $\bar{f}$ = 2.1, suggesting that a part of the F-statistics deviates from the $\bar{f}$ -distribution. This result underscores that these genotypes and diet feeds significantly impact on expression of a portion of genes in rat with respect to stroke.

The numbers of genes whose expression is significantly different among the four groups HS-SHRSPs, LS-SHRSPs, HS-SHRs, and LS-SHRs, are found to be 392, 145, and 107 by our RAF under estimates of FDR of 4.8, 0.7, and <7.0% (see Table 3), respectively. These 107 genes with FDR<0.7% are listed in the Additional file 1. Among these 107 identified probes, 31 belong to the expressed

Table 2: Comparison between SAM and RAF in finding genes differentially expressed among four classes in a simulated data set of small sample size (n = 4)

SAM				RAF				
Delta	Number of significances	Number of false positives	Estimated FDR	Delta	Number of significances	Number of false positive	Estimated FDR	True FDR
0.037534	10	5.6	0.56	0.01253	16	6	0.375	0.125
0.044668	10	5.6	0.56	0.37608	13	4	0.308	0.077
0.045738	10	5.6	0.56	0.74013	13	3	0.231	0.077
0.050144	9	4.7	0.52	1.10516	12	2	0.167	0.083
0.052423	9	4.7	0.52	1.47167	11	1	0.091	0
0.055937	9	4.7	0.52	1.84017	10	1	0.100	0
0.059564	9	4.7	0.52	2.58527	9	0	0	0
0.060798	9	4.7	0.52					
0.062046	9	4.7	0.52					
0.063305	0	0	0					

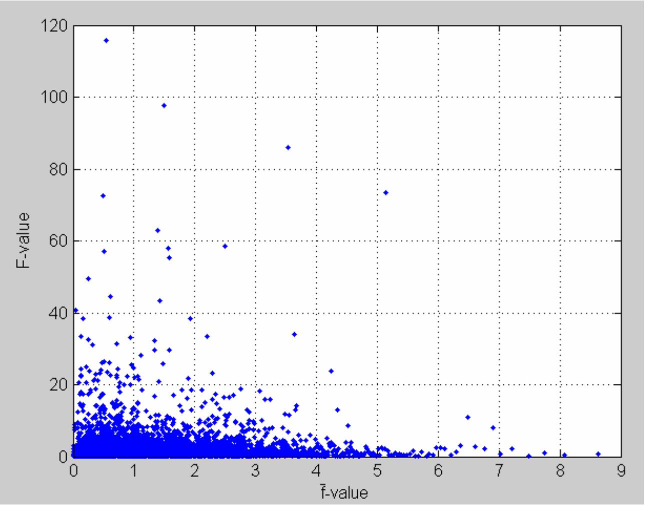

Figure 3

The scatter plot of *F*-values versus $\bar{f}$-values. *F*-values were observed from real microarray data set and $\bar{f}$-values yielded by random splitting approach are an estimate of null *f*-distribution.

