Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 39

Selecting 'Significant' Differentially Expressed Genes from the Combined Perspective of the Null and the Alternative

Beatrijs Moerkerke* Els Goetghebeur†

*Ghent University, Beatrijs.Moerkerke@UGent.be

†egoetghe@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper39

Selecting 'significant' differentially expressed genes from the combined perspective of the Null and the Alternative

Moerkerke B.[1], **Goetghebeur E.**[1,2]

[1] Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 - S9, 9000 Gent, Belgium - Beatrijs.Moerkerke@UGent.be
[2] Department of Biostatistics, Harvard School of Public Health
655 Huntington Avenue, Boston, MA 02115, USA - egoetghe@hsph.harvard.edu

ABSTRACT

In the search for genes associated with disease, statistical analysis yields a key towards reproducible results. To avoid a plethora of type I errors , classical gene selection procedures strike a balance between magnitude and precision of observed effects in terms of p-values. Protecting false discovery rates recovers some power but still ranks genes according to classical p-values. In contrast, we propose a selection procedure driven by the concern to detect well-specified important alternatives. By summarizing evidence from the perspective of both the null and such an alternative hypothesis, genes line up in a substantially different order with different genes yielding powerful signals. A cutoff point for a measure of relative evidence which balances the standard p-value, p_0, with its counterpart, p_1, derived from the perspective of the target alternative, determines our gene selection. We find the cutoff point that maximizes an expected specific gain. This yields an optimal decision which exploits gene-specific variances and thus involves different type I and type II errors across genes. We show the dramatic impact of this alternative perspective on the detection of differentially expressed genes in hereditary breast cancer. Our analysis does not rely on parametric assumptions on the data.

Key words: alternative p-values, balanced testing, gene expression

1. INTRODUCTION

Statistical analysis of gene expression data is often aimed at detecting genes, which are differentially expressed between diseases or experimental conditions. Typically, a vast number of genes are being tested of whom only a small proportion is hoped to truly differentiate. The challenge not only lies in detecting the most promising genes without selecting too many non-differentiating genes, but also in ruling out some genes as being differentially expressed. This plays more generally in association studies such as for instance, those involving SNP's to asses the relative risk of disease. In line with classical individual test procedures, null-minded selection criteria avoid a flood of false alarms by controlling an experimentwise type I error (Hochberg and Tamhane, 1987) while accepting to sacrifice important findings due to limited power.

The philosophy behind controlling the false discovery rate (FDR), that is the expected proportion of true null hypotheses among the rejected ones (Benjamini and Hochberg, 1995), is one of willingness to accept a small number of type I errors relative to the number of rejected hypotheses. A range of procedures have been developed from this principle (see for example Benjamini and Hochberg, 1995; Benjamini and Yekutieli 2001; Genovese and Wasserman 2002; Storey, 2002 & 2003; Storey and Tibshirani, 2003; Wacholder, 2004; Fernando et al., 2004).

Most of these methods turn out to rank the genes in exactly the same order as the p-values which are driven by the perspective of the null hypothesis of no differential expression. Such methods implicitly assume that the most extreme p-values point to the biologically most relevant genes (von Heydebreck et al., 2004). Bickel (2004) however recognizes that biologists often seek a minimum level of differential expression. Delongchamp et al. (2004) express the need to reliably eliminate the unaffected genes. They

therefore not only examine the FDR but also its counterpart, the false non-discovery rate (FNR, Genovese and Wasserman, 2003), and the fraction of genes not selected among the affected genes (FNS). They still however do not directly estimate nor control the number of truly (in)active genes that were missed in terms of a worthwhile alternative. Similarly, Taylor et al. (2004) introduce the 'miss rate' as a complement to the FDR. This is the proportion of non-null genes in a given interval below the rejection region. They find that a low FDR can accompany quite a high miss rate.

Problems with statistical power also occur in marker assisted selection (MAS). Hospital et al. (1997) recover power by allowing higher type I error rates and increase the significance level when heritability is low. Schön et al. (2004) call for further research about optimal type I error rates in view of the goals of plant breeders. Moerkerke et al. (2006) address this issue by incorporating a biologically relevant target alternative into a marker-specific decision criterion to balance the null and the alternative when selecting genetic markers for MAS.

In this paper, we develop the methodology for selecting genes focusing on a biologically relevant alternative. The key novelty of the approach is that the target effect, which is typically specified for power and sample size calculations, is directly involved in the decision criterion. To achieve this, we propose a measure of relative evidence against the null of no effect and the specified alternative, which ranks genes in terms of their corresponding promise. To decide on a cutoff point, we balance the classical p-value, p_0, and its counterpart under the alternative, p_1, by optimizing a weighted average of gene-specific type I and type II error rates resulting in a different rejection region for each gene. This procedure is called the balanced test. As gene-specific variances imply different levels of information, this gene-specific manner of optimization provides a ranking that can never

be achieved by classical p_0-values or test statistics. The strategy followed here aims to prevent researchers from finding only modest effects in a second phase because relevant signals were not picked up when screening all genes. As in Delongchamp et al. (2004), we give user-specific weights to the null and the alternative, reflecting the relative cost of false positives and false negatives but although they emphasize the importance of FNS, their decision criteria are still based on p-value cutoffs.

We apply the new strategy to detect the genes that are differentially expressed between two types of breast cancer and study the corresponding experimentwise operating characteristics. This involves three levels of gene-expression: genes stemming from the null, those with an effect of at least the alternative and non-null genes situated in between. Efron (2004) argues why we may indeed need to consider a whole distribution of null effects. We thus select a substantially different set of genes from those selected by more traditional methods and obtain a different ranking of these genes. We find that the genes selected reveal a more striking separation between the distribution of expression levels. In section 2, we present the data, introduce a measure of relative evidence against the null and the alternative and develop a formal two-sided testing procedure which incorporates p_1, complement to the measure of significance. In section 3, the methodology is applied to the publicly available breast cancer data set of Hedenfalk et al. (2001). We derive experimentwise operating characteristics in section 4 and compare our results with the more standard approach of controlling the FDR.

2. PROBLEM SETTING AND METHODOLOGY

2.1 Data

The search for genes associated with hereditary susceptibility to breast cancer has led to the identification of the BRCA1 and BRCA2 genes. Detecting genes that are differentially expressed between these types of tumors allows to discriminate between both cancers based on gene expression profiles and has the potential to further extend our understanding of gene expression of various cancer cells (Hedenfalk et al., 2001).

In this paper, we develop methodology for powerful gene selection and analyze data on gene expression profiles of BRCA1- and BRCA2-mutation-positive tumors of Hedenfalk et al. (2001) also considered by Storey and Tibshirani (2003). Data for 3 226 genes is available through 7 arrays with the BRCA1 mutation, 8 arrays with the BRCA2 mutation and 7 arrays of sporadic breast cancer. As in Storey and Tibshirani (2003), we restrict the data to 3 170 genes that have no measurements exceeding 20, which is several interquartile ranges away from the interquartile range of all data. The sporadic breast cancer samples are not considered. Expression values are analyzed on the $\log_2$-scale. Information on the data is available on `http://research.nhgri.nih.gov/microarray/NEJM_Supplement/`. We analyze the expression data to identify genes that are differentially expressed between BRCA1- and BRCA2-mutation-positive tumors using evidence against the null but also against a specified alternative.

2.2 Notation

We adopt the following notation:

- Δ_j is the population contrast of interest between the outcome distributions F_{1j} and F_{2j} of gene j in the BRCA1 and BRCA2 group ($j = 1, \ldots, 3\,170$). In general, this may be a difference in means, a relative risk or a ratio of variances. We will consider the absolute difference in mean $\log_2$ expression values between both tumor groups. Hence, with μ_{kj} the mean $\log_2$ expression level of gene j and $k = 1, 2$ for the BRCA1 and BRCA2 group respectively, $\Delta_j = |\mu_{2j} - \mu_{1j}|$.
- Δ^1 is the predefined target magnitude for Δ_j we wish to detect.
- $n_1 = 7$ and $n_2 = 8$ are the number of arrays in the BRCA1 and BRCA2 group respectively.
- x_{klj} is the expression value on the $\log_2$-scale of gene j in sample l in the BRCAk group ($l = 1, \ldots, n_k$; $k = 1, 2$) with sample mean $\overline{x}_{kj} = (1/n_k) \sum_{l=1}^{n_k} x_{klj}$ and sample variance $s^2_{kj} = (1/(n_k - 1)) \sum_{l=1}^{n_k} (x_{klj} - \overline{x}_{kj})^2$. The corresponding population variance is denoted as σ^2_{kj}.
- fc_j is defined as $\overline{x}_{2j} - \overline{x}_{1j}$, the observed fold change for each gene j.

Basic for the balanced test is the definition of a target alternative. In gene-expression studies, a twofold change, that is a mean difference of 1 in $\log_2$ expression values ($|\text{fc}_j| = 1$), is often considered of interest. Since original values are $\log_2$-transformed, this change implies a geometric means ratio of 2. We will target $\Delta^1 = 1$ for the **true** underlying effect Δ_j for each gene j ($j = 1, \ldots, 3\,170$).

