# An Investigation on Performance of Significance Analysis of Microarray (SAM) for the Comparisons of Several Treatments with one Control in the Presence of Small-variance Genes

Peer-reviewed author version

LIN, Dan; SHKEDY, Ziv; BURZYKOWSKI, Tomasz; Ion, R.; Goehlmann, H.W.H.; De Bondt, A.; Perera, T.; Geerts, T.; Van den Wyngaert, I. & BIJNENS, Luc (2008) An Investigation on Performance of Significance Analysis of Microarray (SAM) for the Comparisons of Several Treatments with one Control in the Presence of Small-variance Genes. In: 5th International Conference on Multiple Comparison Procedures. BIOMETRICAL JOURNAL, 50(5) Spec. iss, p. 801-823..

DOI: 10.1002/bimj.200710467
Handle: http://hdl.handle.net/1942/9006

# An Investigation on Performance of Significance Analysis of Microarray (SAM) for the Comparisons of Several Treatments With one Control in the Presence of Small-variance Genes

**D. Lin**[*1], **Z. Shkedy**[1], **T. Burzykowski**[1], **R. Ion**[2], **H. W. H. Göhlmann**[2], **A. De Bondt**[2], **T. Perera**[2], **T. Geerts**[2], **I. Van den Wyngaert**[2], and **L. Bijnens**[2]

[1] I-BioStat, Universiteit Hasselt, Agoralaan 1, B3590 Diepenbeek, Belgium, and Katholieke Universiteit Leuven, Belgium

[2] Johnson & Johnson PR&D, Turnhoutseweg 30, 2340, Beerse, Belgium

Received 00 00 0000, revised 00 00 0000, accepted 00 00 0000
Published online 00 00 0000

*Summary*

One of multiple testing problems in drug finding experiments is the comparison of several treatments with one control. In this paper we discuss a particular situation of such an experiment, i.e., a microarray setting, where the many-to-one comparisons need to be addressed for thousands of genes simultaneously. For a gene-specific analysis, Dunnett's single step procedure is considered within gene tests, while the FDR controlling procedures such as Significance Analysis of Microarrays (SAM) and Benjamini and Hochberg (BH) False Discovery Rate (FDR) adjustment are applied to control the error rate across genes. The method is applied to a microarray experiment with four treatment groups (three microarrays in each group) and 16,998 genes. Simulation studies are conducted to investigate the performance of the SAM method and the BH-FDR procedure with regard to controlling the FDR, and to investigate the effect of small-variance genes on the FDR in the SAM procedure.

*Key words:* Dunnett's single step procedure; Benjamini and Hochberg procedure; False Discovery Rate (FDR); microarray; multiple testing; SAM.

## 1 Introduction

Microarray technology allows for simultaneous monitoring of expression levels of a large number of genes. Microarrays are used to determine changes in mRNA content of samples from different origins (e.g., treated versus untreated cells). Statistical methods are then applied to analyze intensity levels of each gene across conditions/treatments and to evaluate the statistical significance of differences between conditions/treatments.

In a drug finding experiment, Dunnett's test is frequently used to compare several treatments with one control. In this paper, we discuss the situation of many-to-one comparisons in the context of a microarray experiment for drug discovery. The aim of such an experiment is to find genes whose expression levels differentiate between any of treatments and the control, and to find the treatments that regulate the expression levels for a set of targeted genes. This type of study is important for finding informative genes as well as finding potential active compounds (treatments).

The prevalent issue related to gene expression profiling is the adjustment for the large number of comparisons that need to be made. Multiple testing procedures controlling for the Family-wise Error Rate (FWER), that is the probability to reject erroneously at least one true null hypothesis, such as Bonferroni (Hochberg, 1995) or Holm (Holm, 1979) procedure, are conservative and lead to a small number of rejections (Hochberg and Tamhane, 1987). For a microarray experiment, the aim of the study is to find

* Corresponding author: e-mail: dan.lin@uhasselt.be

differentially expressed genes while controlling the number of false positive discoveries. Thus, recently intensive research on False Discovery Rate (Benjamini and Hochberg, 1995) has been conducted, in which the FDR is defined as the expected proportion of false rejections among all rejections. Controlling the FDR has gained its popularity in the microarray setting (Tusher *et al.*, 2001, Storey and Tibshirani, 2001, and Reiner *et al.*, 2003). Compared with multiple testing procedures that control the FWER, the FDR procedures are less stringent and lead to a larger number of rejections. In this paper, Benjamini and Hochberg (1995) step-up procedure for controlling the FDR is considered. Moreover, we discuss in detail the Significance Analysis of Microarrays (SAM), a resampling-based procedure (Tusher *et al.*, 2001) which corrects the test statistics by adding a constant (so-called a fudge factor) to the observed standard errors and controls the FDR empirically. The advantage of the SAM is that it does not rely on assumptions about the asymptotic distribution of the test statistic, which in microarray experiments is usually problematic due to a small sample size.

For a single gene, Dunnett's (1955) single step procedure tests for the many-to-one comparisons simultaneously. To address the combined two-dimensional testing problem comparing several treatments with the control for each of thousands of genes, we consider the following four approaches: (1) single step Dunnett's $p$-values adjusted by the Benjamini-Hochbergh procedure (BH-FDR), (2) resampling-based $p$-values (obtained using permutations) adjusted by the BH-FDR, (3) the SAM procedure without the fudge factor, and (4) the SAM procedure with the fudge factor.

As a case study, we consider a microarray experiment with four experimental conditions (three treatments and one control) in triplicate biological samples that are hybridized to individual chips. This results in a data set with 12 arrays and 16,998 probe sets on each array. For simplicity, we refer to probe sets as genes throughout the paper.

The content of this paper is organized as follows. Section 2 describes the procedure followed to obtain the data. Section 3 briefly discusses the methods used for normally distributed data and Dunnett's procedure for comparing several treatments with one control (Dunnett, 1995). Section 4 presents resampling-based algorithms that control the FDR. In Section 5, the SAM procedure is briefly described. Section 6 presents the results of the application of the different analysis strategies to the data. In Section 7 we conduct simulation studies in order to investigate the performance of the SAM when comparing between several treatments with the control for thousands of genes simultaneously. The paper is completed with a discussion in Section 8.

## 2 Data acquisition

The case study data used in this paper come from an experiment where the human epidermal squamous carcinoma cell line A431 was grown in Dulbecco's modified Eagle's medium, supplemented with L-glutamine (20 mM), Gentamycin (5 mg/ml) and 10% fetal bovine serum. The cells were pretreated with three different compounds. RNA was harvested using RLT buffer (Qiagen). In total of 12 microarrays were used under four conditions (three treatments and one control) with three arrays per group and 16,998 genes on each array.

All microarray-related steps including the amplification of total RNAs, labeling, hybridization and scanning were carried out as described in the GeneChip Expression Analysis Technical Manual, Rev.4 (Affymetrix, Santa Clara, CA, 2004). Biotin-labeled target samples were hybridized to human genome arrays U133 A 2.0 containing probe sets interrogation approximately 22,000 transcripts from the UniGene database (Build 133). Hybridization was performed using 15 $\mu$g of cRNA for 16 h at $45^0$C under continuous rotation at 60 rpm. Arrays were stained in Affymetrix Fluidics stations using streptavidin/phycoerythrin staining. Thereafter, arrays were scanned with the Affymetrix scanner 3000, and images were analyzed using the GeneChip Operating System v1.1 (GCOS, Affymetrix). The collected data was quantile normalized in two steps: first within each sample group, and then across all sample groups. (Bolstad *et al.*, 2002)

## 3 Comparing Several Treatments With the Control

Multiple comparisons in the considered case study arise from comparing several treatments with a control and from testing thousands of hypotheses (genes) simultaneously. To identify genes differentiating between several treatment conditions and the control, an ANOVA type of model is a possible option. Kerr *et al*., (2001) formulated a general ANOVA model for the log-transformed gene-expression measurements. Wolfinger *et al.*, (2001) discussed a "gene by gene" modeling approach, in which gene-specific linear mixed effects models are used to determine the significance and the magnitude of treatment effects independently for each gene. In this paper we follow the "gene by gene" modeling approach and specify gene-specific linear model in the following way. Let $X_{ijk}$ be the *i*th gene expression (*i=1,...m*) on array *j* (*j=1,...n*) in treatment group *k* (*k=0,...3*). The gene-specific linear model is given by

$$X_{ijk} = \mu_{ik} + \varepsilon_{ijk}; \;\; \varepsilon_{ijk} \sim N(0, \sigma_i^2), \tag{1}$$

where $\mu_{ik}$ is the mean expression level for treatment *k* for gene *i*, and $\mu_{i0}$ is the mean expression level for the control group for gene *i*. Inference is made by using the estimates of the means with their variabilities.

