BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 6 2007, pages 739–746
doi:10.1093/bioinformatics/btl664

Gene expression

Quick calculation for sample size while controlling false discovery rate with application to microarray analysis

Peng Liu[1,2,*] and J. T. Gene Hwang[3,4]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA, [2]Department of Statistics, Iowa State University, Ames, IA 50011, USA, [3]Department of Mathematics and Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA and [4]Department of Statistics, National Cheng Kung University, Tainan, Taiwan

Received on October 12, 2005; revised on December 11, 2006; accepted on December 26, 2006

Advance Access publication January 19, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Sample size calculation is important in experimental design and is even more so in microarray or proteomic experiments since only a few repetitions can be afforded. In the multiple testing problems involving these experiments, it is more powerful and more reasonable to control false discovery rate (FDR) or positive FDR (pFDR) instead of type I error, e.g. family-wise error rate (FWER). When controlling FDR, the traditional approach of estimating sample size by controlling type I error is no longer applicable.

Results: Our proposed method applies to controlling FDR. The sample size calculation is straightforward and requires minimal computation, as illustrated with two sample *t*-tests and *F*-tests. Based on simulation with the resultant sample size, the power is shown to be achievable by the q-value procedure.

Availability: A Matlab code implementing the described methods is available upon request.

Contact: pliu@iastate.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray and proteomic experiments are becoming popular and important in many biological disciplines, such as neuroscience (Mandel *et al.*, 2003), pharmacogenomics, genetic disease and cancer diagnosis (Heller, 2002). These experiments are rather costly in terms of both materials (samples, reagents, equipments, etc.) and laboratory manpower. Many microarray experiments employ only a small number of replicates (2–8) (Yang and Speed, 2003). In many cases, the sample size is not adequate to achieve reliable statistical inference, resulting in wastage of resources. Therefore, scientists often ask the following question. How big should the sample size be?

To answer this question, we will calculate sample size that controls some error rate and achieves a desired power. When calculating sample size for a single test, the error rate to control is traditionally the type I error rate, the probability of concluding a false positive by rejecting the true null hypothesis. However, we are simultaneously testing a huge number of hypotheses, each relating to a gene. Hence, multiple testing is commonly applied in the analysis of microarray data. There are several kinds of error rates to control in this context, such as family-wise error rate (FWER) or false discovery rate (FDR). Assume there are m genes on microarray chips and each gene is tested for the significance of differential expression. The test outcomes are summarized in Table 1, where, for example, V is the number of false positives and R is the number of rejections among the m tests (Benjamini and Hochberg, 1995).

The FWER is defined to be the probability of making at least one false positive error: $\mathrm{FWER} = \Pr(V \geq 1)$. Rejecting each individual test with a type I error rate of α/m guarantees, by Bonferroni's type of argument, that FWER is controlled at level α in the strong sense, i.e. $\mathrm{FWER} \leq \alpha$ for any combinations of null and alternative hypotheses. Benjamini and Hochberg (1995) proposed another type of error to control—FDR, which is defined to be the expected proportion of false positives among the rejected hypotheses:

$$\mathrm{FDR} = E[Q] \quad \text{and} \quad Q = \begin{cases} V/R & \text{if } R>0 \\ 0 & \text{if } R=0 \end{cases} \tag{1}$$

Storey (2002) proposed to control positive FDR (pFDR), i.e.

$$\mathrm{pFDR} = E\left(\frac{V}{R}\,\middle|\, R>0\right) = \frac{\mathrm{FDR}}{\Pr(R>0)}. \tag{2}$$

In many cases of genomic data such as microarray, it was argued in Storey and Tibshirani (2003) to be more reasonable and more powerful to control FDR or pFDR instead of FWER. However, the sample size has been traditionally calculated with a certain type I error rate and cannot be directly applied with FDR control.

Several articles have addressed the problem of sample size calculation in microarray experiments (Hwang *et al.*, 2002; Lee and Whitmore, 2002; Warnes and Liu, 2006. Lee and Whitmore (2002) calculated the sample size table with an ANOVA model when controlling the number of false positives ($E[V]$). Hwang *et al.* (2002) proposed a method that first identifies

*To whom correspondence should be addressed.

Table 1. Outcomes when testing m hypothesis

Hypothesis	Accept	Reject	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

differentially expressed genes and then calculates the power and sample size on a space reduced by Fisher discriminant analysis. Warnes and Liu (2006) proposed a method with accumulative plot to visualize the trade-off between power and sample size. Some articles have addressed the sample size calculation problem in different designs (Dobbin and Simon, 2005) or specific settings such as classification (Hua *et al*., 2005). The above methods control type I error and not FDR.

Recently, a few articles investigated the need to calculate sample size while controlling FDR and proposed ways to pursue this goal. Yang *et al*. (2003) applied several inequalities to get a type I error rate that corresponds to the controlled level of FDR. Due to the inequalities applied, the sample size is likely overestimated. Pawitan *et al*. (2005) investigated several operating characteristic curves to visualize the relationship between FDR, sensitivity and sample size. Although their approach can be useful in calculating the sample size, no simple direct algorithm was provided. Jung (2005) derived a formula which relates FDR and the type I error rate. Then FDR is controlled by an appropriate level of type I error rate. Pounds and Cheng (2005) proposed an algorithm to iteratively search for the sample size at which the desired power and controlled level of FDR can be achieved. Since FDR controlling procedure is gaining popularity in multiple testings for many problems including microarray analysis, it is important to be able to calculate sample size needed to control the FDR when designing the experiment.