2.3 An additional measure of significance

As in Moerkerke et al. (2006), we start by performing a one-sided test for $\tilde{\Delta}_j = \mu_{2j} - \mu_{1j}$ of $H_{0j} : \tilde{\Delta}_j = 0$ versus $H_{1j} : \tilde{\Delta}_j = \tilde{\Delta}^1 > 0$ for a given gene j with $\tilde{\Delta}^1$ the target alternative. When outcomes are normally distributed in both tumor groups, $\left(\bar{X}_{2j} - \bar{X}_{1j}\right) \overset{|H_{0j}}{\sim} N\left(0, \sqrt{\sigma^2_{2j}/n_2 + \sigma^2_{1j}/n_1}\right)$ and $\left(\bar{X}_{2j} - \bar{X}_{1j}\right) \overset{|H_{1j}}{\sim} N\left(\tilde{\Delta}^1, \sqrt{\sigma^2_{2j}/n_2 + \sigma^2_{1j}/n_1}\right)$ with $\sqrt{\sigma^2_{2j}/n_2 + \sigma^2_{1j}/n_1}$ the (known) standard error of the difference in sample means $\bar{X}_{2j} - \bar{X}_{1j}$ (fold change), where $\overset{|H_{kj}}{\sim}$ indicates 'has conditional distribution given H_{kj}'. We can derive $p_{0j} = P(\bar{X}_{2j} - \bar{X}_{1j} > \text{fc}_j | H_{0j})$ and $p_{1j} = P(\bar{X}_{2j} - \bar{X}_{1j} \leq \text{fc}_j | H_{1j})$. p_{0j} represents the classical p-value calculated from the perspective of the null while p_{1j} is an alternative p-value for testing H_{1j} versus H_{0j}; it is called a measure of impotence. Genes with small p_{0j} **and** large p_{1j} are of interest. The calculation of both types of p-values are depicted in Figure 1. When performing multiple tests, the gene-specific standard errors will eventually lead to gene-specific rejection regions for $\bar{X}_{2j} - \bar{X}_{1j}$.

Figure 1 about here

To extend the methodology to a two-sided test of H_{0j} versus $H_{1j} : |\mu_{2j} - \mu_{1j}| = \Delta^1$ for each gene j, several considerations must be made. Corresponding p-values are

- $p_{0j} = P(|\bar{X}_{2j} - \bar{X}_{1j}| > |\text{fc}_j| \,\big|\, H_{0j})$, the classical two-sided p-value for testing $H_{0j} : \mu_{2j} - \mu_{1j} = 0$ versus $H^{(0)}_{1j} : \mu_{2j} - \mu_{1j} \neq 0$.
- $p_{1j_>} = P(\bar{X}_{2j} - \bar{X}_{1j} \leq \text{fc}_j \,\big|\, H^{(1)}_{1j})$, the p_1-value for testing $H^{(1)}_{1j} : \mu_{2j} - \mu_{1j} = \Delta^1$ versus $H^{(1)}_{0j} : \mu_{2j} - \mu_{1j} < \Delta^1$.
- $p_{1j_<} = P(\bar{X}_{1j} - \bar{X}_{2j} \leq -\text{fc}_j \,\big|\, H^{(2)}_{1j})$, the p_1-value for testing $H^{(2)}_{1j} : \mu_{1j} - \mu_{2j} = \Delta^1$ versus $H^{(2)}_{0j} : \mu_{1j} - \mu_{2j} < \Delta^1$.

Unlike the traditional p_0-value, the alternative p_1-values are one-sided as they measure evidence against the given alternative in the direction of the null only. Ultimately, only 2 p-values per gene are useful: its p_0-value and the maximum of $p_{1j_>}$ and $p_{1j_<}$, which we call the p_1-value, as the performed tests are two-sided. This means that p_{1j} is $p_{1j>}$ when $\text{fc}_j > 0$ and $p_{1j<}$ otherwise.

In practice, we must allow for unknown variances in the computation of the p-values and given our small sample sizes ($n_1 = 7$ and $n_2 = 8$), we do not rely on distributional assumptions but use a permutation distribution as in Storey and Tibshirani (2003). After permuting the group indicators for BRCA1 and BRCA2 over all samples we find a standard p_{0j}-value as the appropriate tail probability of the permutation distribution of

$$T_{0j} = \frac{\overline{X}_{2j} - \overline{X}_{1j}}{\sqrt{\frac{S_{1j}^2}{n_1} + \frac{S_{2j}^2}{n_2}}}. \tag{1}$$

Under the alternative hypothesis $H_{1j}^{(1)}$ ($H_{1j}^{(2)}$), the values $X_{2j} - \Delta^1$ (X_{2j}) and X_{1j} ($X_{1j} - \Delta^1$) are exchangeable and we permute them to obtain alternative p-values as corresponding tail probabilities of the test statistics

$$T_{1j} = \frac{\overline{X}_{2j} - \overline{X}_{1j} - \Delta^1}{\sqrt{\frac{S_{1j}^2}{n_1} + \frac{S_{2j}^2}{n_2}}} \tag{2}$$

$$T_{2j} = \frac{\overline{X}_{1j} - \overline{X}_{2j} - \Delta^1}{\sqrt{\frac{S_{1j}^2}{n_1} + \frac{S_{2j}^2}{n_2}}}. \tag{3}$$

Technical details about deriving the permutation based p-values are given in appendix A.

Although the distributions of the test statistics T_{kj} could be gene-specific, we pool the permutation distributions over all genes and use a common distribution to derive p-values for each gene (cfr. Storey and Tibshirani, 2003 for the distribution of (1)). In

that way the distributions of (1), (2) and (3) become mixtures of all corresponding gene-specific distributions. As in Taylor et al. (2004), the distribution of T_{0j} can then be seen as the null distribution of a typical inactive gene. Following the same reasoning, the distributions of T_{1j} and T_{2j} are the null distributions of back transformed active genes with effect $\mu_{2j} - \mu_{1j} = \Delta^1$ and $\mu_{1j} - \mu_{2j} = \Delta^1$, respectively.

Eyeballing the joint distribution of the (p_0, p_1)-values on a scatter plot helps to explore promising genes in an efficient and informative way. Through the joint measures we can avoid dismissing a possibly winning gene because of lack of convincing information in terms of the classical p_0-value. The more traditional volcano plot shows significance versus the magnitude of the observed effect size (Jin et al., 2001). Its effect measure does not account for imprecision however. p_0 ignores the target effect size but p_1 formally incorporates both.

2.4 The balanced test and a relative measure of evidence

The one-sided balanced test of Moerkerke et al. (2006) determines a gene-specific decision criterion by maximizing a gain function which is a weighted average of the gene-specific type I and type II error rates:

$$A_j \times P(\text{Accept } H_{0j}|H_{0j}) + B_j \times P(\text{Accept } H_{1j}|H_{1j}) \qquad j = 1, \ldots, 3\,170, \qquad (4)$$

where A_j and B_j represent the weights given to a correct decision under the null and the alternative, respectively. Consider first testing $H_{0j} : \mu_{2j} - \mu_{1j} = 0$ versus $H_{1j}^{(1)} : \mu_{2j} - \mu_{1j} = \Delta^1 > 0$ $(j = 1, \ldots, 3\,170)$. Maximizing (4) or in this case

$$A_j \times P(\overline{X}_{2j} - \overline{X}_{1j} \leq c_{\text{fc}_j}|H_{0j}) + B_j \times P(\overline{X}_{2j} - \overline{X}_{1j} > c_{\text{fc}_j}|H_{1j}^{(1)})$$

leads to the optimal cutoff c_{fc_j} for gene j on the scale of the fold change. The decision procedure for each gene j then becomes:

$$\text{Accept } H_{0j} \text{ if } \text{fc}_j \leq c_{\text{fc}_j} \text{ and accept } H_{1j}^{(1)} \text{ otherwise.} \tag{5}$$

This decision criterion depends not only on Δ^1 but as in to Delongchamps et al. (2004), also on the relative importance of the null and the alternative as expressed through the weight ratio A_j/B_j. In practice, A_j/B_j can be defined taking under consideration the (relative) costs of type I and type II errors and the odds of the null and the alternative. If a study is conducted to rule out genes having not enough effect and to select a pool of promising genes, the focus will primarily be on the alternative and the ratio A_j/B_j will typically be less than 1. If only a few genes can be further investigated, protecting the null becomes more important resulting in a ratio larger than 1. Less prevalent alternative or null genes can possibly also influence the cost of false negatives or positives and hence the weight ratio. Defining Δ^1 and A_j/B_j is inherent for any good study as it is equivalent with outlining the ultimate goals and corresponding cost analysis.