In particular, we focus on the comparison of the treatments versus the control group. Hence the alternative hypotheses of primary interests, as considered by Dunnett (1995), are:

$$\begin{array}{ll} H_{01i} : \mu_{i0} - \mu_{i1} = 0, & H_{11i} : \mu_{i0} - \mu_{i1} \neq 0, \\ H_{02i} : \mu_{i0} - \mu_{i2} = 0, & H_{12i} : \mu_{i0} - \mu_{i2} \neq 0, \\ H_{03i} : \mu_{i0} - \mu_{i3} = 0, & H_{13i} : \mu_{i0} - \mu_{i3} \neq 0. \end{array} \tag{2}$$

The test statistics for the hypotheses can be written as

$$t_{ik} = \frac{\overline{X}_{ik} - \overline{X}_{i0}}{s_i \sqrt{\frac{1}{n_k} + \frac{1}{n_0}}} \quad i = 1, \ldots, m, \; k = 1, 2, 3. \tag{3}$$

where $s_i$ is the pooled variance for gene $i$, $\bar{X}_{ik}$ is the estimated mean gene expression for the *k*th treatment group for gene $i$, $\bar{X}_{i0}$ is the estimated mean gene expression for the control for gene $i$, $n_k$ is the number of arrays for treatment $k$, and $n_0$ is the number of arrays in the control group.

For comparing the treatment means with the control mean, Dunnett (1955, 1964) proposed the following set of $1 - \alpha$ level simultaneous confidence intervals:

$$\mu_{ik} - \mu_{i0} \in \overline{X}_{ik} - \overline{X}_{i0} \pm |d| S_i \sqrt{1/n_k + 1/n_0}, i = 1, \ldots, m, k = 1, 2, 3, \tag{4}$$

where $|d|$ is the two-sided upper $\alpha$ point of the $k$-variate equicorrelated $t$-distribution with common correlation and degrees of freedom ($\nu = \sum_{k=0}^{3} n_k - 2$). The values of $|d|$ for the balanced one-way model have been tabulated in Bechhofer and Dunnett (1988). For the unbalanced design in the general linear model, the factor analytic method discussed by Hsu (1996) has been implemented in SAS to find the value of $|d|$. From this set of confidence intervals, $\mu_{ik} > \mu_{i0}$ ($\mu_{ik} < \mu_{i0}$) can be concluded for treatment $k$ satisfying $t_{ik} > |d|$ ($t_{ik} < |d|$). The probability of all such statements being correct is no less than the confidence level $1 - \alpha$.

The Dunnett's procedure is more powerful as compared to the Bonferroni procedure which does not take the correlation of the test statistics into account. Note that Dunnett's (1955) procedure is a single step procedure, in which the null hypotheses in (2) for gene $i$ are tested simultaneously by considering the multivariate setting of the test statistics.

# 4 Resampling Based Multiple Testing

## 4.1 Multiplicity

The aim of the microarray analysis is to identify differentially expressed genes without too many false positives. Hence, we expect more than one false positive, but we do not want too many in proportion to true positives. In this section we briefly discuss the procedure that allows to control the number of false positives by controlling the FDR (Benjamini and Hochberg, 1995).

Consider the case in which $m$ hypotheses, from which $m_0$ are true null hypotheses and $m_1$ are false null hypotheses, need to be tested. Let $V$ be the number of true null hypotheses that we wrongly reject and $R$ be the total number of rejected hypotheses.

The FDR, introduced by Benjamini and Hochberg (1995), is defined as the expected proportion of false rejection among the rejected hypotheses, $FDR = E(Q)$ where $Q = V/R$ when $R > 0$, and $Q = 0$ otherwise. Benjamini and Hochberg (1995) proposed the following multiple testing procedure to control the FDR.

Let $P_{(1)} \leq P_{(2)} \ldots \leq P_{(m)}$ be the ordered $p$-values and let $H_{(1)}$,$H_{(2)}$,$\ldots$, $H_{(m)}$ be the corresponding hypotheses. The procedure rejects $H_{(1)}$,$H_{(2)}$,$\ldots$,$H_{(\ell)}$, where $\ell$ is the largest value of $i$, for which $P_{(i)} \leq \frac{i}{m}\alpha$. The BH-FDR adjusted $p$-values (Ge *et al.* 2003) are given by

$$\tilde{P}_i = min_{k=i,\ldots,m}\left[min(\frac{m}{i}P_{(i)}, 1)\right]. \tag{5}$$

Thus, the null hypothesis $H_{(i)}$ is rejected if $\tilde{P}_{(i)} \leq \alpha$.

## 4.2 Permutation $p$-values

In a microarray setting, resampling methods are often used (Kerr *et al.*, 2001, Reiner *et al.*, 2003, Tusher *et al.*, 2001, Westfall and Young, 1993, and Ge *et al.*, 2003). The main motivation is to avoid inference based on the asymptotic distribution of the test statistics, which, within the microarray setting, can be problematic because of either typically small sample sizes or because of departure from the assumption about the distribution of the response. Also, in some cases the asymptotic distribution of the test statistic is unknown (Tusher *et al.*, 2001). The resampling approach requires permutation of the sample labels and calculation of the test statistic for each permutation. Matrix of the values of the test statistic for $m$ genes, obtained by using $B$ permutations, is referred to as the permutation matrix $\boldsymbol{T}$ under the null distribution.

The permutation matrix $\boldsymbol{T}$ can be symbolically written as

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1B} \\ t_{21} & t_{22} & \ldots & t_{2B} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ t_{m1} & t_{m2} & \ldots & t_{mB} \end{pmatrix} \tag{6}$$

where $B$ is the number of permutations and each element $t_{ib}$ in matrix $\boldsymbol{T}$ is the test statistic for the *i*th gene in the *b*th permutation. Once the raw $p$-values are calculated from

$$P_i = \frac{\#(b : |t_{ib}| \geq |t_i|)}{B}, \tag{7}$$

where $t_i$ is the observed value of the test statistic for gene $i$, inference can be made based on the $p$-values using the BH-FDR procedure described in (5) (Ge *et al.*, 2003).

# 5 Significance Analysis of Microarrays

## 5.1 Method

Dunnett's single step tests, discussed in Section 3 can be based on either the distribution of the test statistics under the null hypotheses or on the approximation of the distribution of the test statistic under the null hypothesis using resampling methods (Kerr et al., 2001, Tusher *et al.*, 2001, Reiner *et al.*, 2003, and Ge *et al.*, 2002). In this section we briefly discuss an alternative resampling method, a procedure widely used in the microarray setting, namely, Significance Analysis of Microarrays.

The SAM (Tusher *et al.*, 2001, Storey and Tibshirani, 2001) is a testing procedure for microarray analysis, which estimates the FDR using permutations under the assumption that all null hypotheses are true. The procedure consists of three components: (1) the adjusted test statistic, (2) approximation of the distribution of the test statistic based on permutations, and (3) the control of the FDR.

The $t$-test statistic is modified in the SAM procedure as follows:

$$t_{ik}^{SAM} = \frac{\bar{X}_{ik} - \bar{X}_{i0}}{s_{ik} + s_0} = \frac{\Delta_{ik}}{s_{ik} + s_0} \tag{8}$$

where

$$\bar{X}_{i0} = \frac{\Sigma_{j=1}^{n_0} x_{ij0}}{n_0}, \quad \bar{X}_{ik} = \frac{\Sigma_{j=1}^{n_k} x_{ijk}}{n_k},$$

and

$$s_{ik} = \sqrt{(\frac{1}{n_k} + \frac{1}{n_0}) \frac{\Sigma_{j=1}^{n_k} (x_{ijk} - \bar{x}_{ijk})^2 + \Sigma_{j=1}^{n_0} (x_{ij0} - \bar{x}_{ij0})^2}{n_k + n_0 - 2}}.$$

The constant $s_0$ in (8) is called the fudge factor. It is calculated as the percentile of the gene-wise standard errors that minimizes the coefficient of variation of the SAM test statistics. This modification is used to overcome bias for genes with expression difference $\Delta_{ik}$ close to zero, which have a large value of the test statistic due to a small sample variance. By using an inflated standard error of the test statistics, the SAM addresses the problem of the dependence of the value of the test statistic on the variance of expression levels for a particular gene.

The SAM is a resampling-based procedure, which uses permutations to approximate the null distribution of the test statistics. Thus, the choice of the test statistic does not have an effect on the SAM procedure. This distribution-free property in the SAM procedure allows to include in it any forms of test statistics.