Here, we propose a procedure to calculate the sample size for multiple testing while controlling FDR. First, for any estimate of the proportion of non-differentially expressed genes and the level of FDR to control, we find a rejection region for each sample size. Then power is calculated for the selected rejection region for each sample size. According to the desired power, a sample size is finally decided.

Jung's approach (2005), which was known to us after we had finished our first draft, is more related to our proposed approach than others. Both Jung's and our approaches are based on the same model assumptions which lead to the same FDR expression. The FDR expression is then controlled by studying its relationship to a quantity, which is the type I error rate for Jung and the critical value (the rejection region) for us. Jung provided formulas for Z-tests and t-tests. When applying our approach to Jung's setting, it yields the same result. Our approach, however, is more graphical than Jung's. This allows the visualization of the trade-off between power and sample size and provides quick answer when the user-defined quantities such as power are modified.

In spite of the similarity, this article extends the approach further to several different directions and we find our approach very satisfactory. First, we apply our approach to F-tests which are widely used in microarray data analysis (Cui *et al*., 2005). Second, we study our approach carefully for the case when the means and variances for expression levels vary among genes, an important and practical setting for microarray. Third, we also show by simulation, that the q-value procedure for controlling FDR proposed by Storey *et al*. (2004) using our suggested sample size achieves the target power to a satisfactory degree. This answers the question positively as to whether there would be any statistical procedure that can realize the target power claimed by the proposed method. Finally, we also compare our approach with Yang *et al*. (2003) and Pounds and Cheng (2005) which provide more well-defined algorithms than other articles. Our simulation demonstrates that our proposed method is superior.

The article is organized as follows. Section 2 describes our proposed method illustrated with two-sample t-tests and F-tests. In Section 3, we report the result of simulation studies that compare the power based on proposed method to the actual result from q-value procedure. Section 4 summarizes our results.

Codes for the proposed method in Matlab are available to implement the method.

2 METHOD

In this section, we first illustrate our idea and then show how to apply the proposed method for two designs of microarray experiment.

2.1 Proposed method

The proposed method is derived from the definition of pFDR. Let $H=0$ if null hypothesis is true and $H=1$ if alternative hypothesis is true. In a microarray experiment, $H=1$ represents differential expression for a gene whereas $H=0$ represents no differential expression. We assume as in Theorem 1 of Storey (2002) that all tests are identical, independent and Bernoulli distributed with $\Pr(H=0)=\pi_0$, where π_0 is interpreted as the proportion of non-differentially expressed genes. By Storey's theorem,

$$\mathrm{pFDR}(\Gamma)=\Pr(H=0\mid T\in\Gamma), \tag{3}$$

where T denotes the test statistic and Γ denotes the rejection region. Because the number of genes is large, typically ranging from 5000 to 30 000, the probability of no significant findings is close to zero (Storey and Tibshirani, 2003). Therefore our result also applies to controlling FDR because $\mathrm{FDR}=\mathrm{pFDR}\cdot\Pr(R>0)$ and $\Pr(R>0)$ is nearly one. Suppose the level of FDR is chosen to be α, the following relationship is derived via simple algebra (see Appendix A).

$$\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}=\frac{\Pr(T\in\Gamma\mid H=0)}{\Pr(T\in\Gamma\mid H=1)}. \tag{4}$$

For simplicity in notation, we will denote

$$\Lambda=\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}. \tag{5}$$

In order to achieve an FDR level to be α (or less), we choose the rejection region Γ so that the right-hand side of Equation (4) is equal to (or smaller than) Λ (see Appendix A).

2.2 Applications of proposed method

Microarray experiments are usually set up to find differentially expressed genes between different treatments. The data of scanned intensity for microarray usually go through quality control, transformation and normalization, as reviewed in Smyth *et al*. (2003) and Quackenbush (2002). We assume that data first go through those steps before statistical tests are applied. Before the experiment, we have no observations to check the distribution. It seems reasonable to make a convenient assumption that the distribution of the pre-processed data is normal and hence two-sample *t*-tests and *F*-tests are applicable. The same assumption is also made by other proposed methods to calculate sample size (Dobbin and Simon, 2005; Jung, 2005; Hua *et al*., 2005; Hwang *et al*., 2002).

