Published in final edited form as:
J Biopharm Stat. 2012 January ; 22(1): 30–42. doi:10.1080/10543406.2010.500066.

Power and Sample Size Calculation for Microarray Studies

SIN-HO JUNG and
Department of Biostatistics and Bioinformatics, Duke University 2424 Erwin Road, 11070 Hock Plaza, Suite 1102, DUMC Box 2721, Durham, NC 27710, U.S.A, 919-668-8658

S. STANLEY YOUNG
National Institute of Statistical Sciences, Research Triangle Park, NC 27709, U.S.A

SIN-HO JUNG: sinho.jung@duke.edu

SUMMARY

Microarray is a technology to screen a large number of genes to discover those differentially expressed between clinical subtypes or different conditions of human diseases. Gene discovery using microarray data requires adjustment for the large-scale multiplicity of candidate genes. The family-wise error rate (FWER) has been widely chosen as a global type I error rate adjusting for the multiplicity. Typically in microarray data, the expression levels of different genes are correlated because of coexpressing genes and the common experimental conditions shared by the genes on each array. To accurately control the FWER, the statistical testing procedure should appropriately reflect the dependency among the genes. Permutation methods have been used for accurate control of the FWER in analyzing microarray data. It is important to calculate the required sample size at the design stage of a new (confirmatory) microarray study. Because of the high dimensionality and complexity of the correlation structure in microarray data, however, there have been no sample size calculation methods accurately reflecting the true correlation structure of real microarray data. We propose sample size and power calculation methods that are useful when pilot data are available to design a confirmatory experiment. If no pilot data are available, we recommend a two-stage sample size recalculation based on our proposed method using the first stage data as pilot data. The calculated sample sizes are shown to accurately maintain the power through simulations. A real data example is taken to illustrate the proposed method.

Keywords

Co-regulation; Family-wise error rate; Permutation; Sample size recalculation; Single step procedure

1 INTRODUCTION

Because of the huge dimensionality of microarray data, multiplicity of the statistical testing is a critical issue in designing and analyzing such type of studies. The family-wise error rate (FWER) is the probability that one or more false rejections are committed. Despite its well-known conservatism, Bonferroni test has been one of most popular methods in analyzing microarray data for controlling the FWER. Although Holm (1979) and Hochberg (1998) improve upon such conservatism by devising multi-step testing procedures, they do not take into account the dependency of the test statistics and consequently the resulting improvement is often minor. Westfall and Young (1989, 1993) proposed a state-of-the-art, step-down manner using a simulation or resampling method by which dependency among test statistics is effectively incorporated. Westfall and Wolfinger (1997) 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, e.g. Dudoit, Yang and Callow (2000), Dudoit, Shaffer and Boldrick (2003), Ge, Dudoit and Speed (2003).

Suppose that there are two groups to be compared using gene expression data. We assume that group k(= 1, 2) has n_k subjects ($n = n_1 + n_2$) and each subject contributes data of one microarray. For patient i in group k ($1 \le i \le n_k$; k = 1, 2), we observe expression data from m genes, $(x_{ki1}, \ldots, x_{kim})$. We assume that, within group k, $\{(x_{ki1}, \ldots, x_{kim}), i = 1, \ldots, n_k\}$ are IID random vectors from an unknown distribution with means $\mu_{kj} = \mathrm{E}(x_{kij})$, variances $\sigma_j^2 = \mathrm{var}(x_{kij})$ and correlation coefficients $\Sigma = (\rho_{j,j'})_{1 \le j,j' \le m}$. The large sample theory requires $\max_{1 \le j \le m} \sigma_j^2 \le M(<\infty)$.

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. $\bar{H}_j : \mu_{1j} \ne \mu_{2j}$ for each gene. We consider rejecting H_j (or discover gene j) if the absolute value of two-sample t-test statistic

$$\bar{T}_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{s_j \sqrt{n_1^{-1} + n_2^{-1}}}$$

is large, where $\bar{x}_{kj} = n_k^{-1} \sum_{i=1}^{n_k} x_{kij}$ is the sample mean of group k and $s_j^2 = (n_1 + n_2 - 2)^{-1} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (x_{kij} - \bar{x}_{kj})^2$ the pooled variance for gene j.

