False Discovery Rate-Controlled Test Decisions under Correlation in Gene Expression Studies

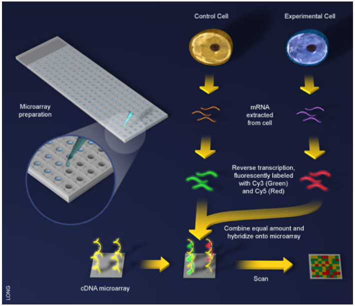
Michael G. Schimek[1,2] Tomáš Pavlík[2]

[1] Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, 8036 Graz, Austria

[2] Masaryk University, Department of Applied Mathematics and Institute of Biostatistics and Analyses, 66295 Brno, Czech Republic

Mathematical Biosciences Institute, Ohio State University,
Statistical Genetics Journal Club Spring 2007

Introduction


Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
cDNA microarray
Scan
LONG

The microarray technology

- It allows to study expression (’activity’) in thousands of genes simultaneously
- We are interested in differences between experimental conditions

Motivation: For many research tasks involving classification and prediction it is necessary to preselect a set of differentially expressed genes

Gain: Preselection helps improving the performance of the classifier or predictor

Task: Selection of a set of differentially expressed genes

Challenges

- High-dimensionality (thousands of genes)
- Small sample sizes (due to limited availability of cases)
- Genes are co-regulated, hence **differential expression can be substantially correlated** (Klebanov et al. 2006)
- Insufficient biological background (pathways etc.)
- **Stability of gene selection** is an issue of increasing importance (Qiu et al. 2006)

When identifying differentially expressed genes via statistical tests we are confronted with a **multiple comparison problem:**

- Using the usual type I error rate α forces the number of false positives to grow enormously
- **Need to control the type I error** $\Rightarrow$ the **false discovery rate** (*FDR*) approach due to Benjamini & Hochberg (1995) is most popular and quite useful

Parametric test statistics for gene *i*

Assume that n genes ($i = 1, \dots, n$) have been measured over two experimental conditions ($j = 1, 2$) on K_1 arrays of condition 1 and K_2 arrays of condition 2 and $K_1 + K_2 = K$

$\bar{x}_{i1}$ and $\bar{x}_{i2}$ mean gene expression for gene i under conditions 1 and 2

Standard test statistic

$$t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i}$$

where s_i the pooled standard deviation for gene i

$$s_i = \sqrt{\left(\frac{1}{K_1} + \frac{1}{K_2}\right) \frac{\sum_{k_1=1}^{K_1}(x_{ik_1} - \bar{x}_{i1})^2 + \sum_{k_2=1}^{K_2}(x_{ik_2} - \bar{x}_{i2})^2}{K - 2}}$$

Modified test statistic

$$d_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i + s_0},$$

where s_0 is a ’correcting’ constant (also called ’fudge factor’)

The correcting constant s_0

Motivation: It should make d_i approximately constant as a function of s_i

Detrimental effect: For a given confidence level the constant s_0 can dramatically affect the number of selected genes

- There is the following empirical evidence (Grant et al., 2005):
  - For $s_0 = 0$ (i.e. standard t_i) the d_i is large for a gene with small variance
  - For $s_0 > 0$ this effect is reduced
  - For s_0 too large, expressed genes with small mean difference and/or small variance are obscured in the overall noise
- The effect of s_0 is unknown for co-regulated genes

When nonparametric alternatives are used (e.g. a rank-sum statistic) no s_0 specification needed, the results however are less powerful (Schimek and Pavlik, 2006)

Medizinische Universität Graz

What is the false discovery rate (*FDR*)?

