Sample size calculation for microarray experiments with blocked one-way design
Sin-Ho Jung*1, Insuk Sohn1, Stephen L George1, Liping Feng2 and Phyllis C Leppert2

Address: 1Department of Biostatistics and Bioinformatics, Duke University Medical Center, North Carolina 27710, USA and 2Center for Fibroid Biology and Therapy, Duke University Medical Center, North Carolina 27710, USA

Email: Sin-Ho Jung* - sinho.jung@duke.edu; Insuk Sohn - insuk.sohn@duke.edu; Stephen L George - stephen.george@duke.edu; Liping Feng - liping.feng@duke.edu; Phyllis C Leppert - phyllis.leppert@duke.edu

* Corresponding author

Abstract

Background: One of the main objectives of microarray analysis is to identify differentially expressed genes for different types of cells or treatments. Many statistical methods have been proposed to assess the treatment effects in microarray experiments.

Results: In this paper, we consider discovery of the genes that are differentially expressed among K (> 2) treatments when each set of K arrays consists of a block. In this case, the array data among K treatments tend to be correlated because of block effect. We propose to use the blocked one-way ANOVA F-statistic to test if each gene is differentially expressed among K treatments. The marginal p-values are calculated using a permutation method accounting for the block effect, adjusting for the multiplicity of the testing procedure by controlling the false discovery rate (FDR). We propose a sample size calculation method for microarray experiments with a blocked one-way design. With FDR level and effect sizes of genes specified, our formula provides a sample size for a given number of true discoveries.

Conclusion: The calculated sample size is shown via simulations to provide an accurate number of true discoveries while controlling the FDR at the desired level.

Background

Clinical and translational medicine have benefited from genome-wide expression profiling across two or more independent samples, such as various diseased tissues compared to normal tissue. DNA microarray is a high throughput biotechnology designed to measure simultaneously the expression level of tens of thousands of genes in cells. Microarray studies provide the means to understand the mechanisms of disease. However, various sources of error can influence microarray results [1]. Microarrays also present unique statistical problems because the data are high dimensional and are insufficiently replicated in many instances. Methods of adjustment for multiple testing therefore become extremely important. Multiple testing methods controlling the false discovery rate (FDR) [2] have been popularly used because they are easy to calculate and less strict in controlling the false positivity compared to the family-wise error rate (FWER) control method [3].
Numerous sample size calculation methods have been proposed for comparing independent groups while controlling the FDR in designing microarray studies. Lee and Whitmore [4] considered comparing multiple groups using ANOVA models and derived the relationship between the effect sizes and the FDR using a Bayesian approach. Their power analysis does not address the multiple testing issue. Muller et al. [5] chose a pair of testing errors, including FDR, and minimized one while controlling the other at a specified level using a Bayesian decision rule. Jung [6] proposed a closed form sample size formula for a specified number of true rejections while controlling the FDR at a desired level. Pounds and Cheng [7] and Liu and Hwang [8] proposed similar sample size formulas which can be used for comparison of K independent samples. These methods are for the FDR-control methods based on independence or a weak dependency assumption among test statistics. Recently, Shao and Tseng [9] introduced an approach for calculating sample sizes for multiple comparisons accounting for dependency among test statistics.

In some studies, specimens for K treatments are collected from the same subject and means are compared across treatment groups. In this case, the gene expression data for the K treatments may be dependent since they share the same physiological conditions. For example, Feng et al. [10] conducted a study to discover the genes differentially expressed between center (C) and edge (E) of the uterine fibroid and the matched adjacent myometrium (M). In this study, specimens are taken from the three sites for each patient. The patients are blocks and the three sites (K = 3), C, E and M, are treatments (or groups) to be compared.

Since a set of K specimens are collected from each patient, we require a much smaller number of patients than a regular unblocked design. Furthermore, the observations within each block tend to be positively correlated, so that a blocked design requires a smaller number of arrays than the corresponding unblocked design just as a paired two-sample design with a positive pairwise correlation requires a smaller number of observations than a two independent sample design. The more heterogeneous the blocks are, the greater the savings in number of arrays for the blocked design.

In this paper, we consider a non-parametric blocked F-test statistic to compare the gene expression level among K dependent groups. We adjust for multiple testing and control the FDR by employing a permutation method. We propose a sample size calculation method for a specified number of true rejections while controlling the FDR at a specified level. Through simulations, we show that the blocked F-test accurately controls the FDR using the permutation resampling method and the calculated sample size provides an accurate number of true rejections while controlling the FDR at the desired level. For illustration, the proposed methods are applied to the fibroid study [10] mentioned above.

Methods
Non-parametric block F-test statistic
Suppose that we want to discover genes that are differentially expressed among K sites (treatments or groups). For each of n patients (blocks), a specimen is collected from each site for a microarray experiment on m genes. In this case, the gene expression data from the K sites tend to be correlated. Let Yijk denote the expression level of gene i (= 1,..., m) from treatment k (= 1,..., K) of block j (= 1,..., n). We consider the blocked one-way ANOVA model

\[ Y_{ijk} = \mu_i + \delta_{ik} + \gamma_{ij} + \epsilon_{ijk}, \]

where, for gene i, \( \mu_i \) is the population mean, \( \delta_{ik} \) is a fixed treatment effect and the primary interest, \( \gamma_{ij} \) is a random block effect, and \( \epsilon_{ijk} \) is a random error term. We assume that \( \sum_{k=1}^{K} \delta_{ik} = 0 \), \( \gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in} \) are independent and identically distributed (IID) with mean 0 and variance \( \sigma^2 \), and error terms and block effects are independent. The standard ANOVA theory using parametric F distributions to test the treatment effect assumes a normal distribution for \( \epsilon_{ijk} \). However, in this paper, we avoid the normality assumption by using a permutation resampling method in testing and a large-sample approximation in sample size calculation.

