STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1960–1972
Published online 12 March 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3237

Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments

Tiejun Tong[1] and Hongyu Zhao[2, 3, *, †]

[1]*Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, U.S.A.*
[2]*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, U.S.A.*
[3]*Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, U.S.A.*

SUMMARY

One major goal in microarray studies is to identify genes having different expression levels across different classes/conditions. In order to achieve this goal, a study needs to have an adequate sample size to ensure the desired power. Owing to the importance of this topic, a number of approaches to sample size calculation have been developed. However, due to the cost and/or experimental difficulties in obtaining sufficient biological materials, it might be difficult to attain the required sample size. In this article, we address more practical questions for assessing power and false discovery rate (FDR) for a fixed sample size. The relationships between power, sample size and FDR are explored. We also conduct simulations and a real data study to evaluate the proposed findings. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: false discovery rate; gene expression data; power; sample size; T-statistic

1. INTRODUCTION

The development of microarray technologies has revolutionized biomedical research. Microarrays allow scientists to simultaneously estimate the expression levels of thousands of genes in a given sample. Such high-dimensional data demand and motivate novel statistical approaches to the experimental design, data analysis and interpretation. One main objective of microarray studies is to identify genes with different expression levels between two or more conditions. Because thousands of tests are conducted simultaneously, there is a significant multiple comparison issue and it is not appropriate in general to control the false positives on a per comparison basis. Instead, the errors

*Correspondence to: Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, U.S.A.
†E-mail: hongyu.zhao@yale.edu

Contract/grant sponsor: NIH; contract/grant numbers: GM-59507, N01-HV-28186, P30-DA-18343
Contract/grant sponsor: NSF; contract/grant number: DMS-0241160

Received 14 July 2006
Accepted 16 January 2008

need to be controlled at a more stringent level. Two commonly used approaches are controlling the family-wise error rate (FWER) and the false discovery rate (FDR). Control of FWER in microarray studies can be conservative and usually has low power. In contrast, FDR measures the proportion of false positives among the identified genes. It is less conservative and more commonly used in microarray data analysis.

The power of a specific study is affected by the following factors: the proportion of differentially expressed genes, the distribution of effect sizes, the variation across samples in each condition and the sample size [1]. Sample size is probably the most important factor in the study design. Due to its importance, a number of approaches to sample size calculation have been proposed in the literature [2–11]. These studies have found that to achieve good power while controlling the false positives at a low level (e.g. low FDR), a large sample size is usually required. For example, under the rule that the number of genes called significant is the same as the number of nonnull genes in the population, Tibshirani [12] observed that the sample size should be increased to 100 in order to get the FDR down to 5 per cent, depending on the proportion of genes truly changed at two-fold. Similar ideal sample sizes were also reported by Jung [4], Pounds and Cheng [8] and others. However, due to the cost and/or experimental difficulties in obtaining biological materials, it might be difficult to attain such a large sample size.

In this article, in addition to re-addressing the sample size calculation problem, we present some practical guidelines for assessing power and FDR for a fixed sample size (or for the largest sample size available within the budget) in microarray experiments by studying the following two questions:

(1) For a fixed sample size and a desired FDR level, what is the maximum power achievable?
(2) For a fixed sample size and a desired power level, what is the minimum FDR achievable?

In addition, we address the relationship between power and FDR and the determination of an appropriate FDR level to employ in practice. As an alternative to the common practice of declaring all genes with corresponding test statistics above a given threshold as significant, we introduce the concept of quantile thresholding where a fixed proportion of the genes that have the largest test statistics are declared significant. The remainder of this paper is organized as follows. In Section 2, we introduce the notation, describe the model and briefly review the history of FDR and its new developments. We re-address the sample size calculation problem in Section 3 and study practical questions for assessing power and FDR in Sections 4 and 5 with extensive simulations. Finally, we analyze a real data set to evaluate the proposed findings in Section 6 and conclude the article in Section 7 with some discussion.

2. T-STATISTICS AND FDR

Given a microarray experiment with m genes, one major goal is to identify differentially expressed genes between different conditions. Let x_{ij} $(j=1,\ldots,n_1)$ and y_{ij} $(j=1,\ldots,n_2)$ denote the observed expression levels of gene i under conditions 1 and 2, respectively. With proper normalization, we assume that x_{ij} and y_{ij} are normally distributed with means μ_{i1} and μ_{i2} and standard deviations σ_{i1} and σ_{i2}. To identify differentially expressed genes is then equivalent to testing H_{i0}: $\mu_{i1}=\mu_{i2}$ *versus* H_{i1}: $\mu_{i1}\neq\mu_{i2}$. We consider the following two-sample t-tests:

$$T_i=\frac{\bar{x}_i-\bar{y}_i}{s_i\sqrt{1/n_1+1/n_2}},\quad i=1,\ldots,m$$

DOI: 10.1002/sim

where $\bar{x}_i = \sum_{j=1}^{n_1} x_{ij}/n_1$ and $\bar{y}_i = \sum_{j=1}^{n_2} y_{ij}/n_2$, and s_i is the pooled standard deviation defined as $s_i = ((\sum_{j=1}^{n_1}(x_{ij}-\bar{x}_i)^2 + \sum_{j=1}^{n_2}(y_{ij}-\bar{y}_i)^2)/(n_1+n_2-2))^{1/2}$. Note that although we have assumed a common variance for simplicity of exposition, the following analysis applies to unequal variances or other t-tests (e.g. paired t-tests) as well.