Once the gene-specific cutoffs c_{fc_j} are obtained, we use them not only to base the decision on but also to rank the genes according to a measure of relative evidence, the R-ratio ($j = 1, \ldots, 3\,170$):

$$\begin{aligned} R_j &= \frac{P(\overline{X}_{2j} - \overline{X}_{1j} > c_{\text{fc}_j} | H_{1j}^{(1)})}{P(\overline{X}_{2j} - \overline{X}_{1j} \leq c_{\text{fc}_j} | H_{0j})} \times \frac{P(\overline{X}_{2j} - \overline{X}_{1j} \leq \text{fc}_j | H_{0j})}{P(\overline{X}_{2j} - \overline{X}_{1j} > \text{fc}_j | H_{1j}^{(1)})} \\ &= \frac{P(\overline{X}_{2j} - \overline{X}_{1j} > c_{\text{fc}_j} | H_{1j}^{(1)})}{P(\overline{X}_{2j} - \overline{X}_{1j} \leq c_{\text{fc}_j} | H_{0j})} \times \frac{1 - p_{0j}^{+}}{1 - p_{1j}}, \end{aligned}$$

where $p_{0j}^{+} = P(\overline{X}_{2j} - \overline{X}_{1j} > \text{fc}_j | H_{0j})$ is the p_0-value corresponding to the one-sided test. (5) is then equivalent to:

Accept H_{0j} if $R_j \leq 1$ and accept $H_{1j}^{(1)}$ otherwise.

The relative evidence measure positions the observed fold change with respect to the optimal cutoff through the ratio of $(1-p_{0j}^{+})/(1-p_{1j})$ multiplied by a scale correction recognizing the different variance structures and rejection regions across genes. As a larger $(1-p_{0j}^{+})/(1-p_{1j})$ implies a smaller one-sided p_0-value and/or a larger p_1-value, a larger R_j reflects more evidence against the null relative to the evidence against the alternative.

For two-sided tests, we adapt the one-sided test strategy. For genes with $\text{fc}_j = \overline{x}_{2j} - \overline{x}_{1j} > 0$, the optimal cutoff and R-ratio for testing $H_{0j} : \mu_{2j} - \mu_{1j} = 0$ versus $H_{1j}^{(1)} : \mu_{2j} - \mu_{1j} = \Delta^1 > 0$ is determined as described above implying that two-sided p_0-values are replaced by the one-sided counterparts p_{0j}^{+}, again calculated using the permutation distribution of (1). Likewise, the optimal cutoff and R-ratio for genes with $\text{fc}_j = \overline{x}_{2j} - \overline{x}_{1j} \leq 0$ are determined for testing $H_{0j} : \mu_{2j} - \mu_{1j} = 0$ versus $H_{1j}^{(2)} : \mu_{1j} - \mu_{2j} = \Delta^1 > 0$. The H_{0j} and $H_{1j}^{(2)}$ distributions of $\overline{X}_{1j} - \overline{X}_{2j}$ are used and fc_j is replaced by $-\text{fc}_j$. The one-sided p_0-values are now $p_{0j}^{-} = P(\overline{X}_{1j} - \overline{X}_{2j} > -\text{fc}_j | H_{0j})$. Computationally, the optimal cutoffs are obtained on the scale of the test statistics (1), (2) and (3) and are based on permutation distributions (see appendix B for further details).

Note that in a strictly one-sided testing framework $A_j/B_j = 1$ means that the null and the alternative are equally important implying equal probabilities of making a type I and type II error or $\alpha_j = \beta_j$ when testing gene j. By determining the cutoff in the same way for a two-sided test, we actually increase α_j while keeping β_j fixed. It follows that the choice $A_j/B_j = 1$ no longer gives equal weights to the null and alternative.

3. RESULTS

3.1 Descriptive analysis

The left hand side of Figure 2 is a volcano plot for the 3 170 genes in the breast cancer data set showing significance as classical p_0-values against the *observed* effect (observed mean difference or fold change). The right hand side of Figure 2 shows p_1-values ($\Delta^1 = 1$) on the x-axis instead of observed fold changes. As expected, many genes with large p_0-values carry small p_1-values containing no evidence against the null and strong evidence against the alternative. However, we also find that some small (large) p_0-values still correspond to small (large) p_1-values. Indeed, in Figure 3, a large range of p_1-values follows the small p_0-values. The further the (p_0, p_1)-values are in the upper left corner on this plot, the more promising the genes.

Figure 2 about here

Figure 3 about here

3.2 Formal analysis

For this particular analysis with $\Delta^1 = 1$, we have chosen $A_j/B_j = 10$ putting more weight on a correct decision under the null given the expected low proportion of true alternatives present (see section 4). $A_j/B_j = 10$ also results in the selection of a similar number of genes as with standard approaches which facilitates comparison. Investigating the effect of different weight ratios on performance measures or error rates can help choosing an appropriate (possible gene-specific) weight ratio, taking into account the

costs following a wrong decision. This is discussed in section 4 where we compare results following $A_j/B_j = 10$ and $A_j/B_j = 1$.

Here, we illustrate the qualitative and quantitative difference between our procedure and methods based on the classical p_0-value. The q-value described in Storey and Tibshirani (2003) measures the minimum FDR that can be obtained when calling a gene (and all genes corresponding to smaller p_0-values) significant. While q-values are useful to control the FDR at a desired level, they increase in the same order as classical p_0-values. Therefore, decision tools based on a q-value cutoff carry the implicit assumption that smaller p_0-values imply more interesting genes.

The balanced test selects 333 genes with a relative evidence measure R_j ($j = 1, \ldots, 3\,170$) larger than 1. On the other hand, 319 genes have a q-value q_j ($j = 1, \ldots, 3\,170$) smaller than or equal to 10%. Using this as a decision criterion, the FDR is kept below 10%. Storey and Tibshirani (2003) remark that a q-value cutoff is arbitrary and that no typical values can be recommended. We use 10% to facilitate comparison. Table 1 shows how 144 of the genes are selected with one of the two procedures only, confirming the different philosophy of the strategies.

Table 1 about here

Table 2 shows the p_0- and p_1-values, observed fold change and R-ratio for the top 10 genes based on the p_0-rank versus the top 10 genes as ranked by the R-ratio which incorporates p_0 and p_1. Our method does not only select a different subset of genes but also ranks these genes entirely differently. It is for instance striking that the gene with the smallest p_0-value is ranked as number 137 due to its relatively small p_1-value

and absolute fold change. This illustrates that biologically relevant effects may not get a favorable ranking using p_0-values when the observed effect is more variable. This difference in ranking is further elucidated in Figure 4. Boxplots clearly show greater distance in distribution of expression values between both tumor groups for the top R-gene than for the top p_0-gene.

Table 2 about here

Figure 4 about here

In line with the findings of Storey and Tibshirani (2003), most of the genes selected with the balanced test are overexpressed in the BRCA1 group (222 of the 333 genes). Storey and Tibshirani state that for example the MSH2 gene (clone 32790) is the gene with the eighth smallest p_0 (0.51×10^{-4}). They note that this p_0-value reflects evidence against the null but that the q-value allows quantification of the t-test statistic being unlikely for a differentially expressed gene. The estimated q-value is 0.013 for this gene which means that, accounting for the number of tests, 1.3% of the genes with smaller p_0-values are expected to be false positives. The p_1-value of 0.72 for the MSH2 gene is undiluted by the high number of tests and indicates that the observed mean difference is not unlikely to stem from the alternative Δ^1. This is exactly what we are targeting. The p_1-value thus quantifies directly what the q-value is claimed to do in this latter example.

4. EXPERIMENTWISE OPERATING CHARACTERISTICS FOR THE BALANCED TEST

As the balanced test optimizes a gain function for each separate gene, it does not directly protect a single predefined error measure as do other multiple testing procedures. In this section, we estimate experimentwise error rates of the balanced procedure applied to the breast cancer data for two different weight ratios. We underline the different approach followed here, we define multiple testing measures of interest and compare the balanced test with the q-value approach. Although defining an effect of interest Δ^1 should be rather straightforward, choosing an appropriate weight ratio A_j/B_j can be a challenge. The approach in this section to estimate multiple testing measures of interest may also be used to construct an ROC-like curve for these measures corresponding to different weight ratios.

The possible outcomes of the m tests of the sharp null of no differential expression for $H_{0j} : \mu_{2j} - \mu_{1j} = 0$ for each test (gene) j ($j = 1, \ldots, m$) versus any non-null alternative $H_{1j} : \mu_{2j} - \mu_{1j} \neq 0$, are typically summarized as in table 3 where various error rates can be derived. Let F and T be the number of false and true positives and V and W represent the number of true and false negatives.