2.2.1 Two-sample comparison with t-test Suppose we want to find differentially expressed genes between a treatment and a control group using two-sample *t*-tests. The tested hypothesis for each gene is $H_0 : \mu_{T,g} = \mu_{C,g}$ versus $H_1 : \mu_{T,g} \neq \mu_{C,g}$, where $\mu_{T,g}$ and $\mu_{C,g}$ are mean expressions of gth gene for treatment and control group, respectively. Let x_{gj} and y_{gj} denote the observed gene expression levels for treatment and control group respectively for the gth gene and jth replicate. Assuming equal variance between treatment and control group, the test statistic for the gth gene is:

$$T_g = \frac{\overline{x}_g - \overline{y}_g}{\sqrt{S_g^2((1/n_1) + (1/n_2))}}, \tag{6}$$

where $S_g^2 = [1/(n_1 + n_2 - 2)][\sum_{j=1}^{n_1}(x_{gj} - \overline{x}_g)^2 + \sum_{j=1}^{n_2}(y_{gj} - \overline{y}_g)^2]$ is the pooled sample variance, $\overline{x}_g$ and $\overline{y}_g$ are the means of observed expression levels for gene g for the two groups, respectively. The test statistic T_g has a central *t*-distribution under the null hypothesis and non-central *t*-distribution under the alternative hypothesis. We reject the null hypothesis if $|T_g| > c_g$, for which c_g is to be determined. Applying Equation (4), we find critical value c_g that satisfies:

$$\begin{aligned}\Lambda &= \frac{\Pr(|T_g| > c_g \mid H = 0)}{\Pr(|T_g| > c_g \mid H = 1)} \\ &= \frac{2 \cdot T_{n_1+n_2-2}(-c_g)}{1 - T_{n_1+n_2-2}(c_g \mid \theta_g) + T_{n_1+n_2-2}(-c_g \mid \theta_g)},\end{aligned} \tag{7}$$

where $T_d(\cdot|\theta)$ is the cumulative distribution function (c.d.f) of a non-central *t*-distribution with d degrees of freedom and non-centrality parameter θ. Moreover, $T_d(\cdot)$ is $T_d(\cdot \mid \theta)$ for $\theta = 0$. In (7),

$$\theta_g = \frac{\Delta_g}{\sigma_g\sqrt{(1/n_1) + (1/n_2)}} \tag{8}$$

where $\Delta_g = \mu_{T,g} - \mu_{C,g}$ is the true difference between the mean expressions of treatment and control groups and σ_g is the standard deviation for gene g. In this section, we assume a simplified case that Δ_g and σ_g are identical for all genes. Section 2.2.3 deals with the more realistic case when Δ_g and σ_g vary among genes. So the subscript g is dropped in this section.

The right-hand side of (7) is strictly decreasing in c and hence the solution of c is unique when exists. The same comment applies to the two equations (14) and (17) in later sections. See Appendix C for proof. In responding to a referee's question, we discover that the minimum (over c) level of FDR is positive, occurring at $c \to \infty$. This is quite interesting since there is no such positive lower bound for the type I error. The minimum FDR, however, converges to zero very fast as sample size increases. See Figure S1 in appendix.

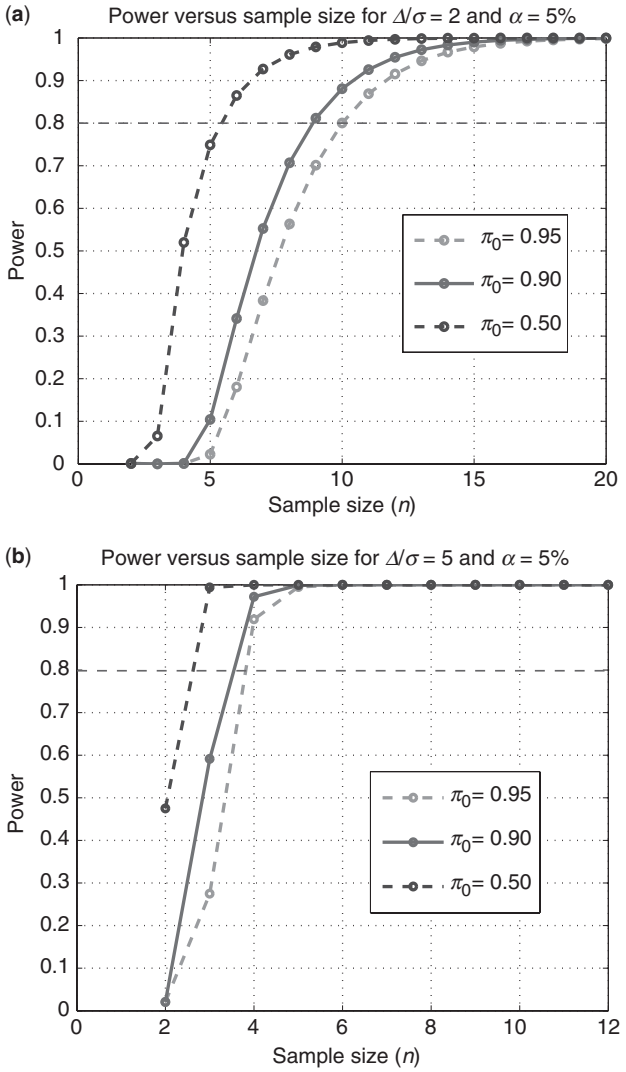

Fig. 1. Plot of power versus sample size for *t*-test. Controlling FDR at 5%, we applied the proposed method to calculate power for each sample size. Panel (**a**) is for $\Delta/\sigma = 2$ and panel (**b**) is for $\Delta/\sigma = 5$.

After finding critical values, power is calculated and sample size will be determined. A special and common case is the balanced design when the two groups have the same sample size:

$$n = n_1 = n_2 \tag{9}$$

Figure 1 plots power versus sample size when FDR is controlled at 5%. As an example, we want to determine the sample size when $\pi_0 = 90\%$. Suppose a 2-fold change is desired

(correspondingly, $\Delta = \log_2(2) = 1$) and $\sigma = 0.5$ from previous knowledge, then $\Delta/\sigma = 2$. Using the middle curve in Figure 1a, a desired power of 80% would require a sample size of 9 for each group.

