MULTIPLE HYPOTHESES TESTING IN SMALL MICROARRAY EXPERIMENTS

by

Meaza Demissie

Licentiate Thesis
Department of Statistics
Örebro university 2008

Abstract

The approach of controlling the false discovery rate (FDR) which is less conservative than controlling the family wise error rate (FWER) has become a new and popular multiple testing procedure in microarray studies. In addition to multiple testing problems the limitation of sample size is another big challenge in this area. This thesis addresses the problem of multiple testing in the context of FDR in small samples. The first paper makes a comparison of three FDR estimation methods and the second paper investigates the effects of unequal variance between groups on the estimation of FDR in small samples using standard methods. This paper also presents a new FDR estimation procedure that deals with the above problem. Simulation results show that the new procedure works well for unequal group variance in small samples and is reliable over wider set of conditions.

Keywords: FDR, FWER, error measure, microarray, small samples, unequal variance.

The thesis consists of this summary and the following papers and supplementary reports:

- Demissie, M. (2008). A comparative review of estimates of FDR in small microarray experiments, Working Paper 2008:?, Department of Statistics, Örebro university.
- Demissie, M, Mascialino, B, Calza, S. and Y. Pawitan (2008). Unequal group variances in microarray data analyses, *Bioinformatics* **24**, 1168-1174.
- Demissie, M, Mascialino, B, Calza, S. and Y. Pawitan (2008). Unequal group variances in microarray data analyses - Supplementary report I, *Bioinformatics* **24**.
- Demissie, M, Mascialino, B, Calza, S. and Y. Pawitan (2008). Unequal group variances in microarray data analyses - Supplementary report II, *Bioinformatics* **24**.

Acknowledgments

First of all I would like to express my deepest thank and respect to the Almighty God for granting me his sustained and loving care, strength, patience and hope to pass through all the challenges during the course of this long and testing work.

I owe my sincere gratitude to my supervisors Prof. Thomas Latila and Prof. Yudi Pawitan without them the accomplishment of this work would not have been possible. Prof. Laitila edited my thesis and paper one and help my summary have a good layout and format. I am grateful to him for giving me all the encouragement and hope to accomplish my work. Thank you Prof. Laitila for all the kind support I got from you.

I am deeply indebted to Prof. Pawitan who has been supervising my work throughout the thesis work. His invaluable scientific guidance, constructive comments and stimulating faith have made my education successful. Thank you Prof. Pawitan for all the support I got from you during my stay at Karoliniska Institute (KI) and for being a strong and challenging supervisor. I have learnt a lot from you.

I am grateful to Prof. Sune Karlsson and Prof. Elizabeth Svensson for their encouragement and constructive advice and their positive wish for my success. My thanks also goes to my colleagues and administrative staff at ESI department of Örebro University for their support. I wish to thank Yen and staff of KI for there valuable support during my stay there.

I wish to express my special thanks to Prof. Dietrich von Rosen for providing me an office and necessary materials so that I would stay near by my family. Thank you Prof. von Rosen for the help I got from you during my studies and for being kind and considerate person to me.

My heart felt and sincere thanks go to my mother, W/ro Yeshareg Kidane, my sisters: Alem and Aster and my brothers: Desalegn and Theodros and my friends Maria, Elizabeth, Tigist, Meseret, Selamawit, Tsegaye and all other friends at EIAR and outside EIAR who encouraged and motivated me to complete my research work.

Last but not least I would like to thank my partner Dr. Haileselassie Yibrah and my lovely son Abel and my lovely daughter Danayit for their endurance to share my life as a student and for their best wishes to my success. I want to thank them from the bottom of my heart as the most important support came from them. In fact, this work is dedicated to my mother and them.

Contents

1 Introduction

A major goal of statistics is to make definite decisions with respect to hypotheses about the probability laws of random variables. Hypothesis testing in statistics deals with problems of uncertainties and amounts to sampling the random variable whose probability law is referred to in the hypothesis and, on the basis of the sample, deciding to accept or reject the stated hypothesis.