To evaluate the performance of the balanced test, we should consider both the sharp null $H_{0j}^S : \mu_{2j} - \mu_{1j} = 0$ versus the broad alternative $H_{1j}^B : \mu_{2j} - \mu_{1j} \neq 0$ and the sharp alternative $H_{1j}^S : |\mu_{2j} - \mu_{1j}| = \Delta_j = \Delta^1$ versus the broad null $H_{0j}^B : |\mu_{2j} - \mu_{1j}| = \Delta_j < \Delta^1$. Standard testing procedures focus on rejecting H_{0j}^S or any non-null difference. As our main concern is to first and foremost detect effects of at least Δ^1 and since genes under H_{1j}^S reflect those effects, our focus is on H_{1j}^S. A true positive then occurs when this target effect is selected while a true negative is the non-selection of a smaller effect. It therefore

makes no sense to evaluate our procedure by means of table 3 and we combine results into a 2×3-table as in table 4. False positives are defined both from the perspective of the sharp null (F_S) and the broad null (F_B). The same distinction is made for the number of true negatives (V_S and V_B) while true positives (T_S) and false negatives (W_S) are only defined from the perspective of H_{1j}^S.

Table 3 about here

Table 4 about here

To estimate the number of strict null genes (m_{0S}), we use the spline procedure of Storey and Tibshirani (2003) which leans on the uniform distribution over $[0, 1]$ of the p_0-values under the strict null to estimate the proportion of true nulls $P(H_{0j}^S)$. The number of genes stemming from the alternative (m_{1S}) is obtained in the same way but by using the bootstrap version (Storey, 2002) and plugging in the $p_{1>}$-values and $p_{1<}$-values calculated from the perspective of $H_{1j}^{(1)}$ ($\mu_{2j} - \mu_{1j} = \Delta^1$) and $H_{1j}^{(2)}$ ($\mu_{1j} - \mu_{2j} = \Delta^1$) to estimate the proportion of true target alternatives $P(H_{1j}^S) = P(H_{1j}^{(1)}) + P(H_{1j}^{(2)})$. We find an estimate of 0.66 for the proportion of true nulls and 0.053 for the proportion of true alternatives.

$E[F_S]/m_{0S}$ is the average type I error rate for H_{0j}^S versus H_{1j}^B. Likewise, $E[W_S]/m_{1S}$ is the average probability of a type II error when considering H_{1j}^S versus H_{0j}^B. These quantities are estimated by calculating the type I and type II error rate for the test associated with each gene and by averaging over all genes. The type I error rates for a q-value approach follow immediately from the corresponding p_0-value cutoff while for

the balanced procedure, each test has a different type I error rate following the different cutoffs. Type II error rates for all decision strategies are obtained in a similar way as the p_1-values of the cutoffs but accounting for the two-sided nature of the tests. Permutation null and alternative distributions are used to derive these error rates. Details are given in appendix C.

Following the construction of 2×3-tables as in table 4 for the different decision strategies, we introduce several performance measures of interest in line with testing H_{1j}^S versus H_{0j}^B:

- TPR: the true positive rate estimated by $\hat{t}_S/m_{1S}$. Note that this corresponds to 1 minus the average type II error rate defined above.
- TNR: the true negative rate estimated as $(\hat{v}_S + \hat{v}_B)/(m_{0S} + m_{0B})$.
- π_T: the proportion of true targets estimated by $\hat{t}_S/N$.
- π_B: the proportion truly below the target estimated as $(\hat{v}_S + \hat{v}_B)/(m - N)$.
- $\pi_{T|\overline{O}}$ the proportion of the target alternative among the non-null features that are selected. This quantity is estimated by $(\hat{t}_S)/(\hat{t}_S + \hat{f}_B)$.

By putting more weight on the alternative or by lowering the cutoff for the relative evidence measure, more genes stemming from the alternative will be selected in trade off with a higher amount of false positives. The effect of the weight ratio is illustrated below. Remember that a cutoff of 1 for R optimizes the gain function. Altering the weights will also change the order in which genes are selected while a different R-ratio cutoff would only change the number of genes selected but not the ordering.

In this section, we investigate the performance of the balanced procedure for $A_j/B_j = 10$ compared to $A_j/B_j = 1$. Data analysis follows several steps:

1. Choose the alternative Δ^1; in our case $\Delta^1 = 1$.
2. Obtain a p_0-value and p_1-value for each gene.
3. Optimize a gene-specific gain function with weight ratio (null versus alternative) A_j/B_j.
4. Calculate relative measures of evidence or R-ratios based on p-values in step 2 and optimal cutoffs from step 3.
5. Select genes with $R_j > 1$.
6. Estimate $P(H_{0j}^S)$ and $P(H_{1j}^S)$. This step is independent from A_j/B_j.
7. Calculate the average type I error (under H_{0j}^S) and type II error rate (under H_{1j}^S).
8. Estimate 2×3-table as in table 4 and the corresponding performance measures.

Both analyses are compared with two q-value strategies, one that selects a similar amount of genes and one with a comparable classical FDR (estimated as $\hat{f}_S/N$).

4.1. Analysis 1: $A_j/B_j = 10$

We compare 3 different decision strategies: select a gene when

- its q-value is smaller than or equal to 0.10. 319 genes are selected and the estimated average type I error and type II error rate equal 0.015 and 0.27, respectively.

- its q-value is smaller than or equal to 0.15. This approach selects 511 genes with an estimated average type I and type II error rate of 0.036 and 0.18.
- the relative evidence measure is larger than 1 (with $A_j/B_j = 10$ for $j = 1, \ldots, m$). 333 genes are selected with an estimated average type I and type II error rate of 0.023 and 0.19.

Results are in table 5 and 7.

Table 5 about here

Slightly better than the first strategy with respect to π_T and comparable with the second strategy with respect to π_B, the balanced test scores convincingly better on $\pi_{T|\overline{O}}$. This results from gathering not only evidence against the null but also against the alternative.

In terms of the classical FDR and FNR derived from testing H_{0j}^S versus H_{1j}^B, our method is situated between the 2 q-value strategies. The FNR is estimated as $(\hat{v}_B + \hat{w}_S)/(m-N)$. The FDR's for the q-value approaches follow immediately from the decision criterion and are equal to 0.10 and 0.15. We obtain an estimated FDR of 0.145 for the balanced test. The estimates for the classical FNR are 0.28, 0.24 and 0.28 respectively for the 3 approaches. Both classical error rates are however of less interest for the approach followed here as false positives include also selected genes with an effect smaller than Δ^1 and false negatives are only the genes with an effect of at least Δ^1 that are not selected.

The main result that shows the discrepancy between the balanced test and procedures based on the ordering of classical p_0-values is the detection of an expected amount of 136 alternative genes on a total of 333. Using the q-value approach, 511 genes need

to be selected to obtain a similar result. One could argue however that our method with $A_j/B_j = 10$ selects a higher expected number of strict null genes than the q-value approach with $q \leq 0.10$ (48.12 versus 31.38) while these genes are of no interest at all. But by doing that, we also select a higher expected number of true targets (136.09 versus 122.65) that we are directly aiming at. This is where the role of the weight ratio steps in: a different weight ratio would reflect a different trade-off. Moreover, the motivation for including the target effect is that genes stemming from the broad null are not the biologically relevant ones. Efron (2004) also addresses the choice of a null hypothesis and estimates a distribution of observed null and alternative effects. Performing many tests allows estimation of an empirical null hypothesis as it is in some cases not realistic to work under the strict null. These issues force us to rethink the definition of an appropriate alternative.

By incorporating the alternative directly in the decision criterion and making the decision criterion gene-specific, it is obvious that we score better than any method based on the classical p_0-values. Shifting a p_0-value cutoff will never achieve the same balance between true positives and true negatives.

4.2. Analysis 2: $A_j/B_j = 1$

We compare 3 different decision strategies: select a gene when

- its q-value is smaller than or equal to 0.234. 834 genes are selected and the estimated average type I error and type II error rate equal 0.093 and 0.10, respectively.
- its q-value is smaller than or equal to 0.299. This approach selects 1 064 genes with an estimated average type I and type II error rate of 0.15 and 0.071.

- the relative evidence measure is larger than 1 (with $A_j/B_j = 1$ for $j = 1, \ldots, m$). 834 genes are selected with an estimated average type I and type II error rate of 0.12 and 0.052.

Results are in table 6 and 7.

Table 6 about here

Table 7 about here

As expected, a weight ratio with more weight on the alternative than in the first analysis results in a higher number of true targets that are selected in trade off with a higher number of false positives. Again, we find that the q-value approach needs to select a higher number of genes to achieve the same number of true alternatives which results in a higher false positive rate. The q-value approach that selects the same amount of genes has a smaller true positive rate. From this follows that our method is the most powerful to detect an effect of 1 while it achieves the smallest number of false positives (effects smaller than 1).

For the balanced test with $A_j/B_j = 1$, equal weights are given to the null and the alternative in the optimization procedure and this corresponds to an equal type I and type II error rate for each gene in the one-sided testing framework. This means $R_j > 1$ is equivalent with selecting a gene when its corresponding one-sided p_0-value (p_{0j}^+ for $\text{fc}_j > 0$ and p_{0j}^- otherwise) is smaller than its p_1-value. However, we have performed the optimization numerically based on permutation distributions (appendix B) because this enables us to calculate type I and type II error rates. The results following numerical

optimization are very similar to those expected theoretically so this approach poses no problem here.