Let $n = n_1 + n_2$ denote the total number of arrays and $\lambda = n_1/n$ the allocation proportion for condition 1. Note that $\lambda = \frac{1}{2}$ represents a balanced design. When n is large, T_i is approximately normally distributed with mean $\sqrt{n\lambda(1-\lambda)}\delta_i$ and variance 1, where $\delta_i = (\mu_{i1}-\mu_{i2})/\sigma_i$ is the so-called effect size. Thus, by assuming a reasonably large n (for example, $n \geqslant 10$) for the moment so that the approximation holds well, we can restate the hypothesis as H_{i0}: $\delta_i = 0$ *versus* H_{i1}: $\delta_i \neq 0$. When n is very small, we recommend the use of the exact t-distribution for the T-statistic rather than a simple normal approximation.

Table I summarizes the various outcomes of testing m hypotheses, of which m_0 represents the total number of true null hypotheses and m_1 represents the total number of false null hypotheses. The quantity V is the number of false positives and R is the total number of rejections. In multiple testing, the ratio of the false discoveries, V/R, is often of interest. Benjamini and Hochberg [13] defined the FDR as

$$\mathrm{FDR} = E\left(\frac{V}{R}\middle| R>0\right) P(R>0) \tag{1}$$

Noting that in microarray studies, the scientists are rarely interested in the situation where no genes are selected (i.e. $R=0$). Storey [14] introduced the positive FDR (pFDR) by removing the term $P(R>0)$ in (1). In the asymptotic setting, FDR and pFDR are equivalent and they possess the same asymptotic properties. This can be observed by noting that $\lim_{m\to\infty} P(R>0) = 1$ for any nontrivial threshold. Both FDR and pFDR consider the expectation of the ratio V/R. Recently, the quantity $E(V)/E(R)$, called the proportion of false positives by Fernando *et al.* [15] or the decisive FDR (dFDR) by Bickel [16], has also been introduced. We refer to it as dFDR for the remainder of the article. dFDR is meaningful as long as $P(R>0) \neq 0$. Although dFDR fails to describe the simultaneous fluctuations in V and R, it has some desirable properties such as it can be optimized using decision theory without the independence assumption [16]. It is also interesting to note that under the *weak dependence* criterion such as finite blocks and ergodic dependence, there is little difference between FDR and dFDR because $\lim_{m\to\infty} |\mathrm{FDR} - \mathrm{dFDR}| = 0$ [17].

Let $\mathscr{M}_0$ (size m_0) denote the set of true null hypotheses and $\mathscr{M}_1$ (size m_1) the set of false null hypotheses. Let $\pi_0 = m_0/m$ denote the proportion of true null hypotheses. Let $H_i = 0$ if the ith null hypothesis is true, and $H_i = 1$ otherwise. We reject the null hypothesis H_{i0} if its corresponding p-value, p_i, is smaller than or equal to a given threshold $\alpha \in (0,1)$. Then it is easy to see that

$$\mathrm{dFDR}(\alpha) = \frac{\sum_{i\in\mathscr{M}_0} P(p_i \leqslant \alpha | H_i = 0)}{\sum_{i\in\mathscr{M}_0} P(p_i \leqslant \alpha | H_i = 0) + \sum_{i\in\mathscr{M}_1} P(p_i \leqslant \alpha | H_i = 1)} \tag{2}$$

Table I. Outcomes when testing m hypotheses.

	Accept	Reject	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	$m = m_0 + m_1$

DOI: 10.1002/sim

Note that (2) holds under any dependence structure among the test statistics, as long as the marginal distribution for each test statistic is maintained. As mentioned above, since there is little practical difference between FDR, pFDR and dFDR when m is large and the dependence between genes is not strong, we will not distinguish them in this article since microarray data usually contain thousands of genes. Let $\beta_i = P(p_i \leqslant \alpha | H_i = 1)$ denote the power for the ith hypothesis test and $\beta = \sum_{i \in \mathscr{M}_1} \beta_i / m_1$ the average power. Because the p-values corresponding to the true null hypotheses are uniformly distributed, we have

$$\mathrm{FDR}(\alpha) = \frac{m_0 \alpha}{m_0 \alpha + \sum_{i \in \mathscr{M}_1} \beta_i} = \frac{\pi_0 \alpha}{\pi_0 \alpha + (1-\pi_0)\beta} \tag{3}$$

3. SAMPLE SIZE CALCULATION

In this section, we derive the required sample size for achieving the desired power β while controlling the level of FDR at γ. On the basis of equation (3), we have

$$\alpha = \frac{\gamma(1-\pi_0)}{(1-\gamma)\pi_0}\beta \tag{4}$$

Recall that for a reasonably large n, T_i is approximately normally distributed with mean $\sqrt{n\lambda(1-\lambda)}\delta_i$ and variance 1. Therefore, $T_i \sim \mathrm{N}(0,1)$ for $i \in \mathscr{M}_0$, and $T_i \sim \mathrm{N}(\sqrt{n\lambda(1-\lambda)}\delta_i, 1)$ for $i \in \mathscr{M}_1$. For two-sided tests, we reject H_{i0} if $|T_i| > z_{1-\alpha/2}$, where z_α is the αth quantile of $\mathrm{N}(0,1)$. Let $\Phi(\cdot)$ denote the cumulative distribution function of $\mathrm{N}(0,1)$. Using the fact that $z_{1-\alpha} = -z_\alpha$, we have $\beta_i(\alpha) = \Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha/2}) + \Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha/2})$ and thus

$$\beta(\alpha) = \frac{1}{m_1} \sum_{i \in \mathscr{M}_1} (\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha/2}) + \Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha/2})) \tag{5}$$

Note that the second term in equation (5) is the minor term as long as the quantity $\sqrt{n\lambda(1-\lambda)}|\delta_i|$ is nontrivial. Combining (5) and (4), we have

$$\alpha = \frac{\gamma(1-\pi_0)}{(1-\gamma)\pi_0} \frac{1}{m_1} \sum_{i \in \mathscr{M}_1} (\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha/2}) + \Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha/2})) \tag{6}$$

The required sample size can then be obtained by solving equation (6), which can be done using numerical methods such as the bisection method [4] or a simple grid search. The R code for implementing the bisection method is available from the authors upon request. In practice, to calculate the required sample size, we need to estimate the proportion of true null hypotheses and the corresponding effect sizes in the set $\mathscr{M}_1$ (see the following section for more details). Throughout this article, we take the smallest integer larger than n whenever necessary.

