Gene expression

Sample size for FDR-control in microarray data analysis

Sin-Ho Jung
Department of Biostatistics and Bioinformatics, CALGB Statistical Center, Hock Plaza, Suite 802, 2424 Erwin Road, Duke University, Durham, NC 27705, USA

Received on December 8, 2004; revised on March 30, 2005; accepted on April 6, 2005
Advance Access publication April 21, 2005

ABSTRACT
Summary: We consider identifying differentially expressing genes between two patient groups using microarray experiment. We propose a sample size calculation method for a specified number of true rejections while controlling the false discovery rate at a desired level. Input parameters for the sample size calculation include the allocation proportion in each group, the number of genes in each array, the number of differentially expressing genes and the effect sizes among the differentially expressing genes. We have a closed-form sample size formula if the projected effect sizes are equal among differentially expressing genes. Otherwise, our method requires a numerical method to solve an equation. Simulation studies are conducted to show that the calculated sample sizes are accurate in practical settings. The proposed method is demonstrated with a real study.

Contact: jung005@mc.duke.edu

1 INTRODUCTION
Microarray method has been widely used for identifying differentially expressing genes, called prognostic genes, in the subjects with different types of disease. Statistical procedures to identify differentially expressing genes involve a serious multiple comparison problem since we perform as many hypothesis testing as the number of the candidate genes in microarrays. If we use a type I error rate \( \alpha \) in each testing, then the probability to reject any hypothesis will greatly exceed the intended overall \( \alpha \) level. In order to avoid this pitfall, two approaches are widely used: false discovery rate (FDR) control and family-wise error rate (FWER) control.

Sample size calculation is a critical procedure when designing a microarray study. There have been several publications on sample size estimation in the microarray context, e.g. Simon et al. (2002). Some focused on exploratory and approximate relationships among statistical power, sample size (or the number of replicates) and effect size (often, in terms of fold-change), and used the most conservative Bonferroni adjustment for controlling FWER (the probability to discover one or more genes when none of the genes under consideration is prognostic) without any attempt to incorporate the underlying correlation structure (Wolfinger et al., 2001; Black and Doerge, 2002; Pan et al., 2002; Cui and Churchill, 2003). Jung et al. (2005) incorporated the correlation structure to derive an accurate sample size when controlling the FWER.

Some researchers proposed a new concept of testing error called FDR, defined as the expected value of the proportion of the non-prognostic genes among the discovered genes (Benjamini and Hochberg, 1995; Storey, 2002). Controlling this quantity relaxes the multiple testing criteria compared with controlling the FWER in general, and consequently increases the number of declared significant genes. Operating and numerical characteristics of FDR are elucidated in recent publications (Genovese and Wasserman, 2002; Dudoit et al., 2003).

Lee and Whitmore (2002) considered multiple group cases, including the two-sample case, using ANOVA models and derived the relationship between the effect sizes and the FDR based on a Bayesian perspective. They discuss a power analysis without involving the multiple testing issue. Müller et al. (2004) chose a pair of testing errors, including FDR, and minimized one while controlling the other at a specified level using a Bayesian decision rule. They proposed a simulation algorithm to demonstrate the relationship between the sample size and the chosen testing errors based on asymptotic results. This approach requires specification of complicated parametric models for prior and data distributions, and extensive computing for the Bayesian simulations. Most of the existing studies for FDR-control do not show the explicit relationship between the sample size and the effect sizes because of different reasons. For example, Lee and Whitmore (2002) and Gadbury et al. (2004) modelled a distribution of \( p \)-values from pilot studies to produce sample size estimates but did not provide an explicit sample size formula. None of the aforementioned studies based on FDR evaluated their sample sizes using simulations.

In this paper, we propose a sample size estimation procedure for FDR-control. We derive the sample size required for a specified number of true rejections (i.e. identifying the prognostic genes) while controlling the FDR at a desired level. As input parameters, we specify the allocation proportions between two groups, the total number of candidate genes, the number of prognostic genes, the effect sizes of the prognostic genes in addition to the required number of true rejections and the FDR level. In general, our procedure requires solving an equation using a numerical method, such as the bisection method. However, if the effect sizes are equal among all prognostic genes, the equation can be solved to give a closed form formula. We review the background of FDR and its estimation method in Section 2, and propose a new sample size method in Section 3. In Section 4, we discuss simulation studies that are conducted to show that the calculated sample sizes are accurate, and demonstrate an application of our method to a real study. van den Oord and Sullivan (2003) considered a similar setting for sample size calculation, but their formulation is so general that they do not provide an explicit formula in any specific case.
be conservative, and the conservativeness increases in conducted in Section 4. We observed between FDR and pFDR. Hence, definitions (1) and (2) are considered to be equal. We accept this argument in this paper and do not distinguish (R > 0) and proposes to control this quantity instead of FDR. Storey (2002) supposes that we conduct m multiple tests, of which the null hypotheses are true for m0 tests and the alternative hypotheses are true for m1 (m = m0 + m1) tests. The tests declare that, of the m0 null hypotheses, A0 hypotheses are null (true negative) and R0 hypotheses are alternative (false rejection, false discovery or false positive). Among the m1 alternative hypotheses, A1 are declared null (false negative) and R1 are declared alternative (true rejection, true discovery or true positive). Table 1 summarizes the outcome of hypothesis tests.