5. DISCUSSION

We have proposed a new selection procedure following analysis of the association between gene expression and phenotype, which balances evidence against the null with evidence against a specified alternative of interest. The methodology is more generally applicable in the context of selecting genes/genetic markers which play an important role in a trait of interest. The approach takes into account the (context-specific) relative importance of type I and type II errors and results in a relative evidence measure R (comparable with a likelihood ratio) according to which genes are ranked. An optimal cutoff for R then determines the selection region. Due to gene-specific variance structures as well as possibly gene-specific losses accompanying type I and type II errors, this entails a different nominal alpha and beta level per gene. As a result, the order in which genes are selected can differ dramatically from the standard p-value generated order. Marker-specific loss functions are also very natural in the marker assisted selection (MAS) context where a highly prevalent marker in the population, leaves little to gain by its introduction in the population. Investigation of performance measures can help to evaluate the weights given to the sharp null and alternative. In this particular data set, very few genes are expected to stem from the alternative, justifying (although not necessary or restrictive) a large weight on the null.

The idea that different null hypotheses are rejected at different significance levels to exploit varying levels of informativeness is not new. It is for instance reflected in the recent work of Wei and Uno (2005), who control global coverage over many association analyses,

when giving separate confidence intervals for different gene effects their own coverage level. von Heydebreck et al. (2004) discuss moderated t-test statistics of Baldi and Long, 2001, Tusher et al., 2001, Lönnstedt and Speed, 2002 and Smyth, 2004. These test statistics recover something of the fold change in the selection criterion by augmenting the gene-specific variance estimator in the denominator by a constant to screen out statistically significant genes with small effects in absolute terms. We have estimated the variances based on relatively few samples. The empirical Bayes approach (see for example Lönnstedt and Speed, 2002 and Smyth, 2004) uses a weighted average of the gene-specific variance and the gene-specific variance for each gene. This may indeed be a good alternative to the procedure followed here.

Concerns about sacrificing power to optimize false discovery rates have also been addressed by Ishwaran and Rao (2003) and Lönnstedt and Speed (2002). The BAM (Bayesian ANOVA for microarrays) technique of Ishwaran and Rao aims to strike a balance between between false rejections and false nonrejections. They consider posterior mean values and Bayesian model selection to assess differential expression. Lönnstedt and Speed motivate their Bayesian approach (B-statistic) by stating that there is in general a willigness to permit more false positives in order to avoid too many false negatives. More in particular, they handle a posterior odds of differential expression. We, on the other hand, incorporate the specified alternative as in classical power calculations, contrasting it with the sharp null of no effect.

While our criterion for selection is cast in terms of R and an optimality criterion, resulting experimentwise type I and type II error rates have been derived. Following our philosophy, false negatives are seen when effects at least as large as the target alternative remain undetected, and similarly false positive results select genes with a truly smaller

effect. Corresponding rates have been estimated under a fixed alternative. The need to specify a target alternative was resolved quite easily in our application. This may be harder in other contexts, for instance when looking for gene-gene interactions. This warrants further research. Very recently, Norris et al. (2006) also consider balanced testing in which they propose to give penalties to false positives and negatives. Contrary to our philosophy, their main concern is that standard FDR controlling procedures often fail to pick up modest effects. Basic differences with our balanced test is their ranking of genes according to classical t-test statistics and the fact that statistical power is defined using the underlying distribution of alternative genes in the study.

We have taken the popular approach of considering one gene at a time, ignoring correlations. As more biological knowledge becomes available, modeling the joint distribution of gene expression becomes feasible. Decision criteria built on this could be adjusted in line with the balanced test for the genes separately to capture a fold change of interest.

In summary we believe the proposed procedure has great promise in providing a semi-automatic selection procedure allowing to screen many genes for their potential impact on phenotype. It makes particular sense to put much greater emphasis on type II error in a first screening round and concentrate more on type I errors later. The ability to balance the power to detect important alternatives with significance at the selection level meets the need of many researchers. We hope to have provided a useful procedure which has this direct focus.

APPENDIX A. PERMUTATION BASED *P*-VALUES

The classical two-sided p-value for each gene j ($j = 1, \ldots, 3\,170$) equals

$$p_{0j} = P(|T_{0j}| > |t_{0j}|\,|H_{0j})$$

where the distribution of T_{0j} under H_{0j} is obtained using permutations and t_{0j} is the observed test statistic in (1). In total 100 permutations ($b = 1, \ldots, 100$) are performed by randomly permuting the labels of the tumor groups. In each permutation step, the 3 170 t-test statistics t_{0j} are re-computed constructing a non-parametric null distribution for each gene conditional on the observed data. The p_0-values are then calculated using the distribution of T_{0j} constructed by these null statistics over all genes ($100 \times 3\,170$ in total):

$$p_{0j} = \sum_{b=1}^{100} \frac{\#\{i : |t_{0i}^b| > |t_{0j}|; i = 1, \ldots, 3\,170\}}{100 \times 3\,170}$$

with t_{0i}^b the permutation based t-test statistic t_{0i} for gene i in permutation step b. This implies that a global null distribution is considered for all genes.

The alternative p_1-values of interest are

$$p_{1j_>} = P(T_{1j} \leq t_{1j}|H_{1j}^{(1)}) \text{ and } p_{1j_<} = P(T_{2j} \leq t_{2j}|H_{1j}^{(2)})$$

where the distribution of T_{1j} under $H_{1j}^{(1)}$ and of T_{2j} under $H_{1j}^{(2)}$ is again obtained using permutations and t_{1j} and t_{2j} are the observed test statistics in (2) and (3), respectively. The distribution of T_{1j} (T_{2j}) is constructed by first subtracting Δ^1 from each expression level in the BRCA2 (BRCA1) group and by randomly permuting the labels of the tumor groups afterwards. Under the hypothesis that $\mu_{2j} - \mu_{1j} = \Delta^1$ ($\mu_{1j} - \mu_{2j} = \Delta^1$) for $j = 1, \ldots, 3\,170$, there is no differential gene expression anymore when subtracting Δ^1 from the observed expression levels in the BRCA2 (BRCA1) group. Hence, the t-test

statistic t_{0j} $(-t_{0j})$ in (1) calculated based on these transformed data after subtracting Δ^1, which we refer to as t^*_{0j} (t^{**}_{0j}), should follow the same distribution regardless of how the group assignments are made. Therefore, the alternative p_1-values are obtained as follows:

$$p_{1j_>} = \sum_{b=1}^{100} \frac{\#\{i : t^{*b}_{0i} \leq t_{1j}; i = 1, \ldots, 3\,170\}}{100 \times 3\,170}$$
$$p_{1j_<} = \sum_{b=1}^{100} \frac{\#\{i : t^{**b}_{0i} \leq t_{2j}; i = 1, \ldots, 3\,170\}}{100 \times 3\,170}$$

with t^{*b}_{0i} (t^{**b}_{0i}) the permutation based t-test statistic t^*_{0i} (t^{**}_{0i}) for gene i in permutation step b. It follows that global alternative distributions are considered for all genes. Note that t^*_{0j} (t^{**}_{0j}) based on the unpermuted data equals t_{1j} (t_{2j}). To determine the distribution of T_{1j} (T_{2j}), Δ^1 has to be subtracted first since group labels then become exchangeable. If labels are permuted and then a permutation based t_{1j} (t_{2j}) in (2) ((3)) is calculated, we will not obtain the appropriate distribution.

APPENDIX B. DETERMINING OPTIMAL CUTOFFS FOR THE BALANCED TEST

As we are interested in testing a two-sided alternative hypothesis for each gene, we introduce the following notation:

$\Delta_{Aj} = \mu_{2j} - \mu_{1j}$ and $a = 1$ when $\text{fc}_j > 0$,

$\Delta_{Aj} = \mu_{1j} - \mu_{2j}$ and $a = 2$ otherwise, $(j = 1, \ldots, 3\,170)$.

For each gene j, we test:

$$H_{0j} : \Delta_{Aj} = 0 \text{ versus } H^{(a)}_{1j} : \Delta_{Aj} = \Delta^1.$$

The final decision criterion is expressed in terms of a cutoff on the scale of the t-test statistics. This optimal cutoff is determined by optimizing

$$A_j \times P(I_{\text{fc}_j} \times T_{0j} \leq c_j | H_{0j}) + B_j \times P(T_{Aj} > c_j^+ | H_{1j}^{(a)})$$

with $I_{\text{fc}_j} = 1$, $T_{Aj} = T_{1j}$ when $\text{fc}_j > 0$ and $I_{\text{fc}_j} = -1$, $T_{Aj} = T_{2j}$ otherwise. The type II error rate implied by the rejection region for gene j is calculated similarly to the p_1-value for c_j^+ but accounting for the fact that the performed tests are two-sided. The type I error rate is found as the p_0-value for c_j. Of course, the criterion $I_{\text{fc}_j} \times T_{0j} > c_j$ is equivalent to $T_{Aj} > c_j^+$. The separate notation is only introduced to stress the different scales we are working under (i.e. the null and the alternative).