Equation (6) suggests that the required sample size is inversely proportional to $\lambda(1-\lambda)$. This implies that the most efficient design to achieve a desired power is the balanced design, i.e. $\lambda = \frac{1}{2}$. It is also easy to see that the required sample size decreases as the effect size increases. For the special case that $|\delta_i| \equiv \delta > 0$ for all $i \in \mathscr{M}_1$, by ignoring the minor term in equation (5) we have $n \approx (z_\beta - z_{\alpha/2})^2/(\lambda(1-\lambda)\delta^2)$, which is the same result found in Jung [4].

DOI: 10.1002/sim

4. POWER CALCULATION

In this section, we calculate the maximum power achievable, denoted by $\beta_{\max}$, when the sample size n and the level of FDR γ are given. For $\beta(\alpha)$ in equation (5), we show in Appendix A that

Lemma 1
(i) $\beta(\alpha)$ is a strictly increasing function of $\alpha \in [0, 1]$ and (ii) $\beta(\alpha)/\alpha$ is a strictly decreasing function of $\alpha \in [0, 1]$.

By (ii) and the fact that $\mathrm{FDR}(\alpha) = \pi_0/[\pi_0 + (1-\pi_0)\beta(\alpha)/\alpha]$, $\mathrm{FDR}(\alpha)$ is a strictly increasing function of $\alpha \in [0, 1]$ as long as $\pi_0 \neq 1$. This suggests that the assigned level of FDR should not be larger than π_0. Because both $\beta(\alpha)$ and $\mathrm{FDR}(\alpha)$ are strictly increasing functions of α, the power is also a strictly increasing function of FDR. Therefore, for any given level of FDR γ, there exists a unique $\beta_{\max}$ such that $\beta_{\max} = \beta(\alpha_{\max})$, where $\alpha_{\max}$ satisfies $\mathrm{FDR}(\alpha_{\max}) = \gamma$. This implies that, to find $\beta_{\max}$, it suffices to find the unique maximum threshold $\alpha_{\max}$. From equation (3), we have

$$\frac{\pi_0 \alpha_{\max}}{\pi_0 \alpha_{\max} + (1-\pi_0)\beta(\alpha_{\max})} = \gamma \tag{7}$$

where

$$\beta(\alpha_{\max}) = \frac{1}{m_1} \sum_{i \in \mathscr{M}_1} (\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha_{\max}/2}) + \Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i| + z_{\alpha_{\max}/2}))$$

By solving $\alpha_{\max}$ in equation (7) using numerical methods mentioned in Section 3, the maximum power achievable can be estimated as

$$\hat{\beta}_{\max} = \beta(\hat{\alpha}_{\max}) = \frac{(1-\gamma)\hat{\pi}_0}{\gamma(1-\hat{\pi}_0)} \hat{\alpha}_{\max} \tag{8}$$

where $\hat{\pi}_0$ is an estimate of the proportion of true null hypotheses.

4.1. Simulation study

To explore the effects of n, λ, π_0 and δ_i on the relationship between $\beta_{\max}$ and FDR, we consider the following four scenarios (plotted in Figure 1): (A) to explore the effect of n, we consider $n = 10, 20, 40, 80$ with $\lambda = 0.5$, $\pi_0 = 0.8$ and $\{\delta_i, i \in \mathscr{M}_1\} \sim \mathrm{U}[0, 2]$; (B) to explore the effect of λ, we consider $\lambda = 0.1, 0.2, 0.3, 0.5$ with $n = 20$, $\pi_0 = 0.8$ and $\{\delta_i, i \in \mathscr{M}_1\} \sim \mathrm{U}[0, 2]$; (C) to explore the effect of π_0, we consider $\pi_0 = 0.4, 0.6, 0.8, 0.95$ with $n = 20$, $\lambda = 0.5$ and $\delta_i \sim \mathrm{U}[0, 2]$ and (D) to explore the effect of δ_i, we draw $\{\delta_i, i \in \mathscr{M}_1\}$ from U[0, 2], U[1, 2], 1, 1.5 with $n = 20$, $\lambda = 0.5$ and $\pi_0 = 0.8$. Note that without loss of generality, by symmetry we have assumed that $\delta_i > 0$ for any $i \in \mathscr{M}_1$. Further, we set $m = 2000$ throughout the simulations since it has little impact on the relationship between $\beta_{\max}$ and FDR.

In all settings, $\beta_{\max}$ increases as FDR increases as expected because power is an increasing function of FDR. For the same FDR level, Panel A shows that although $\beta_{\max}$ increases with the sample size as expected, the increase in power levels off when n becomes larger, which implies that there is little benefit for further increasing the sample size after a certain size. Panel B restates the fact that a balanced design is always more efficient. Panels C and D indicate that $\beta_{\max}$ increases when the proportion of true null hypotheses decreases or when the level of effect size increases.