Hypothesis testing can involve the testing of a single hypothesis or the simultaneous testing of multiple hypotheses. The decision to reject or not reject a hypothesis is characterised by uncertainty and so, as stated in the statistical literature (e.g. Lehmman, 1986), a standard approach of single hypothesis tesing is to specify an acceptable type I error rate (false positive, rejection of a true hypothesis) and base the conclusion by finding tests which minimize the type II error (false negative, non-rejction of a false hypothesis) and thereby maximize the power of the test.

Multiple hypothesis testing involves the testing of two or more single hypotheses at the same time. When more than one hypothesis is tested, the appropriate threshold to declare a test statistic's p value significant becomes complex. Each test has a specified type I error and the chance of drawing at least one false conclusion increases rapidly with the number of tests performed. As stated by Dudiot et al., (2002), a p-value of 0.01 for a single hypothesis among a list of several hypotheses would no longer correspond to a significant finding as there is a high probability of getting a small p-value by chance when considering a large set of hypotheses. This could lead to serious consequences if the set of conclusions must be evaluated as a whole (Shaffer, 1995). This problem is commonly referred to as the multiple testing problem.

The basic concern of the theory of multiple hypothesis testing is to control the type I error by developing methods that would account or adjust for the multiplicity effect. The traditional way of multiple testing procedures has been to control the probability of committing even one type I error, the family wise error rate (FWER) (Hochberg and Tamahane, 1987; Westfall and Young, 1993). A commonly known and used method is the Bonferroni procedure. Other modifications and more powerful procedures of multiple testing problems (e.g. step down procedure of Holm, 1979) are currently available. However, these methods are very conservative and their power is greatly reduced as the number of hypotheses in the family increases leading practitioners to neglect multiplicity control even in medium size problems (Benjamini and Yekutieli, 2001). On the other hand, large-scale simultaneous hypothesis testing problems, with hundreds or thousands of cases considered together with data-sets generated by modern technologies such as light spectroscopy, proteomic devices, flow cytometry, functional resonance imaging, many social science surveys and microarray have become a fact of current-day statistical practice (Efron, 2008).

The multiplicity problem gains a much larger dimension in the case of microarray data, usually several thousands of cases are considered simultaneously. Microarray technology enables researchers to measure the expression levels of thousands of genes in a single experiment (Abul et al., 2004). Microarrays have increasingly been used to address a wide range

of problems in biological and medical research such as classification of tumors, monitoring response to different environmental stress conditions and study of host genomic response to bacterial infections (Alon *et al.*, 1999; Golub *et al.*, 1999; Perou *et al.*, 1999; Alizadeh *et al.*, 2000; Gash *et al.*, 2000; Ross *et al.*, 2000).

This thesis is mainly concerned with the first step in the statistical analysis of gene expression data, the search of differentially expressed genes across two or more kinds of tissue samples (e.g. tissues like normal and cancerous cervical) or samples obtained under different experimental conditions (Dudoit et al., 2003). Testing of the null hypothesis of no differential expression for each gene in microarray study requires inference on several thousands of null hypotheses simultaneously. In light of the large-scale inference problem on microarray data, Benjamini and Hochberg offer another measure of the likelihood of falsely rejecting a number of true null hypotheses, the False discovery rate (FDR).

A number of recent articles have addressed the problem of multiple testing in the context of FDR (e.g. Benjamini and Hochberg, 1995; 2000; Keselman et al, 2002; Pawitan et al, 2005a, 2005b; Ploner et al., 2006; storey 2002; Storey and Tibishirani, 2003). However in addition to the multiple testing problem the limitation of replicated samples is another big challenge in microarray data analyses which leads to problems in estimation of the FDR.

The objective of this thesis is to compare three FDR estimation methods and develop a new FDR estimation procedure which is suitable for unequal group variances in small sample microarray experiments. Paper 1 in the thesis deals with the comparison of FDR estimation methods and paper 2 deals with the development of the new FDR estimation procedure.

In this summary, Section 2 presents some basic notions and ideas behind testing hypotheses which gives the necessary background of multiple hypotheses testing in microarray studies. In Section 3 introduction and description of FDR is summarized. Section 4 presents a brief summary of the two papers included in this thesis. In section 5 a direction to future research is presented.

