BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 13 2005, pages 3017–3024
doi:10.1093/bioinformatics/bti448

Gene expression

# False discovery rate, sensitivity and sample size for microarray studies

Yudi Pawitan[1,*], Stefan Michiels[2,3], Serge Koscielny[3], Arief Gusnanto[1,4] and Alexander Ploner[1]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden, [2]Unit of Biostatistics and Epidemiology and [3]Unit of Functional Genomics, Institute Gustave Roussy, Villejuif, France and [4]Medical Research Council, Biostatistics Unit, Institute of Public Health, Cambridge, CB2 2SR, UK

Received on January 10, 2005; revised on March 23, 2005; accepted on April 10, 2005

Advance Access publication April 19, 2005

**ABSTRACT**

**Motivation:** In microarray data studies most researchers are keenly aware of the potentially high rate of false positives and the need to control it. One key statistical shift is the move away from the well-known $P$-value to false discovery rate (FDR). Less discussion perhaps has been spent on the sensitivity or the associated false negative rate (FNR). The purpose of this paper is to explain in simple ways why the shift from $P$-value to FDR for statistical assessment of microarray data is necessary, to elucidate the determining factors of FDR and, for a two-sample comparative study, to discuss its control via sample size at the design stage.

**Results:** We use a mixture model, involving differentially expressed (DE) and non-DE genes, that captures the most common problem of finding DE genes. Factors determining FDR are (1) the proportion of truly differentially expressed genes, (2) the distribution of the true differences, (3) measurement variability and (4) sample size. Many current small microarray studies are plagued with large FDR, but controlling FDR alone can lead to unacceptably large FNR. In evaluating a design of a microarray study, sensitivity or FNR curves should be computed routinely together with FDR curves. Under certain assumptions, the FDR and FNR curves coincide, thus simplifying the choice of sample size for controlling the FDR and FNR jointly.

**Availability:** R-package OCplus for computing FDR, sensitivity curves and sample size is freely available at http://www.meb.ki.se/ ˜yudpaw

**Contact:** yudi.pawitan@meb.ki.se

## 1 INTRODUCTION

The standard $P$-value was invented for testing individual hypotheses. There is an obvious problem when analyzing gene expression data collected via microarrays, as this usually involves testing from several thousands to tens of thousands of hypotheses simultaneously. When applied in repeated testing, the standard $P$-value is conceptually associated with the specificity of a test, i.e. it is used to control the false positive rate of a test. Declaring a test to be significant when $P$-value $<0.05$ means we are setting specificity to be 0.95. It is well known in diagnostic testing that when the disease prevalence is small, we need a test with very high specificity, as otherwise there are too many false positive results. However, although a number of adjustment procedures are available (Dudoit *et al*., 2002; Shaffer, 1995; Westfall and Young, 1993; Holm, 1979; Hochberg and Tamhane, 1987), it is not immediately clear how small a $P$-value should be to protect against this problem. Furthermore, many adjustments of $P$-value, which are based on controlling the probability of making at least one false positive, are too conservative for microarray studies in that they can lead to low sensitivity (Dudoit *et al*., 2002).

The false discovery rate (FDR) of a test is defined as the expected proportion of false positives among the declared significant results (Benjamini and Hochberg, 1995, 2000; Keselman *et al*., 2002). Because of this directly useful interpretation, FDR is a more convenient scale to work on instead of the $P$-value scale. For example, if we declare a collection of 100 genes with a maximum FDR of 0.10 to be differentially expressed (DE), then we expect a maximum of 10 genes to be false positives. No such interpretation is available from the $P$-value. New methods have been proposed either to transform $P$-value into an FDR or to compute FDR directly (Storey and Tibshirani, 2003; Storey, 2002; Aubert *et al*., 2004; Reiner *et al*., 2003).

The presentation of the FDR in the statistical literature tends to be rather technical, typically with a strong emphasis on the general framework uniting the $P$-value, classical multiplicity adjustment, FDR and recent modifications of the FDR. In this paper, we try to present the FDR as a simple and directly appealing criterion for handling the testing of simultaneous hypotheses. The usual perspective treats the individual genes or hypotheses and their $P$-values separately, and then adjusts the $P$-values for multiplicity after the fact. While this is technically what happens when computing FDRs for a given dataset, it distracts from a more direct way of looking at detecting differential expression and multiple testing.

When controlling the FDR, an experimenter also needs to be aware of the sensitivity or false negative rate (FNR), as he/she does not want to lose too many of the truly DE genes by setting the FDR too low. Thus, the increasing use of FDR needs to be accompanied by the sensitivity or FNR assessment.

At least four factors determine the FDR characteristics of a microarray study: (1) the proportion of truly differentially expressed genes, (2) the distribution of the true differences, (3) measurement variability and (4) sample size. Only the latter is under the experimenter's control. Among other things, the analysis of FDR allows an assessment of sample size needed in microarray experiments. Knowing how many samples are needed has been a problem for many researchers, but no clear recommendation based on the FDR

[*]To whom correspondence should be addressed.

seems to be on offer. Since the $P$-value concept is not suitable for microarray studies, the standard sample size calculation from the traditional hypothesis testing framework, based on controlling the false positive rate (e.g. Dobbin *et al*., 2003; Wang and Chen, 2004; Yang *et al*., 2003; Gadbury *et al*., 2004; Pan *et al*., 2002; Zien *et al*., 2003), is not appropriate. Yang and Speed (2003) even suggested that sample size determination based on classical power consideration was not possible. Hence in this paper we will (1) describe the relationship between the traditional $P$-value and FDR, (2) explain the factors that determine the FDR and sensitivity characteristics of a study and (3) describe a method to compute a sample size for a two-sample comparative study based on controlling the FDR and sensitivity.