Goal

Identify as many differentially expressed genes as possible while incurring a relative low proportion of false positives

Let V be the number of false positives and R be the number of overall rejected hypotheses in a microarray experiment
The *FDR* can be defined as

- expectation of the ratio of V and R (have to account for possibility of $R = 0$)

$$FDR = \mathbf{E}\left(\frac{V}{R}1_{\{R>0\}}\right).$$

However, it can be shown (Storey & Tibshirani, 2003) that $FDR = \mathbf{E}\left(\frac{V}{R}\right) \approx \frac{\mathbf{E}(V)}{\mathbf{E}(R)}$, which is easier to estimate and implement

Medizinische Universität Graz

FDR estimation procedures: `samr`

R version of classical SAM procedure (Tusher et al., 2001)
Let $t_{(1)} \leq t_{(2)} \ldots \leq t_{(g)}$ be the ordered observed test statistics
The expected value for ith rank $\bar{t}_{(i)}$ is estimated via the set of B permutations of the data matrix
Then (Δ arbitrary but fixed) genes satisfying

$$t_{(i)} - \bar{t}_{(i)} \geq \Delta \text{ or } \bar{t}_{(i)} - t_{(i)} \geq \Delta$$

are called **'significant'**

$$\widehat{FDR} = \hat{\pi}_0 \frac{median\ number\ of\ falsely\ called\ genes}{total\ number\ of\ genes\ called},$$

where $\hat{\pi}_0 = \frac{\#\{t_i \in (q25, q75)\}}{g/2}$ is the estimated proportion of truly null hypotheses

Disadvantage:

- **High memory requirements** due to the storage of intermediate results

FDR estimation procedures: `siggenes`

A variant of the SAM procedure implemented in R (Schwender, Krause and Ickstadt, 2003)
Major difference to original SAM: The estimation of the proportion of truly null hypotheses is based on spline smoothing (idea due to Storey & Tibshirani, 2003)

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, g\}}{g(1-\lambda)}, \ \lambda = 0.01, 0.02, \ldots, 0.95$$

The final estimate of π_0 is set to $\hat{\pi}_0 = \hat{f}_{\lambda=1}, \hat{f}$ being a natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ
Advantages:

- **Feasible to use larger number of permutations** without memory allocation problems
- The user can either decide for the **median or the mean** value of falsely significant genes obtained from the set of *B* permutation steps when estimating the *FDR*

FDR estimation procedures: Grant's procedure

- **Idea is to estimate the *FDR* for an adequate set of values that covers the range of observed test statistics, finally picking the value which satisfies the pre-specified α level**

Let k be an arbitrary but fixed value, G_k the set of genes i such that $t_i \geq k$, R_k be the size of G_k, and V_k be the number of truly null genes in G_k

$\mathbf{E}(R_k)$ we then estimate with R_k, $\mathbf{E}(V_k)$ is estimated using the set $\{V_k^1, V_k^2, \ldots, V_k^B\}$ for a fixed k

Taking $\hat{\mu}_k = 1/B \sum_{i=1}^{B} V_k^i$ for $\mathbf{E}(V_k)$ would lead to overestimation; solution due to Grant et al. (2005): **iterative algorithm**

$$\hat{\mu}_k(1) = \frac{\hat{\mu}_k}{g}[g-(R_k-\hat{\mu}_k)] \quad \ldots \quad \hat{\mu}_k(i+1) = \frac{\hat{\mu}_k(1)}{g}[g-(R_k-\hat{\mu}_k(i))]$$

As the final estimate of $\mathbf{E}(V_k)$ we are using $\hat{\mu}_k(n)$, where $\hat{\mu}_k(n) - \hat{\mu}_k(n-1) < 0.0001$

Comparison of the FDR estimation procedures

Procedure	**Grant's**	**siggenes**	**samr**
Estimated formula	$\mathbf{E}(V)/\mathbf{E}(R)$	$\mathbf{E}(V)/\mathbf{E}(R)$	$\mathbf{E}(V)/\mathbf{E}(R)$
Principle of V dist. estimation	permutations	permutations	permutations
Type of test statistic	t or modified t	t or modified t	t or modified t
Automatic s_0 calculation	no	yes	yes
Proportion of truly null genes	not available	available	available
Statistic for falsely called genes	mean	mean / median	median

Questions of interest

- Are there differences in the obtained results (sets of selected genes)?
- Are there differences with respect to power and bias?
- Are there differences in computational costs?
- Can these permutation-based procedures cope well with correlated expression values?