2 Multiple hypothesis testing

2.1 Single hypothesis testing

An important aspect of statistical inference is to specify and test hypothesis. Suppose a set of data generated from some distribution F_θ belonging to a family of distributions indexed by $\theta \in H$ is given. The null hypothesis (H_0), which is some subset of H, is to be tested against an alternative hypothesis (H_1), which is another subset of H. The null hypothesis H_0 is tested based on a statistic T. The statistic T is a function of the data which is used to decide whether $\theta \in H_0$ or $\theta \in H_1$.

In hypothesis testing the sample space of the test statistic T is partitioned into two disjoint parts, called rejection region (critical region) and acceptance region, which are denoted by Υ and Γ respectively. If $T \in \Upsilon$ then H_0 is rejected.

There are two types of errors which might be committed when testing hypothesis. We might reject the hypothesis when it is true or we might accept the hypotheses when it is

false. If we reject the hypotheses when it is true i.e. if T $\in \Upsilon$ when $\theta \in H_0$, we commit a type I error. If we accept the hypotheses when it is false i.e. if T $\in \Gamma$ when $\theta \in H_1$ we commit a type II error.

The probability of making a type I error is usually controlled at some designated level alpha. That is, the researcher controls the probability of rejecting the null hypothesis H_0 when it is in fact true. Correspondingly, the probability of rejecting H_0 when it is false defines the power of the test. The power of a hypothesis test equals the probability of not committing a type II error, that is:
Power = 1 - P(type II error)

2.2 Multiple hypotheses testing

In multiple hypothesis testing one or more single hypotheses are tested simultaneously. The possible outcomes of testing m null hypotheses simultaneously can be summarized by Table 1 below (Benjamini and Hochberg, 1995). In the table, V denotes the number of type I errors (false positives), T denotes the number of type II errors (false negatives) and R denotes the number of rejected hypotheses. In the setting each of the m single hypothesis has possible type I and type II errors

Table 1. Summary of type I and type II error rates of m hypotheses

	Accept	Reject	Total
True null hypotheses	U	V	m0
True alternative hypotheses	T	S	m1
Total	m-R	R	m

In order to measure the errors incurred in multiple hypotheses testing, a variety of generalizations of type I error rates of single hypothesis testing have been proposed. The following are the major ones (Shaffer, 1995).

- Per-comparison error rate (PCER), defined as the expected value of the ratio of the number of false rejections to the number of hypotheses, i.e. PCER=E(V)/m,
- Per-family error rate (PFER), defined as the expected number of false rejections, i.e., PFER=E(V)
- Family-wise error rate (FWER), defined as the probability of at least one type I error, i.e., FWER=pr(V$\geq$1)
- False discovery rate (FDR), defined as the expected proportion of type I errors among the rejected hypotheses (Benjamini and Hochberg ,1995), i.e, FDR=E(V/R) where V/R=0, if R=0

The p-value in single hypothesis testing is the level of the test at which the hypothesis H_0 would just be rejected. An adjusted p-value in multiple hypothesis testing corresponding to the test of single hypothesis is the level of the entire test procedure at which H_i (i=1,...,m) would just be rejected, given the values of all test statistics involved (Shaffer, 1995; Yang and Speed, 2003; Westfall and Young, 1993; Yekutiel and Benjamini, 1999). There are three ways of adjusting p-values: the single step, the step down and and the step up procedures. The single step procedure is a way of adjusting p-values where equivalent multiplicity procedures are performed for all hypotheses, regardless of the ordering of the test statistics or unadjusted p-values, each hypothesis has its own critical value which is not associated to the results of tests of other hypotheses. In a step-down procedure multiplicity adjustments are performed by ordering the unadjusted p-values starting with the most significant. This procedure is less conservative than single-step procedure and hence improves power. In a step-up procedure multiplicity adjustments are performed by ordering the unadjusted p-values starting with the least significant. This procedure is less conservative than the step-down procedure under the assumption of independence of tests.

Many types of multiple testing procedures are available and are to be discussed later. One important question is which testing procedure to choose? The following are some common criteria (Yang and Speed, 2003).