The control of the FDR is performed once the permutation matrix $\boldsymbol{T}$, as defined in (6), is obtained. The SAM procedure requires the test statistics of each permutation to be sorted for all the genes such that the first row of the sorted matrix is the minimum test statistic across permutations and the last row is the maximum, i.e.,

$$\boldsymbol{T}^{SAM} = \begin{pmatrix} t_{(1)1} & t_{(1)2} & \dots & t_{(1)B} \\ t_{(2)1} & t_{(2)2} & \dots & t_{(2)B} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ t_{(m)1} & t_{(m)2} & \dots & t_{(m)B} \end{pmatrix}.$$

In matrix $\boldsymbol{T}^{SAM}$ each element $t_{(i)b}$ is the ordered test statistic in permutation $b$. The expected values of the observed test statistics are approximated by the means of the rows of $\boldsymbol{T}^{SAM}$, $\bar{t}_{(1)}, \bar{t}_{(2)}, \dots, \bar{t}_{(m)}$ that

are constructed in the following way:

$$\boldsymbol{T}^{SAM} = \left(\begin{array}{cccc} t_{(1)1} & t_{(1)2} & \dots & t_{(1)B} \\ \hline t_{(2)1} & t_{(2)2} & \dots & t_{(2)B} \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline t_{(m)1} & t_{(m)2} & \dots & t_{(m)B} \end{array}\right) \Rightarrow \left(\begin{array}{c} \frac{1}{B}\sum_{b=1}^{B} t_{(1)b} \\ \frac{1}{B}\sum_{b=1}^{B} t_{(2)b} \\ . \\ . \\ . \\ \frac{1}{B}\sum_{b=1}^{B} t_{(m)b} \end{array}\right) = \left(\begin{array}{c} \bar{t}_{(1)} \\ \bar{t}_{(2)} \\ . \\ . \\ . \\ \bar{t}_{(m)} \end{array}\right).$$

To call a gene significant, the difference between the observed and expected values of the test statistic needs to be larger than a certain cut-off value $\lambda$ (Parmigiani *et al.*, 2003). For a grid of $\lambda$ values, the corresponding number of significant genes can be listed; at the same time, the number of false positives arising from any permutation matrix $\boldsymbol{T}^{SAM}$ is estimated. Under the null hypotheses, we expect that no differentially expressed genes are present for each permutation. Consequently the median or 90 percentile number of false positives corresponding to $\lambda$ can be obtained from permutation matrix. In this way, the FDR can be calculated for each value of $\lambda$ and an acceptable value of $\lambda$ can be chosen to control the FDR at the desired level. Note that the SAM procedure estimates the proportion of true null hypotheses in obtaining the FDR ($\pi_0 E(V/R) = (m_0/m)E(V/R)$, where $m_0$ is the number of non-differentially expressed genes and estimated from permutations).

In the microarray setting, complications arise not only from the multiple comparisons per gene, but also from the multiple testing for all the genes at the same time. A strength of the SAM is that the null distribution is generated for all the genes at once by permuting the group labels, so that the correlation between test statistics of all the genes is preserved. One can borrow strength across the genes and derive more powerful rejection regions in testing by assuming a statistic from a mixture of the null and alternative distributions, as well as from the pure null distribution (Efron *et al.*, 2001). In our testing strategy, three sets of $t$-test statistics are considered simultaneously. On one hand, the SAM procedure preserves the structural correlation of the three $t$-test statistics per gene. This is an analogue to Dunnett's single step procedure, in which a joint multivariate distribution is estimated to test the order statistics of many-to-one comparisons. On the other hand, the SAM deals with the multiplicity problem due to testing of thousands of genes by estimating the FDR. Within the SAM framework, we rely on the joint distribution estimated from permutations to address these two dimensions of multiple testing simultaneously. Note that the resampling-based methods such as SAM implicitly assume if the marginal distributions of expression levels of the genes are identical across different treatments, then the joint distributions are identical across the treatments as well (Huang *et al.*, 2006, Xu and Hsu, 2007). Whether this assumption is valid in the microarray setting is not known at present.

Adding the fudge factor to the denominator of the test statistics in (8) provides a protection against the false discovery for genes with a relative small expression difference $\Delta_{ik}$ and with a very small variance. However, adding the fudge factor to the denominator of the test statistic leads to the following question: what is the influence of the fudge factor on the other genes, can the SAM test statistics ($t_k^{SAM}$) become too small and as a result, can the truly differentially expressed genes not be detected any more? Moreover, for genes that are not differentially expressed but share the problem of small variance, what is the influence of the fudge factor on those genes, i.e., how is the FDR controlled? In what follows we illustrate this issue graphically.

### 5.2 Graphical Interpretation of the SAM

In order to investigate the effect of the SAM fudge factor on the protection against genes with small variance, we decompose the true null hypotheses into two types (see Figure 1): $m_0^0$ truly non-differentially expressed genes with a moderate to large variance and $m_0^1$ truly non-differentially expressed genes with a relatively small variance (i.e., $\leq 5\%$ percentile of variance in the data). Accordingly, the falsely rejected

true hypotheses $V$ are divided into sets of $V^0$ and $V^1$ for these two types of genes, respectively. Consequently, the FDR can be decomposed as FDR=FDR$^0$+FDR$^1$=$E(V^0/R) + E(V^1/R)$, where $R$ is the total number of genes declared significant.

| Hypothesis | Test result | | |
|---|---|---|---|
| | Accept $H_o$ | Reject $H_o$ | Total |
| $H_o$ True | $U$ | $V^0$ / $V$ / $V^1$ | $m_0^0$ / $m_0$ / $m_0^1$ |
| $H_o$ False | $T$ | $S$ | $m_1$ |
| Total | $W$ | $R$ | $m$ |

$m_0^0$ : genes with small $\Delta$ and $\sigma$ under $H_0$
$m_0^1$ : genes with small $\Delta$, moderate to large $\sigma$ under $H_0$

**Fig. 1** Decision in multiple testing (Benjamini and Hochberg 1995).

Note that it is expected that, without the fudge factor, the small-variance genes will be declared significant and the FDR will increase.

To show the effect of the SAM test statistic, we compare the values of the usual $t$-test statistics and the SAM test statistics and investigate how the SAM test statistic values are affected by the fudge factor. Figure 2 shows the effect size (numerator of $t$-test statistics) versus the absolute values of the SAM $t$-test statistics without (Figure 2$a$) and with the fudge factor (Figure 2$b$). We observe that a large number of genes have large test statistic values with small effect sizes, which are represented by points lying along the zero vertical line (Figure 2$a$). With the introduction of the fudge factor, the points are gathering more around the zero crossing point (Figure 2$b$). However, the values of test statistics for all the genes decrease simultaneously. Figure 3 illustrates how the fudge factor affects genes with different variances.

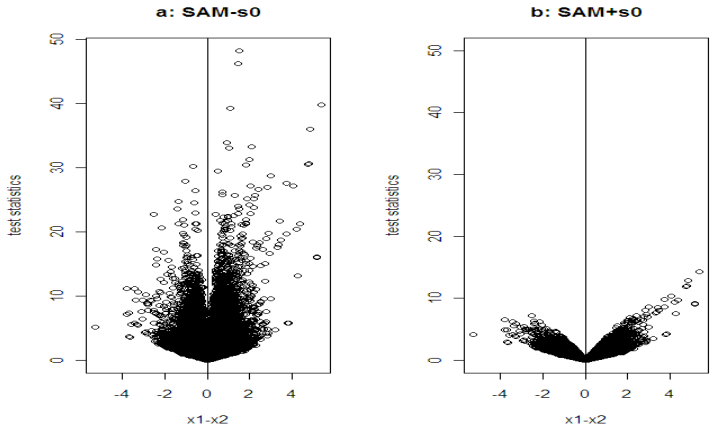


**Fig. 2** Comparison of the SAM test statistics (absolute values) without the fudge factor ($a : SAM - s_0$) and with the fudge factor ($b : SAM + s_0$) using the case study data. $s_0$=0.2422 (60% quantile of the standard errors in the data.)

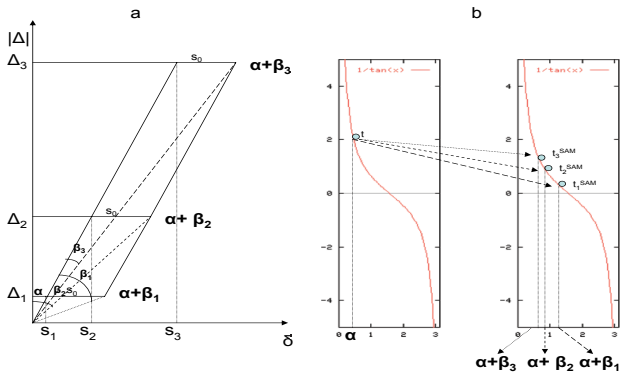


**Fig. 3** Graphical Interpretation of the SAM test statistic: a. the SAM test statistics; b. the cotangent function.

The two axes of Figure 3$a$ represent the numerator (absolute value of effect size) and denominator (standard error) of the $t$-test statistics. The angle $\alpha$ (between the y-axis and the solid line) for the three genes with using small, median and large standard errors ($s_1 < s_2 < s_3$) and corresponding effect sizes ($\Delta_1 < \Delta_2 < \Delta_3$), constitutes the same value of the $t$-test statistic, i.e., $t_1 = \Delta_1/s_1 = t_2 = \Delta_2/s_2 = t_3 = \Delta_3/s_3$. Note that the test statistic value for these three genes is equal to $cot(\alpha)$. When the fudge factor $s_0$ is added in the denominator (extending the standard errors, $s_1$, $s_2$, and $s_3$ by $s_0$, respectively), the new angles are formed by increasing $\beta_1$, $\beta_2$, and $\beta_3$ on the basis of $\alpha$, respectively. The three newly formed angles are provided between the y-axis and the dotted line ($\alpha + \beta_1$), short dashed line ($\alpha + \beta_2$), and the long dashed line ($\alpha + \beta_3$), respectively. Thus, the SAM test statistics for the three genes become $cot(\alpha+\beta_1)$, $cot(\alpha+\beta_2)$, and $cot(\alpha+\beta_3)$, respectively. The values of the SAM test statistics are illustrated by the cotangent function in Figure 3$b$. The left penal of $b$ shows the same $t$-test statistic value of the three genes with angle $\alpha$. However, the introduction of the SAM fudge factor decreases the values of the SAM test statistics for three genes simultaneously, in particular, $t_1^{SAM} < t_2^{SAM} < t_3^{SAM}$ (see the right panel of Figure 3$b$) due to $s_1 < s_2 < s_3$.