Simulation study outline

We evaluated the procedures for the **ordinary** and for the **modified SAM *t*-statistic** with the following values of s_0 ('fudge factor'):

- 0, 0.5, 1, and 5, and $\hat{s}_0$ provided by `siggenes` and `samr`

Power, **bias** and **stability of the number of correctly identified genes** were studied for fixed *FDR* levels of $\alpha = 0.05$ and 0.1

We adopted the following setting:

- Grant's procedure with 10 000 permutation steps
- `siggenes` procedure applying the *mean* with 3 000 permutation steps
- `siggenes` procedure applying the *median* with 3 000 permutation steps
- `samr` procedure with 3 000 permutation steps

Simulation study outline continued

For the purpose of comparison an **empirical Bayes thresholding** (abb. EBT) procedure (no *FDR* control) was used (Johnstone and Silverman, 2004)

- Random thresholding assuming sparse signals (differential expression)
- Prior for each test statistic is mixture of an atom of probability at zero and a double exponential (heavy-tailed) probability
- Minimax squared error properties, hence related to *FDR*

Common features of artificial expression data

Sample size $n = 3000$ genes

Unexpressed genes: simulated from N(0,1)

Expressed genes:

- 100 up-regulated
- 200 down-regulated

in groups of 25 resp. 50

Simulation study outline continued

Correlated data generated from $x_{ij} = \sqrt{\rho} * a_j + \sqrt{(1-\rho)} * y_{ij}$,
where $i = 1, \ldots, 300$, $j = 1, \ldots, 25$,
$\rho = 0.4$ the assumed correlation,
a a random vector for each group,
and y the original vector of simulated values

Model C1 'simple correlated'

- up-regulated from N(2,1)
- down-regulated from N(-2,1)

Model C2 'complex correlated'

- up-regulated from N(1,1), N(1,2), N(2,1), N(2,2) (25 genes each)
- down-regulated from N(-1,1), N(-1,2), N(-2,1), N(-2,2) (50 genes each)

Simulation study outline continued

Model U1 'simple uncorrelated'

- up-regulated from N($\sqrt{0.4} * 2$,1)
- down-regulated from N($\sqrt{0.4} * (-2)$,1)

Model U2 'complex uncorrelated'

- up-regulated from N($\sqrt{0.4}, 1$), N($\sqrt{0.4}, 2$), N($\sqrt{0.4} * 2, 1$), and N($\sqrt{0.4} * 2, 2$) (25 genes each)
- down-regulated from N($-\sqrt{0.4}, 1$), N($-\sqrt{0.4}, 2$), N($\sqrt{0.4} * (-2), 1$), and N($\sqrt{0.4} * (-2), 2$) (50 genes each)