- relevance to the subject under investigation
- type of control: strong or weak; strong control is control of the type I error rate under any combination of true and false hypotheses and weak control is control of the type I error rate when all the null hypotheses are true (Dudoit et al., 2003)
- validity of assumptions
- computability: refers to whether there is numerical or simulation error

The Bonferronni procedure strongly controls the FWER at level α by rejecting any hypothesis H_i with p-value less than or equal to α/m. The corresponding Bonferroni single step adjusted p-values are given by $\tilde{p_i}$=min(mp_i,1). Another procedure which is less conservative than the Bonferronni procedure is the Holm (1979) step down procedure. In this, let the ordered unadjusted p-values be denoted by $p_{r1} \leq p_{r2} \leq ... \leq p_{rm}$ and let $H_{r1}, H_{r2}, ..., H_{rm}$ denote the corresponding null hypotheses. Then the Holm step down adjusted p-values are given by

$$\tilde{p_i} = \max_{k=1,...,i} \{\min((m-k+1)p_{rk}, 1)\}$$

An alternative method which accounts for the dependence structure among the test statistics and which is less conservative than the Holm procedure was proposed by Westfall and Young (1993). Shaffer (1995) and Dudoit *et al.*(2003) have provided an in-depth review of many of these methods.

In the next section, hypothesis testing in the context of a new scientific problem, microarray studies will be presented.

2.3 Multiple hypotheses testing in microarray studies

The approach of controlling for the family-wise type I errors, such as the Bonferonni adjustment, is often too conservative to be useful when large-scale simultaneous testing problems with many thousands of cases are considered together such as in microarray technology. The microarray is a device which enables the measurement of gene expression on a large scale basis. Microarray technologies have extensively been used over the past few years and have the potential to address a wide range of problems in biomedical research such as classification of tumors or the study of host genomic responses to bacterial infections (Boldrick *et al.*, 2002; Golub *et al.*, 1999; Pollak *et al.*, 1999). However there still exists a challenge; how to analyze and interpret large scale data. A common problem is to detect genes with differential expressions under two experimental conditions, which may refer to samples drawn from two types of tissues, tumors or cell lines.

Consider the common set up for a microarray experiment, which produces enormous amounts of data with expression data on m genes (variables) for n samples. Let the gene expression levels be put in an array form as m $\times$ n matrix as Y = (y_{ij}) where rows correspond to gene i =1 to m and columns correspond to individual microarray experiments or samples j, where j=1 to n. In the setting above the gene expression levels, y, are continuous variables and usually this expression is associated with a response or covariate of interest which is recorded for each sample. Suppose this covariate or response be x and the associated random variable X . Then depending on the type of experiment x could be dichotomous, polytomous or continuous. Here we consider x as dichotomous. Let the multiple testing problem to test whether a gene is differentially expressed or not be statistically stated as follows

$$H_i\text{: There is no association between } Y_i \text{ and X}$$

where Y_i is the random variable associated to the expression level for gene i and X is the response variable. Then, as discussed by many authors (e.g. Dudiot *et al.*, 2002; Efron *et al.*, 2000; Golub *et al.*, 1999; Zhao and Pan, 2002; Ploner et al., 2005), the general procedure for handling the biological problem of the question of differential expression as indicated above has two steps. First, computing an appropriate univariate test statistics T for each gene under the set up of the experimental design and then conducting a multiple testing procedure to determine which hypotheses to reject while controlling a suitably defined type I error rate. The first procedure to test the hypothesis that gene i is not differentially expressed is a question of univariate hypothesis testing which has been reviewed in section 1.1 and thoroughly studied and discussed in the statistical literature (e.g. Lehmman, 1986).

As has been discussed in section 1.1 in each of these tests two types of errors can be committed: a false positive (type I) error and a false negative (type II) error. With many thousands of genes measured simultaneously in microarray experiment, we have to infer on many thousands null hypotheses simultaneously. This situation leads to multiple testing problem. When more than one test is performed, the appropriate threshold to declare a test statistic's p value significance becomes complex . Each test has a chance of α to yield a significant result, and the chance of drawing at least one false conclusion increases

rapidly with the number of tests performed. This leads to a question on how to determine the true effect or on how to find for those of the many thousands of p-value that provide enough evidence for a significant change in gene expression. The answer lies in defining an appropriate compound error rate which is relevant to the subject under study, in this case microarray data analyses.