## 2 METHODOLOGY

The most common objective in microarray data analysis is to search for DE genes between two or more conditions. Another common objective is to search for a molecular signature, a set of genes that optimally separates two or more groups. With the former, there is no restriction on the number of DE genes; with the latter, an optimal set of classifier genes usually consists of a small subset selected from the top of the DE gene list. In this paper we are concerned with the first objective.

The statistics of the gene discovery is best seen in a simple two-by-two table (Table 1), where 10 000 genes are classified according to their true status and the test result. In this example, the false positive rate is $B/(A + B) = 475/9500 = 5\%$, and the sensitivity of the test is $D/(C + D) = 400/500 = 80\%$. In conventional terms, we have a test with 95% specificity and 80% sensitivity, but the FDR is $B/(B+D) = 475/875 = 54\%$, i.e. more than half of the 'discovered' genes are bogus. So the standard control of significance level leads to a high rate of false discoveries even when the power of the test would be considered adequate for a single-gene study.

**Table 1.** A simple two-by-two table where 10 000 genes are classified according to their true status and the test result

| | Test result<br>non-DE | DE | Total |
|---|---|---|---|
| True | | | |
| non-DE | $A = 9025$ | $B = 475$ | 9500 |
| DE | $C = 100$ | $D = 400$ | 500 |
| Total | 9125 | 875 | 10 000 |

DE stands for differentially expressed, the rows describe the true state of nature and the columns the test decision based on the experimental data. $A$ is the number of non-DE genes that were correctly classified, and similarly for $B$, $C$ and $D$.

It is immediately obvious that the problem arises from the high proportion of non-DE genes in this example (95%), as the false positive rate controls the percentage of wrong discoveries only relative to the truly non-DE genes. Although it is possible to reduce the FDR by reducing the critical level for the $P$-values, the amount of reduction is determined by the proportion of truly DE genes. Thus a $P$-value control is only an indirect way of controlling the FDR, since, to be more meaningful, it needs to be translated into FDR; hence a more direct approach is preferable.

Another simple message from the table is that the analogous idea of a false non-discovery rate (FNDR) does not have the same appeal as the FDR: in our example, FNDR is very low at $C/(A + C) = 100/9125 = 1\%$. For a small percentage of truly DE genes, as we expect in practice, $C$ will be small compared with $A$, so FNDR will be misleadingly small, which is the same problem as the false positive rate to begin with. In contrast, the standard concept of sensitivity or equivalently the FNR is still useful. The FNR in this example is $C/(C + D) = 100/500 = 20\%$, which is more directly informative of the proportion of truly DE genes missed by the experiment. Thus in microarray studies we believe that it is most meaningful to report or control both the FDR and FNR.

### 2.1 Two-sample comparative studies

To simplify our presentation, we will focus on the common problem of comparison between two independent groups with equal variance, and explain the FDR using a theoretical analysis of this problem. With standard modifications of the statistics involved, the methodology applies to other problems. Thus, we assume a two-group comparison problem with $n$ arrays per group, using the standard $t$-test with pooled variance estimate. The expression values in log2-scale are assumed normally distributed. To simplify the presentation further we will assume for each gene a standard deviation $\sigma = 1$ throughout. This is equivalent to standardizing the expression measurements by their standard deviation, so the fold changes below have a universal scale in standard deviation units. In reality the expression variance varies between genes, but by standardizing the variance we assume that all genes have equal variance.

We do not specify the number of genes, as the method is applicable for any number. If one wants to be more specific we can assume it is of the order of 10 000 genes, so, for example, a list of top 1% genes will have 100 genes. In the analysis we assume the genes are independent, although the results can be expected to hold for weakly dependent genes; see also the Discussion section.

In principle, any formal statistical testing procedure that is applied on a gene-by-gene basis can be characterized as follows: (1) compute the relevant test statistic for each gene, (2) sort the statistics by order and (3) determine a cutoff point beyond which all genes are assumed to be DE. This holds regardless whether the test statistic is the fold change, a conventional $t$-statistic, a modified $t$-statistic, a correlation statistic, a raw or adjusted $P$-value, etc. For the $t$-statistic that is described in this paper, we can compute the FDR for a given scenario explicitly. When explicit formulas are not available, we can always use a simulation study for the scenario of interest and present the results as in Section 3. Later we give an example of FDR and sensitivity computation using permutation-based test.

### 2.2 Assumptions and theory

For clarity, the following list collects and defines all the elements of the current problem, and for some elements the values used in the analysis are stated. These elements are grouped logically rather than alphabetically.

- FDR is the proportion of false positives among the declared DE genes.
- $t$-Statistics is the standard two-sample $t$-statistics with pooled variance.
- Significant result or DE call is declared for $|t\text{-statistics}| > c$. The critical value $c$ is allowed to vary.
- Significance level $\alpha$ of a test is the same as the false positive rate, which is the proportion of false positives among truly non-DE genes.
- Sensitivity is the proportion of truly DE genes which are declared significant and corresponds to the power of the design or 1 minus the FNR.
- $n$ is the sample size per group. For illustration in this paper we will use varying sample sizes from 5 up to 50.
- $p_0$ is the proportion of truly non-DE genes. We will use a range of values $p_0 = (0.9, 0.95, 0.99)$. The latter size is commonly observed in many experiments. For example, if we have 10 000 genes, we expect on the order of 100 truly DE genes.
- $p_1 = 1 - p_0$ is the proportion of truly DE genes. This is assumed to be equally split between down-regulated and up-regulated genes, and the differential expression is assumed to be concentrated at some fold changes.

- Log-fold change is the mean difference in log2-scale and in standard deviation units, so ‘log-fold change = 1’ means a ratio of $2\sigma$ for the mean of Group 1 versus the mean of Group 2.
- Distribution of the true differences shows the log-fold changes of the truly DE genes. The following scenarios are used:
  (1) Log-fold changes at $-1$ and $+1$ (with equal proportions of $0.5 * p_1$ each).
  (2) Log-fold changes at $-2$ and $+2$ (with equal proportions of $0.5 * p_1$ each).