We have included the case when π_0 is relatively small (50%) in Figure 1. When π_0 is small, the microarray data should be normalized with care because the normalization method for microarray typically relies on the assumption of big π_0, i.e. a small number of differentially expressed genes. In this case, we suggest to use housekeeping genes to perform normalization. Our method would still be applicable if the proper estimate of σ (based on appropriately normalized values) is used.

We shall take σ to be 0.2, which is the median of standard deviations of the U133 microarray data set in Warnes and Liu (2006), i.e. the gene expression levels of human smooth muscle cells from healthy volunteers. (One of the referees mentioned to us that the median σ is typically around 0.7 with human samples and U133A arrays. In such a case, we would set σ to be 0.7 instead.) Also in Cui *et al*. (2005), 0.2 is approximately the 90th percentile of residual standard deviations for the granulosa cell tumor microarray data. (Here 90th percentile is a conservative choice in that if we had used a percentage smaller than 90%, the sample size needed would be smaller.) If still a 2-fold change ($\Delta = \log_2(2) = 1$) is considered to be true effect size, then $\Delta/\sigma = 5$. From the middle curve of Figure 1b, corresponding to $\pi_0 = 0.9$, one can determine that a sample size of 4 is needed to obtain at least 80% of power.

2.2.2 Multi-sample comparison with F-test For microarray experiments comparing several treatments, there are different design schemes applied (Yang and Speed, 2003). Suppose without any replication, a design requires s slides. We call the s slides a *set* for this design. For example, we want to compare gene expressions among three independent treatments, such as livers from three genotypes of mice (Horton *et al*., 2003). If we apply a loop design as shown in Figure 2, a 'set' of three slides is needed for two-color microarray experiment. Whether the replicates are different biological samples or different technical repetitions, our method is applicable as long as the appropriate parameter (means and variances) are used in the calculation. We recommend to use different biological samples in the experiment because this would provide more general conclusions. The question is how many sets of the slides are adequate to obtain a sufficient power and a controlled FDR.

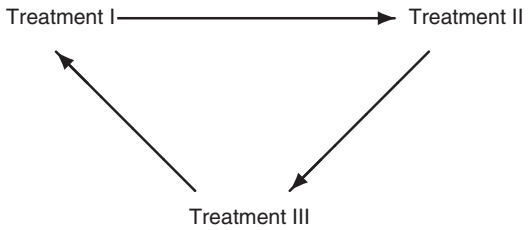

Fig. 2. A design example for microarray experiment to compare gene expressions among three treatments. By convention, each arrow represents one two-color array with the green-labeled sample at the tail and the red-labeled sample at the head of the arrow. This design needs three arrays for one loop.

For each individual gene, the experimental design can be formulated with the same linear model for each set i, $i = 1, 2, \ldots, n$,

$$\mathbf{Y_{g,i}} = X\boldsymbol{\beta}_\mathbf{g} + \boldsymbol{\varepsilon}_\mathbf{g,i}, \tag{10}$$

where $\boldsymbol{\beta}_\mathbf{g}(p \times 1)$ is the vector of parameters for gene g, $\boldsymbol{Y_{g,i}}$ is the observed vector for gth gene in the ith set, X is the design matrix and $\boldsymbol{\varepsilon}_\mathbf{g,i}$ is the error term. It is assumed that the errors are independent across genes and across sets in this section. For the design in Figure 2, $\mathbf{Y_g}$ would be the log-ratio of normalized gene expression levels for gth gene, and two estimable parameters can be the gene expression difference between treatments I and II, and difference between treatments I and III (Yang and Speed, 2003). Then the design matrix is

$$X = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

More complicated models can be constructed for more complex designs and corresponding terms should be added for effects that are not corrected during normalization, such as such array effects, dye effects and block effects. See, for example, Cui *et al*. (2005). For n sets of slides for a design, the least square estimate of $\boldsymbol{\beta}_\mathbf{g}$ is:

$$\hat{\boldsymbol{\beta}}_\mathbf{g} = \sum_{i=1}^{n}(X'X)^{-1}X'Y_{g,i}/n = (X'X)^{-1}X'\sum_{i=1}^{n}\mathbf{Y}_{\mathbf{g},i}/n. \tag{11}$$

With the assumption of normal distribution for the error, $\hat{\boldsymbol{\beta}}_\mathbf{g}$ is also normally distributed,

$$\hat{\boldsymbol{\beta}}_\mathbf{g} \sim N(\boldsymbol{\beta}_\mathbf{g}, \sigma_g^2(X'X)^{-1}/n).$$

We can apply this result and draw statistical inference for these parameters and their linear contrasts.