Optimal cutoffs can be found numerically based on the permutation distributions of T_{0j}, T_{1j} and T_{2j}. For some genes, cutoffs are rather large while very little improvement in expected gain (4) is actually obtained. In such cases, we choose a smaller cutoff for which the expected gain is very close to the maximum. In this application $A_j/B_j = 10$ and the expected gain in (4) is rescaled on a range from 0 to 1 by choosing $A_j = 10/11$ and $B_j = 1/11$. In this way, distances are measured on the same scale for all genes. In the second analysis with $A_j/B_j = 1$, $A_j = B_j = 1/2$.

APPENDIX C. OPERATING CHARACTERISTICS FOR THE BALANCED TEST

C.1. Prevalence of the worthwhile alternative

The estimate for $P(H_{0j}^S)$ equals 0.66 and is obtained based on the classical p_0-values with the algorithm described in Storey and Tibshirani (2003). This procedure leans on

the uniform distribution of the p_0-values under H_{0j}^S. We follow the same reasoning for estimating $P(H_{1j}^S)$ based on the p_1-values. Under $H_{1j}^{(1)} : (\mu_{2j} - \mu_{1j}) = \Delta^1$, the $p_{1>}$-values are uniformly distributed and the $p_{1<}$-values are uniformly distributed under $H_{1j}^{(2)} : (\mu_{1j} - \mu_{2j}) = \Delta^1$. By estimating the prevalence of $H_{1j}^{(1)}$ and $H_{1j}^{(2)}$ based on the $p_{1>}$- and $p_{1<}$-values, we can estimate $P(H_{1j}^S) = P(H_{1j}^{(1)}) + P(H_{1j}^{(2)})$. However, as we are testing H_{1j}^S versus H_{1j}^B, genes having an effect larger than Δ^1 also have large p_1-values. In this respect we are performing a one-sided test as p_1-values are calculated in the direction of the null only. Therefore, we apply the bootstrap procedure of Storey (2002) to obtain estimates for $P(H_{1j}^{(1)})$ and $P(H_{1j}^{(2)})$. Figure 5 shows a density histogram of the 3 170 p_1-values. The horizontal line shows the estimate of 0.053 for $P(H_{1j}^S)$.

Δ^1 represents the alternative we are targeting and not necessarily the true underlying mean of the non-null genes. When many genes stem from an effect larger than Δ^1, this would result in a peak of p_1-values around 1 due to the one-sided test issue mentioned before. If all genes have an effect smaller than Δ^1, there would be no p_1-values close to 1. Both cases may complicate obtaining $P(H_{1j}^S)$. This aspect is the subject of further research. Neither situation seems to be the case here. The histogram density is fairly flat beyond 0.6 and no peak around 1 emerges. The height of this portion is an estimate for the proportion of genes stemming from the alternative. We find that 0.053 seems a reasonable estimate.

Figure 5 about here

C.2. Average type I error rate

When testing $H_{0j}^S : \mu_{2j} - \mu_{1j} = 0$ against $H_{1j}^B : \mu_{2j} - \mu_{1j} \neq 0$ for m genes ($j = 1, \ldots, m$)

the average type I error rate is defined as $E[F_S]/m_{0S}$ or the expected proportion of genes classified under H_{1j}^B among the genes stemming from H_{0j}^S. Using the balanced test, all m marginal tests use a different cutoff for the t-test statistic and hence are performed on a different significance level. If $\mathcal{H}_0$ represents the set of genes stemming from the strict null then

$$\frac{E[F_S]}{m_{0S}} = \frac{\sum_{j\in\mathcal{H}_0}\alpha_j}{m_{0S}}. \tag{6}$$

where α_j represents the significance level for the test for gene j ($j = 1, \ldots, m$).

(6) is estimated as the average significance level over all m genes since

$$\frac{\sum_{j\in\mathcal{H}_0}\alpha_j}{m_{0S}} = \frac{\sum_{j=1}^{m}\alpha_j \times I_{\mathcal{H}_{0j}}}{m_{0S}}$$

with $I_{\mathcal{H}_{0j}} = 1$ if H_{0j}^S is true for gene j. With $\hat{\pi}_0$ an estimator for $P(H_{0j}^S)$, (6) can therefore be estimated as

$$\frac{\sum_{j=1}^{m}\alpha_j\hat{\pi}_0}{\hat{\pi}_0 m} = \frac{\sum_{j=1}^{m}\alpha_j}{m}.$$

The CWER α_j for each gene j is calculated as the p_0-value of the optimal cutoff based on the permutation distribution of T_{0j} which is the permutation distribution of the test statistics under H_{0j}^S. The estimated average type I error rate for the balanced test is 0.023 for $A_j/B_j = 10$ and 0.12 for $A_j/B_j = 1$.

If a single p_0-value cutoff is used for all genes, this cutoff is the average type I error rate. The criterion to select all genes with a q-value lower or equal to 0.10 corresponds to a p_0-value cutoff of 0.015 and a q-value cutoff of 0.15, 0.234 and 0.299 to a p_0-value cutoff of 0.036, 0.093 and 0.15 respectively.

C.3. Average type II error rate

When testing $H_{1j}^S : |\mu_{2j} - \mu_{1j}| = \Delta_j = \Delta^1$ against $H_{0j}^B : |\mu_{2j} - \mu_{1j}| = \Delta_j < \Delta^1$ for all m genes, the average type II error rate is defined as $E[W_S]/m_{1S}$ or the expected proportion of genes classified under H_{0j}^B among the genes stemming from H_{1j}^S. Following the same reasoning as for the average type I error rate, this quantity is estimated as the average type II error rate over all genes:

$$\frac{\sum_{j=1}^{m} \beta_j}{m}$$

with β_j the type II error rate for the test for gene j $(j = 1, \ldots, m)$.

For the balanced test, the type II error rate is based on the optimal cutoff for each gene. For the q-value approaches, the same cutoff is used for all m genes but as the type II error rate also involves the variance structure of the genes, a different type II error rate is obtained for each gene. For the first analysis in section 4, the estimated average type II error rate for the balanced test equals 0.19, 0.27 for the $q \leq 0.10$-approach and 0.18 for the $q \leq 0.15$-approach. For the second analysis, these 3 rates are 0.052, 0.10 and 0.071 respectively.

ACKNOWLEDGEMENT

The authors wish to thank Prof Dr. Ingrid Hedenfalk and co-authors for making their data publicly available.

REFERENCES

Baldi, P. and Long, A.D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**(6),509-519.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **57**(1), 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.

Bickel, D.R. 2004. Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics*, **20**(5), 682-688.

Delongchamp, R.R., Bowyer, J.F., Chen, J.J. and Kodell R.L. 2004. Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics*, **60**(3), 774-782.

Efron, B. 2004. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *J. Am. Stat. Assoc.*, **99**(465), 96-104. Fernando, R.L., Nettleton, D., Southey, B.R., Dekkers, J.C.M., Rothschild, M.F. and Soller, M. 2004. Controlling the proportion of false positives in multiple dependent tests. *Genetics*, **166**(1), 611-619.

Genovese, C. and Wasserman, L. 2002. Operating characteristics and extensions of the false discovery rate procedure.*J. R. Stat. Soc. Ser. B-Stat. Methodol.*,**64**(3), 499-517.

Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent J. 2001. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**(8), 539–548.

Hochberg, Y. and Tamhane, A. 1987. Multiple Comparison Procedures. Wiley & Sons, N. Y.

Hospital, F., Moreau, L., Lacoudre, F., Charcosset, A., Gallais, A. 1997. More on the efficiency of marker-assisted selection. *Theor. Appl. Genet.*, **95**(8), 1181-1189.

Ishwaran, H. and Rao J.S. 2003. Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Am. Stat. Assoc.*, **98**(462), 438-455.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genet.*, **29**(4), 389-395.

Lönnstedt, I. and Speed, T. 2002. Replicated microarray data. *Stat. Sin.*, **12**(1), 31-46.

Moerkerke, B., Goetghebeur, E., De Riek, J., Roldan-Ruiz I. 2006. Significance and Impotence: towards a balanced view of the null and the alternative in marker selection for plant breeding. *J. R. Stat. Soc. Ser. A-Stat. Soc.*, **169**(1), 61-79.

Moreau, L., Charcosset, A., Hospital, F. and Gallais, A. 1998. Marker-assisted selection efficiency in populations of finite size. *Genetics*, **148**(3), 1353-1365.

Norris, A.W., Kahn, C.R. 2006. Analysis of gene expression in pathophysiological states: Balancing false discovery and false negative rates. *Nature Genet.*, **103**(3), 649-653.