Benjamini and Hochberg (1995) define the FDR as

$$\text{FDR} = E \left( \frac{R_0}{R} \right).$$

Note that this expression is undefined if Pr(R = 0) > 0. To avoid this issue, Benjamini and Hochberg (1995) redefine the FDR as

$$\text{FDR} = \Pr(R > 0) E \left( \frac{R_0}{R} \mid R > 0 \right).$$

These two definitions are identical if Pr(R = 0) = 0, in which case we have FDR = E(R0/R|R > 0) (≡pFDR, which will be defined below).

If m = m0, then FDR = 1 by any critical value with Pr(R = 0) = 0. Pointing out this issue, Storey and Tibshirani (2003) defines the second factor in the right-hand side of Equation (2) as pFDR,

$$\text{pFDR} = E \left( \frac{R_0}{R} \mid R > 0 \right)$$

and proposes to control this quantity instead of FDR. Storey (2002) claims that Pr(R > 0) ≈ 1 with a large m, so that pFDR is equivalent to FDR. We accept this argument in this paper and do not distinguish between FDR and pFDR. Hence, definitions (1) and (2) are considered to be equal. We observed R > 0 in all of the simulations conducted in Section 4.

Benjamini and Hochberg (1995) propose a multi-step procedure to control the FDR at a specified level. However, this is known to be conservative, and the conservativeness increases in m0, see, e.g. Storey et al. (2004).

Suppose that, in the j-th testing, we reject the null hypothesis Hj if the p-value pj is smaller than or equal to α ∈ (0, 1). Assuming independence of the m p-values, we have

$$R_0 = \sum_{j=1}^{m} I(H_j \text{ true}, H_j \text{ rejected})$$

$$= \sum_{j=1}^{m} \Pr(H_j \text{ true}) \Pr(H_j \text{ rejected} \mid H_j) + o_p(m),$$

which equals m0α, where m−1o_p(m) → 0 in probability as m → ∞ (Storey, 2002). Ignoring the error term, we have

$$\text{FDR}(\alpha) = \frac{m_0 \alpha}{R(\alpha)}, \quad (3)$$

where R(α) = ∑j=1m I(pj ≤ α). Given α, estimation of FDR by Equation (3) requires estimation of m0.

For the estimation of m0, Storey (2002) assumes that the histogram of m p-values is a mixture of m0 p-values that are corresponding to the true null hypotheses and following U(0, 1) distribution, and m1 p-values that are corresponding to the alternative hypotheses and expected to be close to 0. Consequently, for a chosen constant λ away from 0, none (or few, if any) of the latter m1 p-values will fall above λ, so that the number of p-values above λ, ∑j=1m I(pj > λ), can be approximated by the expected frequency among the m0 p-values above λ from U(0, 1) distribution, i.e. m0/(1 − λ). Hence, given λ, m0 is estimated by

$$\hat{m}_0(\lambda) = \frac{\sum_{j=1}^{m} I(p_j > \lambda)}{1 - \lambda}.$$

By combining this m0 estimator with Equation (3), Storey (2002) obtains

$$\text{FDR}(\alpha) = \frac{\alpha \times \hat{m}_0(\lambda)}{R(\alpha)} = \frac{\alpha \sum_{j=1}^{m} I(p_j > \lambda)}{(1 - \lambda) \sum_{j=1}^{m} I(p_j \leq \alpha)}.$$

For an observed p-value pj, Storey (2002) defines the q-value, the minimum FDR level at which we reject Hj, as

$$q_j = \inf_{\alpha > 0} \text{FDR}(\alpha).$$

This formula is reduced to

$$q_j = \frac{\text{FDR}(p_j)}{\alpha}$$

if FDR(α) is strictly increasing in α, see Theorem 2 of Storey and Tibshirani (2003). Supporting Material shows that this assumption holds if the power function of the individual tests is concave in α, which is the case when the test statistics follow the standard normal distribution under the null hypotheses. We reject Hj (or, equivalently, discover gene j) if qj is smaller than or equal to the prespecified FDR level.

The independence assumption among m test statistics was relaxed to independence only among m0 test statistics corresponding to the null hypotheses by Storey and Tibshirani (2001), and to weak independence among all m test statistics by Storey and Tibshirani (2003) and Storey et al. (2004). These approaches are implemented in the statistical package called SAM (see Storey and Tibshirani, 2003).