Let $s_{(1)}, s_{(2)}, \ldots, s_{(m)}$ be the order statistics of the standard error in the microarray experiment with $m$ genes. Let $s^{(0)}, s^{(1)}, s^{(l)}, \ldots, s^{(100)}$ be the $l$th% quantile of $s_{(1)}, s_{(2)}, \ldots, s_{(m)}$. Let the fudge factor $s_0 = s^{(q)}$, it is easy to see that for gene $i$, the SAM test statistic with the fudge factor ($t_i^{SAM}$) and the $t$-test statistic ($t_i$) have the following relationship:

$$\begin{cases} t_i^{SAM} < 1/2t_i & s_{(i)} < s^{(q)}, \\ t_i^{SAM} = 1/2t_i & s_{(i)} = s^{(q)}, \\ t_i^{SAM} > 1/2t_i & s_{(i)} > s^{(q)}. \end{cases}$$

Hence, the SAM test statistic with the fudge factor is smaller than 1/2 of the $t$-test statistics for genes with their standard errors smaller than the fudge factor. Moreover, the ratio between the SAM with and without the fudge factor is $s_i/(s_i + s_0)$ since $|t|_i^{SAM} = |\Delta|/s_i \times s_i/(s_i + s_0)$. Depending on $s_i$, the standard error of gene $i$, the SAM test statistic becomes smaller by ratio of $s_i/(s_i + s_0)$.

For illustration, let us focus on genes with unadjusted test statistics with values of $8.5 \leq t_i \leq 9.5$ and $15.5 \leq t_i \leq 16.5$. The value of the fudge factor $s_0$ is equal to 0.2422 (i.e., the 60% quantile of $s^{(0)}, s^{(1)}, \ldots, s^{(100)}$). We will elaborate on the choice of $s_0$ in Section 6.

Figure 4$a$ shows the SAM test statistics with and without the fudge factor. For each set, genes with $s_i < s_0$ lying below the line of t'=1/2t reduce the values in test statistics by more than half of the $t$-test statistic; while genes with $s_i > s_0$ lying above the line reduce the values in test statistics by less than half

of the $t$-test statistic. Note that there are 149 genes with $8.5 \leq t_i \leq 9.5$, among which 145 genes (97.3%) have standard error smaller than $s_0$; while there are 24 genes with $15.5 \leq t_i \leq 16.5$, among which 22 genes (91.6%) have standard error smaller than $s_0$. Figure 4$b$ plots the $t_i$ values without the fudge factor versus standard error, where most of genes have their standard error smaller than the fudge factor (at the left side of the vertical line). From Figure 4$c$, we can see that, the relationship between the treatment effect and the standard error. To achieve similar test statistics for each set of genes (either within [8.5,9.5] or [15.5,16.5]), genes with small treatment effects have small standard errors.

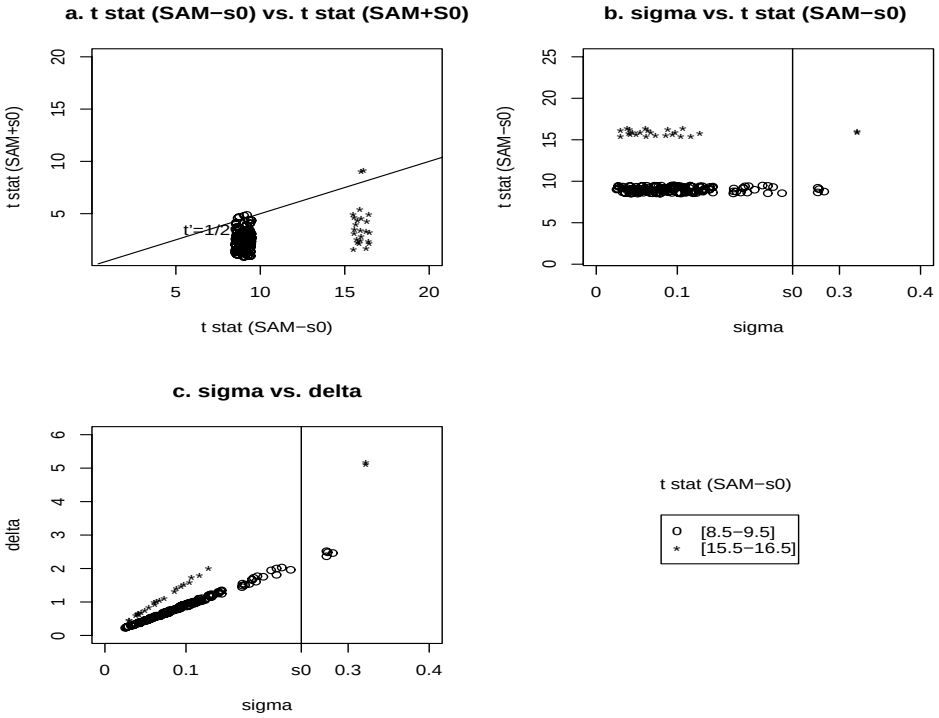


**Fig. 4** Plots of the SAM test statistics with and without the fudge factor, for two sets of genes whose test statistics (without $s_0$) are in the range of [8.5, 9.5] (in stars) and [15.5,16.5] (in pluses). panel a. t stat (SAM-$s_0$) vs. t stat (SAM+$s_0$); panel b. sigma vs. t stat (SAM-$s_0$); panel c. sigma vs. delta.

The SAM procedure introduces the following dilemma. Assuming that $m_0^1 > 0$ (the number of non-differentially expressed genes, for which the variance is relatively small), an analysis without correcting the test statistic using the fudge factor $s_0$ is expected to lead to significant findings of non-differentially expressed genes with small variances. This implies that the FDR will not be controlled. On the other hand, analysis in which the test statistics are corrected using the fudge factor $s_0$ is expected to solve the problem of declaring significant the genes with a small variance, but at the same time the power is reduced. Thus, we need to answer the question how the FDR is controlled for genes with small variances when the SAM procedure is used.

## 6 Application to the Data

In this section, we present results of the application of four procedures, namely, (1) the Dunnett's $p$-values adjusted by using the BH-FDR procedure, (2) the permutation $p$-values adjusted by using the BH-FDR

procedure, and (3) the SAM procedure with the fudge factor, and (4) the SAM without the fudge factor, described in Section 4 and Section 5, to the case study.

### 6.1 Multiple Testing Using Dunnett's $p$-values

First we present the results obtained by using the Dunnett's $p$-values. The single step testing scheme consists of testing all $3 \times 16,998$ tests simultaneously. The gene-wise Dunnett's $p$-values for the three comparisons are obtained using the factor analytic method discussed by Hsu (1996) in SAS, and the BH-FDR procedure is used to control for overall error rate for $3 \times 16,998$ tests. Table 1 (column (1)) presents the results using this approach.

**Table 1** *Number of significant genes identified by using (1) the Dunnett's adjusted p-values adjusted by the BH-FDR procedure, (2) the permutation p-values adjusted by the BH-FDR procedure, (3) the SAM procedure without the fudge factor, and (4) the SAM procedure with the fudge factor.*

| Approach | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| # sign genes | 2319 | 3586 | 5223 | 613 |
| # Comp $1^*$ | 958 | 1555 | 2262 | 262 |
| # Comp $2^*$ | 749 | 1244 | 1514 | 232 |
| # Comp $3^*$ | 612 | 787 | 1447 | 119 |

1*: # of genes declared significant for one treatment compared with the control
2*:# of genes declared significant for two treatments compared with the control
3*:# of genes declared significant for all the three treatments compared with the control

Among the $3 \times 16,998$ tests, the null hypothesis is rejected for 4292 (=$958 + 2 \times 749 + 3 \times 612$) tests, identifying 2319 genes to be significant for at least one comparison between the treatments and the control. The number of genes with one significant comparison is 958; there are 749 genes with two significant comparisons, and 612 genes with all three significant comparisons.

### 6.2 Resampling-based Multiple Testing

In the permutation approach, we use 1000 random permutations of the 12 sample labels and calculate the test statistics for the newly formed treatment groups. The gene-wise permutation $p$-values for each of the three comparisons are obtained using (7). Adjusting the so-obtained $p$-values ($3 \times 16,998$) by the BH-FDR procedure leads to 3586 significant genes, with 1555 genes with one significant comparison, 1244 genes with two significant comparisons, and 787 genes with all three significant comparisons (see column (2) in Table 1).

