Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 26

Testing for Trends in Dose-Response Microarray Experiments: A Comparison of Several Testing Procedures, Multiplicity and Resampling-Based Inference

Dan Lin[*] Ziv Shkedy[†] Dani Yekutieli[‡]

Tomasz Burzykowski[**] Hinrich W.H. Göhlmann[††] An De Bondt[‡‡]

Tim Perera[§] Tamara Geerts[¶] Luc Bijnens[‖]

[*]Hasselt University, dan.lin@uhasselt.be

[†]Hasselt University, ziv.shkedy@uhasselt.be

[‡]Tel Aviv University, yekutiel@post.tau.ac.il

[**]Hasselt University, tomasz.burzykowski@uhasselt.be

[††]Johnson & Johnson PRD, hgoehlma@prdbe.jnj.com

[‡‡]Johnson & Johnson PRD, adbondt@prdbe.jnj.com

[§]Johnson & Johnson PRD, tperera@prdbe.jnj.com

[¶]Johnson & Johnson PRD, tgeerts@prdbe.jnj.com

[‖]Johnson & Johnson PRD, lbijnens@prdbe.jnj.com

Testing for Trends in Dose-Response Microarray Experiments: A Comparison of Several Testing Procedures, Multiplicity and Resampling-Based Inference*

Dan Lin, Ziv Shkedy, Dani Yekutieli, Tomasz Burzykowski, Hinrich W.H. Göhlmann, An De Bondt, Tim Perera, Tamara Geerts, and Luc Bijnens

Abstract

Dose-response studies are commonly used in experiments in pharmaceutical research in order to investigate the dependence of the response on dose, i.e., a trend of the response level toxicity with respect to dose. In this paper we focus on dose-response experiments within a microarray setting in which several microarrays are available for a sequence of increasing dose levels. A gene is called differentially expressed if there is a monotonic trend (with respect to dose) in the gene expression. We review several testing procedures which can be used in order to test equality among the gene expression means against ordered alternatives with respect to dose, namely Williams' (Williams 1971 and 1972), Marcus' (Marcus 1976), global likelihood ratio test (Bartholomew 1961, Barlow et al. 1972, and Robertson et al. 1988), and M (Hu et al. 2005) statistics. Additionally we introduce a modification to the standard error of the M statistic. We compare the performance of these five test statistics. Moreover, we discuss the issue of one-sided versus two-sided testing procedures. False Discovery Rate (Benjamni and Hochberg 1995, Ge et al. 2003), and resampling-based Familywise Error Rate (Westfall and Young 1993) are used to handle the multiple testing issue. The methods above are applied to a data set with 4 doses (3 arrays per dose) and 16,998 genes. Results on the number of significant genes from each statistic are discussed. A simulation study is conducted to investigate the power of each statistic. A R library IsoGene implementing the methods is available from the first author.

KEYWORDS: dose response study, multiple testing, monotonicity, resampling-based procedures

---

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

1 Introduction

Investigation of a dose-response relationship is of primary interest in many drug-development studies. Typically, in dose-response experiments the outcome of interest is measured at several (increasing) dose levels, and the aim of the analysis is to establish the form of the dependence of the response on dose (Agresti 1997). The response can be either the efficacy of a treatment or the risk associated with the exposure to the treatment (in toxicology studies). In a typical dose-response study subjects are randomized to several dose groups, among which there is usually a control group. Ruberg (1995a, 1995b) and Chuang-Stein and Agresti (1997) formulated four main questions usually asked in dose-response studies: (1) Is there any evidence of the drug effect? (2) For which doses is the response different from the response in the control group? (3) What is the nature of the dose-response relationship? and (4) What is the optimal dose?

Within the microarray setting, a dose-response experiment has the same structure as described above. The response is the gene expression at a certain dose level. The dose-response curve, similarly to the dose-response studies, is assumed to be monotone, i.e., the gene activity increases or decreases as the dose level increases. The direction of the relationship is usually unknown in advance.

In this paper we focus on the first question: is there any evidence of the drug effect? To answer this question, we test for the null hypothesis of homogeneity of means (no dose effect) against an ordered alternative. We compare several testing procedures, that take into account the order restriction of the means with respect to the increasing doses and that adjust for multiplicity. In particular, we discuss the testing procedures of Williams (Williams 1971 and 1972), Marcus (Marcus 1976), the global likelihood ratio test (Barlow *et al.* 1972, and Robertson *et al.* 1988), and the M (Hu *et al.* 2005) statistic. Moreover, we propose a novel procedure based on a modification of the estimator of standard error of the M statistic.

Williams (1971, 1972) proposed a step-down procedure to test for the dose effect. The tests are performed sequentially from the comparison between the isotonic mean of the highest dose and the sample mean of the control to the comparison between the isotonic mean of the lowest dose and the sample mean of the control. The procedure stops at the dose level where the null hypothesis (of no dose effect) is not rejected. Marcus (1976) proposed a modification of the Williams procedure, in which the sample mean of the control was replaced by the isotonic mean of the control. A global likelihood ratio test discussed by Bartholomew *et al.* (1961), Barlow *et al.* (1972), and Robertson *et al.*,

(1988) uses the ratio between the variance calculated under the null hypothesis and the variance calculated under an ordered alternative. Recently, Hu *et al.* (2005) proposed a test statistic that was similar to Marcus' statistic, but with the variance estimator calculated under the ordered alternative. The degrees of freedom of the M statistic (the difference between the number of observations and the number of dose levels) are fixed for all the genes and all the arrays. In this paper, we propose a modification for the variance estimator of the M statistic. Namely, the difference between the number of observations and the unique number of isotonic means is used as the degrees of freedom for the variance estimator.

Our goal is to compare the performance of the five test statistics. To this aim we apply them to a case study. The case study data come from a microarray experiment with three microarrays, each containing 16,998 genes, available for each of four dose levels of a drug. When applied to the case study, the five test statistics are adjusted for multiple testing by using resampling-based procedures that control either the Family-Wise Error Rate (FWER) or the False Discovery Rate (FDR). Following the results of the analysis of the case study, we conduct a simulation study to further investigate the performance of the five test statistics.

The paper is organized as follows. Section 2 describes the procedure followed to obtain the case study data. In Section 3 we review the five test statistics. Directional inference to testing isotonic regression and multiplicity issue are discussed in Section 4. In Section 5 we compare the results of the analysis of the case study using the five tests discussed in Section 3. A simulation study conducted to investigate the performance of variance estimators and power of the five test statistics is presented in Section 6. Section 7 completes the paper with a short discussion.

2 Data Acquisition

The human epidermal squamous carcinoma cell line A431 was grown in Dulbecco's modified Eagle's medium, supplemented with Lglutamine (2 mM), Gentamycin (50 mg/ml) and 10% fetal bovine serum. The cells were stimulated with EGF (R&D Systems, 236-EG) at different concentrations (0 ng/ml, 1 ng/ml, 10 ng/ml and 100 ng/ml) for 24h. RNA was harvested using RLT buffer (Qiagen). 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 2004). Biotin-labeled target samples were hybridized to human genome ar-

rays U133 A 2.0 containing probe sets interrogation approximately 22,000 transcripts from the UniGene database (Build 133). Hybridization was performed using 15 μg of cRNA for 16 h at 45^0C 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 were quantile normalized in two steps: first within each sample group, and then across all sample groups obtained (Bolstad *et al.* 2002). The resulting data set consists of 12 samples, for four dose levels and three microarrays at each dose level, with 16,998 probe sets. For simplicity, we refer to probe sets as genes through our paper (Hubbell *et al.* 2002).

3 Testing For Homogeneity of the Means Under Restricted Alternatives

In this section, we discuss several procedures for testing the homogeneity of the means under the restricted alternative. In particular we focus on four existing procedures: Williams' (Williams 1971 and 1972), Marcus' (Marcus 1976), the global likelihood ratio test (Bartholomew 1961, Barlow *et al.* 1972, and Robertson *et al.* 1988), and the M (Hu *et al.* 2005) statistic. Additionally, we introduce a modification to the degree of freedom of the M statistic.