### Table 1. Outcomes of m multiple tests

<table>
<thead>
<tr>
<th>True hypothesis</th>
<th>Accepted hypothesis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Null</td>
<td>Alternative</td>
</tr>
<tr>
<td>Null</td>
<td>A0</td>
<td>R0</td>
</tr>
<tr>
<td>Alternative</td>
<td>A1</td>
<td>R1</td>
</tr>
<tr>
<td>Total</td>
<td>A</td>
<td>R</td>
</tr>
</tbody>
</table>

### 3 SAMPLE SIZE CALCULATION

Let M0 and M1 denote the set of genes for which the null and alternative hypotheses are true, respectively. Note that the cardinalities of M0 and M1 are m0 and m1, respectively. Since the estimated FDR is invariant to the order of the genes, we may rearrange the genes and set M1 = {1, ..., m1} and M0 = {m1 + 1, ..., m}.
By Storey (2002) and Storey and Tibshirani (2001), for large m and under independence (or weak dependence) among the test statistics, we have

\[ R(\alpha) = E[R_0(\alpha)] + E[R_1(\alpha)] + \sigma_\alpha(m) \]

= \( m_\alpha \alpha + \sum_{j \in M_1} \xi_j(\alpha) + \sigma_\alpha(m) \),

where \( R_0(\alpha) = \sum_{j \in M_1} I(p_j \leq \alpha) \) for \( h = 0,1 \), \( \xi_j(\alpha) = P(p_j \leq \alpha) \) is the marginal power of the single \( \alpha \)-test applied to gene \( j \in M_1 \). So, from (3), we have

\[
\text{FDR}(\alpha) = \frac{m_\alpha \alpha}{m_\alpha \alpha + \sum_{j \in M_1} \xi_j(\alpha)}
\]  

(4)

by omitting the error term.

Let \( X_j(2) \) denote the expression level of gene \( j \) for subject \( i \) in group 1 (and group 2, respectively) with common variance \( \sigma_j^2 \). We consider two-sample \( t \)-tests, 

\[
T_j = \frac{\bar{X}_j - \bar{Y}_j}{\hat{\sigma}_j \sqrt{\frac{n_1 + n_2}{n_1 n_2} - 1}}
\]

for hypothesis \( j (= 1, \ldots, m) \), where \( n_j \) is the number of subjects in group \( k (= 1, 2) \), \( \bar{X}_j \) and \( \bar{Y}_j \) are sample means of \( \{X_{ij}, i = 1, \ldots, n_1\} \) and \( \{Y_{ij}, i = 1, \ldots, n_2\} \), respectively, and \( \hat{\sigma}_j^2 \) is the pooled sample variance. We assume a large sample (i.e. \( n_j \rightarrow \infty \)), so that \( T_j \sim N(0, 1) \) for \( j \in M_0 \). Let \( n = n_1 + n_2 \) denote the total sample size, and \( a_0 = n_1/n \) the allocation proportion for group 1.

Let \( \delta_j \) denote the effect size for gene \( j \) in the fraction of its standard error, i.e.

\[
\delta_j = \frac{E(X_j) - E(Y_j)}{\sigma_j}.
\]

At the moment, we consider one-sided tests, \( H_1 : \delta_j > 0 \) against \( H_0 : \delta_j = 0 \), by assuming \( \delta_j > 0 \) for \( j \in M_1 \) and \( \delta_j = 0 \) for \( j \in M_0 \). The two-sided testing case is briefly discussed at the end of this section. Note that, for large \( n \), \( T_j \sim N(\delta_j \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, 1) \) for \( j \in M_1 \), so that we have

\[
\xi_j(\alpha) = \Phi(z_{\alpha/2} - \delta_j \sqrt{\frac{n_1 n_2}{n_1 + n_2}}),
\]

where \( \Phi(\cdot) \) denotes the survivor function and \( z_{\alpha/2} = \Phi^{-1}(\alpha) \) is the upper 100\( \alpha \)-th percentile of \( N(0, 1) \). Hence, Equation (4) is expressed as

\[
\text{FDR}(\alpha) = \frac{m_\alpha \alpha}{m_\alpha \alpha + \sum_{j \in M_1} \Phi(z_{\alpha/2} - \delta_j \sqrt{\frac{n_1 n_2}{n_1 + n_2}})}.
\]  

(5)

From Equation (5), FDR is decreasing in \( \delta_j \), \( n \) and \( |a_1 - 1/2| \). Further, FDR is increasing in \( \alpha \) (see Supporting Material). If the effect sizes are equal among the prognostic genes, FDR is increasing in \( \pi_0 = m_0/m \). It is easy to show that FDR increases from 0 to \( m_0/m \) as \( \alpha \) increases from 0 to 1.

At the design stage of a study, \( m \) is decided by the microarray chips chosen for experiment and \( m_1, |\delta_j|, j \in M_1 \) and \( a_1 \) are projected based on experience or from pilot data if any. The only variables undecided in Equation (5) are \( \alpha \) and \( n \). With all other design parameters fixed, FDR is controlled at a certain level by the chosen \( \alpha \) level. So, we want to find the sample size \( n \) that will guarantee a certain number, say \( r_1 \leq m_1 \), of true rejections with FDR controlled at a specified level \( f \).