Unless stated otherwise, the Scenario 1 is used in all examples. In traditional single-gene studies a true difference of $1\sigma$ is considered a large effect; at 5% significance level it only requires a sample size $n = 16$ samples per group to detect it with 80% power. So, we believe that scenario A is already rather optimistic, but we will see that for microarray studies this scenario leads to large FDR and low sensitivity. A true difference of $2\sigma$ in Scenario 2 is a very large effect, where at 5% level and 80% power, only 4 samples per group are sufficient in single-gene studies.

Under the null hypothesis of no differential expression, the $t$-statistic is distributed according to central $t$-distribution with $2n-2$ degrees of freedom. Under the alternatives, the $t$-statistic has the same distribution, except for a non-centrality parameter. So, the distribution of the observed $t$-statistics is a mixture of the form

$$F(t) = p_0 F_0(t) + p_1 F_1(t),$$

$$F_1(t) = 0.5\{G_1(t) + G_2(t)\},$$

where $F_0(t)$ is the central $t$-distribution with degrees of freedom df $= 2n-2$, and $G_1(t)$ and $G_2(t)$ are non-central $t$-distributions with df $= 2n-2$ and non-centrality parameters $\sqrt{n/2}D/\sigma$ and $-\sqrt{n/2}D/\sigma$, respectively. The parameter $D/\sigma$ is the assumed non-zero log-fold change. In the computation, $D/\sigma$ is, for example, $-1$ or 1. Given a critical value $c > 0$, we can compute the proportion of declared DE genes as $2\{1 - F(c)\}$. The classical significance level is equal to the proportion of false positives, and it is computed as $2\{1 - F_0(c)\}$. The FDR is then given by

$$\text{FDR} = \frac{p_0\{1 - F_0(c)\}}{1 - F(c)}.$$

The sensitivity of the test is computed as $2\{1 - F_1(c)\}$.

Now we show that if we declare the top $(1 - p_0) \times 100\%$ as DE genes, the FNR is the same as the FDR. For the moment, we use the fact that the distributions $F(t)$, $F_0(t)$ and $F_1(t)$ are symmetric around 0. For fixed critical value $c > 0$, the sensitivity is given by

$$2F_1(-c) = (1 - \text{FDR}) \times 2F(-c)/(1 - p_0),$$

where $2F(-c)$ is the proportion of declared DE genes. So if we set $2F(-c) = 1 - p_0$, the sensitivity is $1 -$ FDR and the FNR equals the FDR. In the general asymmetric case, it makes sense to consider the up- and down-regulated genes separately. The same result then holds if we consider the positive and negative values of the test statistics separately.

## 3 RESULTS

### 3.1 Discovery by $t$-statistic

For a particular scenario, each critical value of the $t$-statistic generates a two-by-two table like Table 1, from which we can compute the various rates. Rate curves can then be constructed from a range of critical values. The solid curves in Figure 1 are the FDR as a function of critical value in $t$-statistic scale. Each curve is labeled by the proportion of truly non-DE genes $p_0$. For $n = 5$ arrays per group, the 5%-level two-sided critical value for the $t$-statistic is $c = 2.31$. If the proportion of non-DE genes is $p_0 = 0.9$, and if we declare significance at this 5% level, then we should expect >60% FDR. Using the same sample size $n = 5$, such a criterion will produce ~95% FDR if $p_0 = 0.99$. Hence, in small experiments where we expect a large proportion of non-DE genes, the FDR can be persistently high, even when we use high critical values.

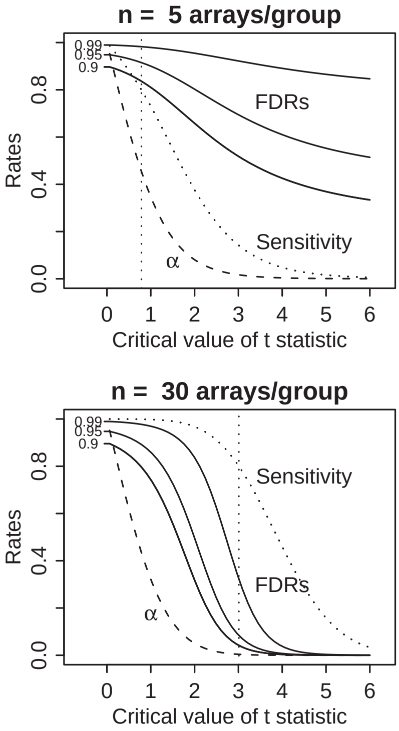

**Fig. 1.** FDR (solid curves), false positive rate $\alpha$ (dashed curves) and sensitivity (dotted curves) as a function of critical value of the $t$-statistic. Each FDR curve is labeled by the proportion of non-DE genes $p_0$. A single sensitivity curve applies for all $p_0$. The dotted vertical line is drawn at 80% sensitivity.

The dashed curve in each plot is the classical significance level or false positive rate as a function of critical value. For $n = 5$, the critical values associated with significance level 0.05, 0.01 and 0.001 are 2.31, 3.36 and 5.04, respectively. Since the significance level is associated with the standard $P$-value cutoffs, this plot shows that standard statistical assessment using $P$-value leads to unacceptably high-FDR. The dotted curve is the sensitivity or power as a function of critical value of the $t$-statistic. For example, when $n = 5$ and significance level is 5%, the sensitivity is ~35%. At 80% sensitivity, the FDR level is persistently above 80%. One must say that this study is seriously underpowered.