In general, assume that the question of interest is to test $H_0 : L'\boldsymbol{\beta}_\mathbf{g} = 0$ versus $H_1 : L'\boldsymbol{\beta}_\mathbf{g} \neq 0$, where L is a $p \times k$ coefficient matrix ($k \leq p$) or $p \times 1$ vector for the linear contrast(s) of interest. For simplicity, we omit the subscript g since we assume that the same test is applied for all genes separately. The F-tests based on n sets can be constructed with the following test statistic:

$$F_n = \frac{(L'\hat{\boldsymbol{\beta}})' \cdot [L'(X'X)^{-1}L/n]^{-1} \cdot (L'\hat{\boldsymbol{\beta}})/k}{\sum_{i=1}^{n}(\mathbf{Y_i} - X\hat{\boldsymbol{\beta}})'(\mathbf{Y_i} - X\hat{\boldsymbol{\beta}})/(d(n))}. \tag{12}$$

Under H_0, F_n follows a F-distribution with k and $d(n)$ degrees of freedom where $d(n)$ is a function of n and depends on the design. For example, $d(n)$ for the design shown in Figure 2 is $3n - 2$. Under H_1, F_n follows a non-central F-distribution with the same degrees of freedom and a non-centrality parameter λ:

$$\lambda = (L'\boldsymbol{\beta})'\Sigma^{-1}(L'\boldsymbol{\beta}), \tag{13}$$

where $\Sigma = \sigma^2 L'(X'X)^{-1}L/n$.

Applying Equation (4), we get

$$\Lambda = \frac{\Pr(F_n > c \mid H = 0)}{\Pr(F_n > c \mid H = 1)} = \frac{1 - F_{k,d(n)}(c)}{1 - F_{k,d(n)}(c \mid \lambda)}, \quad (14)$$

and the same procedure follows to calculate the sample size needed. Here, we choose c to satisfy Equation (14). Similar to (7), solution of c to (14) is unique when exists. See Appendix C. Using such a c, we calculate the power $\Pr(F_n > c \mid H = 1)$ and then plot the power against n. Figure S2 in Appendix shows the resulting curves that are similar to those in Figure 1.

2.2.3 Case for unequal Δ_g and σ_g So far, we have proceeded as if all genes have the same set of parameters. In such cases, the average power across all genes would be the same as the power for individual genes. In reality, each gene may have a different set of parameters. If we use the two-sample comparison as an example, the gene-specific parameters include σ_g, the standard deviation, and Δ_g, the true difference between the mean expressions of the treatment and the control group.

To study the realistic case when Δ_g and σ_g depend on g, we assume that they follow some distribution with the probability density function $\pi(\Delta_g, \sigma_g)$. The distribution can be a parametric or nonparametric one that has been estimated from data of similar experiments. For example, when designing an experiment, a pilot study could be available, based on which the distribution of parameters can be estimated. In this case, our procedure can be extended to calculate a sample size while obtaining an average power across all genes. Here by average power, we mean the power integrated with respect to $\pi(\Delta_g, \sigma_g)$,

$$\Pr(T \in \Gamma \mid H = 1) = \iint \Pr(T \in \Gamma \mid H = 1, \Delta_g, \sigma_g)\pi(\Delta_g, \sigma_g)\,\mathrm{d}\Delta_g\,\mathrm{d}\sigma_g. \quad (15)$$

Using Equation (15) and the argument similar to what leads to Equation (4), we conclude that the FDR is α if

$$\Lambda = \frac{\Pr(T \in \Gamma \mid H = 0)}{\iint \Pr(T \in \Gamma \mid H = 1, \Delta_g, \sigma_g)\pi(\Delta_g, \sigma_g)\,\mathrm{d}\Delta_g\,\mathrm{d}\sigma_g}. \quad (16)$$

where

$$\Lambda = \frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}.$$

When we apply this to the t-tests, similar to Equation (7), Equation (16) becomes

$$\Lambda = \frac{\Pr(\mid T_g \mid > c \mid H = 0)}{\iint \Pr(\mid T_g \mid > c \mid H = 1, \Delta_g, \sigma_g)\pi(\Delta_g, \sigma_g)\,\mathrm{d}\Delta_g\,\mathrm{d}\sigma_g}, \quad (17)$$

where the numerator equals $2 \cdot T_{n_1+n_2-2}(-c)$ and the denominator equals

$$1 - \iint T_{n_1+n_2-2}(c \mid \theta_g)\pi(\Delta_g, \sigma_g)\mathrm{d}\Delta_g\mathrm{d}\sigma_g + \iint T_{n_1+n_2-2}(-c \mid \theta_g)\pi(\Delta_g, \sigma_g)\mathrm{d}\Delta_g\mathrm{d}\sigma_g. \quad (18)$$

Note that θ_g is as defined in (8). As before, $T_d(\cdot \mid \theta)$ denotes the c.d.f. of t-distribution. We then solve for the critical value c and apply the same procedure to get the sample size needed. The solution for c is unique when exists, see Appendix C. The same technique extends to the F-tests or other tests of interest.