In Equation (5), the expected number of true rejections is

\[
E[R_1(\alpha)] = \sum_{j \in M_1} \Phi(z_{\alpha/2} - \delta_j \sqrt{\frac{n_1 n_2}{n_1 + n_2}}).
\]  

(6)

In multiple testing controlling FDR, \( E(R_1)/m_1 \) plays the role of the power of a conventional testing (see Lee and Whitmore, 2002; van den Oord and Sullivan, 2003). With \( E(R_1) \) and the FDR level set at \( r_1 \) and \( f \), respectively, Equation (5) is expressed as

\[
f = \frac{m_\alpha \alpha}{m_\alpha \alpha + r_1}.
\]

By solving this equation with respect to \( \alpha \), we obtain

\[
\alpha^* = \frac{r_1 f}{m_0(1 - f)}.
\]

Given \( m_0, \alpha^* \) is the marginal type I error level for \( r_1 \) true rejections with the FDR controlled at \( f \). With \( \alpha^* \) and \( E(R_1) \) replaced by \( \alpha^* \) and \( r_1 \), respectively, Equation (6) yields an equation \( h(n) = 0 \), where

\[
h(n) = \sum_{j \in M_1} \Phi(z_{\alpha^*} - \delta_j \sqrt{\frac{n_1 n_2}{n_1 + n_2}}) - r_1.
\]  

(7)

We obtain the sample size by solving this equation. In general, solving the equation \( h(n) = 0 \) requires a numerical approach, such as the bisection method:

1. Choose \( s_1 \) and \( s_2 \) such that \( 0 < s_1 < s_2 \) and \( h_1 h_2 < 0 \), where \( h_k = h(s_k) \) for \( k = 1, 2 \). (If \( h_1 h_2 > 0 \) and \( h_1 > 0 \), then choose a smaller \( s_1 \); if \( h_1 h_2 > 0 \) and \( h_1 < 0 \), then choose a larger \( s_1 \).)
2. For \( s_3 = (s_1 + s_2)/2 \), calculate \( h_3 = h(s_3) \).
3. If \( h_1 h_3 < 0 \), then replace \( s_2 \) with \( s_3 \) and \( h_3 \), respectively. Else, replace \( s_1 \) with \( s_3 \) and \( h_3 \), respectively. Go to (2).
4. Repeat (2) and (3) until \( |s_1 - s_2| < 1 \) and \( |h_3| < 1 \), and obtain the required sample size \( n = \{s_3\} + 1 \), where \( \{x\} \) is the largest integer smaller than \( x \).

If we do not have prior information on the effect sizes, we may want to assume equal effect sizes \( \delta_j = \delta (> 0) \) for \( j \in M_1 \). In this case, Equation (7) is reduced to

\[
h(n) = m_1 \Phi(z_{\alpha^*} - \delta \sqrt{\frac{n_1 n_2}{n_1 + n_2}}) - r_1
\]

and, by solving \( h(n) = 0 \), we obtain a closed form formula:

\[
n = \left[ \frac{(z_{\alpha^*} + z_{\delta})^2}{(\alpha^* \delta)^2} + 1 \right]^{-1}.
\]  

(8)

where \( \alpha^* = r_1 f / (m_0(1 - f)) \) and \( \beta^* = 1 - r_1/m_1 \). Note that Equation (8) is the conventional sample size formula when we want to detect an effect size of \( \delta \) with power \( 1 - \beta^* \) while controlling the type I error level at \( \alpha^* \).

In summary, our sample size calculation proceeds as follows:

1. Specify the input parameters:
   (a) \( f \) = FDR level;
   (b) \( r_1 \) = number of true rejections;
Table 2 displays the bisection procedure with starting values $s_1 = 100$ and $s_3 = 200$. The procedure stops after seven iterations and gives $n = 147.7 + 1 = 148$.

### 3.1 Two-sided tests

Suppose one wants to test $H_j: \delta_j = 0$ against $H_j: \delta_j \neq 0$. We reject $H_j$ if $|T_j| > z_{\alpha/2}$ for a certain $\alpha$ level, and obtain the power function $\bar{t}_j(\alpha) = \Phi(z_{\alpha/2} - |\delta_j|/\sqrt{na_1a_2})$. In this case, $\alpha^*$ is the same as that for one-sided test case, i.e.

$$\alpha^* = r_1 f / m_0(1 - f).$$

but Equation (7) is changed to

$$h(n) = \sum_{j \in M_1} \Phi(z_{\alpha/2} - |\delta_j|/\sqrt{na_1a_2}) - r_1. \quad (9)$$

If the effect sizes are constant, i.e. $\delta_j = \delta$ for $j \in M_1$, then we have a closed form formula

$$n = \left[ \frac{(z_{\alpha/2} + z_{\beta/2})^2}{a_1a_2\delta^2} \right]^2 + 1, \quad (10)$$

where $\alpha^* = r_1 f / m_0(1 - f)$ and $\beta^* = 1 - r_1/m_1$.