For gene i(= 1,..., m), the hypotheses for testing the treatment effect are described as

\[ H_i : \delta_{i1} = \delta_{i2} = \ldots = \delta_{iK} = 0 \]

against

\[ \overline{H}_i : \delta_{ik} \neq 0 \text{ for some } k = 1,\ldots, K. \]

We reject \( H_i \) in favor of \( \overline{H}_i \) for a large value of F-test statistic

\[
F_i = \frac{\frac{n}{K} \sum_{k=1}^{K} (\overline{Y}_{1k} - \overline{Y}_{i.})^2 / (K-1)}{\sum_{j=1}^{n} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{ij} - \overline{Y}_{i.} + \overline{Y}_{..})^2 / \{(K-1)(n-1)\}},
\]

(2)
where $\overline{Y}_{i,k} = n^{-1}\sum_{j=1}^{n} Y_{ijk}$, $\overline{Y}_{i,j} = K^{-1}\sum_{k=1}^{K} Y_{ijk}$, and $\overline{Y}_{i,j} = (Kn)^{-1}\sum_{j=1}^{K}\sum_{k=1}^{K} Y_{ijk}$. If the error terms are normally distributed, $F_i$ marginally has the $F_{K-1,(K-1)(n-1)}$ distribution under $H_i$. The normality assumption can be relaxed if $n$ is large.

Without the normality assumption, the joint null distribution of the statistics can be approximated using a block permutation method, where the array data sets for $K$ treatments are randomly shuffled within each block: the permuted data may be represented as $\{(Y_{ij_1},...,Y_{ij_m}), 1 \leq i \leq m, 1 \leq j \leq n\}$, where $(\tilde{i},...,\tilde{K})$ is a random permutation of $(1,...,K)$. Note that there are $(K!)^n$ different permutations, among which $(K!)^{n-1}$ give different $F$-statistic values. The R language package multtest [11] can be used to implement the permutation-based multiple testing procedure for blocked microarray data. We consider adjusting for the multiplicity of the testing procedure 
by controlling the FDR [12,13].

**Permutation-based multiple testing for FDR-control**

(i) Compute the $F$-test statistics $(F_1,...,F_m)$ from the original data, $(f_1,...,f_m)$.

(ii) From the $b$-th permutation data ($b = 1,...,B$), compute the $F$-test statistics $(F_1^{(b)},...,F_m^{(b)})$.

(iii) For gene $i$, estimate the marginal p-value by

$$p_i = \frac{1}{B} \sum_{b=1}^{B} I(F_i^{(b)} \geq f_i),$$

where $I(A)$ is an indicator function of event $A$.

(iv) For a chosen constant $\lambda \in (0, 1)$, estimate the q-value by

$$q_i = \frac{m}{(1-\lambda) \sum_{l=1}^{m} I(p_l \leq p_i)} \sum_{l=1}^{m} I(p_l > \lambda).$$

(v) For a specified FDR level $q^*$, discover gene $i$ (or reject $H_i$) if $q_i < q^*$.

**Sample size calculation**

Let $m_0$ and $m_1$ denote the sets of indices of genes that are equally and differentially expressed, respectively, in $K$ treatments, and $\{ \overline{\delta}_{ik} = \delta_{ik}/\sigma_i, i \in m_1, 1 \leq k \leq K \}$ denote the standardized effect sizes for the differentially expressed genes. Let $m_0$ and $m_1 = m - m_0$ denote the cardinalities of $m_0$ and $m_1$, respectively.

Suppose that we want to discover gene $i$ (or reject $H_i$) if the marginal p-value $p_i$ is smaller than $\alpha \in (0, 1)$. For large $m$ and under the independence assumption or weak dependence among the $F$-test statistics, the FDR corresponding to the cutoff value $\alpha$ is approximated by

$$\text{FDR}(\alpha) = \frac{m_0 \alpha}{m_0 \alpha + \sum_{i \in M_1} \beta_i(\alpha)},$$

where $\beta(\alpha) = P(p_i \leq \alpha)$ is the marginal power of a single $\alpha$-test applied to gene $i \in m$ and $\sum_{i \in M_1} \beta_i(\alpha)$ denotes the expected number of true rejections when we reject $H_i$ for $p_i < \alpha$, see Jung [6].

Now, we derive $\beta(\alpha)$ for gene $i \in m$. By the standard blocked one-way ANOVA theory under the normality assumption for $\delta_{ijk}$

$$\sum_{k=1}^{K} (\overline{Y}_{i,k} - \overline{Y}_{i,j})^2 \sim \frac{\sigma_i^2}{n} \chi^2_{K-1}(n \sum_{k=1}^{K} \overline{\delta}_{ik}^2)$$

and

$$\frac{1}{(K-1)(n-1)} \sum_{j=1}^{n} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{ij} - \overline{Y}_{i,k} + \overline{Y}_{i,j})^2 \sim \sigma_i^2 \chi^2_{n-1}(K-1)(n-1)$$

are independent, where $\chi^2_{\nu}(\eta)$ is the noncentral $\chi^2$-distribution with $\nu$ degrees of freedom and noncentrality parameter $\eta$, and $\chi^2_{\nu} = \chi^2_{\nu}(0)$. Hence, for the $F$-test statistic (2), we have

$$F_i \sim F_{(K-1),(K-1)(n-1)}(n \sum_{k=1}^{K} \overline{\delta}_{ik}^2)$$

where $F_{(K-1),(K-1)(n-1)}$ is the noncentral $F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom, and noncentrality parameter $\eta$. Note that, for $i \in m_0$, $\overline{\delta}_{i1} = \cdots = \overline{\delta}_{iK} = 0$ and $F_i \sim F_{(K-1),(K-1)(n-1)}(0) = F_{(K-1),(K-1)(n-1)}$, the central $F$-distribution.