In the microarray experiment, for each gene, the following ANOVA model is considered

$$Y_{ij} = \mu(d_i) + \varepsilon_{ij},\ i = 0, 1, \ldots, K,\ j = 1, 2, \ldots, n_i, \tag{1}$$

where Y_{ij} is the jth gene expression at the ith dose level, d_i $(i = 0, 1, \ldots, K)$ are the $K+\mathbf{1}$ dose levels, $\mu(d_i)$ is the mean gene expression at each dose level, and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The null hypothesis of no dose effect is given by

$$H_0 : \mu(d_0) = \mu(d_1) = \cdots = \mu(d_K). \tag{2}$$

A one-sided alternative hypothesis of a positive dose effect for at least one dose level (i.e., an increasing trend) is specified by

$$H_1^{Up} : \mu(d_0) \leq \mu(d_1) \leq \cdots \leq \mu(d_K), \tag{3}$$

with at least one strict inequality. When testing the affect of a drug for a positive outcome the researcher can specify a positive effect as the desirable

alternative. However, in the current microarray setting, it seems reasonable to assume that the gene expression levels may increase or decrease in response to increasing doses, but with the direction of the trend not known in advance. Thus we must also consider an additional alternative:

$$H_1^{Down} : \mu(d_0) \geq \mu(d_1) \geq \cdots \geq \mu(d_K), \tag{4}$$

with at least one strict inequality. Testing H_0 against H_1^{Down} or H_1^{Up} requires estimation of the means under both the null and the alternative hypotheses. Under the null hypothesis, the estimator for the mean response $\hat{\mu}$ is the sample mean. Let $\hat{\mu}_0^\star, \hat{\mu}_1^\star, \ldots, \hat{\mu}_K^\star$ be the maximum likelihood estimates for the means (at each dose level) under the ordered alternative. Barlow *et al.* (1972) and Robertson *et al.* (1998) showed that $\hat{\mu}_0^\star, \hat{\mu}_1^\star, \ldots, \hat{\mu}_K^\star$ are the isotonic regression of the observed means.

3.1 Williams' (1971, 1972) and Marcus' (1976) Test Statistics

Williams' procedure defines H_0 as the null hypothesis, and H_1^{Up} or H_1^{Down} as the one-sided alternative. Williams' (1971, 1972) test statistics was suggested for a setting, in which n_i observations are available at each dose level. As all dose levels are compared with the control level, the test statistic is given by

$$t_i = \frac{\hat{\mu}_i^\star - \bar{y}_0}{\sqrt{2S^2/r}}. \tag{5}$$

Here, $\bar{y}_0$ is the sample mean at the first dose level (control), $\hat{\mu}_i^\star$ is the estimate for the mean at the ith dose level under the ordered alternative, r is the number of replications at each dose level, and S^2 is an estimate of the variance. For $\hat{\mu}_i^\star$, Williams (1971, 1972) used the isotonic regression of the observed response with respect to dose (Barlow *et al.* 1972). Williams' test procedure is a sequential procedure. In the first step, $\hat{\mu}_K^\star$ is compared to $\bar{y}_0$. If the null hypothesis is rejected, $\hat{\mu}_{K-1}^\star$ is compared to $\bar{y}_0$, etc.

Marcus (1976) proposed a modification to Williams' test statistic that replaced $\bar{y}_0$ with $\hat{\mu}_0^\star$, the estimate of the first dose (control) mean under ordered restriction. Marcus' test statistic performs closely to Williams' in terms of power (Marcus 1976). Note that, for $K = 1$, Williams' and Marcus' test statistics reduce to the two-sample t-test.

3.2 Likelihood Ratio Test Statistic for Monotonicity (Barlow *et al.* 1972, and Robertson *et al.* 1988)

Williams' and Marcus' procedures are step-down procedures, i.e., the comparison between a lower dose and control is tested only if the test of a higher dose vs. control is significant. The underlying assumption is that there is a monotone dose-response relationship with a known direction.

Testing the equality of ordered means using likelihood ratio tests (when response is assumed to be normally distributed) were discussed by Barlow *et al.* (1972) and Robertson *et al.* (1988). Both authors considered the likelihood ratio test, in which the variance under the null and the alternative were compared. The likelihood ratio test statistic is given by

$$\Lambda_{01}^{\frac{2}{N}} = \frac{\hat{\sigma}_{H_1}^2}{\hat{\sigma}_{H_0}^2} = \frac{\sum_{ij}(y_{ij} - \hat{\mu}_j^\star)^2}{\sum_{ij}(y_{ij} - \hat{\mu})^2}, \tag{6}$$

where $\hat{\sigma}_{H_0}^2$ and $\hat{\sigma}_{H_1}^2$ are the parameter estimates for the variance under the null and the alternative hypothesis, respectively. The null hypothesis is rejected for a "small" value of $\Lambda_{01}^{\frac{2}{N}}$. Equivalently, H_0 is rejected for large value of $\bar{E}_{01}^2$, where

$$\bar{E}_{01}^2 = 1 - \Lambda_{01}^{\frac{2}{N}} = \frac{\sum_{ij}(y_{ij} - \hat{\mu})^2 - \sum_{ij}(y_{ij} - \hat{\mu}_j^\star)^2}{\sum_{ij}(y_{ij} - \hat{\mu})^2}. \tag{7}$$

Estimating the parameters using isotonic regression requires the knowledge of the direction of the trend. In practice, the direction of the trend is often not known in advance. In such a case one can maximize the likelihood twice: for a monotone decreasing trend and for a monotone increasing trend, and choose the trend with a higher likelihood. In practice, we can calculate $\bar{E}_{01}^2$ for each direction and choose the higher value of $\bar{E}_{01}^2$ (Barlow *et al.* 1972). In this paper we use a resampling-based approach to approximate the null distribution for the test statistic, so that the two sided p-values are obtained for inference.

3.3 The M Test Statistic of Hu *et al.* (2005)

Recently, Hu *et al.* (2005) proposed the following test statistic M to test for a monotonic trend:

$$M = \frac{\hat{\mu}_K^\star - \hat{\mu}_0^\star}{\sqrt{\sum_{i=0}^{K}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i^\star)^2/(N-K)}}. \tag{8}$$

Hu *et al.* (2005) discussed a setting, in which the comparison of primary interest is the difference between the highest dose level (K) and the control

dose. The numerator of the M test statistic is the same as of Marcus' statistic, while the denominator is an estimate of the standard error under an ordered alternative. This is in contrast to Williams' and Marcus' approaches that use the unrestricted means to derive the estimate for the standard error.

Hu *et al.* (2005) evaluated the performance of the $\bar{E}^2_{01}$ and M test statistics by comparing the ranks of genes obtained by using both statistics, and reported similar findings for simulated and real life data sets.

3.4 A Modification to the M test Statistic

For the variance estimate, Hu *et al.* (2005) used $N - K$ degrees of freedom (see equation (8)). However, the unique number of isotonic means is not fixed, but changes across the genes. For that reason, we propose a modification to the standard error estimator used in the M statistic by replacing it with $\sqrt{\sum_{i=0}^{K}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i^\star)^2/(N-I)}$, where I is the unique number of isotonic means for a given gene. Such a modification is expected to improve the standard error estimates across all the genes.

4 Directional Inference in Isotonic Regression and Multiplicity

4.1 Multiplicity and Resampling-based Multiple Testing

In microarray experiments a large of number of null hypotheses usually needs to be tested. The FamilyWise Error Rate (FWER, Westfall and Young 1993) and the False Discovery Rate (FDR, Benjamini and Hochberg 1995) are two quantities that are commonly used in controlling the error rate.

FWER is defined as the probability to reject at least one true null hypothesis. FDR, introduced by Benjamini and Hochberg (1995), is defined as the expected proportion of false rejections among the rejected hypotheses. Testing procedures that control FDR tend to gain more power as compared to procedures controlling for FWER.