In the third and fourth approach (i.e., the SAM procedure with and without the fudge factor), the multiplicity issue of the $3 \times 16,998$ tests is adjusted simultaneously in the SAM procedure. The choice of the fudge factor is made based on the algorithm provided in the SAM manual (Tusher *et al.*, 2001). The method chooses the quantile of the standard error which minimizes the CV (coefficient of variation) of the SAM test statistics. Figure 5 plots the CVs versus the quantiles of standard error. Based on the graph, the 60th percentile is chosen. Figure 6 illustrates the selection of threshold $\lambda$ to control the FDR. Panel $a$ shows the relationship between the FDR and $\lambda$, which allows to choose the $\lambda$ corresponding to the desired level of the FDR. For instance, to control the FDR at 0.05, the required $\lambda$ is 0.98. Panels $b$ and $c$ display the number of significant findings and false positives in function of $\lambda$. The last panel shows the observed and expected $t$-test statistics, where the genes lying outside the $\lambda$ band (i.e., the absolute difference between the expected and the observed test statistics larger than a certain $\lambda$) are considered to be significant. With $\lambda$ equal to 0.98, the number of significant genes is 613 and the median number of false positives is 35.

The analysis without the fudge factor $s_0$ leads to 5224 genes declared to be significant (see column (3) in Table 1), while the analysis with the fudge factor $s_0$ reduces the number of significant findings to 613 genes (see column (4) in Table 1).

Figure 7 shows a plot of $t$-test statistics against the gene-specific standard error. Grey points in the middle zone indicate the tests for which the null hypothesis is not rejected. Grey points in the outer zone represent test, for which the null hypothesis is rejected by the SAM with the fudge factor. These significant tests, together with the tests represented by black points are rejected by the SAM without the fudge factor. Note that the use of fudge factor influences the significance of all test statistics simultaneously.

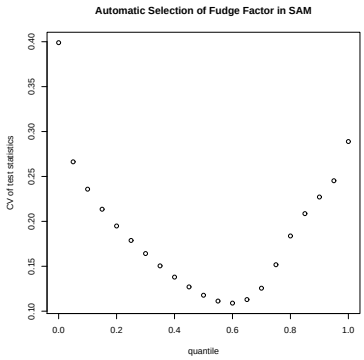


**Fig. 5** Plot of quantiles of the fudge factor vs. the CV (coefficient of variation) of the test statistics: selection of the fudge factor in the SAM.

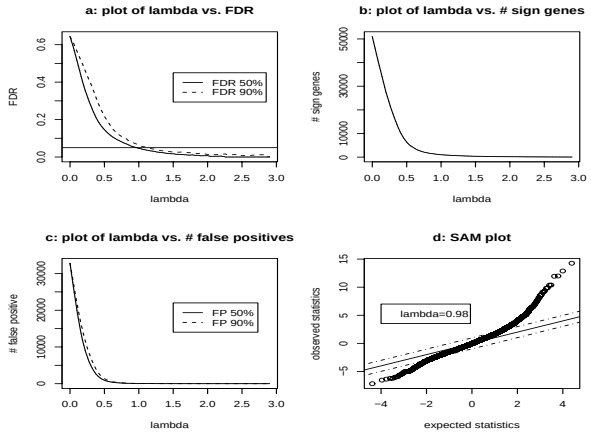


**Fig. 6** $SAM(+s_0)$ plots: a. FDR vs. $\lambda$; b. # of significant gene vs. $\lambda$; c. # of false positives vs. $\lambda$; d. observed vs. expected test statistics.

Note that the number of significant genes found by the Dunnett's $p$-values adjusted by the BH-FDR procedure and the permutation $p$-values adjusted by the BH-FDR procedure lies in-between the numbers for the SAM without and with the fudge factor.

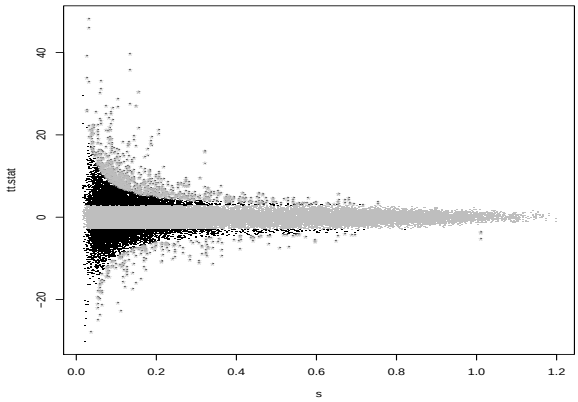


**Fig. 7** $t$-test statistics vs. their standard errors. The grey zone in the middle is non-significant tests; grey points in the outer zone are tests declared significantly by the SAM with and without the fudge factor; black points are tests declared significantly by the SAM without the fudge factor.

## 7 Simulation Study

In this section, we present the results of several simulation studies. The first simulation study investigates the performance of the four testing approaches discussed in Section 6. In the second simulation, we investigate the performance of the automatic choice of the SAM fudge factor, and in particular the effect of the fudge factor on both the $\text{FDR}^0$ and $\text{FDR}^1$. In the third simulation study, we investigate the influence of unequal sample sizes on the $\text{FDR}^0$ and $\text{FDR}^1$. In the fourth simulation study, we investigate the performance of the proposed procedures for the case of non-Gaussian distribution.

### 7.1 Performance of the Four Testing Approaches

#### 7.1.1 Simulation Setting

Data are generated with 16,998 genes per microarray using the gene-specific variances as in the case study. We assume that about 10% of the total genes (i.e., 1700 genes) are truly differentially expressed. Gene expression levels are generated according to the model $x_{ijk} \sim \text{N}(\mu_{ik}, \hat{\sigma}_i^2)$, $i = 1, \ldots, 16,998$, $j = 1, \ldots, n_k$, $k = 0, 1, 2, 3$, where $\hat{\sigma}_i^2$ is the estimated gene-specific variance from the data. The means for the treatment groups are specified in the following way:

$$\mu_{ik} = \begin{cases} \delta_{ik} \times \hat{\sigma}_i, & i \leq 1700 \text{ and } k = 1, \\ \delta_{ik} \times \hat{\sigma}_i, & i \leq 1309 \text{ and } k = 1, 2, \\ \delta_{ik} \times \hat{\sigma}_i, & i \leq 714 \text{ and } k = 1, 2, 3, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here $\delta_{ijk} \times \hat{\sigma}_i$ represents the mean of the treatment group for the differentially expressed genes. It is assumed that $\delta_{ik} \sim U(2.8, 4.5)$. Among the 1700 truly differentially expressed genes, 319 genes are assumed to be differentially expressed for only one treatment; 595 genes are assumed to be differentially expressed for two treatments; and 714 genes are assumed to be differentially expressed for all the three treatments. In total, four settings with the sample size of three, four, five, and six arrays per treatment group are considered. For each setting 100 data sets are generated.

The uniform distribution $\delta_{ik} \sim U(2.8, 4.5)$ for the differentially expressed genes was chosen since we expect that the $t$-test statistics will be in the range of $2.8 \times \sqrt{3/2}$ and $4.5 \times \sqrt{3/2}$. A $t$-test statistic value

in that range is declared significant (without multiple testing) and is expected to be large enough to be declared significant with multiplicity adjustment. The range of the test statistic value above is obtained since $t_{ik} = \hat{\mu_{ik}}/(\hat{\sigma}_i \times \sqrt{2/3}) = \delta_{ik} \times \hat{\sigma}_i/(\hat{\sigma}_{ik} \times \sqrt{2/3}) = \delta_{ik} \times \sqrt{3/2}$ and the pooled variance $\hat{\sigma}_i = \hat{\sigma}_{ik}$ is applied in accordance with Dunnett's procedure.

### 7.1.2 Simulation Results

Using the simulated data, we compare the number of significant genes identified by the Dunnett's $p$-values, the permutation $p$-values adjusted by the BH-FDR procedure, and the SAM approach with and without the fudge factor. The two columns of Figure 8 show the boxplot of the FDR and power for 100 simulated datasets. For both the SAM procedure (with and without the fudge factor), the FDR is well controlled at around 5%. For each simulated dataset, the fudge factor selection is based on the automatic calculation described in the SAM manual.

With three arrays per group, the power obtained for the Dunnett's and permutation $p$-values using the BH-FDR for multiplicity adjustment (32% and 34%, respectively) is much lower than for the SAM approach with the fudge factor (54%) and without the fudge factor (72%). Note that for the setting with three arrays per group the permutation inference using the BH-FDR adjustment yields a slightly higher FDR than 0.05. Increasing the sample size to four, five, or six arrays per group substantially improves the power of the four approaches. It also improves the control of the FDR for the permutation approach. For the SAM with the fudge factor, for example, the power increases to 85% for sample size of 16 arrays, to 96% for 20 arrays, and to almost 99% for 24 arrays.