DOI: 10.1002/sim

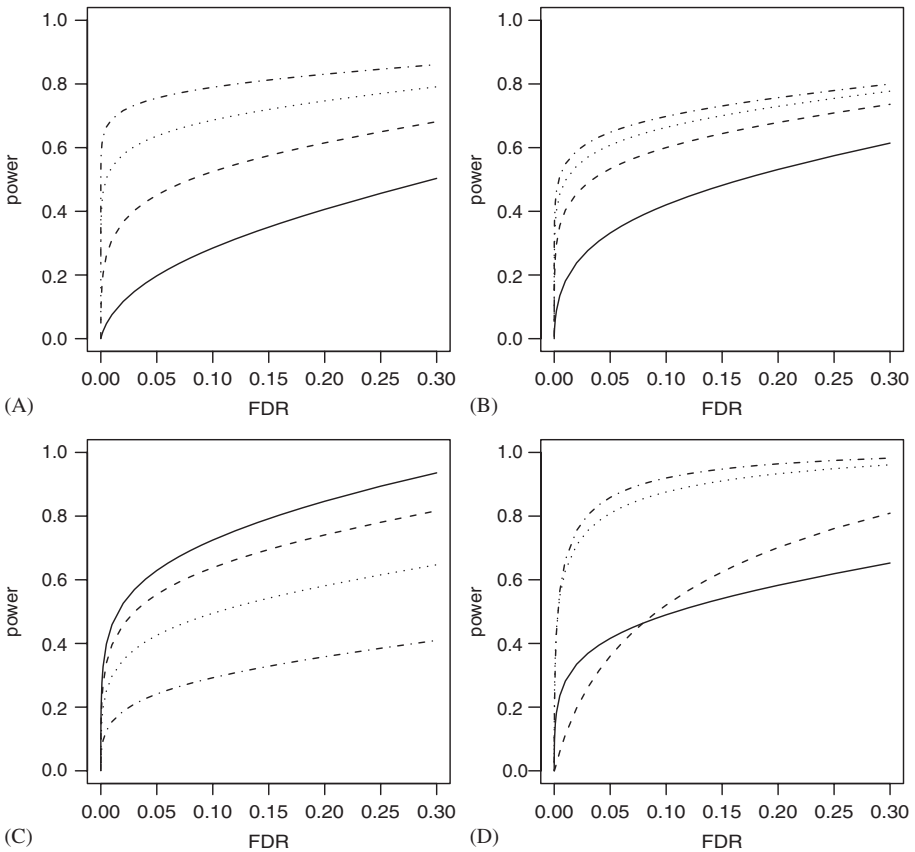

Figure 1. Plots of the power *versus* FDR. Four curves (solid, dashed, dotted, dash-dotted) correspond to four different values of n $(10, 20, 40, 80)$ in panel A (various sample sizes), four different values of λ $(0.1, 0.2, 0.3, 0.5)$ in panel B (various λ), four different values of π_0 $(0.4, 0.6, 0.8, 0.95)$ in panel C (various null proportions) and four different sets of $\{\delta_i, i \in \mathscr{M}_1\}$ (U[0, 2], 1, U[1, 2], 1.5) in panel D (various effect sizes), respectively.

Another interesting finding from these results is that most curves of $\beta_{\max}$ on FDR are concave with a clear elbow, especially when the power is nontrivial. To balance the trade-off between power and FDR, we suggest the use of an FDR near the elbow point, which provides an effective power while controlling FDR at a low level. Accurate determination of the FDR level can be critical. One approach based on decision theory is to maximize the quantity $c_1\beta_{\max}(\text{FDR}) - c_2\text{FDR}$, where $c_1 \geqslant 0$ is the *benefit* of the achieved power and $c_2 \geqslant 0$ is the *cost* paid for the falsely discovered hypotheses. In addition, although m has little impact on the relationship between $\beta_{\max}$ and FDR, in practice, a larger FDR might have to be chosen to accommodate the typical microarray studies having the order of 54 000 genes. More research is required to determine the FDR level for an efficient control.

To check the validation of (8) in practice, we conduct simulations to evaluate the performance of $\hat{\beta}_{\max}$ by reporting $\hat{\beta}_{\max} - \beta_{\max}$ as a measure of accuracy. Set $m = 2000$ as before. Our first

DOI: 10.1002/sim

simulation study investigates the effect of $\hat{\delta}_i$ by assuming that a consistent estimate of π_0 is available at this moment [18]. Noting that we estimate the effect size as $(\bar{x}_i - \bar{y}_i)/s_i$ which is approximately normally distributed with mean δ_i and variance $1/n_1 + 1/n_2$ (see Section 2), we simulate $\hat{\delta}_i$ from $N(\delta_i, 1/n_1 + 1/n_2)$, where the true effect size δ_i is drawn from U[0, 2]. We consider a balanced design for simplicity and consider three different values of n (10, 30, and 50). We further consider two different levels of FDR (0.01 and 0.05) and three different values of π_0 (0.6, 0.8 and 0.95). We run 5000 simulations and report the mean values of $\hat{\beta}_{\max} - \beta_{\max}$ and their standard deviations in Table II. In general, the FDR level has little impact on the accuracy of $\hat{\beta}_{\max}$, whose accuracy decreases as the sample size reduces. We observe little difference between $\hat{\beta}_{\max}$ and $\beta_{\max}$ when the sample size is at least moderately large. For a small sample size, e.g. $n = 10$, $\hat{\beta}_{\max}$ is slightly larger than $\beta_{\max}$. In addition, for each given sample size, $\hat{\beta}_{\max}$ becomes more variable when π_0 is approaching 1.