The marginal powers are expressed as
\[
\beta_i(\alpha) = P\{\sum_{k=1}^{K} \delta_{ik}^2 \geq F_{1-K}\} \geq F_{1-K}\},
\]

where \( F_{1-K}\) denotes the 100(1 - \( \alpha \)) percentile of \( F_{1-K} \) distribution. The marginal powers can be calculated using R, SAS or some other packages. Suppose we want \( r_1 \) true rejections while controlling the FDR at \( q^* \). By combining this with (3) and (4), we obtain two equations

\[
r_1 = \sum_{i \in M_1} P\{\sum_{k=1}^{K} \delta_{ik}^2 \geq F_{1-K}\},
\]

and

\[
q^* = \frac{m_0 \alpha}{m_0 \alpha + r_1}.
\]

Note that \( r_1/m_1 \) denotes the probability of true rejection.

At the design stage of a study, \( m \) is given by the number of genes included in the chips to be used for microarray experiment, \( m_1 \) and \{ \( \delta_{ik}, i \in 1, 1 \leq k \leq K \) \} are projected based on biological knowledge or estimated from pilot data, and \( K, r_1 \) (or \( r_1/m_1 \)) and \( q^* \) are prespecified. The only unknown variables in (5) and (6) are \( \alpha \) and \( n \). By solving (6) with respect to \( \alpha \), we obtain \( \alpha^* = r_1 q^*/(m_0 (1 - q^*)) \) and, by plugging this in (5), we obtain an equation for \( r_1 \) depending only on \( n \),

\[
r_1 = \sum_{i \in M_1} P\{\sum_{k=1}^{K} \delta_{ik}^2 \geq F_{1-K}\} \geq F_{1-K}\},
\]

The marginal power function (4) includes \( n \) in the degrees of freedom of the denominator as well as the noncentrality parameter of the \( F \)-distributions. The impact of the degrees of freedom of the denominator of the \( F \)-statistic on the marginal power is much weaker than that of the noncentrality parameter, so that \( \beta_i(\alpha) \) is a monotone increasing function of \( n \), and consequently equation (7) has a unique solution. Figure 1 demonstrates the relationship between \( n \) and \( \beta_i(\alpha) \) with \( \alpha = 0.05 \) and \( \delta_{ik} = (k - (K + 2)/2)/K \) for \( 1 \leq k \leq K \). This monotone relationship becomes clear for large \( n \) as shown by an approximate sample size formula given below. Note that the variance of block effect \( \nu_i \) has no impact on the sample size and power of the test statistic for treatment effect.

In summary, the sample size (i.e., number of blocks) \( n \) for \( r_1 (\leq m_1) \) true rejections is calculated as follows, assuming that the error terms in model (1) are normally distributed.

**Sample size calculation based on the noncentral \( F \)-distribution**

(i) Specify the input variables:
- \( K = \) number of treatments;
- \( m = \) total number of genes for testing;
- \( m_1 = \) number of genes differentially expressed in \( K \) treatments \( (m_0 = m - m_1) \);
- \{ \( \delta_{ik}, i \in 1, 1 \leq k \leq K \) \} = standardized effect sizes for prognostic genes;
- \( q^* = \) FDR level;
- \( r_1 = \) number of true rejections

(ii) Using the bisection method, solve...
If the error terms have a finite 4-th moment, then, for large samples, applying a normal distribution assumption of the error terms is not a problem. The power functions based on the noncentral F-distribution, the relationship between power and sample size, can be obtained by solving the equation,

\[
r_1 = \sum_{i \in M_1} P \{ F_{(K-1),(K-1),(n-1)} \left( n \sum_{k=1}^{K} \delta_{ik}^2 \right) \geq F_{(K-1),(K-1),(n-1),\alpha} \},
\]

with respect to \( n \), where \( \alpha^* = r_1 q^*/\{m_0(1 - q^*)\} \).

(iii) The required sample size is \( n \) blocks, or \( nK \) array chips.

In the sample size formula based on the noncentral F-distribution, the relationship between \( n \) and the marginal power functions based on the F-distribution is complicated and a normal distribution assumption of the error terms is required. In the large sample case, we can loosen the normality assumption and simplify this relationship. If the error terms have a finite 4-th moment, then, for large \( n \), the distribution of \( F_i \) is approximated by

\[
\frac{1}{(K-1)} \chi^2_{K-1}(n \sum_{k=1}^{K} \delta_{ik}^2).
\]

A proof is given in the Appendix. Similarly, for large \( n \), the distribution of \( F_{(K-1),(K-1),(n-1)} \) is approximated by \( (K-1)^{-1} \chi^2_{K-1} \), so that \( F_{(K-1),(K-1),(n-1),\alpha} \approx \chi^2_{K-1,\alpha} \), where \( \chi^2_{v,\alpha} \) is the 100(1 - \( \alpha \)) percentile of the \( \chi^2 \) distribution with \( v \) degrees of freedom. Hence, the marginal power for \( F_i \) is approximated by

\[
\beta_i(\alpha) = P \{ \chi^2_{K-1}(n \sum_{k=1}^{K} \delta_{ik}^2) \geq \chi^2_{K-1,\alpha} \},
\]

and a sample size based on the \( \chi^2 \)-distribution approximation is obtained by solving

\[
r_1 = \sum_{i \in M_1} P \{ \chi^2_{K-1}(n \sum_{k=1}^{K} \delta_{ik}^2) \geq \chi^2_{K-1,\alpha} \},
\]

with respect to \( n \), where \( \alpha^* = r_1 q^*/\{m_0(1 - q^*)\} \).