Schön, C.C., Utz, H.F., Groh, S., Truberg, B., Openshaw, S. and Melchinger, A.E. 2004. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics*, **167**(1), 485-498.

Smyth, G. 2004. Linear models and Empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3: Article 3.

Storey, J.D. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **64**(3), 479-498.

Storey, J.D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stat.*, **31**(6), 2013-2035.

Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, **100**(16), 9440-9445.

Taylor, J., Tibshirani, R. and Efron, B. 2005. The 'miss rate' for the analysis of gene expression data *Biostatistics*, **6**(1), 111-117.

Tusher, V.G., Tibshirani, R. and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.*, **98**(9), 5116-5121.

Uno, H., Tian, L., Wei, L.J. 2005. The optimal confidence region for a random parameter. *Biometrika*, **92**(4), 957-964.

von Heydebreck, A., Huber, W. and Gentleman, R. 2004. Differential Expression with the Bioconductor Project. *Bioconductor Project Working Papers*, Working Paper 7.

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., Rothman, N. 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J. Natl. Cancer Inst.*, **96**(6), 434-442.

Tables

Table 1: Crosstab comparing the balanced test ($A_j/B_j = 10$) and q-value approach

	$q_j \leq 0.10$	$q_j > 0.10$	Total
$R_j > 1$	254	79	333
$R_j \leq 1$	65	2 772	2 837
Total	319	2 851	3 170

Table 2: Comparing ranks of genes according to classical p_0-values and the relative evidence measure (R-ratio, $A_j/B_j = 10$)

Results for the 10 genes with smallest p_0-values					Results for the 10 genes with largest R-ratios				
$p_0 \times 10^4$	p_1	fc	R-ratio	Rank R-ratio	$p_0 \times 10^4$	p_1	fc	R-ratio	Rank p_0
0.03	0.27	-0.92	1.37	137	0.73	0.99	1.99	106.74	9
0.16	0.37	-0.95	1.58	112	0.41	0.99	-1.72	72.94	6
0.22	0.89	-1.22	9.28	11	0.79	0.97	1.50	27.24	11
0.35	0.82	-1.16	5.54	24	4.35	0.97	1.73	24.42	38
0.38	0.95	-1.36	19.30	6	3.41	0.97	1.67	23.60	31
0.41	0.99	-1.72	72.94	2	0.38	0.95	-1.36	19.30	5
0.44	0.89	-1.25	8.58	13	4.10	0.96	1.59	17.64	35
0.51	0.72	-1.09	3.46	41	1.01	0.93	-1.37	12.79	16
0.73	0.99	1.99	106.74	1	9.75	0.94	1.58	11.54	76
0.73	0.86	-1.21	6.62	21	5.17	0.92	-1.43	9.62	48

Table 3: Possible outcomes of m tests for all $j = 1, \ldots, m$ for $H_{0j} : \mu_{2j} - \mu_{1j} = 0$ versus $H_{1j} : \mu_{2j} - \mu_{1j} \neq 0$

	True null	True non-null	Total
Called significant	F	T	N
Called not significant	V	W	$m - N$
Total	m_0	m_1	m

Table 4: Possible outcomes of m tests for all $j = 1, \ldots, m$ for $H_{0j}^S : \mu_{2j} - \mu_{1j} = 0 \leftrightarrow H_{1j}^B : \mu_{2j} - \mu_{1j} \neq 0$ and $H_{1j}^S : |\mu_{2j} - \mu_{1j}| = \Delta_j = \Delta^1 \leftrightarrow H_{0j}^B : |\mu_{2j} - \mu_{1j}| = \Delta_j < \Delta^1$

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	F_S	F_B	T_S	N
Called not significant	V_S	V_B	W_S	$m - N$
Total	m_{0S}	m_{0B}	m_{1S}	m

Table 5: Estimated 2×3-tables for the 3 decision strategies in analysis 1

Call a gene significant with a q-value ≤ 0.10

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	31.38	164.97	122.65	319
Called not significant	2060.82	744.82	45.36	2 851
Total	2 092.20	909.79	168.01	3 170

Call a gene significant with a q-value ≤ 0.15

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	75.32	297.91	137.77	511
Called not significant	2016.88	611.88	30.24	2 659
Total	2 092.20	909.79	168.01	3 170

Call a gene significant with an R-ratio > 1 ($A_j/B_j = 10$)

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	48.12	148.79	136.09	333
Called not significant	2044.08	761.00	31.92	2 837
Total	2 092.20	909.79	168.01	3 170

Table 6: Estimated 2×3-tables for the 3 decision strategies in analysis 2

Call a gene significant with a q-value ≤ 0.234

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	194.57	488.22	151.21	834
Called not significant	1 897.63	421.57	16.80	2 336
Total	2 092.20	909.79	168.01	3 170

Call a gene significant with a q-value ≤ 0.299

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	313.83	594.09	156.08	1 064
Called not significant	1 778.37	315.70	11.93	2 106
Total	2 092.20	909.79	168.01	3 170

Call a gene significant with an R-ratio > 1 ($A_j/B_j = 1$)

	H_{0j}^S	$H_{0j}^B \cap H_{1j}^B$	H_{1j}^S	Total
Called significant	251.06	423.67	159.27	834
Called not significant	1 841.14	486.12	8.74	2 336
Total	2 092.20	909.79	168.01	3 170

Table 7: Estimated performance measures for the different test strategies

Decision criterion	$q \leq 0.10$	$q \leq 0.15$	$R > 1$ $(A_j/B_j = 10)$
estimated true positive rate (TPR)	0.73	0.82	0.81
estimated true negative rate (TNR)	0.94	0.88	0.93
estimated % true targets (π_T)	0.38	0.27	0.41
estimated % truly below target (π_B)	0.98	0.99	0.99
estimated % of target alternative among selected non-null features ($\pi_{T\|\overline{O}}$)	0.43	0.32	0.48
Decision criterion	$q \leq 0.234$	$q \leq 0.299$	$R > 1$ $(A_j/B_j = 1)$
estimated true positive rate (TPR)	0.90	0.93	0.95
estimated true negative rate (TNR)	0.773	0.70	0.775
estimated % true targets (π_T)	0.18	0.15	0.19
estimated % truly below target (π_B)	0.993	0.994	0.996
estimated % of target alternative among selected non-null features ($\pi_{T\|\overline{O}}$)	0.24	0.21	0.27

Figures

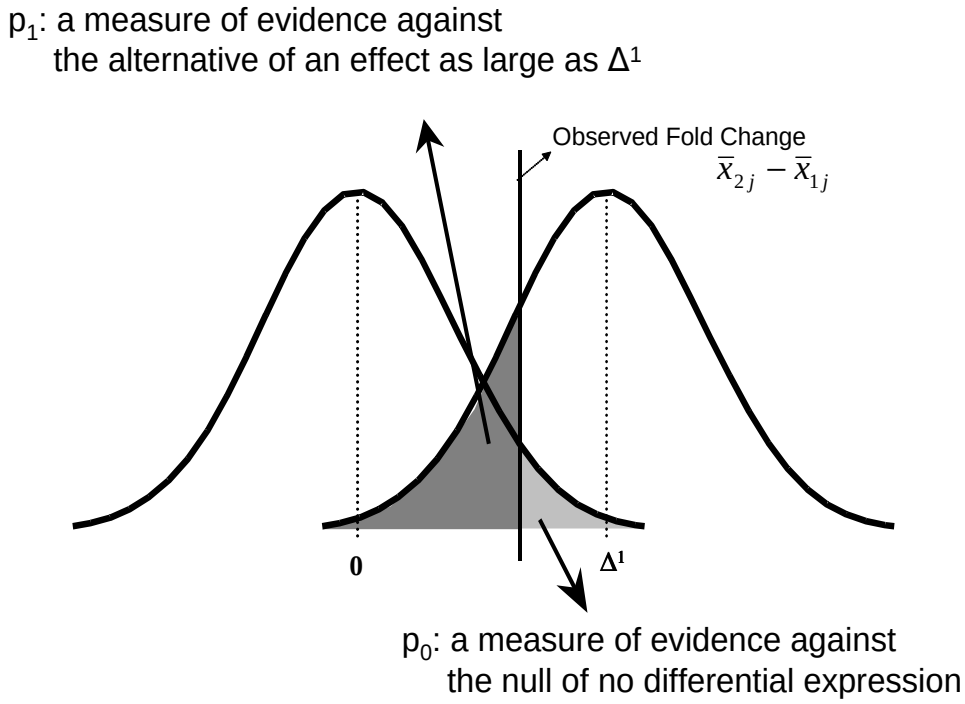

Figure 1: Classical p_0 and alternative p_1: measures of (in)significance and (im)potence