3 False Discovery Rates

A classical and common approach in simultaneous testing has been to construct a procedure that controls the FWER, the probability of committing at least one type-I error within the tested family of hypotheses (Tukey, 1953; Hochberg and Tamahane, 1987). The main problem with such classical procedures is that they tend to have low power. A different and new approach to multiple testing, the false discovery rate (FDR)was proposed by Benjamini and Hochberg (1995). The FDR is the expected proportion of wrongly rejected null hypotheses among the total number of null hypotheses rejected. Following the notation from table 1, FDR is defined as E(Q) , where Q =V/R if R>0 and 0 if R=0. When FDR is equal to the FWER under the complete null, the procedures that control the FDR also control the FWER in the weak sense (Dudoit *et al.*,2003). When some of the tested hypotheses are in fact false, FDR control is less strict than FWER control and thus FDR controlling procedures are potentially more powerful. The approach of controlling false discovery rates which is less conservative than the approach of controlling family wise error rate, has gained popularity in microarray data analyses. However, some applications such as analysis of clinical trials still require FWER control.

Under the assumption of independence of the test statistics, the following linear step-up procedure of Benjamini and Hochberg (1995) leads to strong control of the FDR at level α. Let $p_{r1} \leq p_{r2} \leq ... \leq p_{rm}$ be the observed ordered unadjusted p-values. Calculate $\hat{i} = \max\{i : p_{ri} \leq (i/m)\alpha\}$ then reject the null hypotheses corresponding to p_{r1},...,$p_{r\hat{i}}$. It was shown by Benjamini and Yekuteli (2001) that the above procedure controls the FDR also under special forms of dependence structure such as positive regression dependence. These authors also proposed a conservative modified procedure of controlling FDR which works for general dependence structure.

4 Summary of papers

4.1 Paper I: A comparative review of estimates of FDR in small microarray experiments

This paper shows the comparison of three FDR estimation methods in microarray studies with the commonly posed question of identifying genes that are differentially expressed across two or more types of samples obtained under different experimental conditions. The observed

test statistics has the following form of mixture model

$$F(z) = \pi_0 F_0(z) + (1 - \pi_0) F_1(z)$$

where F is the distribution of the observed t- statistic Z, to test whether a gene is differentially expressed (DE) or not between two groups, π_0 is proportion of truly non-DE genes, F_0 is the distribution of the statistics of non-DE genes and F_1 is the distribution of the statistics of DE genes .

The three FDR estimation methods considered are

1. Direct FDR estimation using p-values (FDR.p) which is given as (Pawitan et al. 2005)

$$\widehat{FDR}(k) = \frac{\hat{\pi}_0 m p_{(k)}}{k}$$

   where $(p_{(k)})$ is the k'th highest p-value; monotonicity is imposed by applying a cumulative minimum over k,...,m.

2. FDR estimation using average of local false discovery rate (FDR.avg) is computed as

$$FDR(z) = E\{fdr(Z) \mid Z \leq z\}$$

   where fdr is local false discovery rate $fdr(z) = \frac{\pi_0 f_0(z)}{f(z)}$ which is computed using Bayes formula from the mixture model (5) of Ploner et al.(2006)

3. FDR estimation using average of multidimensional false discovery rate (FDR.avg.fdr2d). The global FDR can also be found by taking average of the ordered fdr2d values as (Ploner et al. 2006)

$$\text{FDR(R)=E}\{\ fdr(z_1, z_2) \mid R\}$$

   where R represents the rejection region of the two dimensional statistics (z_1, z_2) such that $Z \varepsilon R$ are called DE

Simulation results show that when the proportion of non-differentially expressed genes (p0) in the mixture frame work (5) is set at 0.9, FDR.p and FDR.avg are accurate for large sample size such as n=10 up to n=20 per group. However when p0 is increased to 0.95 and decreased to 0.8 both methods overestimate and underestimate the true FDR respectively, indicating the effect of p0 in estimation of FDR. The comparison of FDR.p and FDR.avg to FDR.avg.fdr2d based on a simulation plot of FDR of the first 10 percent top differentially expressed genes indicates that FDR.avg.fdr2d performs better than the other two methods when one uses reasonably large sample size and sample size as small as 5 subjects per group. Analysis of a real data set also confirm that FDR.avg.fdr2d performs better than the other two methods when sample size as small as 5 subjects per group is used. However, none of the methods performs well when the sample size is reduced down to 3 subjects per group. All methods turn out to be very far from the true FDR. This implies that there is a need to develop a new FDR estimation procedure which is suitable for very small sample sizes.