Equation (8) is especially useful when we want to compare the powers between a blocked one-way design and an unblocked one-way design. Using similar approximations, it is easy to show that an approximate sample size \( N = nK \) for a study with unblocked one-way design with a balanced allocation is obtained by solving

\[
r_1 = \sum_{i \in M_1} P \{ \chi^2_{K-1}(n \sum_{k=1}^{K} \delta_{ik}^2) \geq \chi^2_{K-1,\alpha} \},
\]

with respect to \( n \), where \( \delta_{ik} = \delta_{ik} / \sqrt{v_i + \sigma_i^2} \). The only difference between (8) and (9) is the standardized effect sizes, \( \delta_{ik} = \delta_{ik} / \sigma_i \), and \( \delta_{ik} = \delta_{ik} / \sqrt{v_i + \sigma_i^2} \). The latter is always smaller than the former because of the variance among blocks, \( v_i \). If \( v_i \) is large compared to the variance of experimental errors, \( \sigma_i^2 \), then a blocked one-way design requires much smaller number of arrays than an unblocked one-way design. Let \( n_u \) and \( n_b \) denote the sample sizes \( n \) calculated under an unblocked and a blocked design, respectively. If \( v_i / \sigma_i^2 \) are constant \( f \) among the prognostic genes, then from (8) and (9), we have \( n_u = (1 + f)n_b \). As an example, consider the design of the fibrosis study as described in Background Section and suppose that the variance of the block effects is half of that of measurement errors for the prognostic genes, i.e. \( f = 0.5 \). In this case, if a blocked design requires \( n_b = 100 \) patients and \( 3n_b = 300 \) array chips, then the corresponding unblocked design with a balanced allocation requires \( n_u = 150 \) patients per group or a total 450 patients. For an unblocked design, the number of array chips is identical to that of patients, and compared to the blocked design, the unblocked design requires 1.5 times more chips and 4.5 times more patients.

Results and discussion

Simulations

First, we investigate the accuracy of the FDR control based on blocked one-way ANOVA tests and the sample size formulas via simulations. For the simulations on FDR control, we consider blocked one-way designs with \( K = 3 \) treatments and \( n = 10, 30, \) or 50 blocks. For gene \( i (= 1, \ldots, m) \) from treatment \( k (= 1, \ldots, K) \) of block \( j (= 1, \ldots, n) \), block effect \( \gamma_{ij} \) and error terms \( \epsilon_{ijk} \) are generated from \( N(0, 0.5^2) \) and \( N(0,1) \), respectively. For differentially expressed genes \( i \in I \), the standardized treatment effects are set at \((\delta_{i1}, \delta_{i2}, \delta_{i3}) = (1, 0, -1) \) or \((1, -2, 1) \). We set the total number of genes \( m = 4000 \); the number of differentially expressed genes \( m_1 = 40 \) or 200; and the nominal FDR level \( \alpha = 0.05, 0.1, 0.2, 0.3, 0.4, \) or 0.5. We conducted \( N = 1000 \) simulations under each setting, and the null distribution of the test statistics is approximated from \( B = 1000 \) permutations for each simulation sample. In simulation \( I(= 1, \ldots, N) \), the FDR-control multiple testing method is applied to the simulated data using tuning parameter \( \lambda = 0.95 \) [12] to count the numbers of total rejections \( \tilde{r} \) and
false rejections \( \hat{r}_0 \) and to estimate the FDR, \( \hat{q} = \hat{r}_0 / \hat{r} \).

Then the empirical FDR is obtained as

\[
\hat{q} = \frac{1}{N} \sum_{i=1}^{N} \hat{q}_i.
\]

Table 1 reports the simulation results. The testing procedure controls the FDR accurately, i.e. \( \hat{q} = q^* \approx q^* \), when \( m_1 \) is large (\( m_1 = 200 \)), but tends to be anti-conservative, i.e. \( \hat{q} > q^* \), when \( m_1 \) is small (\( m_1 = 40 \)). Jung and Jang [13] made similar observations for two-sample t-tests and Cox regression.

For the simulations on sample size calculation, we set \( m = 4000; m_1 = 40 \) or 200; number of treatment \( K = 3 \); treatment effects \( (\delta_{11}, \delta_{12}, \delta_{13}) = (1/4, 0, -1/4) \) or \( (1/4, -1/2, 1/4) \) for \( i \in \{1, 2, 3\} \). We want the number of true rejections \( r_1 \) to be 30%, 60% or 90% of \( m_1 \) while controlling the FDR level at \( q^* = 1\% \), 5% or 10%. For each design setting, we first calculate the sample size \( n \) based on the F-distribution or the chi-square approximation, and then generate \( N = 1000 \) samples of size \( n \) under the same setting. From each simulation sample, the number of true rejections are counted while controlling the FDR at the specified level using \( \lambda = 0.95 \). 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 1000 simulation samples.

Table 2 summarizes the simulation results by the two methods. As expected, sample size increases in \( r_1 \) and decreases in \( m_1 \) and \( q^* \). Since the standardized effect sizes for the differentially expressed genes influence the sample size through their sum of squares, the combination of effect sizes \((1/4, 0, -1/4)\) requires a larger sample size than \((1/4, -1/2, 1/4)\). The sample size based on the chi-square approximation is always smaller than that based on the F-distribution. The median \( (Q_{2}) \) of the empirical true rejections \( \hat{r}_1 \) is smaller than the nominal \( r_1 \) for the sample size based on the chi-square approximation, especially with a small \( n \), while the sample size based on the F-distribution is always accurately powered, i.e. \( Q_{2} \approx r_1 \).