Now we derive the relationship between the sample size for one-sided test case and that for two-sided test case. Suppose that the input parameters $m, m_1, a_1$ and $\{\delta_j, j \in M_1\}$ are fixed and we want $r_1$ true rejections in both cases. Without loss of generality, we assume that the effect sizes are non-negative. The only difference between the two cases is the parts of $\alpha^*$ in Equation (7) and $\alpha^*/2$ in Equation (9). Let $f_1$ and $f_2$ denote the FDR levels for one- and two-sided testing cases, respectively. Then, the two formulae will give exactly the same sample size as far as these two parts are identical, i.e.

$$\frac{r_1 f_1}{m_0(1 - f_1)} = \frac{r_1 f_2}{2m_0(1 - f_2)},$$

which yields $f_1 = f_2/(2 - f_2)$. In other words, with all other parameters fixed, the sample size for two-sided tests to control the FDR at $f$ can be obtained using the sample size formula for one-sided tests [Equation (7)] by setting the target FDR level at $f/(2 - f)$. Note that this value is slightly larger than $f/2$. The same relationship holds when the effect sizes for prognostic genes are constant.

Example 3. (Two-sided tests and constant effect sizes) We assume $(m, m_1, \delta, a_1, r_1, f) = (4000, 40, 1, 0.5, 24, 0.01)$ as in Example 1.
but we want to use two-sided tests here. Then

$$\alpha^* = \frac{24 \times 0.01}{3960 \times (1 - 0.01)} = 0.612 \times 10^{-4}$$

and $$\beta^* = 1 - 24/40 = 0.4$$, so that $$z_{\alpha^*/2} = 4.008$$ and $$z_{\beta^*} = 0.253$$. Hence, from Equation (10), the required sample size is given as

$$n = \left[ \frac{(4.008 + 0.253)^2}{0.5 \times 0.5 \times 1^2} \right] + 1 = 73.$$  

By the above argument, we obtain exactly the same sample size using formula (8) and $$f = 0.01/(2 - 0.01)$$, which is 7300. Note that this sample size is slightly larger than $$n = 68$$ which was obtained for one-sided tests in Example 1.

### 3.2 Exact formula based on t-distribution

If the gene expression level, or its transformation, is a normal random variable and the available resources are so limited that only a small sample size can be considered, then one may want to use the exact formula based on t-distributions, rather than that based on normal approximation. In one-sided testing case, Equation (5) will be modified to

$$\text{FDR}(\alpha) = \frac{m_{0}\alpha}{m_{0}\alpha + \sum_{j \in M_1} T_{n-2, \delta_j, \sqrt{m_{0}(n-2, \alpha^*)}} - r_1},$$

where $$T_{n-2, \delta_j, \sqrt{m_{0}(n-2, \alpha^*)}}$$ is the survivor function for the non-central t-distribution with $$v$$ degrees of freedom and non-centrality parameter $$n$$, and $$t_{\alpha}(\alpha)$$ is the upper 100$$\alpha$$-th percentile of the central t-distribution with $$v$$ degrees of freedom. The required sample size $$n$$ for $$r_1$$ true rejections with the FDR controlled at $$\alpha$$ solves $$h_T(n) = 0$$, where

$$h_T(n) = \sum_{j \in M_1} T_{n-2, \delta_j, \sqrt{m_{0}(n-2, \alpha^*)}} - r_1$$

and $$\alpha^* = r_1 f / (m_{0}(1 - f))$$. If the effect sizes are constant among the prognostic genes, then the equation reduces to

$$T_{n-2, \delta_j, \sqrt{m_{0}(n-2, \alpha^*)}} = r_1 / m_{1},$$

but, contrary to the normal approximation case, we do not have a closed form sample size formula since $$n$$ is included in both the degrees of freedom and the non-centrality parameter of the t-distribution functions.

Similarly, the sample size for two-sided t-tests can be obtained by solving $$h_T(n) = 0$$, where

$$h_T(n) = \sum_{j \in M_1} T_{n-2, \delta_j, \sqrt{m_{0}(n-2, \alpha^*)}} - r_1$$

and $$\alpha^* = r_1 f / (m_{0}(1 - f))$$. Note that the sample size for FDR = $$f$$ with two-sided testings is the same as that for FDR = $$f/(2 - f)$$ with one-sided testings as in the testing based on normal approximation.