However, even when the sensitivity at the traditional 5% significance is high enough (e.g. at $n = 20$ arrays per group, sensitivity is ~90%, plot not shown), compared with FDR, the level of significance and the associated $P$-value is too small and not sensible as a control of false positives. At 20 arrays per group, controlling for FDR is feasible, but the resulting FNR might be high.

The situation improves when the sample size is increased to $n = 30$ per group. For example, a critical value of $c = 3$, associated with a $P$-value cut-off of 0.004, leads to <10% FDR if $p_0$ <0.9. If $p_0$ is near 0.99, the FDR is ~32%. At this sample size, a 0.4%-level test is associated with ~80% sensitivity. The high-FDR when $p_0$ is high, indicates that we cannot naively use the usual formula in sample size computation, i.e. requiring 80% sensitivity at 5% significance level.

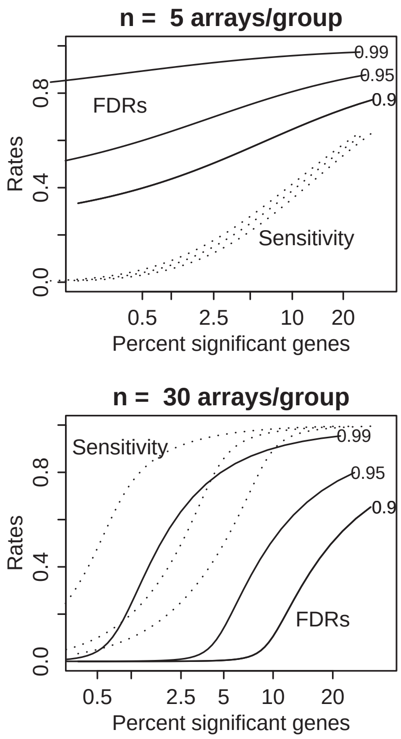

**Fig. 2.** FDR (solid curves) and sensitivity (dotted curves) as a function of percentage significant genes. Each FDR curve is labeled by the proportion of truly non-DE genes $p_0$. The sensitivity curves are in the same order as the FDR curves, for example, the top curve corresponds to $p_0 = 0.99$.

## 3.2 Discovery by percentage of significant results

Sometimes a list of DE genes is determined by taking a number or a proportion of top-ranking genes with largest absolute $t$-statistics. Figure 2 shows the FDR as a function of percentage significant results. For example, if we specify the top 1% of all genes as differentially expressed, using $n = 5$ arrays per group and the proportion of non-DE genes $p_0$ is 0.99, then the FDR >80%. At this low sample size, such a procedure is safe from false discovery only if $p_0$ is smaller than 0.8, that is, there is a very large proportion of truly DE genes, which in practice is an extremely rare occurrence. As we increase the sample size to $n = 30$ arrays per group, the FDR is improved substantially, especially if $p_0$ is not too large. Even with $n = 30$, if $p_0$ is near 0.99, the FDR will still be >20% if we declare the top 1% to be DE.

The sensitivity (dotted) curves in Figure 2 show that the discovery by declaring a small proportion of the top genes to DE can lead to low sensitivity or large FNR. This may or may not be a problem depending on the purpose of the analysis. If the purpose is prediction, sometimes a few top genes are adequate and the lack of information of which genes are DE is not an issue. However, if the purpose is to find as many DE genes as possible, the loss of sensitivity might be less tolerable. Most researchers are probably aware of this problem intuitively, if not quantitatively. It might seem less intuitive that higher sensitivity is obtained with higher $p_0$, but observe that it is achieved at the price of higher FDR.

Regarding sensitivity in this approach, an interesting coincidence occurs, namely, if we declare the top $(1 - p_0) \times 100\%$ as DE genes, the FNR is the same as the FDR. For example, when $p_0 = 0.99$ and we declare the top 1% genes to be DE, then the FDR and FNR are both ~20%. The proof is given in Section 2.

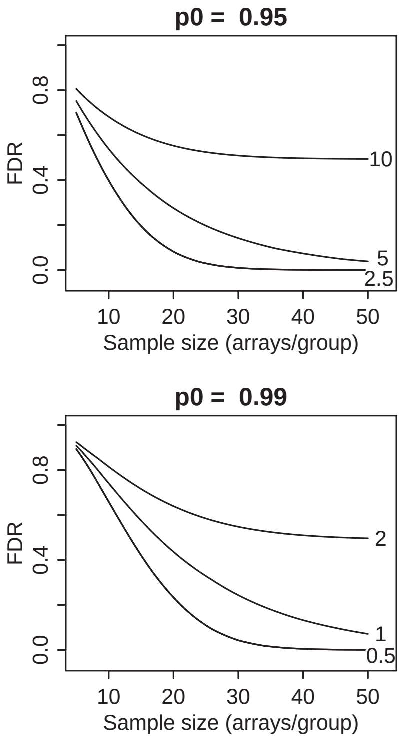

**Fig. 3.** FDR as a function of sample size and percentage significant results. Each curve is for a fixed percentage of significant genes. For example, the label '10' at the end of one curve refers to a fixed 10% of the top genes being declared significant. In this plot, the FNR for $(1 - p_0) \times 100\%$ significant genes coincides with the FDR curve; for example, for $p_0 = 0.95$, the FDR curve for 5% DE genes is the same as the FNR.

The plots in Figure 2 also show that we may have to consider carefully the genes included in an analysis. Genome-wide chips with 45 000 probes are now in common use; if the number of DE genes is fixed, increasing the number of probes will increase the proportion of non-DE genes $p_0$. This will result in larger FDR, sometimes dramatically so, for example, compare the FDR curves for $p_0 = 0.95$ versus $p_0 = 0.99$ for $n = 30$.