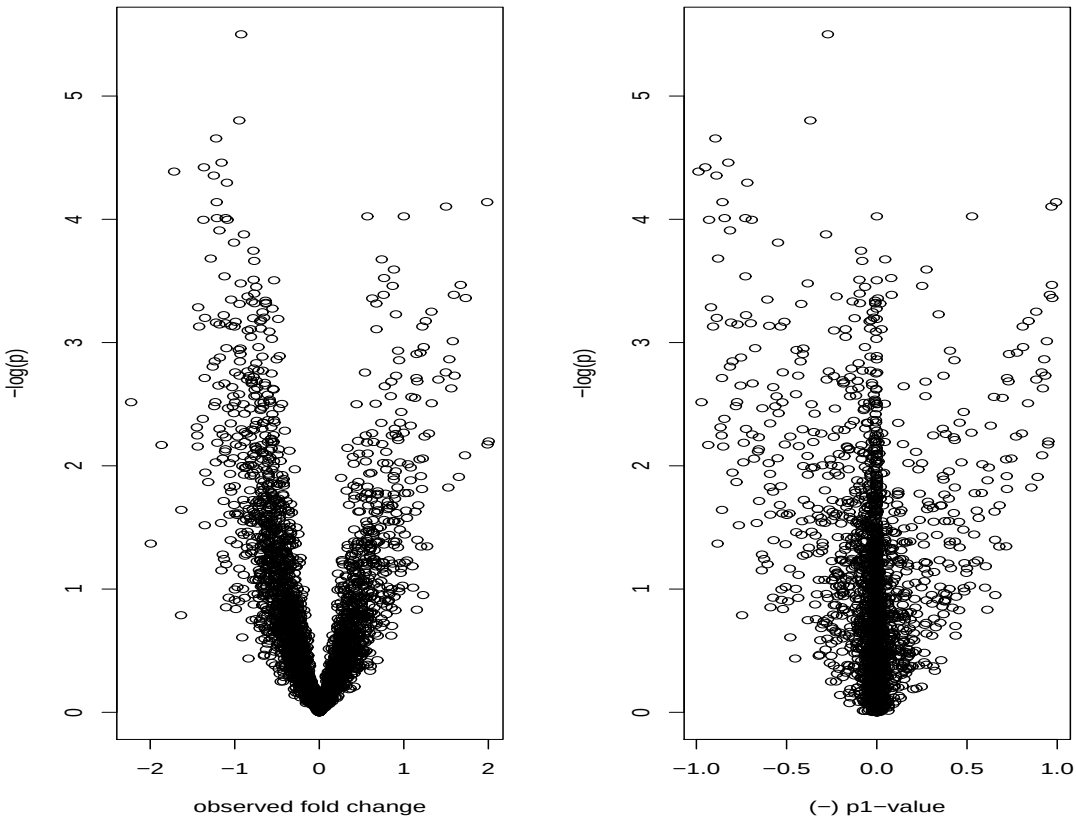

Figure 2: Left: classical volcano plot - Right: volcano plot with p_1 for positive and $-p_1$ for negative observed fold changes on x-axis

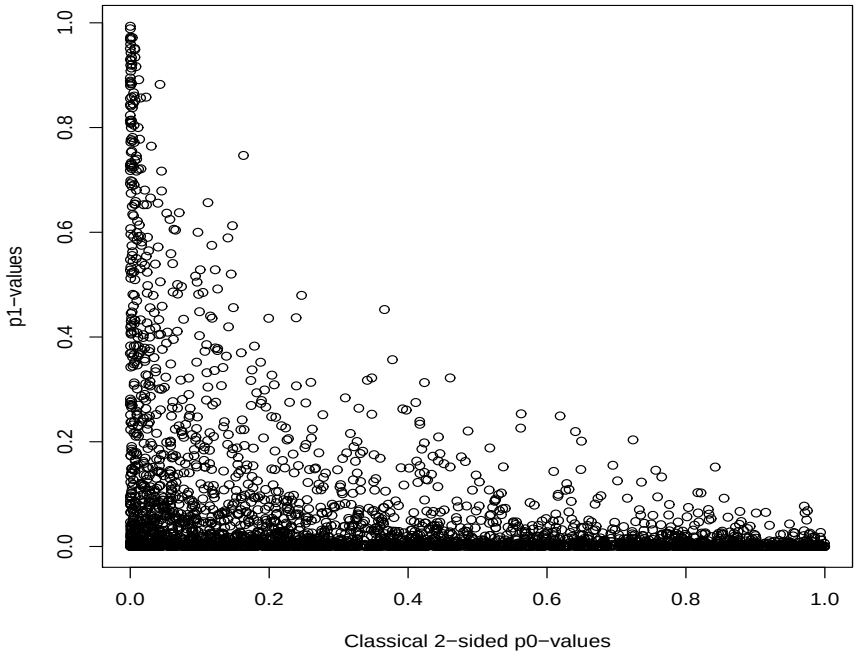

Figure 3: p_1-values versus classical two-sided p_0-values

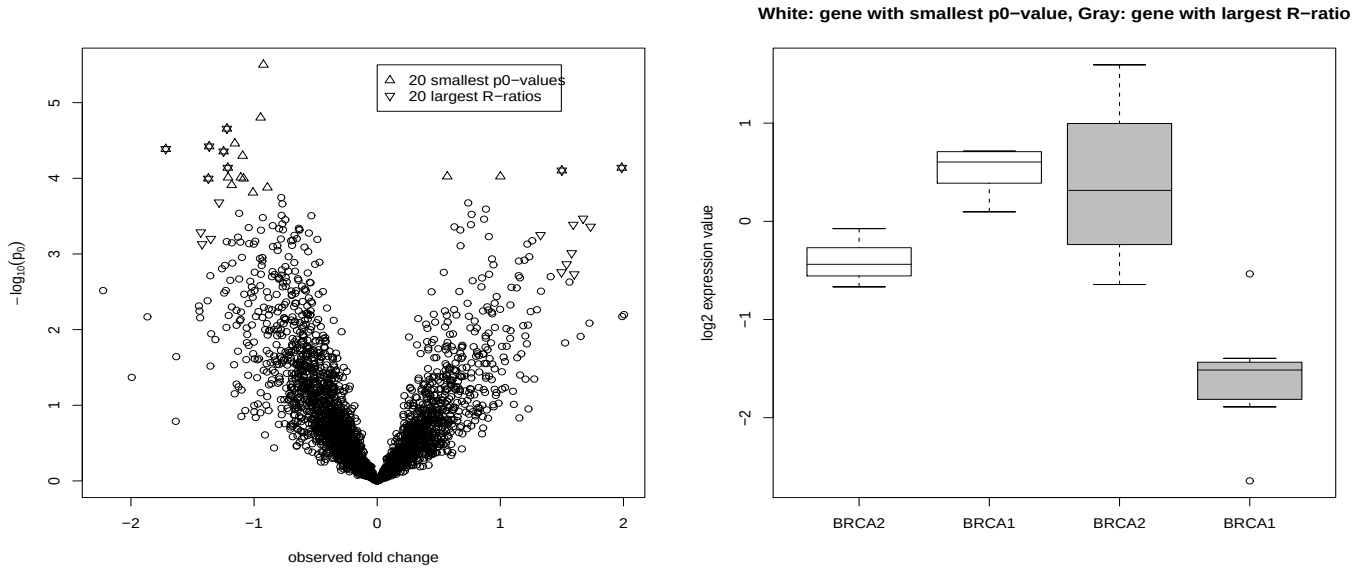

Figure 4: Balanced test ($A_j/B_j = 1$) versus classical p_0-values

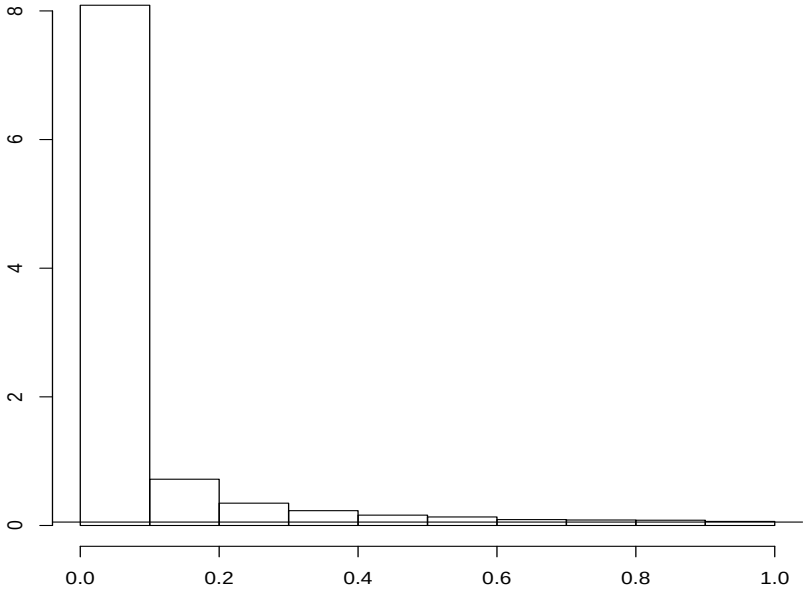

Figure 5: Density histogram of the p_1-values