![Table 1: Empirical FDR from N = 1000 simulations with B = 1000 permutations for each simulation data set](image-url)

Table 1: Empirical FDR from \( N = 1000 \) simulations with \( B = 1000 \) permutations for each simulation data set

| \( m_1 \) | \( (\delta_{11}, \delta_{12}, \delta_{13}) \) | \( q^* \) | 10 | 30 | 50 |
|---|---|---|---|---|---|
| 40 | \( (1, 0, -1) \) | 0.05 | 0.1766 | 0.0921 | 0.0925 |
| | | 0.1 | 0.1819 | 0.1647 | 0.1705 |
| | | 0.2 | 0.2736 | 0.2462 | 0.2506 |
| | | 0.3 | 0.3636 | 0.3478 | 0.3512 |
| | | 0.4 | 0.4546 | 0.4449 | 0.4431 |
| | | 0.5 | 0.5435 | 0.5389 | 0.5399 |
| | \( (1, -2, 1) \) | 0.05 | 0.0936 | 0.0899 | 0.0915 |
| | | 0.1 | 0.1619 | 0.1663 | 0.1665 |
| | | 0.2 | 0.2402 | 0.2498 | 0.2421 |
| | | 0.3 | 0.3373 | 0.3469 | 0.3461 |
| | | 0.4 | 0.4347 | 0.4481 | 0.4421 |
| | | 0.5 | 0.5318 | 0.5446 | 0.5340 |
| 200 | \( (1, 0, -1) \) | 0.05 | 0.0653 | 0.0573 | 0.0603 |
| | | 0.1 | 0.1120 | 0.1093 | 0.1130 |
| | | 0.2 | 0.2076 | 0.2105 | 0.2146 |
| | | 0.3 | 0.3079 | 0.3086 | 0.3176 |
| | | 0.4 | 0.4070 | 0.4056 | 0.4171 |
| | | 0.5 | 0.5051 | 0.5013 | 0.5162 |
| | \( (1, -2, 1) \) | 0.05 | 0.0567 | 0.0554 | 0.0591 |
| | | 0.1 | 0.1080 | 0.1079 | 0.1111 |
| | | 0.2 | 0.2142 | 0.2061 | 0.2116 |
| | | 0.3 | 0.3120 | 0.3052 | 0.3113 |
| | | 0.4 | 0.4124 | 0.4049 | 0.4148 |
| | | 0.5 | 0.5141 | 0.5010 | 0.5162 |
Example

We applied the permutation-based blocked one-way ANOVA and the sample size calculation method to the fibroid study discussed in the Background Section. From each patient, specimens are taken from two sites of fibroid tissue, center (C) and edge (E), and one normal myometrium (M). Five patients are accrued to the study. We regard the three sites as treatments ($K = 3$) and the patients as blocks ($n = 5$). mRNA was amplified and hybridized onto HG-U133 GeneChips according to the protocols recommended by Affymetrix (Santa Clara, CA), and $m = 54675$ probe sets on the array were analyzed. Expression values were calculated using the Robust Multichip Average (RMA) method [14]. RMA estimates are based upon a robust average of background corrected PM intensities. Normalization was done using quantile normalization [15]. We filtered out all "AFFX" genes and genes for which there were 4 or fewer present calls (based on Affymetrix's present/marginal/absent (PMA) calls using mismatch probe intensity, the ratio of PM to MM). That is, a gene is included only if there are at least 3 present calls among the 15 PMA calls. Filtering yielded 30711 genes to be used in the subsequent analyses.

In order to group the samples according to the degree of similarity present in the gene expression data, we first applied a hierarchical clustering analysis to the filtered 30711 gene expression data and generated a dendrogram.

### Table 2: $Q_2 (Q_1, Q_3)/n$, where $n$ is the sample size and $Q_k (k = 1, 2, 3)$ are the $k$-th quartile of the empirical true rejections $r_1$ from $N = 1000$ simulations

| $m_1$ | $r_1$ | $q^*$ = 1% | 5% | 10% |
|-------|-------|-------------|----|-----|
|       |       |             |    |     |

Based on the chi-square approximation

$\left( \overline{\delta}_{11}, \overline{\delta}_{12}, \overline{\delta}_{13} \right) = (1/4, 0, -1/4)$

| $m_1$ | $r_1$ | $Q_1$ | $Q_2$ | $Q_3$ |
|-------|-------|-------|-------|-------|
| 40    | 12    | 9, 13 | 1/4   | 0     |
| 24    | 23    | 20, 26| 1/4   | 0     |
| 36    | 36    | 34, 37| 1/4   | 0     |
| 200   | 60    | 49, 61| 1/4   | 0     |
| 120   | 115   | 109, 120| 1/4 | 0     |
| 180   | 179   | 176, 182| 1/4 | 0     |

Based on the F-distribution

$\left( \overline{\delta}_{11}, \overline{\delta}_{12}, \overline{\delta}_{13} \right) = (1/4, 0, -1/4)$

| $m_1$ | $r_1$ | $Q_1$ | $Q_2$ | $Q_3$ |
|-------|-------|-------|-------|-------|
| 40    | 12    | 6, 10 | 1/4   | 0     |
| 24    | 21    | 19, 23| 1/4   | 0     |
| 36    | 35    | 33, 37| 1/4   | 0     |
| 200   | 60    | 42, 48| 1/4   | 0     |
| 120   | 103   | 98, 109| 1/4 | 0     |
| 180   | 176   | 173, 180| 1/4 | 0     |

$m_1$ is the sample size and $r_1$ is the $r_1$-th quartile of the empirical true rejections $r_1$ from $N = 1000$ simulations.
Suppose that we want to design a new fibroid study using the data analyzed above as pilot data. In the sample size calculation, we set $m = 30,000$. We assume that the $m_1 = 50$ genes which were selected as the top 50 genes in terms of parametric p-value are differentially expressed in the three sites ($K = 3$). From the pilot data, we estimate the standardized treatment effect $\hat{\delta}_{ik}$. For illustration, the effect sizes of these $m_1 = 50$ genes are taken to be $\hat{\delta}_{ik}$. We need $n = 15$ patients (blocks) to discover 90% of the prognostic genes, i.e. $r_1 = 0.9 \times 50 = 45$, while controlling the FDR at $q^* = 5\%$ level. In a simulation study, we generated $N = 1000$ microarray data sets of size $n = 15$ under this design setting. With $q^* = 0.05$, we observed the quartiles $Q_1, Q_3 = 46(45, 47)$ from the empirical distribution of the observed true rejections.
Expression patterns of six genes that are significantly differentially expressed in three sites.