## 3.3 Discovery by other statistics

The gene discovery approaches above can be extended to other statistics, such as fold change, correlation coefficient, rank-based statistics or permutation-based $P$-values and various adjusted $P$-values. Many of these statistics are either exact one-to-one maps or highly correlated with the $t$-statistic, so we can expect roughly similar relationships between false positive rate, FDR, sensitivity and sample size.

## 3.4 How many samples do we need?

If the proportion of truly non-DE genes $p_0$ is not so large (e.g. 0.9), then it makes sense that we should be able to identify a lot more significant genes than if $p_0$ is close to one. If we can guess what $p_0$ is, it may be sensible to declare the top $(1 - p_0) \times 100\%$ genes as significant as shown in Figure 2. In Figure 3, we show that the success of such procedure depends on the sample size and the size of $p_0$. If the sample is too small or if $p_0$ is near one, the FDR may still be too large. If $p_0 = 0.99$ and we want to select the top 1% genes, then the FDR is persistently high at >20%, unless the sample size is more than 35 arrays per group. Around 45 arrays per group is needed if we want to get 10% FDR.

There are currently several methods to estimate $p_0$ (e.g. Efron *et al*., 2001; Storey, 2002), so it might be sensible to declare the top $(1 - p_0) \times 100\%$ genes as DE. If a pilot study is available, one can estimate $p_0$ and the distribution of DE genes, then plan the study better. This is a similar situation as planning of experiments in classical hypothesis testing framework. As previously stated, if we declare the top $(1 - p_0) \times 100\%$ as DE genes, the FNR is the same as the FDR. This means that in Figure 3, the middle FDR curve also functions as the FNR, and controlling FDR automatically controls for sensitivity. For example, 10% FDR corresponds to 90% sensitivity.

Figure 3 shows that the required sample size in an experiment depends on (1) the number and (2) the distribution of the truly differentially expressed genes, and (3) on how much FDR we can tolerate. When the number of such genes is small, or when the fold changes are small, a large sample size is needed to control for the FDR. In small experiments involving say 5 arrays per group, one must hope for genes with quite large fold changes ($>3\sigma$), otherwise the situation is hopeless.

## 3.5 Larger fold changes

We consider Scenario B of truly DE genes with log-fold changes at $\{-2, +2\}$ with proportion $p_1/2$ each. As we note previously, at 5% significance level and 80% power, we only need $n = 4$ samples per group to detect this effect size for one gene. Compared with Figure 1, Figure 4 shows some reduction of FDR, although not very dramatic if $p_0$ is close to 1. Sensitivity is also increased, so at $n = 30$ arrays per group the sensitivity is 100% for critical value $c = 5$. Nevertheless, the false positive rate is still much smaller than the FDR, indicating that the FDR assessment is still better than $P$-value. The same exercise can be repeated at larger fold changes, or with a more complex distribution of fold changes, not only at two values, but also at several values or over a range of values. Our key message is similar, that is, the existence of genes with large fold changes is the only hope that small experiments can still give reasonable results with small FDR.

## 3.6 Genes with tiny effects

Presumably genes with tiny effects will not be of interest to the scientists, partly because the discovery of these genes is not likely to be replicated except in very large studies. This problem was discussed by Efron (2004). These genes have a strong impact on the FDR via two factors: (1) increasing the size of $p_0$ and (2) widening the null distribution of the observed $t$-statistic. Current procedures to estimate $p_0$ assume that we are interested in a sharp null hypothesis of non-DE; in reality, we can imagine that the true log-fold change is distributed around zero, and there will be a fraction of genes whose log-fold change is near zero. If genes of tiny effects are not of interest, then $p_0$ should be extended to include these genes. From previous analyses, we know that increasing $p_0$ will increase the FDR, hence making it harder to find the truly DE genes.

The second effect of these genes is that they widen the null distribution of the statistic, which will also increase the FDR. As an example, in Figure 5, instead of a point mass of probability $p_0$ at 0

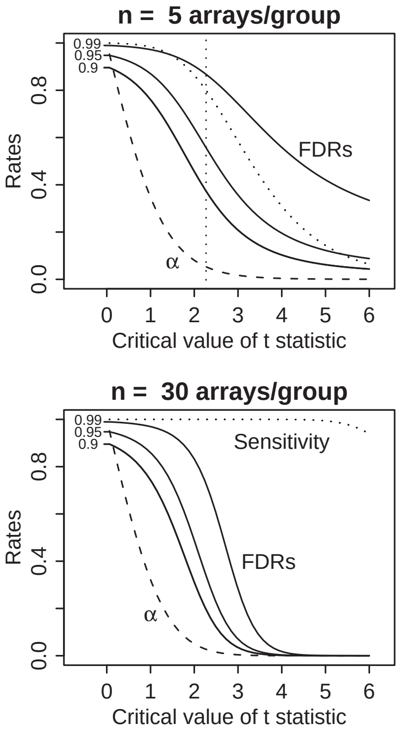

**Fig. 4.** Similar to Figure 1, except the DE genes have log-fold changes at −2 and +2. Compared with Figure 1, the FDR is reduced and the sensitivity is increased.