Note that the mean is shifted for comparability with the correlated models

For each setting the sampling was **replicated 10 times**

Selected simulation results: Fudge factor

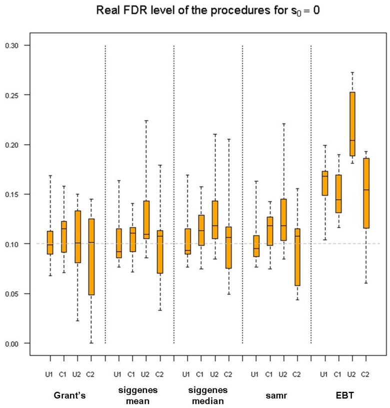


Figure 1a

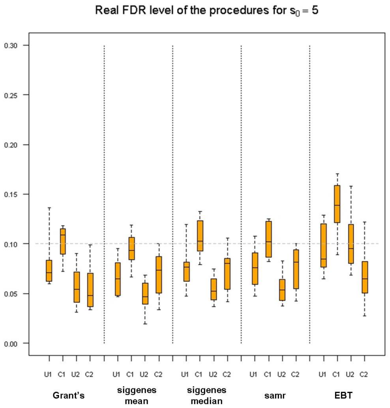


Figure 1b

Medizinische Universität Graz

Selected simulation results: Fudge factor

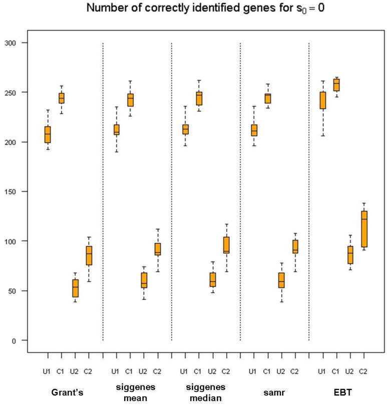


Figure 2a

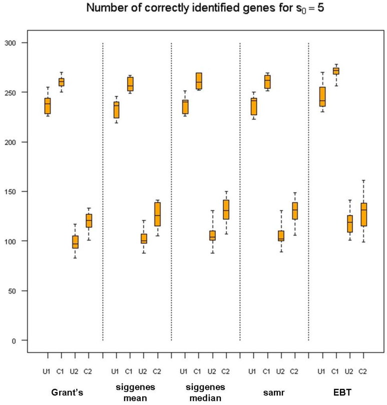


Figure 2b

Selected simulation results: Real *FDR* level for $\alpha = 0.1$

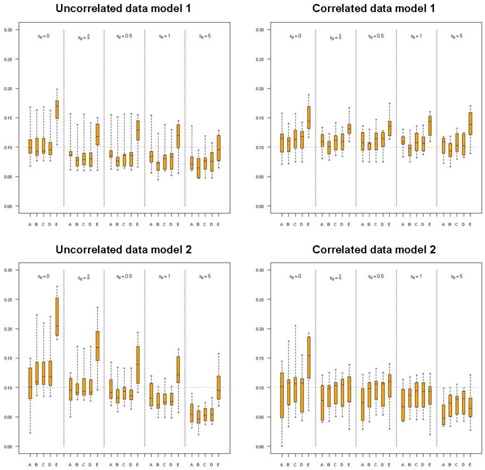


Figure 3

Procedure labels '**A**': Grant, Liu and Stoeckert (2005), '**B**': `siggenes` with *mean*,

'**C**': `siggenes` with *median*, '**D**': `samr`, '**E**': EBT

Selec. sim. results: Correctly picked genes for $\alpha = 0.1$

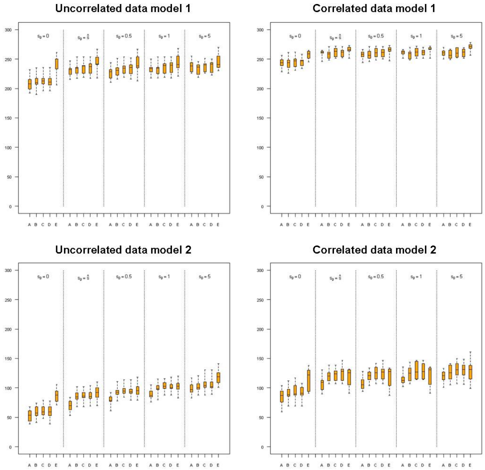


Figure 4

Procedure labels '**A**': Grant, Liu and Stoeckert (2005), '**B**': `siggenes` with *mean*,

'**C**': `siggenes` with *median*, '**D**': `samr`, '**E**': EBT

Medizinische Universität Graz

Selected simulation results: Model U1

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the UNCORRELATED DATA MODEL 1