FWER can be controlled by using, e.g., the Bonferroni, Holm (Holm 1979), Hochberg (1995), or maxT (Westfall and Young 1993) procedures. Hochberg and Benjamini (FDR-BH, 1995) and Benjamini and Yekutieli (FDR-BY, 1999) proposed approaches for controlling FDR.

In a microarray setting, resampling methods to adjust for multiplicity are often used (Kerr and Churchill 2001, Reiner *et al.* 2003, Tusher *et al.* 2001, and Ge *et al.* 2003). The main motivation is to avoid inference based on asymptotic distribution of the test statistics, which, within the microarray setting, can be problematic because of either typically small sample sizes or departure from the assumption about the distribution of the response. Also, in some cases the asymptotic distribution of the test statistics 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 each gene and for each permutation is referred to as the permutation matrix under the null distribution. Further inference is based on the unadjusted p-values obtained from the permutation matrix. For example, maxT procedure proposed by Westfall and Young (1993) to control FWER computes the adjusted p-values from the distribution of maxima of the test statistics over the nested subsets of ordered test statistics calculated under the null hypothesis (by applying the permutation matrix). Alternatively, once the unadjusted p-values of a test statistic are computed (Reiner *et al.* 2003 and Ge *et al.* 2003), they can be adjusted for multiple testing using various procedures such as Bonferroni, Holm, FDR-BH or FDR-BY.

4.2 Directional Inference in Isotonic Regression

The five test statistics discussed in Section 3 should be calculated assuming a particular direction of the ordered alternative. However, the direction of the test is unknown in advance. In this section, we address the issue of how to obtain the two-sided p-value from the five testing procedures, and how to determine the direction of the trend from two-sided p-value afterwards.

We focus on the two possible directions of the alternatives: H_1^{Up} defined in equation (3) and H_1^{Down} defined in equation (4). Let p^{Up} and T^{Up} denote the p-value and the corresponding test statistic computed to test H_0 vs. H_1^{Up}, and let p^{Down} and T^{Down} denote the p-value and the corresponding test statistic computed to test H_0 vs. H_1^{Down}. Barlow *et al.* (1972) showed that, for $K > 2$, a $\bar{\chi}^2$ statistic for testing H_0 may actually yield $p^{Up} < \alpha$ and $p^{Down} < \alpha$. However, $p = 2\min(p^{Up}, p^{Down})$ is always a conservative p-value for the two-sided test of H_0 vs. either H_1^{Up} or H_1^{Down}.

Hu *et al.* (2005) adapted the approach by taking the larger of the likelihoods of H_1^{Up} or H_1^{Down}, i.e., the larger of T^{Up} and T^{Down} is used as the test statistic for two-sided inference. In contrast to Hu *et al.* (2005), we obtain two-sided p-values by taking $p = \min(2\min(p^{Up}, p^{Down}), 1)$, where p^{Up} and p^{Down} are calculated for T^{Up} and T^{Down} using permutations to approximate the null

distribution of these test statistics. We use p^{Up} and p^{Down} to determine the direction of the test.

After rejecting the null hypothesis against the two-sided test there is still a need to determine the direction of the trend. The direction can be inferred by the following procedure. If $p^{Up} \leq \alpha/2$, then reject H_0 and declare H_1^{Up}; if $p^{Down} \leq \alpha/2$, then reject H_0 and declare H_1^{Down}. The validity of this directional inference is based on the following property: under H_1^{Up}, p^{Down} is stochastically larger than $U[0,1]$; and under H_1^{Down}, p^{Up} is stochastically larger than $U[0,1]$. Thus, the probability of falsely rejecting H_0 is $\leq \alpha$, and the probability of declaring a wrong direction for the trend is $\leq \alpha/2$. It is also important to note that the event $p^{Up} < \alpha/2$ and $p^{Down} < \alpha/2$ may be observed. Under H_0, H_1^{Up} or H_1^{Down}, this event is unlikely. However, it is likely if the treatment has a large and non-monotone effect. An example of this unique situation, in which the null hypothesis can be rejected for both directions, is given in Section 5.1.

In order to verify whether the property needed for directional inference applies to the five test statistics, we conduct a simulation study to investigate the distribution of the p^{Up} and p^{Down} values. For each simulation, data are generated under H_1^{Up}: the means are assumed to be equal to $(1,\ 2,\ 3,\ 4)/\sqrt{5}$ for the four doses, respectively, and the variance is equal to $\sigma^2 = 1$. The test statistics T^{Up} and T^{Down} are calculated for the two possible alternatives H_1^{Up} and H_1^{Down}. Their corresponding p^{Up} and p^{Down}-values are obtained using 10,000 permutations.

Figure 1 shows the cumulative distribution of p^{Up} and p^{Down}. Clearly, the simulations show that the cumulative distribution of p^{Down} (the p-value of the test statistics calculated assuming the wrong direction, dotted line in Figure 1) is stochastically higher than $U[0,1]$ (solid line in Figure 1), which is the distribution of the p-values under the null hypothesis. Moreover, the distribution of p^{Up} (the p-value for the test statistics calculated assuming the right direction, dashed line in Figure 1) is, as expected, stochastically smaller than $U([0,1]$. Similar results (not shown) are obtained when the data are generated under H_1^{Down}. The results imply that all the five test statistics process the property required for the directional inference: under H_1^{Up} the distribution of p^{Down} is stochastically greater than $U[0,1]$. Further discussion of the simulation results is given in the supplementary material to this paper.

4.3 Control of the Directional FDR

When FDR controlling procedures are used to adjust for multiplicity in the microarray setting, the set of two-sided p-values computed for each gene is

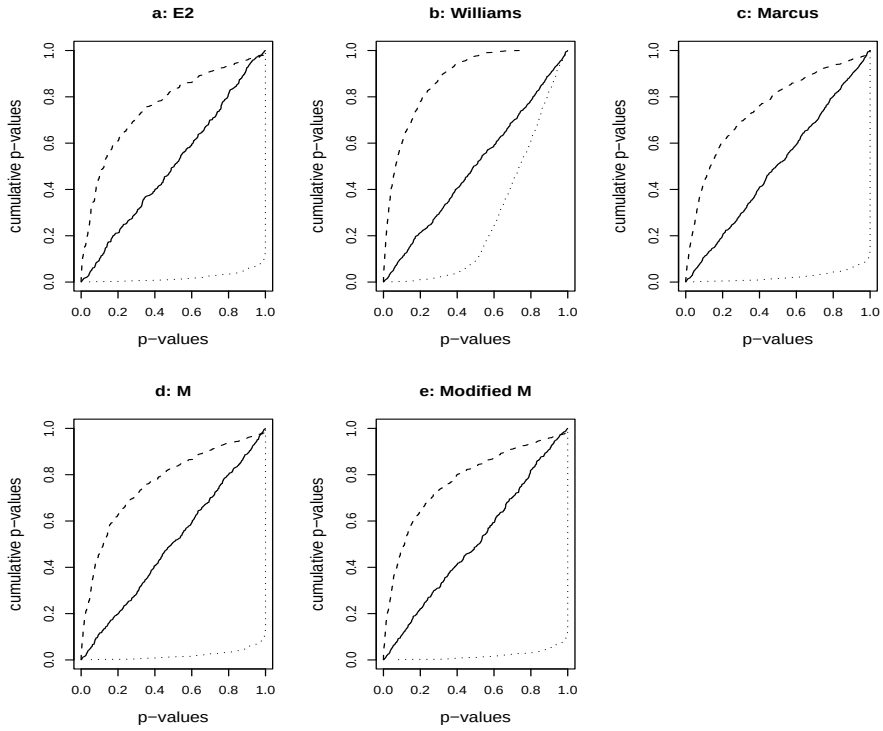

Figure 1: The cumulative distribution of p^{Up}-values (dashed line) and p^{Down}-values (dotted line) for the five test statistics. Data are generated under H_1^{Up} with isotonic means $(1, 2, 3, 4)/\sqrt{5}$ for the four doses. Solid line: cumulative distribution of $H_0 \sim U[0, 1]$.

adjusted using FDR-BH and FDR-BY procedure. A discovery in this case is a rejection of H_0 for some gene; a false discovery is to reject H_0 when H_0 is true. As mentioned before, in a microarray dose-response experiment we are also interested in the direction of the dose-response trend.