4.2 Paper II: Unequal group variances in microarray data analyses: problem and solution

The purpose of this paper is to investigate the effects of unequal variance on the estimation of false discovery rate using standard methods and to present a new procedure that deals with this problem. The test being proposed is a moderated Welch statistics and its form is motivated by the moderated t statistics proposed by Smyth, (2004)

$$W_m = \frac{\bar{y_1} - \bar{y_2}}{se_m}$$

where se_m is the moderated standard error

$$se_m^2 = \frac{d_0 s_0^2 + d_w se_w^2}{d_0 + d_w}$$

and d_0 and s_0^2 are hyper-parameters to be estimated from the data.

Here, se_w^2 is calculated as a weighted average of pooled and unpooled standard errors as

$$se_w^2 = wse_p^2 + (1-w)se_u^2$$

and its distribution is approximated by a scaled χ^2 distribution with degrees of freedom given by $d_w = wd + (1-w)df$, where d and df are the degrees of freedom of the pooled and unpooled standard errors respectively.

The false discovery rate for the k top genes is then estimated by

$$\widehat{FDR}(k) = m\pi_0 p_{(k)}/k$$

where p_k is the kth highest p-value. The proportion π_0 of non-DE genes is estimated by (Storey and Tibishirani, 2003)

$$\hat{\pi_0} = \frac{Numberofp - values > \lambda}{m(1-\lambda)}$$

Properties of the estimator along with other standard methods are studied using simulation and exemplified with real data.

Simulation results show that when there is unequal group variance, standard procedures like ordinary t and moderated (t_m) give biased FDR estimates. The standard statistical test that deals with unequal-variance-the Welch test-lacks power in small samples. Whereas, the moderated form of the Welch test (MWT) performs well for unequal-group variance problem in small samples. When group variances are the same t_m performs similar to MWT. In conclusion, results indicate the reliability of MWT over a wider set of conditions.

5 Future Research

One area for future research is to complement the findings in this thesis with studies of the properties of the estimators under other proportions of non-DE genes. Also, the results show the need for future developments of alternative procedures when the sample size is small.

The multiple testing problem of microarray data is challenging both in its magnitude and in the typical nature of the data; genes being highly correlated. Some of the reasons of dependency are (Pawitan et al 2006; Ploner et al 2005; Reiner et al)

- co-regulation based on genomic locations and gene expression biases based on effects of aneuploidy in studies of cancer
- factors related to the normalization process
- pooled variability estimation and
- RNA resource

While the method (moderated Welch) described in paper II offers an attractive solution to one of the most challenging recent problems in microarray data analysis, it is designed to work under the assumption of independence of genes. That is to say the hierarchical model described in paper II did not assume that the estimators $\bar{y}_d$ and se_w^2 from different genes are dependent and so further research could be conducted under the general assumption of gene dependence.

References

[1] Abul, O, Alhajj, R, Polat, F. and K. Barker (2004). Finding differentially expressed genes for pattern generation. *Bioinformatics*, 21(4), 445-450.

[2] Alizadeh, A.A, Eisen, M.B, Davis, R.E, Ma, C, Lossos, I.S, Rosenwald, A, Boldrick, J.C, Sabet, H, Tran, T, Yu, X, Powell, J.I, Yang, L, Marti, G.E, Moore, T, Hudson, J. Jr, Lu, L, Lewis, D.B, Tibshirani, R, Sherlock, G, Chan, W.C, Greiner, T.C, Weisenburger, D.D, Armitage, J.O, Warnke, R, Levy, R, Wilson, W, Grever, M.R, Byrd, J.C, Botstein, D, Brown, P.O. and L.M. Staudt (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-511.