To illustrate our idea in more detail, we assume that the mean difference expression level of differentially expressed genes, Δ_g, follows a normal distribution and variances of expression levels for all genes follow an inverse gamma distribution:

$$\Delta_g \sim N(\mu_\Delta, \sigma_\Delta^2),$$
$$\sigma_g^2 \sim \text{Inverse Gamma } (a, b),$$

and we use $\pi_1(\Delta_g)$ and $\pi_2(\sigma_g)$ to denote the p.d.f. of Δ_g and σ_g, respectively. Then we solve for c based on Equations (17) and (18) for specified level (α) of FDR and proportion of non-differentially expressed genes (π_0). This involves integrations. To deal with the integration, say in (18), the inner integral equals (see the Appendix B for derivation)

$$\int T_{n_1+n_2-2}\left(c \mid \Delta_g \Big/ \sigma_g\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)\pi_1(\Delta_g)\mathrm{d}\Delta_g = T_{n_1+n_2-2}\left(\frac{c}{\sqrt{\sigma_\Delta^2/(\sigma_g^2(1/n_1+1/n_2))+1}} \,\middle|\, \frac{\mu_\Delta}{\sqrt{\sigma_\Delta^2+\sigma_g^2(1/n_1+1/n_2)}}\right). \quad (19)$$

For the integration with respect to σ_g, we can apply adaptive Lobatto quadrature for numerical integration which allows a stable calculation to get the root of c. The calculation with this numerical integration provides answers instantly. Once we get answers of c for each sample size, we calculate power accordingly and find the needed sample size based on power.

3 SIMULATION

How realistic is the calculated sample size proposed in this article? More specifically, if the desired power is 80%, FDR = 5% and our approach results in a sample size of 9 for the two-sample comparison with t-test, is there a statistical test that would actually achieve all the operating characteristics with 9 slides? To find out, we simulate data with calculated sample size and perform multiple testing with an FDR controlling procedure. Then we checked:

- whether the multiple testing actually results in desired power for the calculated sample size, and
- whether the observed FDR is comparable with the level that we want to control.

If we can find a statistical procedure that achieves the desired FDR and power at the calculated sample size, our procedure is then demonstrated to be practical. This is indeed the case.

There are several procedures to control FDR, such as the q-value procedure proposed by Storey and Tibshirani (2003) and Storey *et al.* (2004), and the procedures proposed by Benjamini and Hochberg (1995, 2000). These procedures all have the FDR conservatively controlled (Storey *et al.*, 2004). For the purpose of simulation study, we apply the q-value procedure as outlined in Storey *et al.* (2004) to control FDR.

Table 2. Parameter values in simulation study

Parameter	Values in simulation
π_0	0.995, 0.95, 0.9, 0.8
σ_Δ	0.2, 1, 2
ρ	0, 0.2, 0.5, 0.8

The earlier version of the manuscript applied the procedure in Storey and Tibshirani (2003) and the results were similar to the report here.

We first test the proposed method when observations (genes) are independent of each other. In a microarray setting, we suppose there are a total of 5000 genes and we have equal sample size for the treatment and the control groups ($n_1 = n_2 = n$). Gene-specific variances, σ_g^2, are simulated from an inverse gamma distribution. Same as in Wright and Simon (2003), we chose $1/\sigma^2 \sim \Gamma(3,1)$ because this distribution approximates well several microarray data sets that we have been analyzing. For the control group, gene expression values are simulated from $N(0, \sigma_g^2)$. For the treatment group, we set Δ_g=0 for non-differentially expressed genes and simulate Δ_g from $N(2, \sigma_\Delta^2)$ for differentially expressed genes, then gene expression values are simulated from $N(\Delta_g, \sigma_g^2)$.

There are several parameters involved for the simulation, π_0 (the proportion of non-differentially expressed genes), σ_Δ (the standard deviation of effect size) and for the dependent case, the correlation coefficient ρ. To evaluate the accuracy of our sample size calculation method, we perform the simulation with a factorial design and the levels (values) of each factor (parameter) are summarized in Table 2. For each of the 48 parameter settings, the FDR is controlled at 5% for multiple testing.

For each parameter setting of independent cases, we calculate the anticipated power for each sample size and generate the power curve as described in Section 2. We also simulate 200 sets of data and perform *t*-tests for each data set with q-value procedure (Storey *et al*., 2004) to control FDR. The observed power is averaged over the 200 simulated data sets and observed proportion of false discoveries is also recorded. Comparing with the simulation results, the anticipated power curves based on our calculation are almost indistinguishable from the simulation results for all parameter settings. Examples are shown in Figure 3a. Hence, our proposed method provides an accurate estimate of sample sizes. The observed FDR is also close to the controlled level (5%) as shown in Figure 3b, justifying the validity of the procedure in Storey *et al*. (2004).

Since many genes may function as groups, it is very likely that dependencies exist in gene expression data. To check the performance of the proposed method when the assumption of independence is violated, gene expression levels are also simulated according to a dependence structure (Ibrahim *et al*., 2002). Then the same procedure of testing as above is applied and the resulting power curves are compared with our calculation.