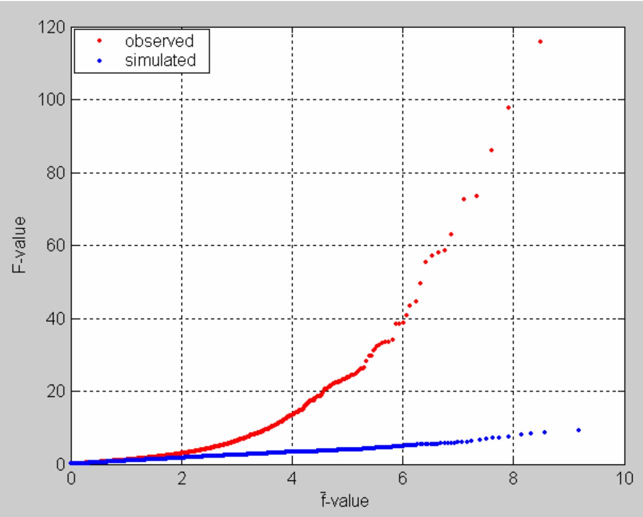

Figure 4

Comparison between the observed (red) and simulated (blue) plots of *F*-values versus $\bar{f}$-values. *F*-values were observed from real (red) and simulated (blue) microarray data sets of 3770 genes and 6 replicates. $\bar{f}$-value yielded by randomly splitting approach is an estimate in null *f*-distribution. *F*-distribution from simulated data set without treatment effects is null distribution. Ranked *F*-values corresponds to ranked $\bar{f}$-values.

sequence tags (ESTs), 76 are unique genes who have known functions in the brain or central nervous system and belong to six major functional classes: (a) neurotransmission such as neurexin III-alpha, Neurodap-1, non-neuronal enolase (NNE), beta isoform of catalytic subunit of cAMP-dependent protein kinase; (b) cell signaling and transportation such as trans-Golgi network integral membrane protein (TGN38), glutamate transporter, alternatively spliced GTP-binding protein alpha subunit intracellular, signal regulatory protein alpha, synaptic vesicle protein 2B (SV2B), L-type amino acid transporter 1 (LAT1), N-ethylmaleimide-sensitive factor (NSF); (c) cell proliferation, differentiation, and apoptosis anti-proliferative factor (BTG1), thyroid hormone receptor a1 (c-erb A α1); (d) metabolism such as stearoyl-CoA desaturase 2, beta isoform of catalytic subunit of cAMP-dependent protein kinase, ATP-citrate lyase. (e) RNA transcript and regu-

Table 3: The results of RAF identifying genes differentially expressed among HS-SHRSPs, LS-SHRSPs, HS-SHRSRs, and LS-SHRSRs.

Delta	Number of genes called significant	Number of false discoveries	Estimated FDR
0.01253	3543	1181	0.333
0.74013	1504	500	0.332
1.10516	1157	173	0.150
1.47167	944	117	0.124
1.84017	794	83	0.105
2.21118	668	59	0.088
2.58527	580	44	0.076
2.96301	515	34	0.066
3.34503	437	24	0.055
3.73199	392	19	0.048
4.12463	370	15	0.041
4.52373	338	12	0.036
4.93017	307	10	0.033
5.34493	269	7	0.026
5.7691	250	6	0.024
6.20391	229	5	0.022
6.65078	209	4	0.019
7.11135	194	3	0.015
7.58753	182	2	0.011
9.13461	145	1	0.007
11.6257	107	0	<0.007

HS, high salt; LS, low salt; SHRSP, stroke-prone SHR/A3 (Heid) rats; SHRSR, stroke-resistant SHR/N (CRiv) rats.

lation such as Zinc finger gene, Jun-D gene, and ribosomal protein genes encoding larger ribosomal subunits L13, L8, and L22; (f) ion channel/pump such as potassium channel-Kv2, electrogenic Na+ bicarbonate cotransporter (NBC), type II Ca2+/calmodulin-dependent protein kinase beta subunit, and protein kinase C-regulated chloride channel.

Independent verification of array findings

Fornage et al (2003) used TagMan assay to measure the relative expressions of 7 genes encoding atrial natriuretic peptide (Anp), the neurotrophin receptor protein tyrosine kinase (TrkB, short), casein kinase 2 (Ck2), complexin 2 (Cplx2), stearoyl CoA desaturase 2 (Scd2), glycerol-3-phosophate acyltransterase (Gpan), and inositol 1,4,5-triphosphate receptor (Itpr1). They found these 7 genes had significantly differentially expressed between SHRSP and SHR strains with $p < 0.05$. Except that genes Anp and Gpan were out of our data, genes for TrkB (short), Cplx2, and Scd2 called significant differential expressions at FDR<0.7%, and for CK2 and Itpr1 at FDR = 0.7% were found among HS-SHRSP, LS-SHRSP, HS-SHR, and LS-SHR strains. Interestingly, Tropea et al [27] also found the genes encoding glutamate receptor (GluR-A) and GABA receptor had significant expression difference between two groups of mice treated by dark rearing and monocular deprivation.