### 4 NUMERICAL STUDIES

In order to investigate the accuracy of the proposed sample size formula, we conducted extensive simulation studies. We set $$m = 4000$$, $$m_1 = 40$$ or 200, constant effect sizes $$\delta = 0.5$$ or 1, and $$\alpha_1 = 0.5$$ or 0.7. We want $$r_1$$ to be 30, 60 or 90% of $$m_1$$ while controlling the FDR level at $$f = 1$$, 5 or 10% using one-sided $$P$$-values. Given a design setting, we first calculate the sample size $$n$$ using formula (8), which is based on normal approximation, and then generate $$N = 5000$$ samples of size $$n$$ from independent normal distributions under the same setting. From each simulation sample, the number of true rejections are counted when controlling the FDR at the specified level using the Storey’s approach discussed in Section 2 with $$\lambda = 0.5$$. The first, second and third quartiles, $$Q_1$$, $$Q_2$$ and $$Q_3$$, of the observed true rejections, $$\hat{r}_1$$, are estimated from the 5000 simulation samples. Table 3 reports $$n$$ and the three quartiles of $$\hat{r}_1$$ for each design setting. We observe that $$n$$ increases in $$|\alpha_1 - 1/2|$$ and $$r_1$$, and decreases in $$\delta$$ and FDR. The median, $$Q_2$$, of $$\hat{r}_1$$ is close to the nominal $$r_1$$ overall except when $$(\alpha_1, m_1, \delta, r_1) = (0.5, 200, 1, 60)$$ or $$(0.7, 200, 1, 60)$$, for which $$n$$ is relatively small and $$r_1$$ tends to be overestimated. With a large $$n$$, $$r_1$$ is very accurately estimated, i.e. $$Q_2$$ is close to $$r_1$$ and the interquartile range $$(Q_3 - Q_1)$$ is small. The interquartile range of $$\hat{r}$$ increases in $$r_1$$, but does not seem to be much dependent on the FDR level.

Figure 1 displays the empirical distribution of $$\hat{r}$$ from 5000 simulations. With $$(\alpha_1, m_1, \delta)$$ fixed at $$(0.5, 0.4, 0.5)$$, the four figures are generated for (1) $$r_1$$, FDR = (12, 0.01), (2) $$r_1$$, FDR = (12, 0.1), (3) $$r_1$$, FDR = (36, 0.01) and (4) $$r_1$$, FDR = (36, 0.1). Note that $$\hat{r}_1$$ is distributed around the nominal $$r_1$$ under each setting. The distributions are truncated by 0 from below and $$m_1 = 40$$ from above, so that they will be skewed to the right if $$r_1$$ is close to 0 and to the left if $$r_1$$ is close to $$m_1$$. As mentioned above, the distribution of $$\hat{r}_1$$ does not seem to depend on FDR, and has less dispersion with a larger $$r_1$$.

Now, we consider a case where we have pilot data. Golub et al. (1999) explored $$m = 6810$$ genes extracted from bone marrow in $$n = 38$$ patients, of which $$n_1 = 27$$ with acute lymphoblastic leukaemia and $$n_2 = 11$$ with acute myeloid leukaemia, in order to identify the susceptible genes with potential clinical heterogeneity in the two subclasses of leukaemia. Suppose that we use the dataset from this study as pilot data in designing a new study with the same study objective. For gene $$j$$ ($$j = 1, \ldots, 6810$$), we calculated the sample means $$\bar{x}_j, \bar{y}_j$$ and the sample variances

$$s_j^2 = \frac{\sum_{i=1}^{n_1}(x_{ij} - \bar{x}_j)^2 + \sum_{i=1}^{n_2}(y_{ij} - \bar{y}_j)^2}{n_1 + n_2 - 2}$$

and estimated the effect sizes

$$\delta_j = \frac{\bar{x}_j - \bar{y}_j}{\sqrt{n_1^{-1} + n_2^{-1}}}$$

from the pilot data. In order to reflect the variability of the estimated effect sizes and for a slightly conservative sample size, we multiply 0.6 with the observed effect sizes, i.e. $$\hat{\delta}_j = 0.6\delta_j$$, in the following sample size calculation. We assume that the top $$m_1 = 50$$ genes with the largest effect sizes in absolute value are prognostic. Suppose that we want to identify 60% of the prognostic genes, i.e. $$r_1 = 0.6 \times 50 = 30$$, while controlling the FDR at $$f = 1$$% level using two-sided $$P$$-values. Based on the pilot data, we set $$\alpha_1 = 0.7(=27/38)$$ and $$m = 7000$$. In this case, we have

$$\alpha^* = \frac{30 \times 0.01}{6970 \times (1 - 0.01)} = 0.436 \times 10^{-4},$$

so that $$z_{\alpha^*/2} = 4.088$$. From Equation (9), we solve

$$\sum_{j=1}^{50} \Phi(4.088 - \delta_j \sqrt{0.7 \times 0.3 \times n}) = 30$$
### Table 3. Sample size \( n \) for \( r_1 (=30, 60 \) or \( 90\% \) of \( m_1 \)) true rejections at FDR=1, 5 or 10% level by one-sided tests when \( m = 4000, m_1 = 40 \) or \( 200, \delta = 0.5 \) or \( 1, a_1 = 0.5 \) or \( 0.7 \)