Benjamini and Yekutieli (2005) provide a framework for addressing the multiplicity problem when attempting to determine the direction of multiple parameters: a discovery is to declare the sign of a parameter as either being positive or negative. Three types of false discoveries are possible: declaring a zero parameter either as negative or as positive, declaring a negative parameter as positive, and declaring a positive parameter as negative. The FDR corresponding to these discoveries is termed the Mixed Directional FDR. In the current setting the Mixed Directional FDR is the expected value of the number of genes, for which H_0 is true, that are erroneously declared to have either a positive or negative trend plus the genes with a monotone trend but

the direction of the declared trend is wrong, divided by the total number of genes declared to have a trend. Benjamini and Yekutieli (2005) prove that if p-values pose the directional property described in Section 4.2, then applying the BH procedure at level q to the the set of two-sided p-values computed for each gene, and declaring the direction of the trend corresponding to the smaller one-sided p-value, controls the Mixed Directional FDR at level $q/2 \cdot (1+m_0/m)$, where m is the total number of genes and m_0 is the number of genes, for which H_0 holds.

In general, directional inference is a more general setting than hypotheses testing (Benjamini and Yekutieli, 2005). Nevertheless, as a false discovery is made based on the p-value that is stochastically larger than $U[0,1]$, then the resampling-based methods that control FDR (Yekutieli and Benjamini, 1999) also control the Mixed Directional FDR. This is achieved by simply applying the resampling-based procedure to test H_0, and if H_0 is rejected, declaring the direction of the trend according to the minimum one-sided p-value. For each rejected null hypothesis it is also advisable to examine if the larger p-value is $\leq \alpha$. If this is the case, this may serve as an indication of a non-monotone dose-response relationship.

5 Results

In this section, we present results of an application of the five testing procedures to the case study. We compare the performance of each of five test statistics in combination with the Bonferroni, Holm, maxT, and FDR-BH multiple-testing adjustment procedures. In Section 5.1 we examine the number of significant genes for all the testing procedures. In Section 5.2 we make a comparison between the global likelihood ratio test $\bar{E}_{01}^2$ and the two t-test type statistics: M and the modified M.

5.1 Number of Significant Findings for Each Statistic Using Different Multiple Testing Adjustment

The testing procedures discussed in the previous sections are applied to the case study data. For each test statistic, p^{Up} and p^{Down} are obtained based on the permutation matrix, in which the null distribution of the test statistics (T^{Up} and T^{Down}) are approximated using 1000 permutations. The inference is made based on the two-sided p-values obtained using the method described in Section 4.2.

Table 1 shows the number of rejected hypotheses using several multiplicity adjusting methods and the five test statistics that are tested at the significance level of 0.05. Figure 2 shows the adjusted p-values for the five test statistics. Clearly, the adjusted p-values for maxT, Bonferroni, and FDR-BY are larger than the adjusted p-values obtained for FDR-BH. For instance, for $\bar{E}_{01}^2$, without adjusting for multiple testing, we reject the null hypothesis for 5457 genes. With Bonferroni, Holm, and FDR-BY adjustment procedures we obtain the same number of significant genes, i.e., 1814. Using maxT for controlling FWER seems to be the most conservative approach with only 224 genes declared significant.

Table 1: *Number of rejected null hypotheses for various testing procedures at the significance level of 0.05.*

Method	$\bar{E}_{01}^2$	Willams	Marcus	M	Modified M
Unadjusted	5457	5238	5465	5449	5451
maxT	224	215	223	265	251
Bonferroni	1814	1592	1669	1755	1745
Holm	1814	1592	1669	1755	1745
FDR-BH	3613	3209	3533	3562	3567
FDR-BY	1814	1592	1669	1755	1745

Note that the number of significant genes obtained for each test statistic for a given multiple testing adjustment is similar. For example, for the FDR-BH adjustment, we find 3613, 3562, and 3567 significant genes for $\bar{E}_{01}^2$, M, and the modified M statistic, respectively. This method yields more liberal results as compared to the other multiple testing adjustment procedures. For that reason, FDR adjustment for multiplicity is commonly used within the microarray framework (Ge *et al.* 2003, Tusher *et al.* 2001, Storey and Tibshirani 2003). Moreover, FDR-BH controls for the directional FDR (as discussed in Section 4.3). Therefore, in what follows, we use FDR-BH procedure to investigate the performance of the considered test statistics.

As we argue in Section 4.2, there is a possibility (although unlikely), that the null hypothesis is rejected for both directions (i.e., $p^{Up} \leq \alpha/2$ and $p^{Down} \leq \alpha/2$). For the analysis discussed above, only five genes are rejected by Marcus' statistic with p^{Up} and p^{Down} smaller than the rejection threshold (with multiple testing adjustment), suggesting a non-monotonic trend. The five genes are shown in Figure 3.

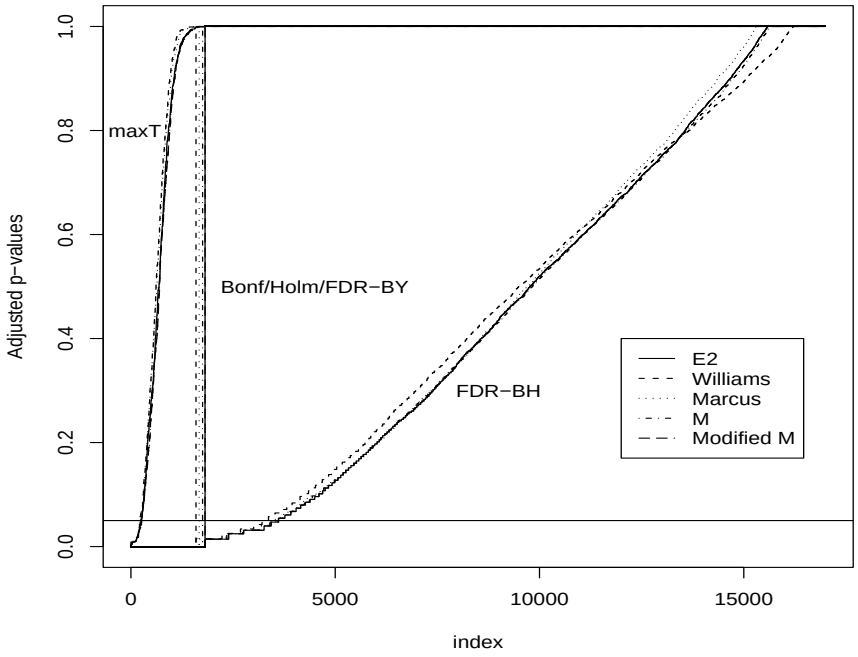

Figure 2: Adjusted p-values using Bonferroni, BH(FDR) and maxT for the five test statistics.

As can be observed from Figure 3, for the five genes the data reveal a non-monotonic pattern. For Marcus' statistic, the large values of T^{Up} and T^{Down} are obtained from the large difference between the isotonic mean of the highest and control doses relative to the variance calculated under the unrestricted alternative. Instead, $\bar{E}^2_{01}$, M, and the modified M use the variance estimator calculated under the ordered alternative, that results in small test statistic values. Hence, using these test statistics, the null hypothesis is not rejected. If the difference between the highest isotonic mean and control sample mean exists, Williams' test statistic will tend to reject the null hypothesis as well.

In particular, for the five genes, the estimates of σ^2 (of Williams' and Marcus' test statistics) calculated under the unrestricted alternative are equal, respectively, to 0.0414, 0.0075, 0.0204, 0.0145, and 0.0232. They are smaller than the estimates for σ^2 (of $\bar{E}^2_{01}$, M, and the modified M) calculated under the ordered alternative H_1^{Up}, that are equal, respectively, to 0.2995, 0.1788, 0.3277, 0.3317, and 0.2437, and under H_1^{Down}, that are equal, respectively, to 0.2608, 0.1868, 0.4679, 0.4401, and 0.2065.