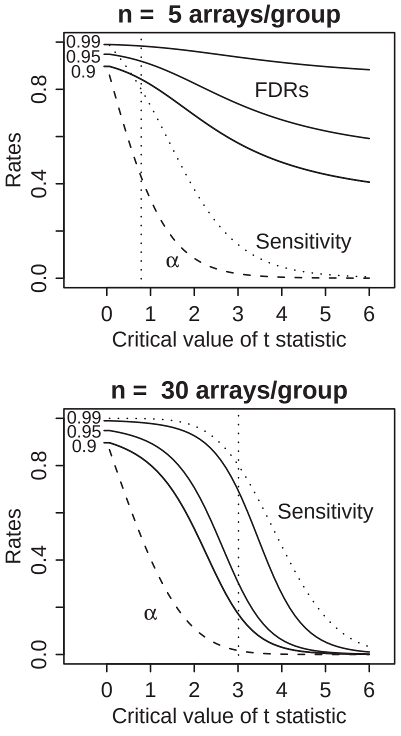

**Fig. 5.** Similar to Figure 1, except the non-DE genes are diffused around zero-fold change. Compared with Figure 1, here the FDR is generally increased.

log-fold change, suppose we assume a mixture of log-fold changes at $-0.25$, 0 and $+0.25$ with equal probability (all equal to $p_0/3$). Compared with Figure 1, for the same $p_0$, the FDR is increased. For example, from Figure 5, to get 80% sensitivity with $n = 30$ arrays per group, for $p_0$ equal to 0.90, 0.95 and 0.99 the FDR is 17, 30 and 69%, respectively. From Figure 1, the corresponding values are 4, 8 and 32%. Hence the impact of genes of small effects on the FDR assessment can be substantial.

### 3.7 Data analysis

Assessment of sample size for a study requires knowledge of the distribution of fold changes, but estimation of such distribution from a dataset is beyond the scope of this paper. Instead here we will only describe briefly how the concepts of FDR and sensitivity can be applied to real data. There is a growing literature on the estimation of FDR, but not of sensitivity. We consider the analysis of 240 cases of diffuse large B-cell lymphoma data from Rosenwald *et al.* (2002). The average follow-up was 4.4 years, and 138 patients died during this period. To simplify, we will ignore the censoring information and compare the survivors and non-survivors. A 'lymphochip' cDNA microarray was used, containing 12 196 probes, but after various quality controls, 7399 probes were used for analysis. The computation of FDR follows the same mixture model

$$F(t) = p_0 F_0(t) + p_1 F_1(t),$$

where the null distribution $F_0(t)$ is computed using the permutation of the group labels (e.g. Efron *et al.*, 2001), and the test uses the two-sample statistic with pooled variance. The permutation-based result is valid without assuming log-normality or equal variance. (If the sample size is large enough, the independence assumption is not required, where in the computation the permutation of the group labels is performed only once and applied to all the genes simultaneously. Commonly used permutation tests permute the labels for each gene separately, as is done in this example.) Since $F(t)$ is observed and $p_0$ can be estimated from the data (e.g. Efron *et al.*, 2001), we can estimate $F_1(t)$. The estimates of FDR and sensitivity are computed using the formulae in Section 2.2, and they are constrained to be monotone.

The top plot in Figure 6 shows the FDR and sensitivity curves for a random subsample of $n = 60$ patients per group. Had the study been done at this sample size, the sensitivity would be much too low and there would not be any gene with low-FDR. With the full dataset we are comparing $n = 138$ non-survivors versus 102 survivors, the FDR curve is now low enough, although at a critical value of, say, $c = 2.5$, the sensitivity is quite low at $\sim$0.2.

### 3.8 Robustness of the assumptions

In the previous illustrations we have made strong assumptions regarding, for example, the log-normality and equal variance. To check the robustness of these assumptions, we performed the following study using the lymphoma data from above:

(1) Take a random subsample of $n$ arrays from each group.

(2) Replace the observed gene-wise means by

$$m_{ij} = \mu_{ij} + t_{(n-1),ij} s_j/\sqrt{n}, \quad i = 1, 2; \quad j = 1, \ldots, p,$$

where $\mu_{ij}$ is the hypothetical true mean of gene $j$ from group $i$, and $t_{(n-1),ij}$ is a random realization from the $t$-distribution with $(n - 1)$ degrees of freedom, $s_j$ is the standard deviation of gene $j$ in the survivor group. The true means will be set according to the groups, and, to mimic the theoretical model, a proportion of $p_1$ of the genes is allowed a shift of $\pm D$ (in standard deviation units):

$$\mu_{1j} \equiv 0,$$
$$\mu_{2j} \equiv D s_j \times B_j,$$

where $B_j$ takes random value $\pm 1$ with probability $p_1/2$. If $y_{ijk}$ is the log-expression of gene $j$ in the $k$-th array of group $i$, this step involves computing

$$y_{ijk} \leftarrow y_{ijk} - \bar{y}_{ij} + m_{ij},$$

where $\bar{y}_{ij}$ is the observed mean.

(3) Perform the permutation-based analysis on the new dataset as described above to compute the FDR and sensitivity.

The second step is based on the standard theory that

$$\frac{\text{sample mean} - \text{true mean}}{\text{std dev}/\sqrt{n}}$$

has $t$-distribution with $n - 1$ degrees of freedom. In effect, we generate a dataset that has all the properties of real microarray data,

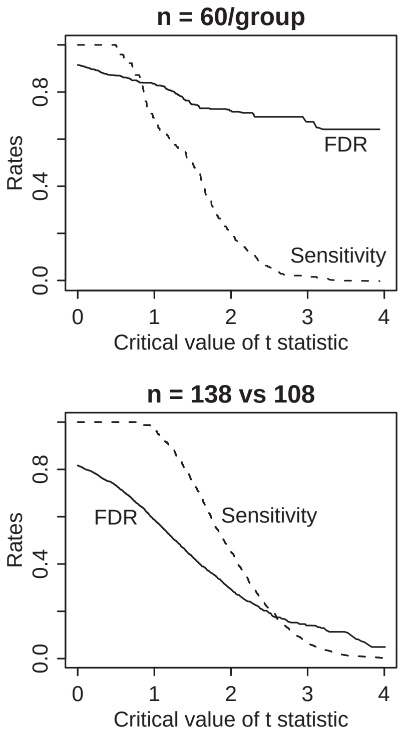

**Fig. 6.** The FDR (solid) and sensitivity (dashed) curves for the diffuse large B-cell lymphoma data from Rosenwald *et al.* (2002). The two groups being compared are the non-survivors versus survivors during follow-up. The left plot shows that a sample size of 60 per group is not adequate since the FDR is too high and the sensitivity too low. The right plot is based on the full study samples, comparing 138 non-survivors versus 108 survivors.