Our second simulation study does not assume the existence of a consistent estimate of π_0. We set $\pi_0 = 0.8$ for illustration. Noting that most existing estimators for π_0 in the literature are conservative and overestimate π_0 [19, 20], we consider two different values of $\hat{\pi}_0$ at 0.85 (i.e. $\hat{m}_1 = 300$) and 0.9 (i.e. $\hat{m}_1 = 200$). In each simulation, we draw $\hat{m}_1$ samples from $\{\hat{\delta}_i, i \in \mathscr{M}_1\}$ without replacement. All other settings are the same as before. Let $\tilde{\beta}_{\max}$ denote the estimated power, and we report the mean values of $\tilde{\beta}_{\max} - \beta_{\max}$ and their standard deviations in Table III. We also list in Table III the corresponding results for $\hat{\pi}_0 = 0.8$ from the previous simulation study for comparison. Similar to previous results, the FDR level has little impact on the accuracy of $\tilde{\beta}_{\max}$. We observe that when

Table II. Simulation results for the mean values of $\hat{\beta}_{\max} - \beta_{\max}$ and their standard deviations (parentheses) in various settings.

π_0	FDR	$n = 50$	$n = 30$	$n = 10$
0.6	0.01	0.0001 (0.007)	0.004 (0.008)	0.075 (0.009)
	0.05	0.0000 (0.007)	0.002 (0.009)	0.066 (0.011)
0.8	0.01	0.0003 (0.010)	0.007 (0.012)	0.069 (0.011)
	0.05	−0.0001 (0.010)	0.003 (0.012)	0.074 (0.014)
0.95	0.01	0.0004 (0.020)	0.013 (0.024)	0.051 (0.018)
	0.05	0.0003 (0.021)	0.006 (0.024)	0.070 (0.022)

Table III. Simulation results for the mean values of $\tilde{\beta}_{\max} - \beta_{\max}$ and their standard deviations (parentheses) in various settings.

$\hat{\pi}_0$	FDR	$n = 50$	$n = 30$	$n = 10$
0.8	0.01	0.0003 (0.010)	0.007 (0.012)	0.069 (0.011)
	0.05	−0.0001 (0.010)	0.003 (0.012)	0.074 (0.014)
0.85	0.01	−0.0003 (0.016)	0.006 (0.017)	0.069 (0.013)
	0.05	−0.0002 (0.016)	0.002 (0.018)	0.074 (0.017)
0.9	0.01	−0.0001 (0.024)	0.006 (0.025)	0.069 (0.017)
	0.05	−0.0007 (0.023)	0.002 (0.025)	0.075 (0.022)

DOI: 10.1002/sim

the inconsistency of $\hat{\pi}_0$ increases (i.e. $\hat{\pi}_0-\pi_0$ increases), the mean difference between $\tilde{\beta}_{\max}$ and $\beta_{\max}$ stays similar but the variation increases. Overall, $\tilde{\beta}_{\max}$ is fairly robust to the choice of $\hat{\pi}_0$.

5. FDR CALCULATION

In this section, we calculate the minimum level of FDR achievable, denoted by $\mathrm{FDR}_{\min}$, when the sample size n and the desired power β are given. Similar arguments to those in Section 4 indicate that FDR is also a strictly increasing function of power; thus, there exists a unique $\mathrm{FDR}_{\min}$ such that the desired power β is achieved. An easy way to calculate $\mathrm{FDR}_{\min}$ is through $\mathrm{FDR}_{\min}=\mathrm{FDR}(\alpha_\beta)=\pi_0\alpha_\beta/[\pi_0\alpha_\beta+(1-\pi_0)\beta]$, where α_β is the unique solution of

$$\sum_{i\in\mathscr{M}_1}(\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\alpha/2})+\Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\alpha/2}))=m_1\beta$$

Under this setting, α_β can be solved using numerical methods as before. The relationship between $\mathrm{FDR}_{\min}$ and power is also readable from Figure 1. In general, to achieve a desired power, $\mathrm{FDR}_{\min}$ increases with the proportion of true null hypotheses but decreases with the effect sizes and sample size.

Let $\alpha_i, i\in\mathscr{M}_1$, denote the threshold of the ith hypothesis test so as to have a detection power β and α_β denote the common threshold such that the average power equals β. For the special case that $|\delta_i|\equiv\delta>0$, the α_i are all the same and thus $\alpha_\beta=\sum_{i\in\mathscr{M}_1}\alpha_i/m_1$. In general, this relationship does not hold between α_β and $\{\alpha_i,\ i\in\mathscr{M}_1\}$ when the $|\delta_i|$ are not all the same. By ignoring the minor term in equation (5) (or similarly for a one-sided test), α_i has an explicit form as $\alpha_i/2=\Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_\beta)$. Define $\bar{\alpha}=\sum_{i\in\mathscr{M}_1}\alpha_i/m_1$, $\beta_i(\bar{\alpha})=\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\bar{\alpha}/2})$ and $\beta(\bar{\alpha})=\sum_{i\in\mathscr{M}_1}\beta_i(\bar{\alpha})/m_1$. In Appendix B we show that

Lemma 2
(i) When $\beta+\bar{\alpha}/2\leqslant 1$, we have $\beta(\bar{\alpha})\geqslant\beta$ and thus $\bar{\alpha}\geqslant\alpha_\beta$; (ii) when $\beta+\bar{\alpha}/2>1$, we have $\beta(\bar{\alpha})<\beta$ and thus $\bar{\alpha}<\alpha_\beta$.