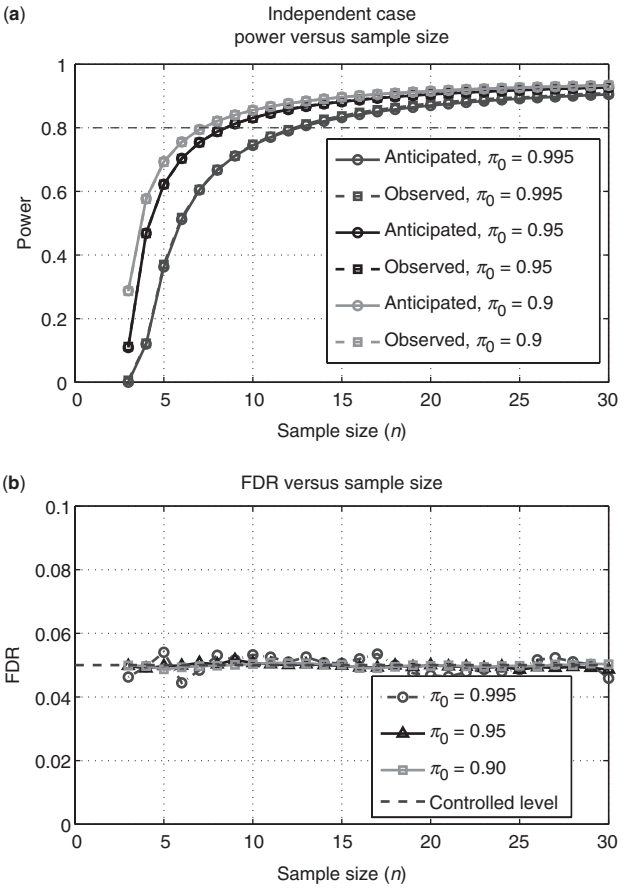

Fig. 3. Simulation results. (**a**) Observed power curves are plotted with dashed lines while the anticipated power curves based on our calculation are plotted with solid lines for different π_0's. For all three π_0's, the difference between the anticipated and observed power are almost indistinguishable. (**b**) Observed false discovery rates (FDRs) for the three parameter settings corresponding to (**a**) are plotted. The controlled level of 5% is indicated with the dashed line.

More specifically, gene expression levels for differentially expressed genes are simulated in blocks of 25 according to the following hierarchical structure described in Section 4 of Ibrahim *et al*. (2002):

$$\begin{aligned}
\mu_X &\sim N(0, v_0^2)\\
\mu_Y &\sim N(2, v_0^2)\\
\mu_{Xg} \mid \mu_X &\sim N(\mu_X, \tau^2)\\
\mu_{Yg} \mid \mu_Y &\sim N(\mu_Y, \tau^2)\\
\sigma_g^2 &\sim \text{Inverse Gamma}(3,1)\\
X_{gi} \mid \mu_{Xg} &\sim N(\mu_{Xg}, \sigma_g^2)\\
Y_{gi} \mid \mu_{Yg} &\sim N(\mu_{Yg}, \sigma_g^2),
\end{aligned}$$

where X_{gi} and Y_{gi} ($g = 1, 2, \ldots, G$, $i = 1, 2, \ldots, n$) are the gene expression levels for the control group (indexed with X) and treatment group (indexed with Y), respectively. For non-differentially expressed genes, we simulate μ_{Xg} the same as above and set $\mu_{Yg} = \mu_{Xg}$, based on which we simulate the gene expression levels X_{gi} and Y_{gi}. Please note that the

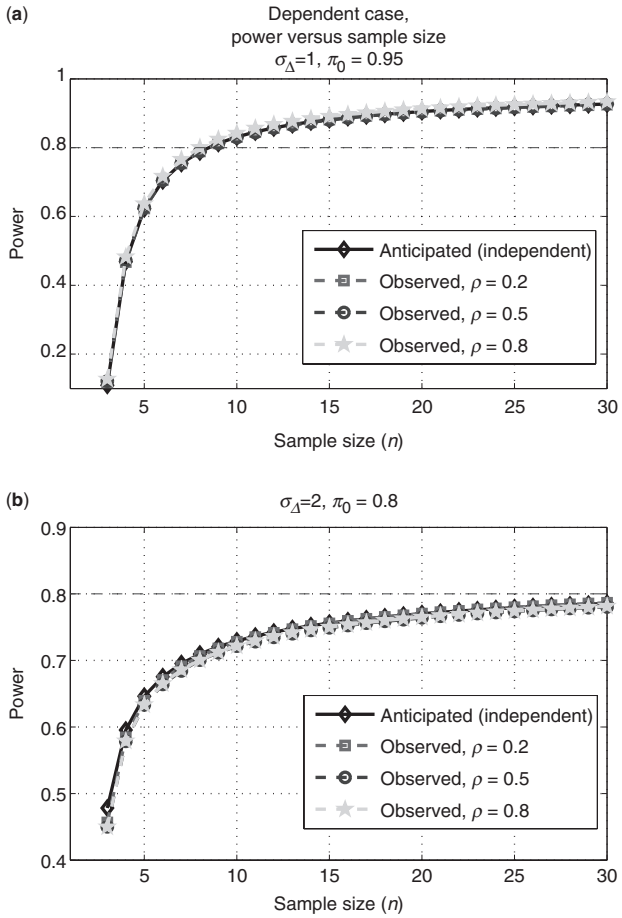

Fig. 4. Simulation results. Observed power curves are plotted with dashed lines while the anticipated power curve based on our calculation is plotted with solid lines for different parameter settings in (**a**) and (**b**).

correlation coefficient, ρ, equals $v_0^2/(v_0^2+\tau^2)$ and $\sigma_\Delta^2 = 2(v_0^2+\tau^2)$ with $\Delta_g = \mu_{Yg} - \mu_{Xg}$. Examples of power curves are presented in Figure 4. For all 36 parameter settings of the dependent case, 34 of them show results similar as in Figure 4a. This demonstrates that the anticipated power approximates really well to the actual power. There are two settings that the discrepancy between anticipated power and calculation is relatively larger than others. Figure 4b includes the worse one ($\rho = 0.8$) of the two. Even for this case, the anticipated power based on our calculated sample size is very close to the simulation results.