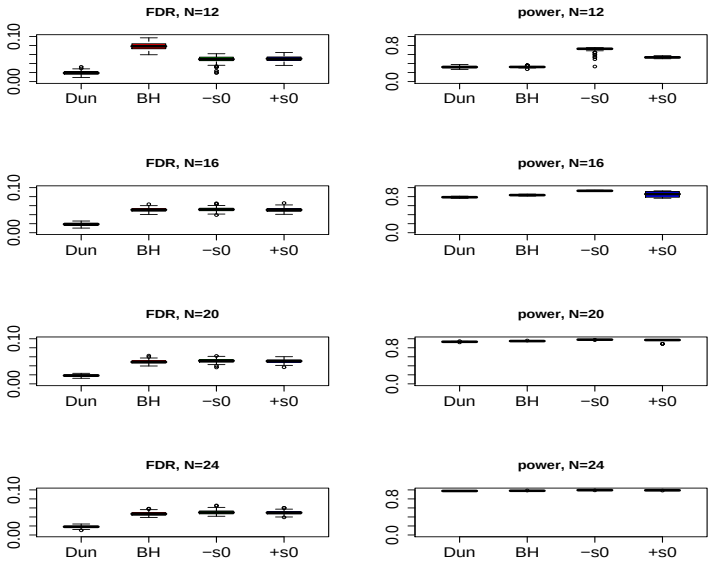


**Fig. 8** Simulation with equal sample sizes. The boxplots in each panel of the figure present the results for the Dunnett's approach (Dun), the permutation approach (BH-FDR), the SAM without the fudge factor ($-s_0$), and the SAM with the fudge factor ($+s_0$). Plots in the left panel show the FDR achieved for three, four, five, and six arrays per group. Plots in the right panel show the power.

In Table 2 we present the number of genes with significant comparisons between the treatments and the control. As expected, the same ordering of the number of significant findings for each procedure can be observed. The number of genes with one, two, and three significant comparisons increases in a similar way as compared to the true number used for simulation. For example, for the SAM procedure with the

fudge factor, and three arrays per group, the mean number of genes with one significant comparison is large, i.e., 560.5 genes, among which on average only 205 genes have a true difference in expression for one treatment; among 368.5 genes with two significant comparisons, on average 209 genes have a true difference in expression for two treatments; and among 271.5 genes with three significant comparisons, on average 269 genes have a true difference in expression for three treatments. We can conclude from the table that, although the number of significant findings decreases as the sample size increases, the average number of truly significant findings corresponds to the number simulated in the setting. Note that results for genes with three significant comparisons contain almost no false positives.

For the first setting (three arrays per group) the SAM procedure controls the FDR empirically at 5% significance level as well. The SAM without the fudge factor seems to outperform the SAM with the fudge factor with respect to power. This is because non-differentially expressed genes with a small variance are not generated in this simulation. In this case there is no need to introduce the fudge factor, as it diminishes the power of the procedure.

**Table 2** *Simulation results. Number of significant comparisons using (1) Dunnett's adjusted p-values adjusted by BH-FDR , (2) permutation p-values adjusted by BH-FDR, (3) SAM procedures without the fudge factor, and (4) SAM with the fudge factor. The number of true significant comparisons are given in parentheses.*

| | | | (1) | (2) | (3) | (4) |
|---|---|---|---|---|---|---|
| 391 | | 1 | 334(120) | 442(111) | 619.5(280) | 560.5(205) |
| 595 | N=12 | 2 | 197.5(127) | 254.5(136) | 529.5(357) | 368.5(209) |
| 714 | | 3 | 166.5(165) | 120(117) | 387.5(382) | 271.5(269) |
| 391 | | 1 | 523(296) | 623(321) | 581(360) | 605(334.5) |
| 595 | N=16 | 2 | 495(394) | 610(445) | 628(519) | 564.5(443) |
| 714 | | 3 | 495(492.5) | 477(473) | 603.5(598) | 534(529.5) |
| 391 | | 1 | 484(357) | 569(369) | 552.5(380) | 562(378) |
| 595 | N=20 | 2 | 573(516) | 632(544) | 624(570) | 620(560) |
| 714 | | 3 | 639.5(635) | 632(629) | 683(677) | 670(666) |
| 391 | | 1 | 457.5(377.5) | 546.5(383) | 541.5(385) | 544.5(385) |
| 595 | N=24 | 2 | 592(564.5) | 622(575) | 616(585) | 616(581) |
| 714 | | 3 | 690(687) | 683(679) | 708(703) | 701(698) |

1: # of genes with one significant comparison
2:# of genes with two significant comparisons
3:# of genes with three significant comparisons

### 7.2 The Effect of the Fudge Factor on the FDR and Power

In this simulation study, we focus on the SAM procedure, and in particular we focus on the automatic selection of fudge factor and the influence of the fudge factor on the FDR. As discussed in Section 5, the fudge factor makes it more difficult to declare significant for genes with a small standard error while ensuring the control of the FDR. Here, the question of primary interest is to what degree the fudge factor can cope with the increasing number of genes with a small standard error. The questions of secondary interest are what are the effect of unequal sample sizes and of distributional assumption on the FDR and power?

#### 7.2.1 Simulation Setting

Similar to the simulations discussed in Section 7.1, 100 data sets, each with 16,998 genes, among which 10% of genes (1700) are truly differentially expressed, are generated. However, in order to study the effect

of the fudge factor on the genes with a small treatment effect and small variance, we simulate four settings with different proportions of these genes: (1) none, (2) 5% (850), (3) 10% (1700), and (4) 20% (3400) of total gene number. The aim of the simulation study is to investigate how the SAM procedure deals with the proportionally increasing number of small-variance genes in terms of controlling the FDR.

The means of treatment groups for genes with a small variance are generated from normal distribution $N(0, 0.3)$. As a result, $E(\mu_1) = E(\mu_2) = E(\mu_3) = E(\mu_4)$, what ensures the homogeneity of expected means for the four treatment groups, but what generates the possibility of small mean differences. The variance of these genes is set equal to 0.0036, which is the 5% quantile of the standard errors observed in the case study data. In total 12 arrays (three arrays per groups) are generated for four treatments.

### 7.2.2 Simulation Results Using Equal Sample Sizes

Figure 9 compares the results for the four settings described above. Figure 9$a$ shows the relationship between the power and the FDR, where the power is estimated as the mean true discovery proportion and the FDR is estimated as the mean false discovery proportion across 100 simulated datasets. The four lines (representing four settings) show the power and the FDR obtained using no fudge factor, and 1%, ..., 100% centile of the standard error distribution as the fudge factor. The result of automatic choice of SAM is indicated using the capital letter (A, B, C, D) for the four settings, respectively. For the last three settings, where the number of non-differentially expressed genes with small variances increases, the FDR is no longer controlled at the desired level. Also, the power of the procedure decreases, because a larger quantile of standard error is used as the fudge factor.

Let us decompose the FDR into two parts, i.e., FDR=FDR$^0$+FDR$^1$, where FDR$^0$ is estimated as the mean proportion of false positives among $m_0^0$ non-differentially expressed genes with moderate to large variances, while FDR$^1$ can be estimated as the mean proportion of false positives among $m_0^1$ non-differentially expressed genes with small variances. From Figure 9$b$, we can see that the FDR$^0$ for the SAM procedure is maintained around 5% regardless of the setting. However, the FDR$^1$ in Figure 9$c$ is not controlled at the desired level, implying that the SAM procedure fails to remove genes with a small variance from the significant comparison list. The problem of not controlling the FDR remains unless a much higher quantile of standard error is used as the fudge factor. However, that results in a great loss of power.

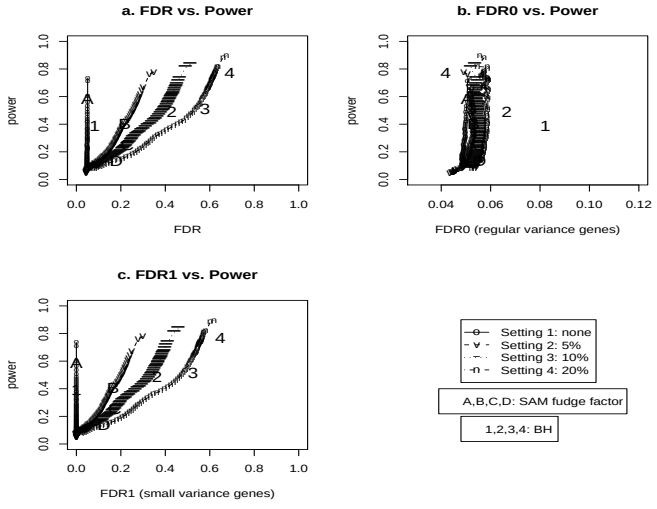


**Fig. 9** Simulation with equal sample sizes. a: Power vs. FDR; b: Power vs. FDR$^0$ (small-variance genes); c: Power vs. FDR$^1$.

On the other hand, the approach based on the permutation $p$-values adjusted by the BH-FDR procedure (indicated by 1, 2, 3, and 4 in Figure 9) seems to yield lower power as compared to the SAM without the fudge factor. From Figure 9c, we can see that the $\text{FDR}^1$ is slightly higher than 0.05 for the first two settings (indicated by 1 and 2), while it is getting close to 0.05 for the last two settings (indicated by 3 and 4). When the proportion of small-variance genes increases, the problem with controlling the $\text{FDR}^1$ becomes more severe. Since the permutation approach adjusted by the BH-FDR procedure is not intended to protect against the small-variance genes, it yields high value of the $\text{FDR}^1$, similarly to the SAM approach without the fudge factor.