Furthermore, we have $\mathrm{FDR}(\bar{\alpha})\geqslant\mathrm{FDR}(\alpha_\beta)$ if $\beta+\bar{\alpha}/2\leqslant 1$ and $\mathrm{FDR}(\bar{\alpha})<\mathrm{FDR}(\alpha_\beta)$ if $\beta+\bar{\alpha}/2>1$ by noting that $\mathrm{FDR}(\alpha)$ is an increasing function of α. In practice, it is common that scientists are more interested in validating a small number of genes (e.g. top 10 or top 50 genes) than all the genes inferred to be differentially expressed. This implies that using a quantile threshold can also be of interest. Let $\alpha_{(k)}$ denote the kth smallest value in $\{\alpha_i, i\in\mathscr{M}_1\}$, $\beta_i(\alpha_{(k)})=\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\alpha_{(k)}/2})+\Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\alpha_{(k)}/2})$, then $\#(\beta_i(\alpha_{(k)})\geqslant\beta)=k$. That is, when $\alpha_{(k)}$ serves as a common threshold for all the tests, the detection power is at least β for each of the top k significant genes with $\alpha_i\leqslant\alpha_{(k)}$. For the special case that $\alpha_{\mathrm{med}}=\mathrm{median}\{\alpha_i, i\in\mathscr{M}_1\}$, we have $P(\beta_i(\alpha_{\mathrm{med}})\geqslant\beta)=0.5$, which implies that the detection power is at least β for half of the genes with $\delta_i\geqslant\mathrm{median}\{\delta_i, i\in\mathscr{M}_1\}$.

6. REAL STUDY

We use a well-studied data set, the colon cancer data set [21] containing $n_1=22$ normal colon tissue samples and $n_2=40$ colon tumor samples with expression levels from 2000 genes in each sample to

DOI: 10.1002/sim

evaluate the proposed findings. The main objective of this colon cancer study is to identify important genes that can distinguish colon tumors from normal tissues. We follow the same normalization steps as those in Huang and Pan [22]. That is, to remove possible array effects, we standardize the data in each array by subtracting the median expression level of the array and then dividing this value by the difference in its third quantile and its first quantile of the expression levels.

We use Storey's method [20] to estimate the proportion of true null hypotheses. Noting that the alternative p-values are more likely to be small, for a reasonably large $p_0 \in (0, 1)$, the majority of p-values larger than p_0 should correspond to the true null hypotheses. This suggests a conservative estimate of π_0 as $\hat{\pi}_0 = \#\{p_i > p_0\}/m(1-p_0)$. In this study, p_0 is chosen as the median of all p-values as in Ge *et al.* [23]. Now since the sample sizes for both tissues are at least moderate, for simplicity, we calculate the p-values using the normal approximation. From the data set, we obtain $\hat{\pi}_0 = 0.616$ and thus $\hat{m}_1 = 768$. We then treat the largest 768 values of observed $|\hat{\delta}_i|$ as the true $\{\delta_i, i \in \mathscr{M}_1\}$, where $\hat{\delta}_i = (\bar{x}_i - \bar{y}_i)/s_i$.

The plots of $\hat{\beta}_{\max}$ on FDR for various combinations of (n_1, n_2) are presented in Figure 2. It is clear that the power increases as the total sample size increases when the ratio n_1/n_2 stays the same. The dash-dotted line with $n1 = n2 = 31$ is always above the solid line. This demonstrates again that a balanced design is most efficient. It is also interesting to note that under the same level of FDR, the power with $(n1, n2) = (6, 56)$ is even lower than that with $(n1, n2) = (11, 20)$, which indicates that an extremely unbalanced design is not recommended in practice unless necessary. As an illustration to determine the FDR level in the solid line, if we choose $c_1 = 1 - \hat{\pi}_0$ and $c_2 = \hat{\pi}_0$, an efficient control suggests an FDR level at 0.098.

Figure 3 displays the pattern of FDR on power when the quantile threshold, $\alpha_{(k)}$, is employed. We consider $\alpha_{(100)}$, α_{med}, $\alpha_{(500)}$ and α_β, where α_β is the threshold such that the average power is exactly β (Section 5). As expected, FDR($\alpha_{(k)}$) increases with k for any given power level. It is interesting to see that α_{med} has similar performance as α_β, which suggests that α_{med} can serve as a proxy of α_β in practice. In addition, when only a small number of top genes, e.g. $k = 100$, are of interest to scientists, we can control FDR at a very satisfactory level.

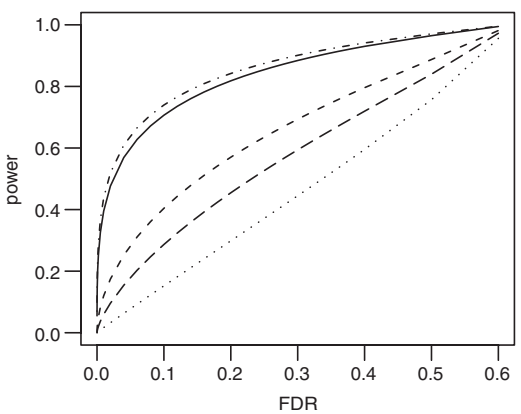

Figure 2. Plots of power *versus* FDR for the colon cancer data set. Five curves (solid, dashed, dotted, dash-dotted and long-dashed) correspond to five different pairs of (n_1, n_2): $(22, 40)$, $(11, 20)$, $(6, 10)$, $(31, 31)$ and $(6, 56)$, respectively.

DOI: 10.1002/sim

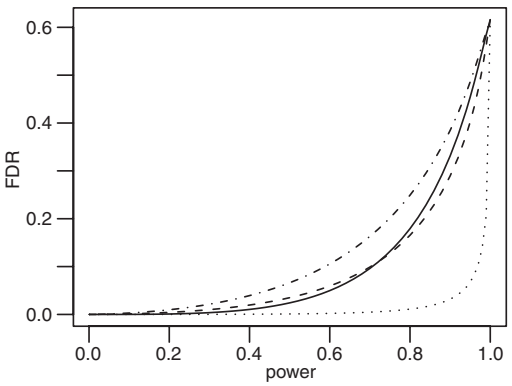

Figure 3. Plots of FDR *versus* the power when the quantile thresholds are used. Four curves (solid, dashed, dotted and dash-dotted) correspond to four different thresholds (α_β, α_{med}, $\alpha_{(100)}$ and $\alpha_{(500)}$).