Discussion

To our knowledge, the ranking analysis of *F*-statistics for identifying differentially expressed genes among multiple groups (classes) has not been reported. There are two main difficulties to be overcome: (a) estimate of the null *F*-distribution and (b) estimate of FDR. In conventional statistical methods, permutation is very popular to generate empirical distributions as estimates of the null distributions. However, the permutation approach may not be suitable for microarray data [10-13] because in general microarray experiments have a small sample size due to cost, as a result, treatment effect residues that cannot be removed are amplified in permutation distribution and resulting estimated null distribution has a heavier tail compared to true null distribution [12]. This would results in two consequences: (a) the estimated null distribution is not stable, which, as seen in Table 3, leads to low conservativeness of estimate of FDR, and (b) low power. Our RAF method is successful because the $\bar{f}$ -distribution obtained by applying the RS approach [14] does not contain treatment effects and hence is a desirable estimate of the null *F*-distribution, which is supported by the fact that the observed and simulated results agree well with those expected by theory. Therefore, the number (M) of splits is much smaller than that of permutations for estimate of the null F-distribution. Simulation results showed that 50 splits are enough to obtain a stable and smooth $\bar{f}$ -distribution. In addition, since the $\bar{f}$ -distribution is generated from all the genes detected on microarrays and does not contain treatment effect residences, impact of sample size on the $\bar{f}$ -distribution is very weak. However, we also noted that the $\bar{f}$ -distribution would underestimate the null *F*-distribution when sample sizes are smaller than 4. In this situation, Eq. (7) should be changed to $\sigma^2(\bar{e}_g^J) = \sum_{i=1}^{n} 4(\bar{e}_{gi}^J - \bar{e}_g^J)^2/(n-1)$.

FDR is often used to control the error rate in the BH procedure [7], the BL procedure [8], and in SAM [9]. In practice, for a ranking test, it is necessary to obtain an accurate estimate of FDR. In SAM, FDR is estimated through the permutation approach in which fluctuations around expectation occur among permutated samples. The fluctuations would be dependent on the data itself, i.e., sample size, treatment effect, and data noise. In addition, as indicated above, permutation fails to remove the treatment effects in the data permuted from the microarray data with a small sample size so that the fluctuations are not purely due to random errors. Thus, this approach may give a biased estimate of FDR for a given threshold. The RAF estimator is based on a two-simulation strategy and hence avoids these problems of the SAM estimator, that is, its accuracy is not affected by sample size, treatment effect, and noise. As a result, the number (B) of simulations is also relatively small. Our simulation study indicates that more than 40 simulated data sets (B ≥ 40) would produce stable estimates of FDR across all given thresholds.

Our current RAF method can be readily extended to other test statistics such as Brown-Forsythe test statistic [28], Welch test statistic [29], and Cochran test statistic [30] by replacing F-statistic with the respective statistics.

Conclusion

We developed a new statistical method that is suitable for analyzing microarray data to identify differentially expressed genes among multiple groups, especially, when sample size is small.

Authors' contributions

YDT participated in the method development, performed the statistical analysis and drafted the manuscript. MF participated in the design of the study, carried out the microarray experiment. HX conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Genes with significant expression changes among four groups. The data provided are the tables containing detailed information of the 107 genes with significant expression changes among the four groups with FDR < 0.7%.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-142-S1.pdf]

Acknowledgements

This research was supported by grants from the U.S. National Institutes of Health (HL69126) to MF.