In order to examine the variability in the estimated FDR and power, Figure 10 shows the boxplots of the FDR (panel $a$), $\text{FDR}^0$ (panel $b$), and $\text{FDR}^1$ (panel $c$) obtained for the SAM with and without the fudge factor and for permutation $p$-values with the BH-FDR adjustment. Figure 11 shows the boxplots of the power of the three approaches. The same conclusion can be drawn from these figures as from Figure 9, but additionally the distribution of the FDR can be examined. Note that the variability of the FDR and power for the SAM procedure with the fudge factor seems to be larger than for the other two procedures.

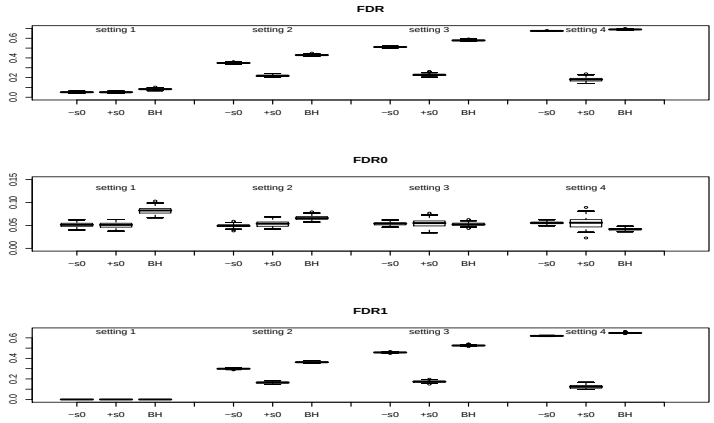


**Fig. 10** Simulation with equal sample sizes. a: Boxplots of the FDR using the SAM without the fudge factor (-s0) and with the fudge factor (+s0), and permutation approach (BH); b: boxplots of the $\text{FDR}^0$ (small-variance genes); c: boxplots of the $\text{FDR}^1$.

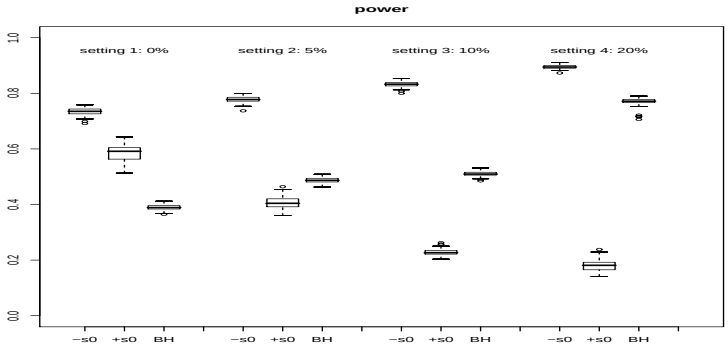


**Fig. 11** Simulation with equal sample sizes. Boxplots of power using the SAM without the fudge factor (-s0) and with the fudge factor (+s0), and permutation approach (BH).

### 7.2.3 Results of Simulation Study Using Unequal Sample Sizes per Group

In order to investigate the influence of unequal sample sizes per group, the simulation study discussed in Section 7.2.2 is repeated with three arrays for the control, and four, five, and six arrays for each of the three treatment groups, respectively.

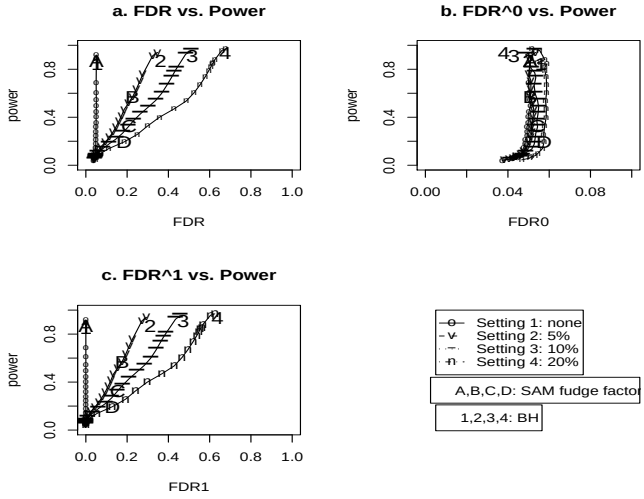


**Fig. 12** Simulation with unequal sample sizes. a: Power vs. FDR; b: Power vs. $\text{FDR}^0$ (small-variance genes); c: Power vs. $\text{FDR}^1$.

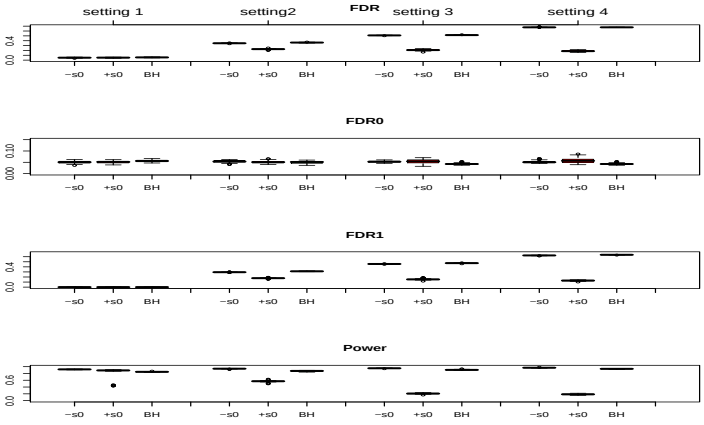


**Fig. 13** Simulation with unequal sample sizes. a: Boxplots of the FDR using the SAM without the fudge factor (-s0) and with the fudge factor (+s0), and permutation approach (BH); b: boxplots of the $\text{FDR}^0$ (small-variance genes); c: boxplots of the $\text{FDR}^1$; d: boxplots of the power.

Figure 12 shows the relationship between the power and the FDR. For the permutation approach, the power of the test increases because the number of arrays increases from 12 (three arrays per group) in the study described in Section 7.2.1 to 18 (3+4+5+6) in the current simulation. Despite that, similar findings can be reported as in the previous simulation study. The overall FDR (in panel $a$) can not be controlled as the proportion of small-variance genes increases. The $\text{FDR}^0$ (in panel $b$) retains its mean value around

0.05 and the FDR$^1$ (in panel $c$) shows the proportion of false positives from non-differentially expressed genes with small variances around 0.15.

Figure 13 shows the boxplots of the FDR, FDR$^0$, FDR$^1$, and power. The conclusions are similar as those in the studies described in the previous sections.

### 7.2.4 Results of Simulation Study Using a Non-Gaussian Distribution

The simulation studies discussed above assume a Gaussian distribution for the expression levels. In order to study the effect of non-Gaussian distribution of microarray data, the Gaussian distribution is replaced by the $t$-distribution with $n-1$ degrees of freedom in the simulation setting, where $n$ is the total number of arrays. We consider both equal sample sizes (i.e., three arrays per groups) and unequal sample sizes (i.e., and three, four, five, and six for the four treatments, respectively).

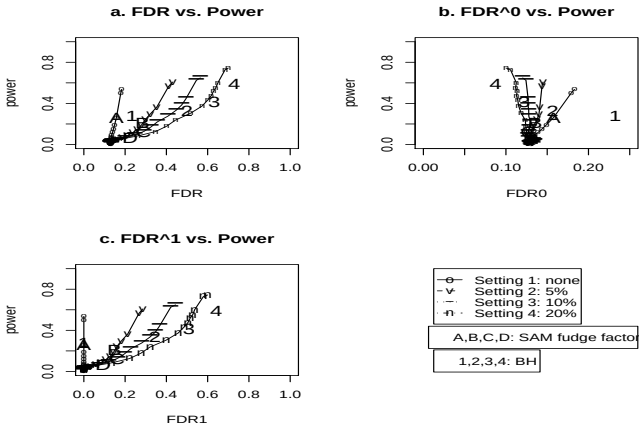


**Fig. 14** Simulation with $t$-distribution and equal sample sizes. a: Power vs. FDR; b: Power vs. FDR$^0$ (small-variance genes); c: Power vs. FDR$^1$.

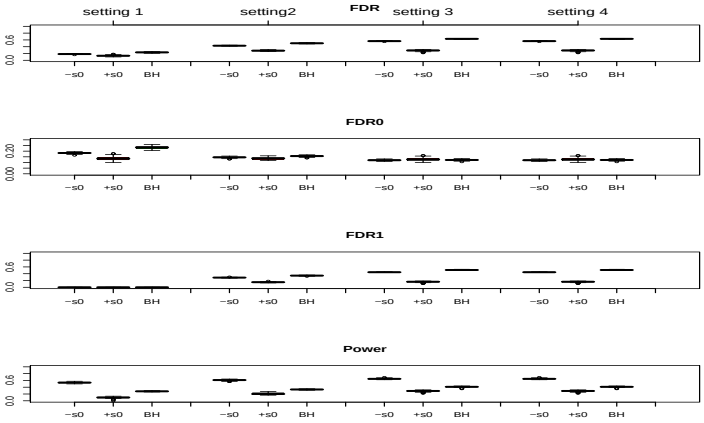


**Fig. 15** Simulation with $t$-distribution and equal sample sizes. a: Boxplots of the FDR using the SAM without the fudge factor (-s0) and with the fudge factor (+s0), and permutation approach (BH); b: boxplots of the FDR$^0$ (small-variance genes); c: boxplots of the FDR$^1$; d: boxplots of the power.