220273_at, p-value= 0.0000

210255_at, p-value= 0.0000

229160_at, p-value= 0.0000

1553194_at, p-value= 0.0000

204620_s_at, p-value= 0.0000

217287_s_at, p-value= 0.0000

**Figure 3**
Expression patterns of six genes that are significantly differentially expressed in three sites.
Table 3: The result of uterine fibroid tissue and adjacent myometrium microarray experiment

| probe_set_id | Gene_Descriptor | parametric | non-parametric |
|--------------|----------------|------------|----------------|
| C-020273     | interleukin 17B | 0.0000     | 0.0000         | 0.0008 0.0131 |
| 131479        | RADS5-like 1 (S. cerevisiae) | 0.0000 | 0.0000 0.0015 0.0144 |
| 210255        | prostate androgen-regulated transcript 1 | 0.0000 | 0.0000 0.0077 0.0219 |
| 229160        | melanoma associated antigen (mutated) 1-like 1 | 0.0000 | 0.0000 0.0008 0.0131 |
| 210817        | calcium binding and coiled-coil domain 2 | 0.0000 | 0.0000 0.0015 0.0144 |
| 1553194       | neuronal growth regulator 1 | 0.0000 | 0.0000 0.0008 0.0131 |
| 202965        | calpain 6 | 0.0000 | 0.0000 0.0108 0.0239 |
| 204620        | chondroitin sulfate proteoglycan 2 (versican) | 0.0000 | 0.0000 0.0054 0.0196 |
| 217287        | transient receptor potential cation channel, subfamily C, member 6 | 0.0000 | 0.0000 0.0008 0.0131 |
| 227875        | kelch-like 13 (Drosophila) | 0.0000 | 0.0000 0.0023 0.0156 |
| 205286        | transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma) | 0.0000 | 0.0000 0.0046 0.0189 |
| 242737        | RADS5-like 1 (S. cerevisiae) | 0.0000 | 0.0000 0.0062 0.0206 |
| 209965        | RADS5-like 3 (S. cerevisiae) | 0.0000 | 0.0000 0.0008 0.0131 |
| 202007        | nidogen 1 | 0.0000 | 0.0000 0.0015 0.0144 |
| 221731        | chondroitin sulfate proteoglycan 2 (versican) | 0.0000 | 0.0000 0.0077 0.0219 |
| 224813        | RADS5-like 1 (S. cerevisiae) | 0.0000 | 0.0000 0.0015 0.0144 |
| 201310        | chromosome 5 open reading frame 13 | 0.0000 | 0.0000 0.0008 0.0131 |
| 210258        | regulator of G-protein signalling 13 | 0.0000 | 0.0000 0.0008 0.0131 |
| 202589        | thymidylate synthetase | 0.0000 | 0.0000 0.0054 0.0196 |
| 228766        | gb-AW299226 | 0.0000 | 0.0000 0.0054 0.0196 |
| 218380        | NLR family, pyrin domain containing 1 | 0.0000 | 0.0000 0.0008 0.0131 |
| 201417        | SRY (sex determining region Y)-box 4 | 0.0000 | 0.0000 0.0015 0.0144 |
| 219772        | Prostate androgen-regulated transcript 1 | 0.0000 | 0.0000 0.0093 0.0231 |
| 212942        | KIAA1199 | 0.0000 | 0.0000 0.0046 0.0189 |
| 202966        | calpain 6 | 0.0000 | 0.0000 0.0108 0.0239 |
| 205943        | tryptophan 2,3-dioxygenase | 0.0000 | 0.0000 0.0015 0.0144 |
| 213668        | SRY (sex determining region Y)-box 4 | 0.0000 | 0.0000 0.0015 0.0144 |
| 219454        | EGF-like-domain, multiple 6 | 0.0000 | 0.0000 0.0008 0.0131 |
| 235503        | ankyrin repeat and SOCS box-containing 5 | 0.0000 | 0.0000 0.0069 0.0212 |
| 222834        | guanine nucleotide binding protein (G protein), gamma 12 | 0.0000 | 0.0000 0.0008 0.0131 |
| 210198        | proteolipid protein 1 (Pelizaeus-Merzbacher disease, spastic paraplegia 2, uncomplicated) | 0.0000 | 0.0000 0.0015 0.0144 |
| 220565        | chemokine (C-C motif) receptor 10 | 0.0000 | 0.0000 0.0008 0.0131 |
| 237671        | RADS5-like 1 (S. cerevisiae) | 0.0000 | 0.0000 0.0093 0.0231 |
| 201220        | C-terminal binding protein 2 | 0.0000 | 0.0000 0.0039 0.0180 |
| 217771        | golgi phosphoprotein 2 | 0.0000 | 0.0000 0.0015 0.0144 |
| 224002        | FK506 binding protein 7 | 0.0000 | 0.0000 0.0008 0.0131 |
| 213770        | glutathione peroxidase 7 | 0.0000 | 0.0000 0.0008 0.0131 |
| 211980        | collagen, type IV, alpha 1 | 0.0000 | 0.0000 0.0031 0.0167 |
| 211981        | collagen, type IV, alpha 1 | 0.0000 | 0.0000 0.0031 0.0167 |
| 212282        | transmembrane protein 97 | 0.0000 | 0.0000 0.0008 0.0131 |
| 2013090       | chromosome 5 open reading frame 13 | 0.0000 | 0.0000 0.0015 0.0144 |
| 211917        | prolactin receptor//prolactin receptor | 0.0000 | 0.0000 0.0008 0.0131 |
| 212281        | transmembrane protein 97 | 0.0000 | 0.0000 0.0008 0.0131 |
| 231930        | ELMO/CED-12 domain containing 1 | 0.0000 | 0.0000 0.0123 0.0248 |
| 205347        | thymosin-like 8 | 0.0000 | 0.0000 0.0015 0.0144 |
| 223571        | C1q and tumor necrosis factor related protein 6 | 0.0000 | 0.0000 0.0015 0.0144 |
| 204619        | chondroitin sulfate proteoglycan 2 (versican) | 0.0000 | 0.0000 0.0046 0.0189 |
| 231741        | endothelial differentiation, sphingolipid G-protein-coupled receptor, 3 | 0.0000 | 0.0000 0.0054 0.0196 |
Conclusion
We have considered studies where microarray data for K treatment groups are collected from the same subjects (blocks). We discover the genes differentially expressed among K groups using non-parametric F-statistics for blocked one-way ANOVA while controlling the FDR. We employ a permutation method to generate the null distribution of the F-statistics without a normal distribution assumption for the gene expression data. The permutation-based multiple testing procedure can be easily modified for controlling the familywise error rate, see e.g. Westfall and Young [28] and Jung et al. [29].