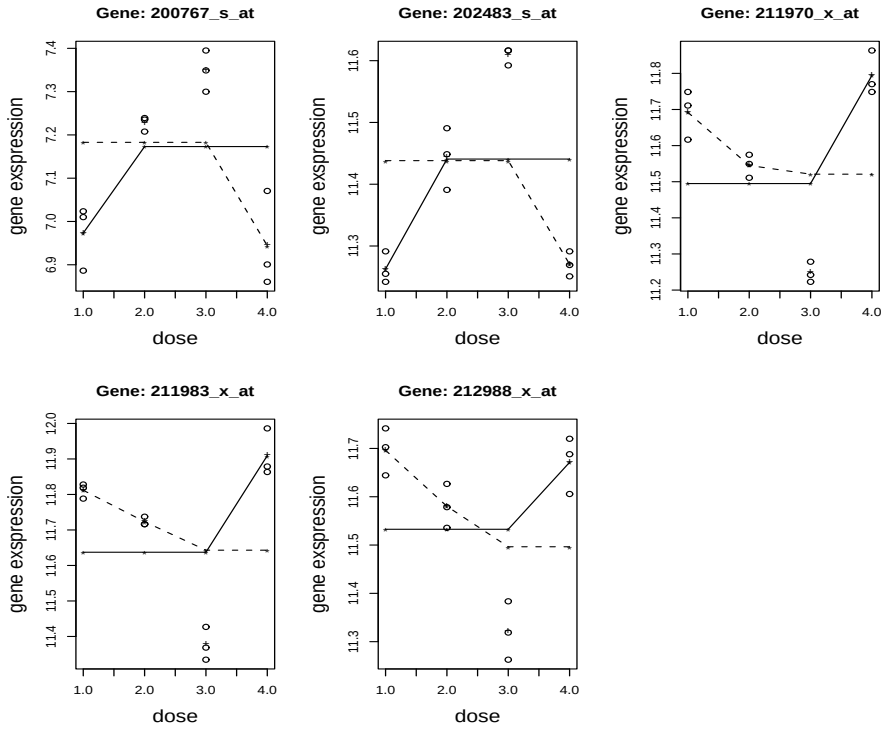

Figure 3: Five genes rejected by Marcus' statistics with both p^{Up} and p^{Down} values smaller than the rejection threshold. Solid line: the isotonic means obtained for testing H_0 against H_1^{Up}. Dashed line: the isotonic means obtained for testing testing H_0 against H_1^{Down}.

5.2 Comparison Between $\bar{E}_{01}^2$, M, and the Modified M Test Statistics

Although in our case study, the number of significant genes obtained for the five testing procedures is very similar, there are some discrepancies. In this section, we investigate the subset of genes not commonly found by $\bar{E}_{01}^2$, M, and the modified M statistics, respectively.

First we compare genes identified as significant or non-significant by M and $\bar{E}_{01}^2$. The logarithm of two-sided p-values for these genes is shown in Figure 4. Among the total of 16,998 genes, 3420 genes are found significant for monotonic trends for both statistics. However, 193 genes are found to be significant for $\bar{E}_{01}^2$ and non-significant for M-test statistic, while for 142 genes the reversed order is observed. These genes account for 8.9% ((193+142)/(3420+193+142)) of the total significant findings for both test statistics, which is not negligible.

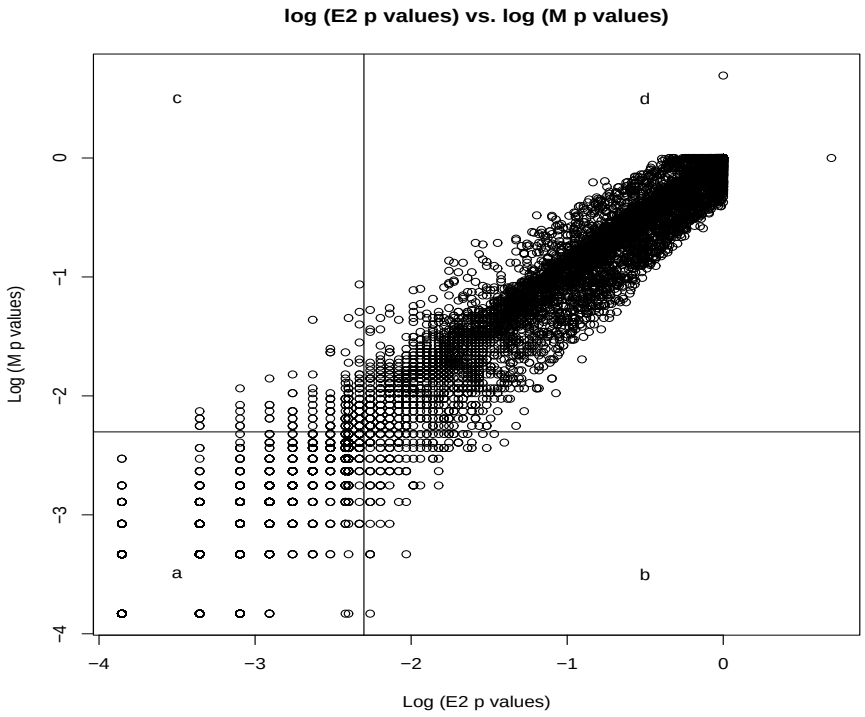

Figure 4: Logarithm of p-values (two sided) for $\bar{E}^2_{01}$ and M. Panel a: 3420 genes rejected by both $\bar{E}^2_{01}$ and M statistics; panel b: 142 genes are rejected by M statistic only; panel c: 193 genes in are rejected from $\bar{E}^2_{01}$ only; panel d: 13,244 genes are not rejected by either statistic.

Similar to Hu *et al.* (2005), we compare the ranking of M and $\bar{E}^2_{01}$ of all the genes. In both Hu *et al.* (2005) and our example the correlation of the ranks is equal to 0.99. Based on their observation, Hu *et al.* (2005) concluded that the two statistics perform similarly. However, in our data, the correlation of ranks of 142 genes found significant only for the M statistic (panel c of Figure 5) is 0.92, while the correlation of ranks of 193 genes significant only for $\bar{E}^2_{01}$ (panel b) is 0.85. Both are somewhat lower than the correlation for genes in panel a (3420 genes significant for both statistics, correlation of 0.98) and in panel d (genes non-significant by either statistic, correlation of 0.99). The discrepant conclusions (rejecting the null only for one of statistic) can be explained by the fact that the M statistic looks for the mean difference between the highest dose and the control. On the other hand, $\bar{E}^2_{01}$ is a global test for the monotonic trend.