Figure 14 plots the power vs. the FDR (including the $\text{FDR}^0$ and $\text{FDR}^1$). Compared with Figure 9, an increase of the FDR and $\text{FDR}^0$ can be observed. The increase of the overall FDR is due to the increase of $\text{FDR}^0$, which is no longer controlled at 0.05 as in the previous simulation studies. With the simulated $t$-distributed data, the $\text{FDR}^0$ increases and ranges between 0.1 and 0.23 (the largest value obtained for the setting in which there are no small variance genes).

Figure 15 shows the distribution of the estimated FDR, $\text{FDR}^0$, $\text{FDR}^1$ and power, respectively. The variability of the distribution is comparable to that reported in Sections 7.2.2 and 7.2.3.

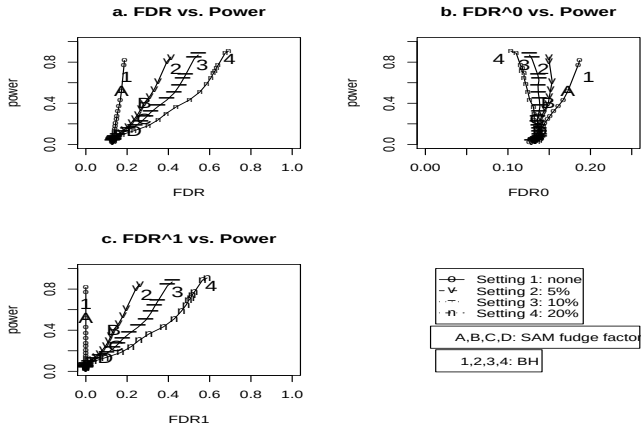


**Fig. 16** Simulation with $t$-distribution and unequal sample sizes. a: Power vs. FDR; b: Power vs. $\text{FDR}^0$ (small-variance genes); c: Power vs. $\text{FDR}^1$.

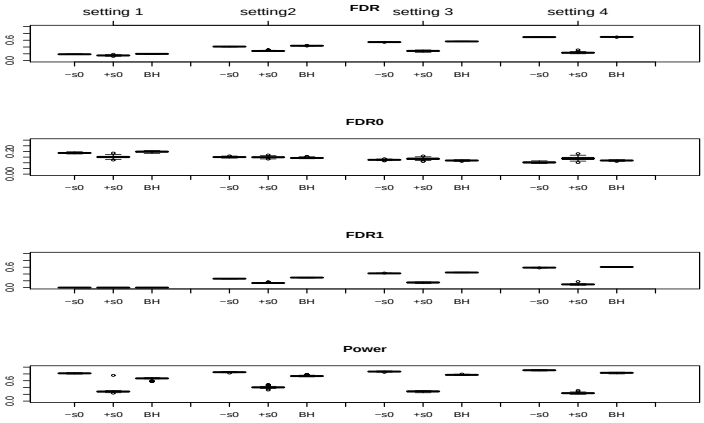


**Fig. 17** Simulation with $t$-distribution and unequal sample sizes. a: Boxplots of the FDR using the SAM without the fudge factor (-s0) and with the fudge factor (+s0), and permutation approach (BH); b: boxplots of the $\text{FDR}^0$ (small-variance genes); c: boxplots of the $\text{FDR}^1$; d: boxplots of the power.

Similar results are observed (Figure 16 and 17) for the setting with the unequal sample sizes (i.e., three, four, five, and six arrays per treatment group).

## 8 Discussion

The aim of the microarray experiment presented in this paper was to find genes whose expression levels differentiated between any of treatments and the control. Such genes were useful as indicators for the active treatment effect. In terms of the multiplicity adjustment, such an experiment required an adjustment for comparisons within a gene (treatment versus control) and an adjustment for testing of thousands of genes. In this paper, we considered four approaches, which addressed the two dimensional testing problem simultaneously. The analysis of the case study revealed substantial differences between the different methods used for the multiplicity adjustment. The SAM procedure with the fudge factor led to the least number of significant findings as compared to the other three procedures, while the SAM without the fudge factor resulted in the largest number of significant discoveries. This difference motivated the simulation studies discussed in Section 7.

The performance of the Dunnett's approach, permutation approach, and the SAM method with and without the fudge factor were compared in the first simulation study (for the case where genes with small variance were not present). We showed that with a small sample size (three arrays per treatment group) the SAM approach without the fudge factor preformed better with respect to power and identification of a larger number of truly significant comparisons between several treatments with the control. When sample size increased, the FDR obtained for the four approaches was well controlled and the power obtained by all approaches was comparable.

The second question of interest was the capability of the SAM procedure to control the FDR for the non-differentially expressed genes with small variances. We showed in the simulation study that the overall FDR cannot be controlled even when the proportion of such genes was relatively small. Moreover, we showed that when the FDR was decomposed to the $\text{FDR}^0$ and $\text{FDR}^1$, there was no problem to control the $\text{FDR}^0$ (regardless of the proportion of non-differentially expressed genes with a small variance), but the $\text{FDR}^1$, associated with the small-variance genes, was not well controlled. When the proportion of the small-variance genes increased, the SAM with the fudge factor was either no longer having the same power as the SAM without the fudge factor or was not controlling the FDR at the desired level. Thus, the automatic selection of the fudge factor did not guarantee the power and the FDR of the SAM procedure at the desired level.

Moreover, we investigated the effect of unequal sample sizes on the FDR and the power of the considered approaches. For the permutation approach, the power of the test increased due to the increased number of simulated arrays (three, four, five and six arrays for four treatment groups, respectively). Similar conclusion can be drawn based on the simulation results.

Finally, we investigated the influence of non-Gaussian distribution of expression levels on the FDR and the power of the considered approaches. The expression levels were generated using the $t$-distribution. We noted that the $\text{FDR}^0$ increased from 0.05 to the range between 0.1 and 0.23 (depending on the setting), while the $\text{FDR}^1$ remained between 0.15 and 0.19, which was comparable in all the simulation studies. This pattern is a topic for further investigation.

Different ways of selecting the fudge factor are discussed by Wu (2005), Broberg (2003), Efron *et al.* (2001), and Efron and Tibshirani (2002). In addition, a mixture model for the variance can be used to detect genes with small variance and to estimate the proportion $m_0^1/m$. A comparison between the methods for selecting the fudge factor is currently under investigation and will be presented in a separate paper.

## Acknowledgement

Financial support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

## Reference


Affymetrix (2003) *GeneChip Expression Analysis Technical Manual, Rev.4.* Santa Clara, CA, available at http://www.affymetrix.com/support/technical/manual/expression_manual.affx

Bechhofer, R.E. and Dunnett, C.W. (1988) Percentage points of multivariate Student t distributions. *In: Selected Tables in Mathematical Statistics, vol. 11. American Mathematical Society, Providence, RI.*

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: Series B* **57**, 289-300.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2002) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185-193.

Broberg, P. (2003) Statistical methods for ranking differentially expressed genes. *Genome Biology* **4**, R41.

Dunnett, C.W. (1955) A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096-1121.

Dunnett, C.W. (1964) New tables for multiple comparisons with a control. *Biometrics* **20**, 482-491.

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

Ge, Y., Dudoit, S., and Speed, P.T. (2003) Resampling based multiple testing for microarray data analysis. *University of Berkeley, technical report 633*.

Hochberg, Y. (1995). A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple Comparison Procedures*, John Wiley & Sons.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hsu, J. C. (1996) *Multiple Comparisons: Theory and methods*, Chapman & Hall.

Huang, Y., Xu, H., Calian, V., and Hsu, J. C. (2006) To permute or not to permute. *Bioinformatics* **22(18)**, 2224–2248.

Hubbell, E., Liu, W.M., and Mei, R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18(12)**, 1585-1592.

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J., Churchill, G.A., (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**, 203-217.

Parmigiani, G., Garrett, E.S., Irizarry, R.A. and Zeger, S.L. (2003) *The analysis of gene expression data: methods and software*, Springer.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19(3)**, 368-375.

Storey, J.D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report* **2001-28,** Department of Statistics, Stanford University.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *The Proceedings of the National Academy of Sciences* **98**, 5116-5121.

Westfall, P.H. and Young, S.S. (1993) *Resampling based multiple testing*, Willy.

Wolfinger, RD., Gibson, G., Wolfinger, ED., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, RS. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8(6)**, 625-637.

Wu, B. (2005) Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics* **21**,1565-1571.

Xu., H and Hsu., J.C. (2007) Applying the generalized partitioning principle to control the generalized familywise error rate. *Biometrical Journal* **49**, 52-67.