We propose a simple sample size calculation method to estimate the required number of subjects (blocks) given the total number of genes m, number of differentially expressed genes m_i and their standardized effect sizes (\( \delta_{hi} \), 1 ≤ i ≤ m, 1 ≤ k ≤ K) and the number of true rejections r_i at a specified FDR level q*. Through simulations and analysis of a real data set, we found that the permutation-based analysis method controls the FDR accurately and the sample size formula performs accurately. While the number of hypotheses m is small. However, it is not appropriate for designing a microarray study with a large number of dependent test statistics.

A sample size calculation program in R is available from http://www.duke.edu/~is29/BlockANOVA/.

Appendix
We want to prove that \( F_i \) converges to 
\[
(K - 1)^{-1} \sigma_i^2 X_{K-1}^2 \left( n \sum_{k=1}^{K} \delta^2_{ik} \right) \]
Distribution regardless of the normal distribution assumption on \( i \) and \( j \). We only assume that \( E(X_{ijk}^4) < \infty \). The following is one of key lemmas used to derive the distribution of the F-statistics in the standard ANOVA theory, see e.g. Section 3b.4 of Rao [33].

Lemma: Suppose that, for \( k = 1, ..., K \), \( z_k \) are independent \( N(\mu_k, 1) \) random variables and \( A \) is an idempotent \( K \times K \) matrix with rank \( v \). Let \( z = (z_1, ..., z_K)^T \) and \( \mu = (\mu_1, ..., \mu_K)^T \). Then, 
\[
z^T A z \sim \chi^2_v(\mu^T A \mu)
\]
We have
\[
\frac{1}{(K-1)(n-1)} \sum_{i=1}^{K} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{ij} - \overline{Y}_{ik} + \overline{Y}_{..})^2 = \frac{1}{(K-1)(n-1)} \sum_{i=1}^{K} \sum_{k=1}^{K} (\delta_{ik} - \tau_i - \tau_k + \tau_{..})^2
\]
where \( \tau_i = K^{-1} \sum_{k=1}^{K} z_{ijk} \), \( \tau_k = n^{-1} \sum_{j=1}^{n} z_{ijk} \) and 
\( \tau_{..} = (Kn)^{-1} \sum_{j=1}^{K} \sum_{k=1}^{K} z_{ijk} \). By the strong law of large numbers, we have 
\[
n^{-1} \sum_{j=1}^{n} z_{ijk} \rightarrow \sigma_i^2, \quad n^{-1} \sum_{j=1}^{n} z_{ijk}^2 \rightarrow \sigma_i^2 / K, \quad \tau_{ik} \rightarrow 0 \quad \text{and} \quad \tau_{..} \rightarrow 0 \quad \text{almost surely (a.s.)}.
\]
Hence,
\[
\frac{1}{(K-1)(n-1)} \sum_{j=1}^{K} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{ij} - \overline{Y}_{ik} + \overline{Y}_{..})^2 \rightarrow \sigma_i^2 \quad \text{a.s.}
\]

Let 
\[
z_k = \sigma_i^{-1} \sqrt{n} (\delta_{ik} + \tau_{ik})
\]
and \( \Sigma = K^{-1} \sum_{k=1}^{K} z_k = (K\sigma_i)^{-1} \sum_{k=1}^{K} \tau_{ik} \). Then, \( z_1, ..., z_K \) are independent and, by the central limit theorem, \( z_k \) is approximately \( N(\delta_{ik}, 1) \). Let \( I \) be the \( K \times K \) identity matrix, \( 1 = (1, ..., 1)^T \) the \( K \times 1 \) vector with components 1, \( z = (z_1, ..., z_K)^T A = I \cdot K^{-1} 11^T \). Note that \( A \) is an idempotent matrix with rank \( K - 1 \) and \( \mu^T A \mu = n \sum_{k=1}^{K} \delta_{ik}^2 \), where 
\[
\mu = \sqrt{n} (\delta_{11}, ..., \delta_{ik})^T
\]
is approximately distributed as \( \chi^2_{K-1}(n \sum_{k=1}^{K} \tau_{ik}^2) \) by the lemma. Since 
\[
\sum_{k=1}^{K} (\overline{Y}_{ik} - \overline{Y}_{i..})^2 = \sigma_i^2 \sum_{k=1}^{K} (z_k - \Sigma)^2
\]
\[
\sum_{k=1}^{K} (\overline{Y}_{ik} - \overline{Y}_{i..})^2 \quad \text{is approximately distributed as} \quad \chi^2_{K-1}(n \sum_{k=1}^{K} \tau_{ik}^2)
\]
By combining this result with (A.1) using the Slutsky's theorem, we complete the proof.