7. DISCUSSION

Our work is motivated by the fact that the sample size required for achieving good power and low FDR in microarray studies is usually large and hard to obtain, due to the cost and/or other experimental difficulties. Instead of sample size calculation under FDR control, we have studied several practical questions for assessing power and FDR for a fixed sample size. The relationships between power, sample size and FDR are explored. We hope our methods can help with microarray study designs. Our methods can be further used in the post-experimental stage, such as to determine the appropriate level of FDR control or to use a quantile threshold to follow up the top genes in the validation study.

For simplicity of exposition, we have focused on experiments with two conditions. Our methods can be generalized to more than two conditions or to more general settings. We have further assumed that there is little difference between FDR, pFDR and dFDR. The relationship between power and FDR is explored through the concept of dFDR. We note that the results in Sections 3–5 are still valid when the test statistics are correlated with each other by noting that (2) holds in general. When the *weak dependence* of Storey *et al.* [17] does not hold, we cannot claim that FDR and dFDR are asymptotically equivalent; thus, the relationship between power and FDR may no longer hold. Further research is needed to explore the relationship between power and FDR (or pFDR) in the situations where the data are strongly correlated. In addition, we have assumed that the sample size is reasonably large (e.g. $n>10$), such that the T-statistics are approximately normally distributed. When only a small sample size is available, we recommend the use of the exact t-distribution for the T-statistic rather than a simple normal approximation. Simulations (not shown) indicate that the patterns are also similar.

APPENDIX A: PROOF OF LEMMA 1

(1) The first result is made trivial by noting that both $\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\alpha/2})$ and $\Phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\alpha/2})$ are increasing functions of $\alpha\in[0,1]$.

DOI: 10.1002/sim

(2) Let $\phi(\cdot)$ denote the density function of N(0, 1). By Hung *et al.* [24] or Sackrowitz and Samuel-Cahn [25], it is easy to see that the density function of p_i is

$$f(p_i)=\frac{\phi(\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{p_i/2})}{2\phi(z_{p_i/2})}+\frac{\phi(-\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{p_i/2})}{2\phi(z_{p_i/2})}$$
$$\propto \exp(-\sqrt{n\lambda(1-\lambda)}|\delta_i|z_{p_i/2})+\exp(\sqrt{n\lambda(1-\lambda)}|\delta_i|z_{p_i/2})$$

Note that $f(p_i)\equiv 1$ for any $i\in\mathscr{M}_0$. When $i\in\mathscr{M}_1$, $f(p_i)$ is a strictly decreasing function of $p_i\in[0,1]$ by noting that $z_{p_i/2}$ is a strictly increasing function of $p_i\in[0,1]$ and $g(x)=e^x+e^{-x}$ is strictly decreasing on $x\in(-\infty,0)$. This implies that $\beta_i(\alpha)/\alpha$ is a strictly decreasing function of $\alpha\in[0,1]$ for $i\in\mathscr{M}_1$, and so is $\beta(\alpha)/\alpha$.

APPENDIX B: PROOF OF LEMMA 2

We prove result (i) first. For ease of notation, denote $\xi_i=\alpha_i/2$ and $\bar{\xi}=\sum_{i\in\mathscr{M}_1}\xi_i/m$. Result (i) is then equivalent to proving that $\sum_{i\in\mathscr{M}_1}\Phi(\sqrt{n\lambda(1-\lambda)}|\delta_i|+z_{\bar{\xi}})\geqslant m_1\beta$ when $\beta+\bar{\xi}\leqslant 1$. Denote $\xi_i=\bar{\xi}+d_i$ where $d_i\in(-\bar{\xi},1-\bar{\xi})$. Noting that $\sqrt{n\lambda(1-\lambda)}|\delta_i|=z_\beta-z_{\xi_i}$, to prove (i), it suffices to prove that

$$\sum_{i\in\mathscr{M}_1}\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d_i})\geqslant m_1\beta$$

When $m_1=2$, let $g(d)=\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d})+\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}-d})-2\beta$ and we need to prove that $g(d)\geqslant 0$ for any $d\in(-\bar{\xi},1-\bar{\xi})$. Without loss of generality, we assume $d\geqslant 0$. Denote $z_{\bar{\xi}+d}=y$, then $d=\Phi(y)-\bar{\xi}$. We have

$$\frac{\partial d}{\partial y}=\frac{\partial}{\partial y}\int_{-\infty}^{y}\phi(t)\,\mathrm{d}t=\phi(y)$$

This implies that $(\partial/\partial d)z_{\bar{\xi}+d}=1/\phi(z_{\bar{\xi}+d})$ and similarly $(\partial/\partial d)z_{\bar{\xi}-d}=-1/\phi(z_{\bar{\xi}-d})$. Thus,

$$\frac{\partial}{\partial d}g(d)=-\frac{\phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d})}{\phi(z_{\bar{\xi}+d})}+\frac{\phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}-d})}{\phi(z_{\bar{\xi}-d})}$$
$$\triangleq\frac{e^{-B_1/2}-e^{-B_2/2}}{2\pi\phi(z_{\bar{\xi}+d})\phi(z_{\bar{\xi}-d})}$$

where $B_1=(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}-d})^2+z_{\bar{\xi}+d}^2$ and $B_2=(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d})^2+z_{\bar{\xi}-d}^2$. Noting that $z_\beta+z_{\bar{\xi}}\leqslant 0$ since $\beta+\bar{\xi}\leqslant 1$, we have

$$B_1-B_2=2(z_\beta+z_{\bar{\xi}})(z_{\bar{\xi}+d}-z_{\bar{\xi}-d})\leqslant 0$$

since $z_{\bar{\xi}+d}\geqslant z_{\bar{\xi}-d}$ for any $d\geqslant 0$. Therefore, $(\partial/\partial d)g(d)\geqslant 0$ and thus $g(d)\geqslant g(0)=0$.