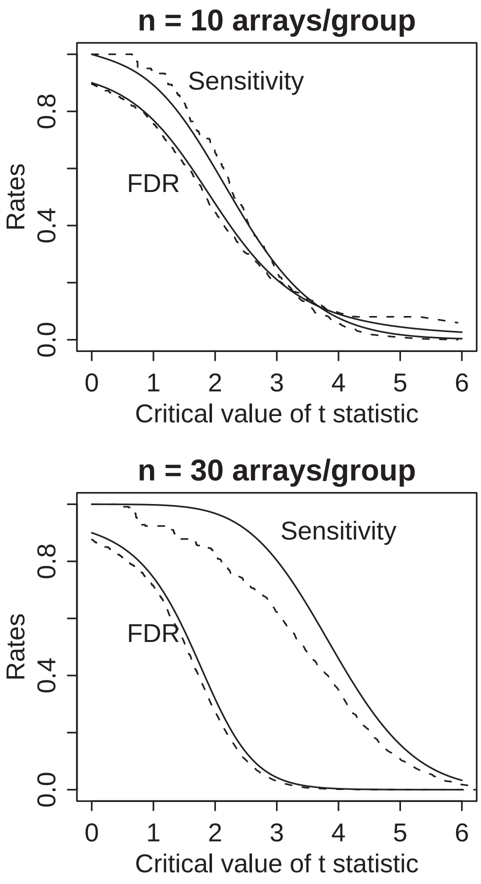

**Fig. 7.** Comparison of the theoretical FDR and sensitivity curves (both in solid lines) versus the permutation-based curves (dashed lines). The data are random subsamples from the lymphoma data, where the means are recomputed to allow theoretical comparisons.

including non-normality, unequal variance and dependence, but the fold changes have been set to have particular values conforming to our theoretical model. Figure 7 compares the theoretical FDR and sensitivity curves as previously computed and the corresponding curves computed using the permutation test. In the computation we use $p_0 = 0.9$ and $D = 1$. The FDR curve computed using the strict assumptions is remarkable robust for samples as small as $n = 10$ per group. For larger $n$ the theoretical sensitivity has positive bias, which means that the true sensitivity in many studies is somewhat worse than our projection. Further investigation is needed to understand and correct the bias of the theoretical sensitivity.

### 3.9 Software

An R package called `OCplus` for the computations of the operating characteristic (OC) curves in this paper is available at http://www.meb.ki.se/∼yudpaw, containing functions to compute

- theoretical OC curves, such as FDR, false positive rate and sensitivity as a function of critical values. For example, Figure 1 of this paper can be generated by the command

  TOC(p0 = c(0.9, 0.95, 0.99), D = 1, n = 5)

- FDR as a function of sample size. Figure 3 of this paper can be generated by the command

  samplesize(p0 = 0.99, D = 1, crit.style = “top percentage”,

  crit = c(0.005,0.01,0.02))

The TOC() command also allows general null hypotheses and general alternatives. The empirical versions of the OC curves for real data are available using the command EOC().

## 4 DISCUSSION

Technological progress in genome-wide measurements has changed the scientific discovery process to more data-driven rather than hypothesis-driven approaches. Our motivation comes mainly from RNA expression microarray data analysis, where scientists are facing a flood of such discoveries. It is not unusual to hear an investigator declare with desperation that he/she has several thousand ‘significant’ genes, which make further steps in the experiment anything but clear. Characterization of a gene list in terms of FDR is useful, so it is important that more scientists understand the FDR concept, certainly in addition to the standard $P$-value. We have chosen to present the FDR mostly from a conceptual perspective, as we believe it is simpler to understand it that way. It is worth highlighting the useful analogy between gene discovery and population screening for a relatively rare disease. In the latter context it is well known that the false positive rate is not sufficient to describe the performance of a procedure. We have illustrated, however, that all the concepts in the paper can be readily applied to real data analysis.

There are a lot of recent works on the extensions of FDR. Genovese and Wasserman (2002) investigated the properties of the FNDR, which is also similar to the ‘miss rate’ in Taylor *et al*. (2005). From the simple table in the Methodology section, we have seen that when the probability $p_0$ of truly non-DE genes is high, as is usually the case in practice, FNDR will be misleadingly small. For the same reasons we prefer FDR over the $P$-value, we believe that the FNR is more meaningful than the FNDR.

It is arguable how much FDR one should tolerate. If gene discoveries require laborious clinical or biological validations, then one might argue for a low-FDR. On the other hand, low-FDR means high-FNR, which might not be acceptable either. Our point here is that it is important to know first the FDR and FNR characteristics of a study, so one has a realistic expectation about the results. The decision about how low the FDR should be can be left open for the investigators. One advantage of the FDR–sample size relationship we show in Figure 3 is that FDR control is automatically a control of FNR. This simplifies the conceptual planning of sample size of a microarray study.

Many early microarray studies involved small numbers of arrays (e.g. DeRisi *et al*., 1997; White *et al*., 1999; Lee *et al*., 2000; Smid-Koopman *et al*., 2000; Tusher *et al*., 2001; Tibshirani *et al*., 2002; Efron *et al*., 2001; Lock *et al*., 2002). All microarray studies of cancer prognosis published between 1995 and April 2003 had a median of 25 patients (Ntzani and Ioannidis, 2003). The analysis here shows that such studies are susceptible to large FDR, unless there is a large proportion of truly DE genes, or there are some genes with very large effects. A meta-study of 16 studies comparing two groups and done between 1999 and 2002 (Pavlidis *et al*., 2003) used progressive re-sampling from the existing data to study the effect of sample size on selected gene lists. The authors suggested no less than 5 and 10–15 replicates per group as optimal, but our analysis now shows that in general one cannot provide such an assurance, as the optimal sample size is a function of underlying parameters of the study.