References

1. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer.** *Nature genetics* 1996, **14(4):**457-460.
2. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278(5338):**680-686.
3. Kim YD, Sohn NW, Kang C, Soh Y: **DNA array reveals altered gene expression in response to focal cerebral ischemia.** *Brain research bulletin* 2002, **58(5):**491-498.
4. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7(6):**819-837.
5. Holm S: **A simple sequentially rejective multiple test procedure.** *Scandinavian Journal of Statistics* 1979, **6:**65-70.
6. Hochberg Y: **A sharper Bonferroni procedure for multiple tests of significance.** *Biometrika* 1988, **75(4):**800-802.
7. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B Methodological* 1995, **57(1):**289-300.
8. Benjamini Y, Liu W: **A step-down multiple hypotheses tesing procedure that controls the false discovery rate under independence.** *Journal of Statistical Planning and Inference* 1999, **82:**163-170.
9. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98(9):**5116-5121.
10. Zhao Y, Pan W: **Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments.** *Bioinformatics (Oxford, England)* 2003, **19(9):**1046-1054.
11. Pan W: **On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression.** *Bioinformatics (Oxford, England)* 2003, **19(11):**1333-1340.
12. Xie Y, Pan W, Khodursky AB: **A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data.** *Bioinformatics* 2005, **21(23):**4280-4288.
13. Gao X: **Construction of null statistics in permutation-based multiple testing for multi-factorial microarray experiments.** *Bioinformatics (Oxford, England)* 2006, **22(12):**1486-1494.
14. Tan YD, Fornage M, Fu YX: **Ranking analysis of microarray data: a powerful method for identifying differentially expressed genes.** *Genomics* 2006, **88(6):**846-854.
15. Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome biology* 2003, **4(4):**210.
16. Li H, Wood CL, Liu Y, Getchell TV, Getchell ML, Stromberg AJ: **Identification of gene expression patterns using planned linear contrasts.** *BMC Bioinformatics* 2006, **7:**245.
17. Chen D, Liu Z, Ma X, Hua D: **Selecting genes by test statistics.** *Journal of biomedicine & biotechnology* 2005, **2005(2):**132-138.
18. Tsai PW, Lee ML: **Split-plot microarray experiments: issues of design, power and sample size.** *Applied bioinformatics* 2005, **4(3):**187-194.
19. Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates.** *Biostatistics (Oxford, England)* 2005, **6(1):**59-75.
20. Fornage M, Swank MW, Boerwinkle E, Doris PA: **Gene expression profiling and functional proteomic analysis reveal perturbed kinase-mediated signaling in genetic stroke susceptibility.** *Physiological genomics* 2003, **15(1):**75-83.
21. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al.*: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nature biotechnology* 1996, **14(13):**1675-1680.
22. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Annals of Statistics* 2001, **29(4):**1165-1188.
23. Storey JD: **A direct approach to false discovery rates.** *Journal of the Royal Statistical Society Series B Methodological* 2002, **64(3):**479-498.
24. Storey JD, Taylor JE, Siegmund D: **Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach.** *Journal of the Royal Statistical Society Series B Methodological* 2004, **66(1):**187-205.
25. Pounds S, Cheng C: **Robust estimation of the false discovery rate.** *Bioinformatics (Oxford, England)* 2006, **22(16):**1979-1987.
26. Cho H, Lee JK: **Bayesian hierarchical error model for analysis of gene expression data.** *Bioinformatics (Oxford, England)* 2004, **20(13):**2016-2025.
27. Tropea D, Kreiman G, Lyckman A, Mukherjee S, Yu H, Horng S, Sur M: **Gene expression changes and molecular pathways mediating activity-dependent plasticity in visual cortex.** *Nature neuroscience* 2006, **9(5):**660-668.

28. Brown MB, Forsythe AB: **The small sample behavior of some statistics which test the equality of several means.** *Technometrics* 1974, **16:**129-132.
29. Welch BL: **On the comparison of several mean values: An alternative approach.** *Biometrika* 1951, **38:**330-336.
30. Cochran WG: **Problems arising in the analysis of a series of similar experiments.** *Journal of Royal Statistics Society Serial C Applied Statistics* 1937, **4:**102-118.