The logarithm of the two sided p-values for the genes identified as sig-

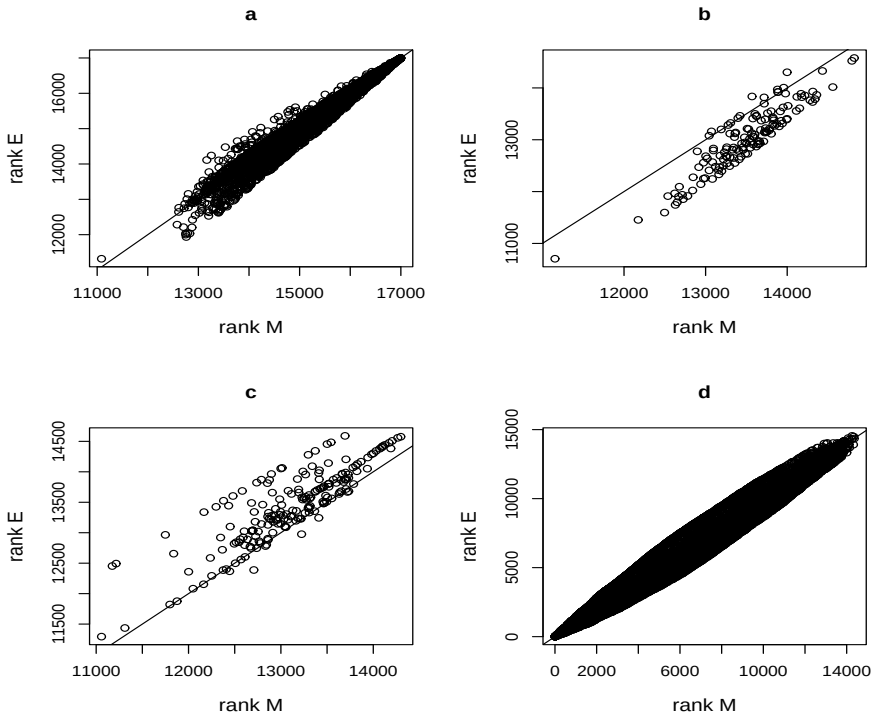

Figure 5: Correlation between $\bar{E}^2_{01}$ and M. Panel a: correlation (0.98) between rankings of 3420 genes rejected both from $\bar{E}^2_{01}$ and M. Panel b: correlation (0.92) between rankings of 142 genes rejected only from M. Panel c: correlation (0.85) between rankings of 193 genes rejected only from $\bar{E}^2_{01}$ and Panel d: correlation (0.99) between rankings of 13,244 genes not rejected from $\bar{E}^2_{01}$ and M.

nificant or non-significant by the M and modified M statistics is shown in Figure 6. Among the total of 16,998 genes, 3478 genes are found significant for monotonic trends for both tests. However, 86 genes are found to be significant for the M statistic and non-significant for the modified M test, while for 89 genes the reverse is true. These genes account for about 4.8% $((86 + 89)/(86 + 89 + 3478))$ of the total significant findings for both test statistics.

The overall correlation between the ranks of genes obtained for M and the modified M test statistics is 0.99. The correlation between genes in each panel of Figure 7 is also very high, with 0.97 (in panel b) for genes rejected only by the modified M, 0.98 (in panel c) for genes rejected only by M, 0.99 (in panel a) for genes rejected by both of the test statistics, and 0.998 (in panel c) for

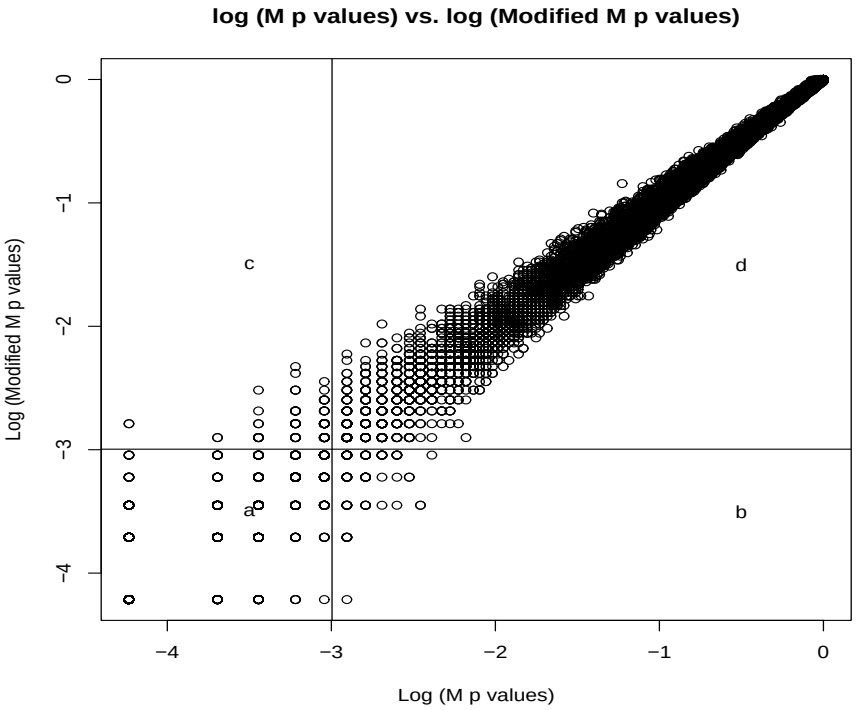

Figure 6: Logarithm of p-values (two sided) for the M and the modified M. Panel a: 3478 genes are rejected by both M and the modified M statistics; panel b: 86 genes are rejected by M statistic only; panel c: 89 genes are rejected from the modified M only; panel d: 13,345 genes are not rejected by either statistic.

genes rejected by neither of the test statistics. The difference between the two statistics lies in the adjustment of the degrees of freedom in the standard error estimator of the modified M test statistic. Nevertheless, the discrepancy found is not substantial.

6 Simulation Study

We conduct a simulation study to investigate the performance of the five test statistics. In Section 6.1, we compare the three estimators for the variance of Williams' and Marcus' (which is the same), M, and modified M test statistics. In Section 6.2 we investigate the power of the five statistics for a single gene, while in Section 6.3, the power of the tests with the multiple testing adjustment

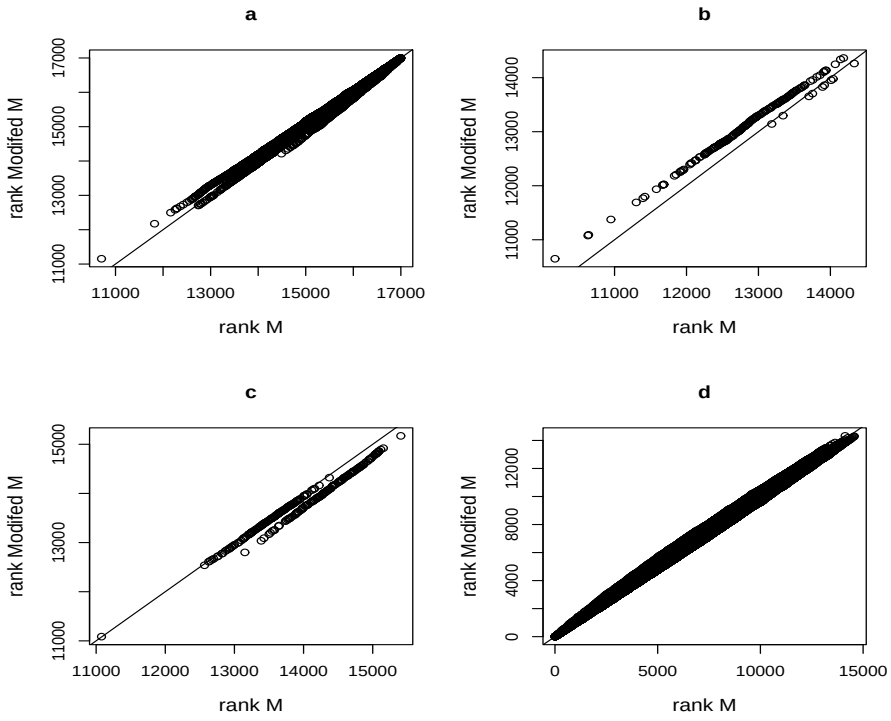

Figure 7: Correlation between M and the modified M. Panel a: correlation (0.99) between rankings of 3478 genes rejected both from M and the modified M. Panel b: correlation (0.97) between rankings of 89 genes rejected only from the modified M. Panel c: correlation (0.98) between rankings of 86 genes rejected only from M and Panel d: correlation (0.998) between rankings of 13,345 genes not rejected from the M and modified M.

is evaluated.

6.1 Standard Error Comparison

As a base for the simulations, the ANOVA model (1) is assumed. With four dose levels, the order-restricted alternative hypothesis (3) can be classified into seven possible trends. Table 2 defines the mean structure and assumed parameter values for these seven models, and for the null model (H_0) used in the simulations. The scale parameter λ controls the magnitude of the isotonic means. The larger λ, the larger distance between the means. In this set of simulations it is chosen to equal 1 and 3 based on the settings considered by Marcus (1976).

For each model, $L = 10,000$ datasets are generated. Each dataset contains three arrays per each of four dose levels, i.e., 12 arrays (observations) in total are generated, with variance σ^2=1.