Authors' contributions
SJ proposed the research project and wrote the manuscript. IS performed statistical analysis. SLG supported the
research and participated in the writing of the manuscript. PCL was responsible for the study design, conduct and oversight of the experiments and interpretation of results. She contributed to the preparation of the manuscript. LF was responsible for preparing the tissue samples for microarray analysis and interpretation of results and in manuscript preparation. The authors are solely responsible for the content of this study. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Holly Dressler and the staff of the Duke Microarray Facility for their assistance in the conduct of the genome-wide expression profiling. Funding in part was provided by NIH 1UL1-RR-024128 and Department of Obstetrics and Gynecology Duke University.

References

1. Catherino WH, Leppert PC, Segars JH: The promise and perils of microarray analysis. Am J Obstet Gynecol 2006, 195:389-393.
2. Benjamin Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 1995, 57:289-300.
3. Westfall PH, Wolfinger RD: Multiple tests with discrete distributions. American Statistician 1997, 51:3-8.
4. Lee MLT, Whitmore GA: Power and sample size for DNA microarray studies. Stat Med 2002, 21:3543-3570.
5. Muller P, Parmigiani G, Robert C, Rousseau J: Optimal sample size for multiple testing: the case of gene expression microarrays. J Am Stat Assoc 2004, 99:990-1001.
6. Jung SH: Sample size for FDR-control in microarray data analysis. Bioinformatics 2005, 21:3097-3103.
7. Pounds S, Cheng C: Sample size determination for the false discovery rate. Bioinformatics 2005, 21:4263-4271.
8. Liu P, Hwang JTG: Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. Bioinformatics 2007, 23:739-746.
9. Shao Y, Tseng CH: Sample size calculation with dependence adjustment for FDR-control in microarray studies. Stat Med 2007, 26:4219-4237.
10. Feng L, Walls M, Sohn I, Bebera M, Jung SH, Leppert PC: Novel approach to the analysis of genome-wide expression profiling. Reproductive Sciences 2008, 15:298A.
11. Ge Y, Dudoit S, Speed TP: Resampling-based multiple testing for microarray data hypothesis. Test 2003, 12:1-44.
12. Storey JD: A direct approach to false discovery rates. J Roy Stat Soc Ser B 2002, 64:479-498.
13. Jung SH, Jung W: How accurately can we control the FDR in analyzing microarray data? Bioinformatics 2006, 22:1730-1736.
14. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics 2003, 4:249-264.
15. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 2003, 19:185-193.
16. Hartigan JA: Clustering Algorithms. Wiley: New York; 1975.
17. Catherino WH, Prupas C, Tsibris JC, Leppert PC, Payson M, Nieman LK, Segars JH: Strategy for elucidating differentially expressed genes in leiomyomata identified by microarray technology. Fert Steril 2003, 80:282-290.
18. Skubitz K, Skubitz AP: Differential gene expression in uterine leiomyoma. Journal of Laboratory and Clinical Medicine 2003, 141:278-308.
19. Leppert PC, Catherino WH, Segars JH: A new hypothesis about the origin of uterine fibroids based on gene expression profiling with microarrays. Am J Obstet Gynecol 2006, 195:415-420.
20. Rogers R, Norian J, Malik M, Christman G, Abu-Asab M, Chen F, Korecki C, Ixartidis J, Catherino WH, Dhillon N, Leppert P, Segars JH: Mechanical homeostasis is altered in uterine leiomyoma. Am J Obstet Gynecol 2008, 198:474.e1-11.
21. Malik M, Webb J, Catherino WH: Retinoic acid treatment of human leiomyoma cells transformed the cell phenotype to one strongly resembling myometrial cells. Clin Endocrinol (Oxf) 2008 in press.
22. Yagi Y, Andoh A, Inatomi O, Tsujikawa T, Fujiyama Y: Inflammatory responses induced by interleukin-17 family members in human colonic subepithelial myofibroblasts. J Gastroenterol 2007, 42:746-753.
23. Cortez DM, Feldman MD, Mumordi S, Valente AJ, Steffensen B, Vincenti M, Barnes J, Chandrasekar B: IL-17 stimulates MMP-1 expression in primary human cardiac fibroblasts via p38 MAPK- and ERK 1/2-dependent C/EBP-α, NF-κB, and AP-1 activation. Am J Physiol Heart Circ Physiol 2007, 293:3356-3365.
24. Schoenmakers EFPM, Huysmans C, Van den Hooft JP: Allelic knockout of novel splice variants of human recombination repair gene RAD51B in t(12;14) uterine leiomyomas. Cancer Res 1999, 59:19-23.
25. Reeves R: Molecular biology of HMGA proteins: hubs of nuclear function. Gene 2001, 277:63-81.
26. Gatas GJ, Quade BJ, Nowak RA, Morton CC: HMGC expression in human adult and fetal tissues and in uterine leiomyomata. Genes Chromosomes and cancer 1999, 25:316-322.
27. Peng Y, Laser J, Shi G, Mitulak J, Melamed J, Lee P, Wei JY: Antiproiferative effects by Let-7 repression of high-mobility group A2 in uterine leiomyoma. Mol Cancer Res 2008, 6:663-673.
28. Westfall PH, Young SS: Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. Wiley: New York; 1993.
29. Jung SH, Bang H, Young S: Sample size calculation for multiple testing in microarray data analysis. Bioinformatics 2005, 6:157-169.
30. Hu J, Zou F, Wright FA: Practical FDR-based sample size calculations in microarray experiments. Bioinformatics 2005, 21:3264-3272.
31. Jarstad TS, Midefart H, Bones AM: A mixture model approach to sample size estimation in two-sample comparative microarray experiments. BMC Bioinformatics 2008, 9:117.
32. Glueck DH, Mandel J, Kaminour-Fard A, Hunter L, Muller K: Exact calculations of average power for the Benjamini-Hochberg procedure. International Journal of Biostatistics 2008, 4:Article 11.
33. Rao CR: Linear Statistical Inference and Its Applications. Wiley: New York; 1965.