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( m_1 )</th>
<th>( \delta )</th>
<th>( r_1 )</th>
<th>FDR = 1%</th>
<th>5%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>40</td>
<td>0.5</td>
<td>12</td>
<td>12 (9, 15)/195</td>
<td>12 (9, 14)/152</td>
<td>12 (8, 14)/133</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>24 (22, 26)/269</td>
<td>24 (21, 26)/216</td>
<td>24 (21, 26)/192</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>36 (35, 37)/404</td>
<td>36 (35, 37)/337</td>
<td>36 (35, 37)/306</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>12</td>
<td>13 (10, 16)/49</td>
<td>13 (10, 16)/38</td>
<td>14 (11, 17)/34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>25 (22, 27)/68</td>
<td>24 (22, 27)/54</td>
<td>24 (22, 27)/48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>36 (35, 37)/101</td>
<td>36 (35, 37)/85</td>
<td>36 (35, 37)/77</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>60</td>
<td>62 (56, 68)/152</td>
<td>61 (55, 68)/110</td>
<td>62 (55, 69)/92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>120</td>
<td>121 (115, 126)/216</td>
<td>120 (114, 126)/163</td>
<td>121 (115, 127)/140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>180</td>
<td>180 (177, 183)/337</td>
<td>180 (177, 183)/268</td>
<td>180 (177, 183)/236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>60</td>
<td>67 (61, 73)/58</td>
<td>71 (64, 78)/28</td>
<td>72 (65, 78)/23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>120</td>
<td>121 (115, 127)/54</td>
<td>122 (117, 128)/41</td>
<td>123 (117, 129)/35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>180</td>
<td>180 (177, 183)/85</td>
<td>179 (176, 182)/67</td>
<td>180 (176, 183)/59</td>
</tr>
<tr>
<td>0.7</td>
<td>40</td>
<td>0.5</td>
<td>12</td>
<td>12 (9, 14)/232</td>
<td>11 (9, 14)/181</td>
<td>11 (8, 14)/158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>24 (22, 26)/320</td>
<td>24 (21, 26)/257</td>
<td>24 (21, 26)/228</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>36 (35, 37)/481</td>
<td>36 (35, 37)/401</td>
<td>36 (35, 37)/364</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>12</td>
<td>13 (10, 15)/58</td>
<td>13 (10, 15)/46</td>
<td>14 (11, 16)/40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>24 (22, 27)/80</td>
<td>24 (22, 27)/65</td>
<td>24 (22, 27)/57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>36 (35, 37)/121</td>
<td>36 (35, 37)/101</td>
<td>36 (35, 37)/91</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>60</td>
<td>62 (55, 68)/181</td>
<td>61 (55, 68)/131</td>
<td>62 (55, 69)/110</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>120</td>
<td>121 (115, 127)/257</td>
<td>120 (114, 126)/194</td>
<td>119 (114, 126)/166</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>180</td>
<td>180 (177, 183)/401</td>
<td>180 (177, 183)/319</td>
<td>180 (177, 183)/281</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>60</td>
<td>65 (59, 72)/46</td>
<td>64 (57, 70)/33</td>
<td>71 (65, 78)/28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>120</td>
<td>122 (116, 128)/65</td>
<td>121 (114, 126)/49</td>
<td>122 (115, 128)/42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>180</td>
<td>180 (177, 183)/101</td>
<td>180 (177, 183)/80</td>
<td>180 (177, 183)/71</td>
</tr>
</tbody>
</table>

Each cell consists of \( Q_1, Q_2, Q_3)/n \), where \( n \) is the required sample size, and \( Q_1, Q_2 \) and \( Q_3 \) are the first, second and third, respectively, quartiles of the observed number of true rejections from 5000 simulations.

![Graphs](https://academic.oup.com/bioinformatics/article-abstract/21/14/3097/266798/fig1)
using the bisection method, and obtain $n = 58$, or $(n_1, n_2) \approx (41, 17)$. We generated 5000 simulation samples of size $n = 58$ under the design setting, and observed the quartiles $Q_2(Q_1, Q_3) = 30(28, 33)$ from the empirical distribution of the observed true rejections. Note that the median $Q_2$ exactly matches the projected $r_1$ in this case. Table 4 reports the sample sizes under different settings: $m_1 = 50$ or 100; $r_1/m_1 = 0.3$ or 0.6; and FDR = 1, 5 or 10%.