[3] Alon, U, Barkai, N, Notterman, D.A, Gish, K, Ybarra, S, Mack, D. and A.J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12), 6745-6750.

[4] Benjamini, Y. and Y. Hochberg (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.

[5] Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple hypothesis testing under dependency, *Annals of Statistics*, 29(4), 1165-1188.

[6] Boldrick, J.C, Alizadeh, A.A, Diehn, M, Dudoit, S, Liu, C.H, Belcher, C.E, Botstein, D, Staudt, L.M, Brown, P.O. and D.A. Relman (2002). Stereotyped and specific gene expression programs in human innate immue responses to bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2), 972-977.

[7] Dudoit, S, Yang, Y.H, Callow, M. and T. Speed (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistical Science*, 12, 111-139.

[8] Dudoit, S, Shaffer, J.P. and J.C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18, 71-103.

[9] Efron, B, Tibshirani, R, Goss, V. and G. Chu (2000). Microarrays and their use in a comparative experiment. Mimeo, Department of Statistics, Stanford University.

[10] Efron, B. (2005). Local false discovery rates. Mimeo, Department of Statistics, Stanford University.

[11] Gasch, A.P. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the cell*, 11, 4241-4257.

[12] Golub, T.R, Slonim, D.K, Tamayo, P, Huard, C, Gaasenbeek, M, Mesirov, J.P, Coller, H, Loh, M.L, Downing, J.R, Caligiuri, M.A, Bloomfield, C.D. and E.S. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

[13] Hochberg, Y. and A.C. Tamahane (1987). *Multiple comparison procedures*. Wiley, New York.

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

[15] Lehmann, E.L. (1986). *Testing Statistical Hypotheses, 2nd edn*. Springer-Verlag, New York.

[16] Pollack, J.R, Perou, C.M, Alizadeh, A.A, Eisen, M.B, Pergamenschikov, A, Williams, C.F, Jeffrey, S.S, Botstein, D. and P.O. Brown (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23, 41-46.

[17] Pawitan,Y, Michiels, S, Koscielny, S, Gusnanto, A. and A. Ploner (2005a). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21, 3017-3024.

[18] Pawitan, Y, Karuturi, R, Murthy, K, Michiels S. and A. Ploner (2005b). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, 21, 3865-3872.

[19] Perou, C.M, Jeffrey, S.S, van de Rijn, M, Rees, C.A, Eisen, M.B, Ross, D.T, Pergamenschikov, A, Williams, C.F, Zhu, S.X, Lee, J.C.F, Lashkari, D, Shalon, D, Brown, P.O. and D. Botstein (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 9212-9217.

[20] Ploner, A, Calza, S, Gusnanto, A. and Y. Pawitan (2006). Multidimensional local false discovery rate for micorarray studies. *Bioinformatics*, 22, 556-565.

[21] Ross, D.T, Scherf, U, Eisen, M.B, Perou, C.M, Rees, C, Spellman, P, Iyer, V, Jeffrey, S.S, van de Rijn, M, Waltham, M, Pergamenschikov, A, Lee, J.C, Lashkari, D, Shalon, D, Myers, T.G, Weinstein, J.N, Botstein, D. and P.O. Brown (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24, 227-234.

[22] Shaffer, J.P. (1995). Multiple hypothesis testing. *Annals of Review Psychology*, 46, 561-584.

[23] Tukey, J.W. (1953). The problem of multiple comparisons. Unpublished manuscript. In the collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983 1-300. Chapman and Hall, New York.

[24] Yang, Y.H. and T. Speed (2003). Design and analysis of comparative microarray experiments. In Speed, T. (ed.), *Statistical Analysis of Gene Expression Microarray Data*, Chapman and Hall/CRC, Boca Raton, FL, pp. 35-92.

[25] Westfall, P.H, Zaykin, D.V. and S.S. Young (2002). Multiple tests for genetic effects in association studies. In S.W. Looney (ed.), *Biostatistical Methods*, Human Press Inc, Totowa, NJ, pp. 143-168.

[26] Yekutieli, D. and Y. Benjamini (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, *Journal of Statistical Planning and Inference*, 82, 171-196.

[27] Zhao, Y. and W. Pan (2003). Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 19(9), 1046-1054.