Table 2: *Simulation settings:* μ_i *is the mean response of dose level* i, $i = 1, 2, 3, 4$, *and* $\lambda = 1$ *or* 3.

Model	Mean Structure	μ_1	μ_2	μ_3	μ_4	
g_1	$\mu_1 = \mu_2 = \mu_3 < \mu_4$	(1	1	1	2)	$\times 2\lambda/\sqrt{3}$
g_2	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	(1	1	2	2)	$\times \lambda$
g_3	$\mu_1 < \mu_2 = \mu_3 = \mu_4$	(1	2	2	2)	$\times 2\lambda/\sqrt{3}$
g_4	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	(1	2	2	3)	$\times \lambda/\sqrt{2}$
g_5	$\mu_1 = \mu_2 < \mu_3 < \mu_4$	(1	1	2	3)	$\times 2\lambda/\sqrt{11}$
g_6	$\mu_1 < \mu_2 < \mu_3 = \mu_4$	(1	2	3	3)	$\times 2\lambda/\sqrt{11}$
g_7	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	(1	2	3.5	4)	$\times \lambda/\sqrt{5}$
Null	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	(0	0	0	0)	$\times \lambda$

The performance of the standard error estimators for the Williams and Marcus, M, and the modified M test statistics is evaluated.

For Williams' statistic the estimator is $\sqrt{2/3}S^2$=$\sqrt{2/3}\hat{\sigma}_1$, where

$$\hat{\sigma}_1 = \sqrt{\sum_{i=0}^{3}\sum_{j=1}^{3}(y_{ij} - \bar{y}_i)/(12-4)},$$

and where y_{ij} is the gene expression at dose level i and array j, while $\bar{y}_i$ is the sample mean of gene expression levels at dose i.

The estimator of the M statistic, proposed by Hu *et al.* (2005), is given by

$$\hat{\sigma}_2 = \sqrt{\sum_{i=0}^{3}\sum_{j=1}^{3}(y_{ij} - \hat{\mu}_i^{\star})^2/(12-4)}.$$

Moreover, we consider the standard error estimate of the modified M, denoted as

$$\hat{\sigma}_3 = \sqrt{\sum_{i=0}^{3}\sum_{j=1}^{3}(y_{ij} - \hat{\mu}_i^{\star})^2/(12-I)},$$

where I is the number of unique isotonic mean levels obtained in the isotonic regression model.

First, we evaluate the mean squared error (MSE) of $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\sigma}_3$. The squared bias is estimated by $\hat{b}^2_{\hat{\sigma}} = (\bar{\hat{\sigma}} - \sigma)^2$, with $\bar{\hat{\sigma}} = \sum_{j=1}^{L} \hat{\sigma}_j / L$. The empirical variance is estimated by $\hat{v}_{\hat{\sigma}} = \sum_{j=1}^{L} (\hat{\sigma}_j - \bar{\hat{\sigma}})^2 / L$, leading to the simulation estimate of the MSE given by $\hat{\mathrm{MSE}}_{\hat{\sigma}} = \hat{b}^2_{\hat{\sigma}} + \hat{v}_{\hat{\sigma}}$.

Table 3 shows the squared bias, variance, and the MSE estimates of the three standard error estimators under the null hypothesis and under the seven alternative hypotheses. The smallest MSE values are obtained for $\hat{\sigma}_3$. Note that although $\hat{\sigma}_3$ tends to have the highest squared bias, its mean square error is the smallest due to the small variability of this estimator.

Table 3: *Squared bias, variance and MSE for $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\sigma}_3$. The numbers in the table are on 10^{-3} scale.*

Bias^2	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\sigma_3}$
g_1	0.739	1.593	0.954
g_2	1.079	0.831	1.382
g_3	0.805	1.548	1.02
g_4	1.185	0.122	1.984
g_5	1.115	0.199	1.982
g_6	0.917	0.304	1.751
g_7	0.706	0.117	1.843
Null	0.739	1.086	2.062
Variance	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\sigma_3}$
g_1	60.143	61.702	52.806
g_2	61.001	61.883	53.308
g_3	60.254	61.787	52.608
g_4	57.691	58.198	51.585
g_5	59.264	59.172	52.394
g_6	60.8	60.795	53.321
g_7	60.131	60.513	54.102
Null	60.143	59.686	51.092
MSE	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\sigma_3}$
g_1	60.881	63.295	53.76
g_2	62.08	62.714	54.69
g_3	61.059	63.335	53.628
g_4	58.876	58.321	53.569
g_5	60.379	59.371	54.376
g_6	61.717	61.099	55.072
g_7	60.837	60.63	55.945
Null	60.881	60.772	53.154

6.2 Power Study for a Single Gene Setting

Another simulation study is conducted to evaluate the power of the five test statistics, for a single gene setting. Similarly, as in the study presented in Section 6.1, datasets of 12 arrays are generated under the seven order-restricted models and the null model (Table 2). For each model (except for the null model), 5000 datasets are generated with an increasing and a decreasing trend, respectively. For the null model, 10,000 datasets in total are simulated for the comparison of the error rates. The isotonic means of the seven alternatives are specified in Table 2 with variance $\sigma^2 = 1$.

For each dataset and each test, p-values are obtained from 10,000 permutations. The results are summarized by the proportion of significant tests (with permutation-based p-values ≤ 0.05) that correctly classify the increasing or decreasing trend. For each λ the power and Type I error are shown in Table 4. The standard error estimate of the power can be obtained by $\sqrt{\hat{p}(1-\hat{p})/10,000}$ (Marcus 1976) where $\hat{p}$ is the estimate for the power.

The estimated Type I error probability is around 5% for all the tests. The power of the tests depends on the alternative. In general, regarding $\bar{E}^2_{01}$, Williams' and Marcus' tests, we arrive at the same conclusion as Marcus (1976), that the tests yield similar power. We can additionally observe that M and the modified M tests perform similarly as the other three. Hence, for a single gene setting, no test is uniformly better than the others across the set of the considered alternative hypotheses.

6.3 Power Study Under Multiple Testing Adjustment

We have also investigated the power of the considered test statistics when dealing with the multiple testing problem. Micorarrays with 5000 genes per microarray are generated. For each of the seven alternative models (see Table 2) a set of 100 genes (1400 genes in total) with an increasing and a decreasing trend is included. For the remaining 3600 genes no dose effect is assumed (the null model). The p-values for the considered test statistics are obtained using 10,000 permutations, and the multiplicity adjustment is provided by using the FDR-BH procedure.

In total, 100 datasets are generated for settings with $\lambda = 1$ and $\lambda = 3$. Table 5 shows the power and FDR with their simulation-based standard error estimates.

For $\lambda = 1$ the power of all the tests is very low. Moreover, FDR is not controlled at the desired level of 5%. This is related to the multiplicity adjustment procedure: the total number of rejected hypothesis is small, and

Table 4: *Power of the five test statistics for a single gene setting when data are generated under the eight models in Table 2.*

		$\bar{E}_{01}^2$	Williams	Marcus	M	Modified M
	g_1	0.2261	0.1882	0.2173	0.2299	0.1996
	g_2	0.2772	0.2196	0.2404	0.2371	0.2331
	g_3	0.2245	0.2189	0.199	0.2259	0.2096
$\lambda = 1$	g_4	0.2602	0.2943	0.2706	0.3046	0.3177
	g_5	0.3271	0.2684	0.2873	0.3134	0.301
	g_6	0.2662	0.2454	0.2345	0.2604	0.2819
	g_7	0.2953	0.2866	0.2744	0.3053	0.3231
	g_1	0.9739	0.9369	0.961	0.9669	0.9169
	g_2	0.9761	0.9058	0.9289	0.9462	0.887
	g_3	0.9772	0.9773	0.9678	0.9773	0.9416
$\lambda = 3$	g_4	0.9787	0.9914	0.9873	0.993	0.994
	g_5	0.9871	0.9624	0.9761	0.9844	0.9822
	g_6	0.9684	0.9706	0.9579	0.9747	0.9856
	g_7	0.9803	0.9826	0.978	0.9883	0.9936
	Null	0.0556	0.0584	0.0579	0.059	0.0564

Table 5: *Power study of the five test statistics under multiple testing adjustment.*

$\lambda = 1$	$\bar{E}_{01}^2$	Williams	Marcus	M	Modified M
Power	0.0354	0.0287	0.0289	0.0306	0.0309
SE(Power)	(0.0049)	(0.0046)	(0.0046)	(0.0048)	(0.0048)
FDR	0.1944	0.2077	0.2135	0.1835	0.1907
Se(Power)	(0.0507)	(0.0579)	(0.0568)	(0.0534)	(0.0534)
$\lambda = 3$	$\bar{E}_{01}^2$	Williams	Marcus	M	Modified M
Power	0.9112	0.8454	0.8477	0.8905	0.8928
SE(Power)	(0.0074)	(0.0099)	(0.0096)	(0.0082)	(0.0079)
FDR	0.0404	0.0424	0.0426	0.0399	0.0401
SE(Power)	(0.0053)	(0.0054)	(0.0053)	(0.0052)	(0.0052)

the proportion of wrong rejections is not well estimated, i.e, FDR is not well controlled.