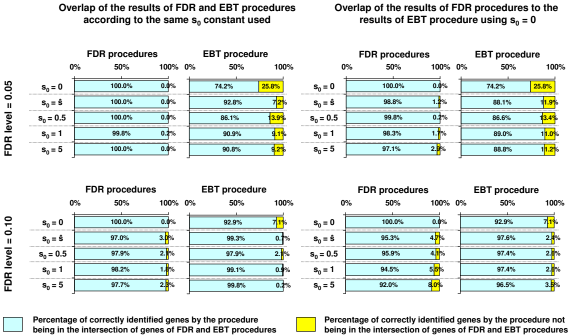
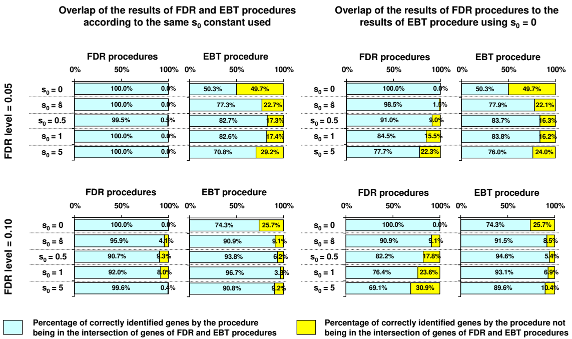


Figure 5

Selected simulation results: Model U2

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the UNCORRELATED DATA MODEL 2



Overlap of the results of FDR and EBT procedures according to the same s_0 constant used

Overlap of the results of FDR procedures to the results of EBT procedure using $s_0 = 0$

□ Percentage of correctly identified genes by the procedure being in the intersection of genes of FDR and EBT procedures

■ Percentage of correctly identified genes by the procedure not being in the intersection of genes of FDR and EBT procedures

Figure 6

Medizinische Universität Graz

Selected simulation results: Model C1

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the CORRELATED DATA MODEL 1

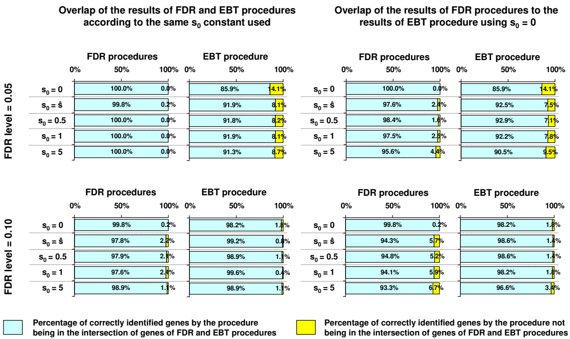


Figure 7

Medizinische Universität Graz

Selected simulation results: Model C2

Distribution of the correctly identified genes with respect to the overlap of the FDR procedures and EBT procedure evaluated for FDR levels 0.05 and 0.10 in the CORRELATED DATA MODEL 2

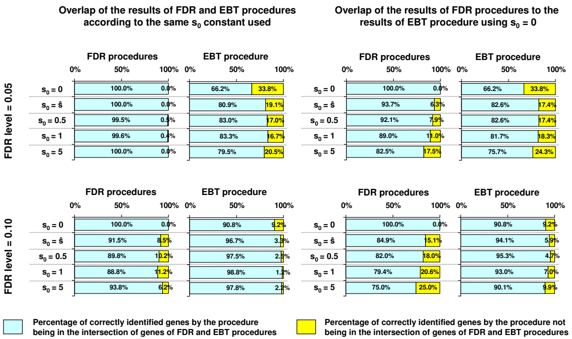


Figure 8

Summary of results and conclusions

- The behaviour of the **SAM procedures** with respect to power and bias is **quite uniform**
- An **adequate choice of the correcting constant** s_0 can **improve** the gene selection process, at least for the simple data models
- The automatic SAM choice of s_0 **can be far from optimal**
- The complexity of the data is definitely more relevant than the presence of correlation
- **Empirical Bayes thresholding tends to outperform the SAM procedures** at the cost of too large real *FDR* levels
- The behaviour of empirical Bayes thresholding can be further improved (bias reduction) for $s_0 > 0$

Summary of results and conclusions continued

- **Grant's procedure requirers substantially more permutation steps** compared to the other techniques and cannot be recommended
- The permutation-free **empirical Bayes thresholding** procedure is by far the **most efficient one** (recommended for huge data sets and screening purposes)
- The **original SAM procedure performs reasonably well for the simple data model**, even under correlation, but **not for the complex data model**
- The number of correctly identified genes interacts with the type of procedure and the specified *FDR* level ($\alpha = 0.1$ recommended, EBT approximates this value)

Medizinische Universität Graz