The proportion of non-DE genes $p_0$ turns out to be a key parameter that determines the FDR characteristics of a microarray study. For a given dataset there are several procedures to estimate this quantity (Storey and Tibshirani, 2003; Storey, 2002; Efron and Tibshirani, 2002; Efron *et al*., 2001). For some experiments, for example, those involving knockouts, biologically we might not expect a large number of DE genes, so one must be rather cautious in setting critical values for DE. In general, the key message from our analysis is that relatively large microarray studies, larger than many current studies, are required to control for FDR while maintaining reasonable sensitivity.

Previous suggestions on sample size for microarray studies had been based on classical significance level and power (e.g. Pan *et al*., 2002; Dobbin *et al*., 2003; Yang *et al*., 2003; Zien *et al*., 2003; Gadbury *et al*., 2004; Wang and Chen, 2004; Dobbin and Simon, 2005). In contrast, we present a sample size computation purely based on FDR control. The latter approach is advantageous, since we have argued here that FDR is a more natural scale to work on rather than the $P$-value. Lee and Whitmore (2002) proposed sample size computations based on controlling the absolute number of false positives. Such a scale is better than the classical significance level, but again it is not as appealing as the direct control of the FDR. Müller *et al*. (2004) considered theoretically the question of optimal sample size, by maximizing the number of DE genes for a given FDR.

In our analyses we have assumed that the genes are independent. For the conceptual understanding of FDR, this is not an issue. However, for the sample size computation, this might be a problem. The current theory of FDR (Storey and Tibshirani, 2003; Storey, 2002; Efron, 2004) indicates that independence is not necessary, and similar results can be expected to hold for weakly dependent genes, as would be expected for genes connected in biological pathways. Our work in normalization of microarray data (Ploner *et al*., 2005) shows that a large collection of genes—that have been properly normalized—are on the average uncorrelated. Further understanding of the genome-wide dependence structure is needed for parsimonious modeling that can be used to improve the sample size computation.

In summary, the simultaneous testing of thousands of hypotheses in microarray data analysis allows a stronger assessment of false positives in terms of FDR, so conceptual understanding of FDR is becoming a necessity. The standard $P$-value is suited to single hypotheses, and does not give a proper sense of uncertainty when there are many tests performed. Theoretical connections among FDR, sensitivity and sample size allow us to plan studies with reasonable FDR characteristics.

## REFERENCES

Aubert,J. *et al*. (2004) Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, **5**, 125–133.

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*, **57**, 289–300.

Benjamini,Y. and Hochberg,Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Edu. Behav. Stat.*, **25**, 60–83.

DeRisi,J.L. *et al*. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Dobbin,K. and Simon,R. (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Dobbin,K. *et al*. (2003) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J. Natl Cancer Inst.*, **95**, 1362–1369.

Dudoit,S. *et al*. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.*, **99**, 96–104.

Efron,B. and Tibshirani,R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, **23**, 70–86.

Efron,B. *et al*. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Gadbury,G.L. *et al*. (2004) Power and sample size estimation in high dimensional biology. *Stat. Methods Med. Res.*, **13**, 325–338.

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

Hochberg,Y. and Tamhane,A. (1987) *Multiple Comparison Procedures*. Wiley, NY.

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

Keselman,H.J. *et al*. (2002) Controlling the rate of type I error over a large set of statistical tests. *Br. J. Math. Stat. Psychol.*, **55**, 27–39.

Lee,M.L. and Whitmore,G.A. (2002) Power and sample size for DNA microarray studies. *Stat. Med.*, **21**, 3543–3570.

Lee,C.-K. *et al*. (2000) Gene-expression profile of the ageing brain in mice. *Nat. Genet.*, **25**, 294–297.

Lock,C. *et al*. (2002) Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nat. Med.*, **8**, 500–508.

Müller,P. *et al*. (2004) Optimal sample size for multiple testing: the case of gene expression microarray. *J. Am. Stat. Assoc.*, **99**, 990–1001.

Ntzani,E.E. and Ioannidis,J.P. (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*, **361**, 1439–1444.

Pan,W. *et al*. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.*, **3**, research0022.1–research0022.10.

Pavlidis,P. *et al*. (2003) The effect of replication on gene expression micorarray experiments. *Bioinformatics*, **19**, 1620–1627.

Ploner,A. *et al*. (2005) Using correlations to evaluate low-level analysis procedures for high-density oligonucleotide microarray data. *BMC Bioinformatics*, **6**, 80.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Rosenwald,A. *et al*. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Shaffer,J.P. (1995) Multiple hypotheses testing. *Annu. Rev. Psychol.*, **46**, 561–584.

Smid-Koopman,E. *et al*. (2000) Gene expression profiles of human endometrial cancer samples using a cDNA-expression array technique: assesment of an analysis method. *Br. J. Cancer*, **83**, 246–251.

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

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Taylor,J. *et al*. (2005) The 'miss rate' for the analysis of gene expression data. *Biostatistics*, **6**, 111–117.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tusher,V. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,S.-J. and Chen,J.-J. (2004) Sample size for identifying differentially expressed genes in microarray experiments. *J. Comput. Biol.*, **11**, 714–726.

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

White,K.P. *et al*. (1999) Microarray analysis of *Drosophila* development during metamorphosis. *Science*, **286**, 2179–2184.

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

Yang,M.C.K. *et al*. (2003) Microarray experimental design: power and sample size considerations. *Physiol. Genomics*, **16**, 24–28.

Zien,A. *et al*. (2003) Microarrays: how many do you need? *J. Comput. Biol.*, **10**, 653–667.