With $\lambda = 3$ the power of the test statistics is greatly improved and FDR

is well controlled. $\bar{E}_{01}^2$ seems to provide a slightly higher power compared to the other tests. This can be explained by good performance in power under individual seven models. Note that the power obtained using the modified M test statistic is comparable. When multiplicity is taken into account, $\bar{E}_{01}^2$, M, and the modified M have higher power compared to Williams' and Marcus' tests (0.9112, 0.8905, and 0.8928 compared to 0.8454 and 0.8477, respectively).

7 Discussion

In this paper, we evaluate several test statistics for testing monotonic trend in the relationship of gene expression and doses in a microarray context. In particular, we consider Williams' step down procedure (Williams 1971, 1972), Marcus' procedure (Marcus 1976), likelihood ratio statistic (Robertson *et al.* 1988), M (Hu *et al.* 2005), and the modified M test statistic. Directional inference using these statistics is discussed for the situation when the direction of the trend is unknown in advance. To avoid inference based on asymptotic theory, we consider the use of permutation tests. Accordingly, several multiplicity adjustment methods including directional FDR are applied. BH procedure controlling FDR provides the most powerful approach as compared to the other methods (Tusher *et al.* 2001, Ge *et al.* 2003, Storey and Tibshirani 2003).

For the analysis discussed above, we observe comparable results for the five test statistics. However, a difference in the results between M and $\bar{E}_{01}^2$ is observed. Modifying the number of degrees of freedom for the M statistics improves the MSE of the estimate of the standard error. However, the simulation study investigating power of the five test statistics under multiple testing adjustment shows that the M and the modified M have a similar power.

As we argue in Section 4.2, a two sided inference can result in rejecting the null hypothesis in both directions ($p^{Up} < \alpha/2$ and $p^{Down} < \alpha/2$). This implies, as illustrated in Section 5.1, a non-monotone dose-response relationship. The difference between the four t-type test statistics (Williams', Marcus', M, and the modified M) is due to the estimates of the standard error. Williams and Marcus used the unbiased estimator calculated under the unrestricted ordered alternative, while M and the modified M use an estimator calculated under the ordered alternative. Williams' and Marcus' tests tend to reject genes when the difference calculated for the numerator exists and the standard error calculated under the unrestricted alternative is small. In particular, when the true means follow a simple tree (i.e., $\mu_1 \leq [\mu_2, \mu_3, \mu_4]$), a unimodal partial ordering (i.e., $\mu_1 \leq \mu_2 \leq \mu_3 \geq \mu_4$) or a simple loop (i.e., $\mu_1 \leq [\mu_2, \mu_3] \leq \mu_4$)

(Robertson *et al.* 1988), Williams' and Marcus' tests are more likely to reject the null hypothesis of homogeneity of means (no dose effect) in favor of the simple ordered alternative (H_1^{Up} or H_1^{Down}) than M and the modified M test statistics. We have shown that for a single gene the power of the four t-type test statistics is comparable (Table 4). However, the power after adjusting multiplicity obtained for M and the modified M is higher than those obtained for Williams' and Marcus'.

For a single gene the power obtained for $\bar{E}_{01}^2$ is comparable to the power obtained for the four t-type test statistics. Moreover, after adjustment for multiplicity, the power obtained for $\bar{E}_{01}^2$ is only slightly higher than M and the modified M tests (shown in Table 5). In our opinion, if the question of primary interest is the comparison between the highest and the lowest dose levels, $\bar{E}_{01}^2$, M, and the modified M tests are comparable (in terms of FDR controlling and power). However, if the question of primary interest is to detect a monotone trend, the global test $\bar{E}_{01}^2$ is to be preferred.

In this paper, we focus on testing the null hypothesis against a simple ordered alternative. Whenever the null hypothesis is rejected, the primary interest is to identify the dose-response curve shape. For a dose-response experiment with $K+1$ dose levels, there is a finite number of isotonic models which can be fitted to the data. For example, for an experiment with four dose levels there are seven upward monotone models (given in Table 2) and seven downward monotone models, which can be fitted to the data. The testing procedures discussed in this paper allows us to identify genes, for which the dose response curve is monotone, but not to identify the dose-response curve shape. The latter can be done using a model selection procedure, based on information criteria. Such a procedure will be presented in a future paper.

The R library implementing the methods presented in this paper is available from the first author.

References

Affymetrix (2004) GeneChip Expression Analysis Technical Manual, Rev.4. *Santa Clara, CA,* available at *http://www.affymetrix.com/support/technical/manual/expression_manual.affx.*

Agresti, A. (1997) *Statistical Methods for the Social Sciences*, Finlay.

Barlow, R.E., Bartholomew, D.J., Bremner, M.J. and Brunk, H.D. (1972) *Statistical Inference Under Order Restriction*, New York: Wiley.

Bartholomew, D.J. (1961) Ordered tests in the analysis of variance, *Biometrika*, **48**, 325-332.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. B*, **57**, 289-300.

Benjamini Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency, *ANN STAT*, **29(4)**, 1165-1188.

Benjamini Y. and Yekutieli D. (2005a) False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters, *Journal of the American Statistical Association*, **100**, 71-81.

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.

Chuang-Stein, C. and Agresti, A. (1997) Tutorial in biostatistics: A review of tests for detecting a monotone dose-response relationship with ordinal response data, *Statistics in Medicine*, **16**, 2599-2618.

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.

Holm, S. (1979) A simple sequentially rejective multiple test procedure, *Scand. J. Statist.*, **6**, 65–70.

Hu, J., Kapoor, M., Zhang, W., Hamilton3, S.R., and Coombes1,K.R. (2005) Analysis of dose response effects on gene expression data with comparison of two microarray platforms, *Bioinformatics*, **21(17)**, 3524–3529.

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

Kerr, M.K., Churchill, G.A. (2001) Statistical analysis of a gene expression microarray experiment with replication, *Biostatistics*, **2**, 183-201.

Marcus, R. (1976) The powers of some tests of the quality of normal means against an ordered alternative, *Biometrika*, **63**, 177-83.

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

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988), *Order Restricted Statistical Inference*, Wiley.

Ruberg, S.J. (1995a) Dose response studies. I. Some design considerations, *J. Biopharm. Stat.*, **5(1)**, 114.

Ruberg, S.J. (1995b) Dose response studies. II. Analysis and interpretation, *J. Biopharm. Stat.*, **5(1)**, 1542.

Storey, JD and Tibshirani R. (2003) Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences*, **100**, 9440-9445.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrys applied to the ionizing radiation response, *PNAS*, **98** , 5116-5121.

Westfall, P.H. and Young, S.S. (1993) *Resampling Based Multiple Testing*, Willy.

Williams, D.A. (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control, *Biometrics*, **27**, 103–117.

Williams, D.A. (1972) The comparision of several dose levels with a zero dose control, *Biometrics*, **28**, 519–531.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics, *J. Stat. Plan. Infer.*, **82**, 171-196.