When Δ_g and σ_g^2 are the same for all genes, simulation shows that our method can provide accurate sample size estimation both for independent genes and dependent data similarly as the simulation results shown above.

There are several articles addressing the question of calculating sample size while controlling FDR. Among these articles, Yang *et al.* (2003) and Pounds and Cheng (2005) provided clearly defined algorithms. We have compared our approach with these methods in the context of two-sample *t*-test for fixed Δ_g and σ_g^2. Table 3 shows that the calculated sample size based on our proposed approach agrees with the actual sample size needed based on simulation result. Yang's approach results in similar answers as ours except that in some case, it is a little conservative. Answers from Pounds and Cheng's algorithm are too liberal in one situation (when $\Delta/\sigma = 1$) and deviate from the right answer a lot more than the other two methods.

Table 3. Comparison of sample size calculation methods including Yang's approach, Pounds and Cheng's approach (PC), the proposed method in this paper (LH) with the actual simulation result (Simu)

$\Delta/\sigma = 2$	Yang's	PC	LH	Simu
$\pi_0 = 0.5$	8	7	6	6
$\pi_0 = 0.9$	10	10	9	10
$\pi_0 = 0.95$	11	11	11	11
$\Delta/\sigma = 1$	Yang's	PC	LH	Simu
$\pi_0 = 0.5$	22	12	18	18
$\pi_0 = 0.9$	30	16	29	30
$\pi_0 = 0.95$	34	18	33	33

The sample size is selected based on desired power of 80% and FDR at 5%.

4 DISCUSSION

The number of arrays included in microarray experiments directly affects the power of data analysis. It is critical to have a guideline to select a sample size. Because of the huge dimensionality associated with those data sets, controlling FWER is very conservative in many cases (Storey and Tibshirani, 2003). Instead, FDR proposed by Benjamini and Hochberg (1995) and Storey (2002) seem to be a more appropriate error rate to control and has been widely applied to microarray analysis. Therefore, it is important to obtain a method to give the sample size that would control the FDR and guarantee a certain power.

The method is straightforward to apply as described in Section 2 for *t*- and *F*-tests. The proposed method can be generalized to other tests, as long as there is an explicit form to calculate the type I error and power of an individual test. The method presented in this article allows calculation for an accurate sample size with minimum effort when designing an experiment.

ACKNOWLEDGEMENTS

The authors thank the two reviewers and Dr Gregory R. Warnes for insightful comments and suggestions. We also thank Dr Chong Wang for pointing out the Lobatto Quadrature for numerical integration.

REFERENCES

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

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

Cui,X., Hwang,J.T.G., Qiu,J., Blades,N.J. and Churchill,G.A. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

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

Heller,M.J. (2002) DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.*, **4**, 129–153.

Horton,J.D., Shah,N.A., Warrington,J.A., Anderson,N.N., Park,S.W., Brown,M.S. and Goldstein,J.L. (2003) Combined analysis of oligonucleotide microarray data from transgenic and knockout mice identifies direct SREBP target genes. *Proc. Natl. Acad. Sci. USA*, **100**, 12027–12032.

Hua,J., Xiong,Z., Lowey,J., Suh,E. and Dougherty,E.R. (2005) Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, **21**, 1509–1515.

Hwang,D., Schmitt,W.A., Stephanopoulos,G. and Stephanopoulos,G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**, 1184–1193.

Ibrahim,J.G., Chen,M. and Gray,R.J. (2002) Bayesian models for gene expression with DNA microarray data. *J. Am. Stat. Assoc.*, **97**, 88–99.

Jung,S.-H. (2005) Sample size for FDR-control in microarray data analysis. *Bioinformatics*, **21**, 3097–3104.

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

Mandel,S., Weinreb,O. and Youdim,M.B.H. (2003) Using cDNA microarray to assess Parkinson's disease models and the effects of neuroprotective drugs. *Trends Pharmacol. Sci.*, **24**, 184–191.

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

Pounds,S. and Cheng,C. (2005) Sample size determination for the false discovery rate. *Bioinformatics*, **21**, 4263 –4271.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet. Suppl.*, **32**, 496–501.

Smyth,G.K., Yang,Y.H. and Speed,T. (2003) Statistical issues in microarray data analysis. *Method Mol. Biol.*, **224**, 111–136.

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

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

Storey,J.D., Taylor,J.E. and Siegmund,D. (2004) Strong control, conservative point estimation and simultaneous rates: a unified approach. *J. R. Stat. Soc. B*, **66**, 187–205.

Warnes, G. R., Liu, P. (2006) Sample Size Estimation for Microarray Experiments, Technical Report 06/06, Department of Biostatistics and Computational Biology, University of Rochester.

Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.

Yang,Y.H. and Speed,T. (2003) Design and analysis of comparative microarray experiments. In *Statistical Analysis of Gene Expression Microarray Data*. Chapman&Hall/CRC press, p. 51.