A referee raised a question about the accuracy of our sample size estimate when the gene expression data are correlated or have other distributions than normal distributions. In order to address this issue, we first consider normal gene expression data with a block compound symmetry (CS) correlation structure: there are 400 independent blocks, and each block consists of 10 dependent genes with a CS structure and correlation coefficient $\rho = 0.6$. The first half of Table 5 reports the distribution of the empirical true rejections under $(a_1, m_1, \delta) = (0.5, 40, 1)$ and $r_1 = 12$ or 24. We assume that the prognostic genes belong to the first four blocks. Note that the estimated sample sizes are given in Table 3 under the same design settings. From Table 5, we observe that the median $Q_{2}$ of the observed true rejections is close to the nominal $r_1$ as in the independent data case. However, the interquartile range is almost doubled from that under independence, from $\sim 5$ to $\sim 10$. In the second set of simulations, we generate gene expression data from a correlated asymmetric distribution: for $b = 1, \ldots, 400$ and $10(b - 1) + 1 \leq j \leq 10b,$

$$X_j = \delta_j + (e_{1j} - 2)(1 - \rho)/\sqrt{4} + (e_{2j} - 2)\sqrt{\rho}/\sqrt{4}$$

$$Y_j = (e_{2j} - 2)(1 - \rho)/\sqrt{4} + (e_{2j} - 2)\sqrt{\rho}/\sqrt{4},$$

where $\rho = 0.6$ and $e_{11}, \ldots, e_{4000}, e_{41}, \ldots, e_{400}$ are i.i.d. random variables from the $N(0,1)$-distribution with 2 degrees of freedom. Note that both $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_m)$ have marginal variances 1, and the same block CS correlation structure as in the above correlated normal data case. The second half of Table 5 reports the simulation results. We observe almost the same results as in the correlated normal data case. Benjamini and Yekutieli (2001) investigate general distributional assumptions for the control of FDR.

5 DISCUSSION

Microarray has been a major high-throughput assay method to display DNA or RNA abundance for a large number of genes concurrently. Discovery of the prognostic genes should be made taking multiplicity into account, but also with enough statistical power to identify important genes successfully. Owing to the costly nature of microarray experiments, however, often only a small sample size is available and the resulting data analysis does not give reliable answers to the investigators. If the findings from a small study look promising, a large-scale study may be developed to confirm the findings using appropriate statistical tools. Our sample size formula will play the role in the design stage of such a confirmatory study. It can be used to check the statistical power, $r_1/m_1$, of a small-scale pilot study too.

The proposed method is to calculate the sample size for a specified number of true rejections (or the expected number of true rejections given a sample size) while controlling the FDR at a given level. The input variables to be pre-specified are total number of genes for testing $m$, projected number of prognostic genes $m_1$, allocation proportions $a_i$ between groups and effect sizes for the prognostic genes. The method does not require any heavy computation, such as Monte Carlo simulations, so that we get a sample size in a second. Especially, if the effect sizes among the prognostic genes are the same, we have a closed form formula that can be calculated using a scientific calculator and a normal distribution table. The proposed method can be used to design a new study based on the parameter values estimated from the pilot data.

It is shown through simulations that the formula based on normal approximation works well overall, even when the expression levels are weakly correlated or have skewed distributions. If there exists dependency among the genes, the observed number of true rejections tends to have a wide variation around the nominal $r_1$. The computer program for sample size calculation is available from the author.

ACKNOWLEDGEMENTS

The author wants to thank the two reviewers for their valuable comments.

REFERENCES


Table 4. Sample size $n$ for $r_1$ (=30 or 60% of $m_1$) true rejections at FDR=1, 5 or 10% level by two-sided $P$-values when $m = 7000$, $m_1 = 50$ or 100, $a_1 = 0.7$

<table>
<thead>
<tr>
<th>$m_1$</th>
<th>$r_1/m_1$</th>
<th>FDR = 1%</th>
<th>5%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.3</td>
<td>39</td>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>58</td>
<td>47</td>
<td>42</td>
</tr>
<tr>
<td>100</td>
<td>0.3</td>
<td>47</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>69</td>
<td>56</td>
<td>50</td>
</tr>
</tbody>
</table>

The effect sizes are estimated from Golub et al. (1999) data.

Table 5. Simulation results from normal or mixture of $\chi^2$-distributions with 400 independent blocks and CS correlation structure with $\rho = 0.6$ within each block of size 10

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>$\chi^2$-mixture</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r_1$</td>
<td>FDR = 1%</td>
<td>5%</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11 (6, 16)</td>
<td>12 (6, 18)</td>
<td>13 (7, 19)</td>
<td>12 (7, 17)</td>
<td>15 (9, 21)</td>
</tr>
<tr>
<td>24</td>
<td>20 (16, 25)</td>
<td>22 (17, 27)</td>
<td>23 (17, 28)</td>
<td>20 (15, 24)</td>
<td>22 (17, 27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 or 60%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other parameters are set at $(a_1, m_1, \delta) = (0.5, 40, 1)$ and $r_1 = 12$ or 24. Each cell consists of $(Q_2(Q_1, Q_3))$ of the observed number of true rejections from 5000 simulations. The sample sizes are given in Table 3 under the same setting for $(a_1, m_1, \delta, r_1)$. 

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/14/3097/266798 by guest on 15 February 2019