For $m_1=3$, we define $g(d_1,d_2)=\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d_1})+\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d_2})+\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}-(d_1+d_2)})-3\beta$. Without loss of generality, we assume $(d_1\geqslant 0, d_2\geqslant 0)$ or $(d_1<0, d_2<0)$. For $d_1d_2<0$, we can

DOI: 10.1002/sim

classify it to case 1 by reordering the terms of $g(d_1, d_2)$ when $d_1+d_2>0$, or case 2 otherwise. Similar arguments as above lead to $g(d_1, d_2)\geqslant 0$ for both $(d_1\geqslant 0, d_2\geqslant 0)$ and $(d_1<0, d_2<0)$.

Denote $\mathscr{M}_1^+=\{i: d_i\geqslant 0, i\in\mathscr{M}_1\}$ and $\mathscr{M}_1^-=\{i: d_i<0, i\in\mathscr{M}_1\}$. And let $L^+=\#\{\mathscr{M}_1^+\}$ and $L^-=\#\{\mathscr{M}_1^-\}$. Clearly, $L^++L^-=m_1$. Thus, by the fact that $\sum_{i\in\mathscr{M}_1^+} d_i+\sum_{i\in\mathscr{M}_1^-} d_i=0$, we have

$$\begin{aligned}\sum_{i\in\mathscr{M}_1}\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d_i}) &= \sum_{i\in\mathscr{M}_1^+}\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d_i})+\sum_{i\in\mathscr{M}_1^-}\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+d_i})\\ &\geqslant (L^+-1)\beta+\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+\sum_{i\in\mathscr{M}_1^+} d_i})\\ &\quad +(L^--1)\beta+\Phi(z_\beta+z_{\bar{\xi}}-z_{\bar{\xi}+\sum_{i\in\mathscr{M}_1^-} d_i})\\ &\geqslant m_1\beta\end{aligned}$$

Therefore, $\beta(\bar{\alpha})\geqslant\beta$.

The proof of result (ii) is shown to be essentially the same as in (i) by noting that $z_\beta+z_{\bar{\xi}}>0$ for $\beta+\bar{\xi}>1$, and thus omitted.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants GM-59507, N01-HV-28186 and P30-DA-18343, and NSF grant DMS-0241160. The authors thank the editor, the associate editor, two reviewers and Matthew Holford for their constructive comments and suggestions that have led to a substantial improvement in the article.

REFERENCES

1. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 2005; **21**:3017–3024.
2. Yang MCK, Yang JJ, McIndoe RA, She JX. Microarray experimental design: power and sample size considerations. *Physiological Genomics* 2003; **16**:24–28.
3. Gadbury GL, Page GP, Edwards J, Kayo T, Prolla TA, Weindruch R, Permana PA, Mountz JD, Allison DB. Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research* 2004; **13**:325–338.
4. Jung SH. Sample size for FDR-control in microarray data analysis. *Bioinformatics* 2005; **21**:3097–3104.
5. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* 2005; **6**:27–38.
6. Li SS, Bigler J, Lampe JW, Potter JD, Feng Z. FDR-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine* 2005; **24**:2267–2280.
7. Hu J, Zou F, Wright FA. Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics* 2005; **21**:3264–3272.
8. Pounds S, Cheng C. Sample size determination for the false discovery rate. *Bioinformatics* 2005; **21**:4263–4271.
9. Tsai CA, Wang SJ, Chen DT, Chen JJ. Sample size for gene expression microarray experiments. *Bioinformatics* 2005; **21**:1502–1508.
10. Ferreira JA, Zwinderman A. Approximate sample size calculations with microarray data: an illustration. *Statistical Applications in Genetics and Molecular Biology* 2006; **5**:Article 25.
11. Liu P, Hwang JTG. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics* 2007; **23**:739–746.
12. Tibshirani R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics* 2006; **7**:106.

DOI: 10.1002/sim

13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society*, *Series B* 1995; **57**:289–300.
14. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* 2003; **31**:2013–2035.
15. Fernando RL, Nettledon D, Southey BR, Dekkers JC, Rothschild MF, Soller M. Controlling the proportion of false positives in multiple dependent tests. *Genetics* 2004; **166**:611–619.
16. Bickel DR. Error-rate and decision-theoretic methods of multiple testing: which genes have high objective probabilities of differential expression? *Statistical Applications in Genetics and Molecular Biology* 2004; **3**:Article 8.
17. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rate: a unified approach. *Journal of Royal Statistical Society*, *Series B* 2004; **66**:187–205.
18. Zhang CH, Tang W. Bayes and empirical Bayes approaches to controlling the false discovery rate. *TR 2005-004*, Department of Statistics, Rutgers University, 2005.
19. Langaas M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of Royal Statistical Society*, *Series B* 2005; **67**:555–572.
20. Storey JD. A direct approach to false discovery rate. *Journal of Royal Statistical Society*, *Series B* 2002; **64**:479–498.
21. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of National Academy of Sciences of the United States of America* 1999; **96**:6745–6750.
22. Huang X, Pan W. Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Functional and Integrative Genomics* 2002; **2**:126–133.
23. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Test* 2003; **12**:1–77.
24. Hung HM, O'nell RT, Bauer P, Köhne K. The behavior of the P-value when the alternative hypothesis is true. *Biometrics* 1997; **53**:11–22.
25. Sackrowitz H, Samuel-Cahn E. P values as random variables—expected P values. *The American Statistician* 1999; **53**:326–331.

DOI: 10.1002/sim