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 $\bar{T}_j$ so that the probability of rejecting one or more H_j's is controlled below a specified level under H_0. A single-step procedure uses a common critical value c to reject H_j in favor of $\bar{H}_j$ when $|\bar{T}_j| > c$. In this case, the FWER fixed at α is defined as

$$\alpha = \mathrm{P}(\max_{1 \le j \le m} |\bar{T}_j| > c | H_0). \quad (1)$$

In order to derive the critical value $c = c_\alpha$, we approximate the null distribution of $\max_{1 \le j \le m} |\bar{T}_j|$ using the permutation method proposed by Westfall and Young (1993). Jung, Bang and Young (2005) claim that the single-step procedure has exactly the same probability of at least one true rejections, global power in this paper,

$$1 - \beta_0 = \mathrm{P}(\max_{1 \le j \le m} |\bar{T}_j| > c | H_a)$$

as the Westfall and Young's (1993) step-down procedure.

Large confirmatory microarray studies are becoming more common recently. Sample size calculation is a critical step in designing such studies. There have been several publications on sample size estimation for FWER-based multiple testing procedures without examining the accuracy of their estimates. 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 conservative Bonferroni adjustment without any attempt to incorporate

underlying correlation structure, e.g. Witt, Elston and Cardon (2000), Wolfinger *et al*. (2001), Black and Doerge (2002), Lee and Whitmore (2002), Pan, Lin and Le (2002), Cui and Churchill (2003). Showing that an ostensibly similar but incorrect choice of sample size ascertainment could cause considerable underestimation of required sample size, Jung, Bang and Young (2005) propose a sample size calculation for the Westfall and Young's permutation method under a hypothetical correlation structure called block compound symmetry. Tsai et al. (2005), Tibshirani (2006) and Shao and Tseng (2007) propose sample size methods for different multiple testing methods assuming specific correlation structures. If the specified correlation structure is correct, then their sample size will be accurate. Because of high dimensionality and complicated dependency of the expression data among genes genes, it is almost impossible to model the true correlation structure accurately. As a result, there have been no sample size methods reflecting the true correlation structure of gene expression data.

In this paper, we propose sample size and power calculation methods for the FWER-based multiple testing that reflect the correlations among the genes under consideration. If pilot data are available, we approximate the true correlation structure from the pilot data. Otherwise, we recommend a two-stage sample size recalculation method using the first stage data as pilot data. Through simulations, we show that the calculated sample sizes accurately maintain the specified power. We also propose a new type of statistical power for multiple testing procedures. A real data example is used to illustrate the proposed methods.

2 POWER AND SAMPLE SIZE CALCULATION

We want to calculate the sample size for a new study whose data will be analyzed by controlling the FWER. In this paper, we use a large sample approximation based on large n assuming that $n_k/n = a_k \in (0, 1)$. As Jung, Bang and Young (2005) show, the power of FWER-based multiple testing methods depends on the standardized effect sizes $\delta_j = (\mu_{1j} - \mu_{2j})/\sigma_j$ under H_a and the correlation coefficients of expression data among genes. The correlation coefficients usually are nuisance parameters in the multiple testing procedures. We may specify δ_j for some candidate prognostic genes based on some prior biological knowledge, but it will be difficult to specify the many correlation coefficients. In order to tackle this problem, we assume that historical or 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 the assumption that the null distribution of the test statistics to be calculated from the future study can be approximated by that from the pilot data set, $\{(x_{ki1}, \ldots, x_{kim}), i = 1,\ldots, n_k, k = 1, 2\}$. Let $\bar{X}_{kj}$ and s_j^2 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 ($a_1 + a_2 = 1$). Naturally, we assume $N_k > n_k$. 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

$$T_j = \frac{\bar{X}_{1j} - \bar{X}_{2j}}{S_j \sqrt{N_1^{-1} + N_2^{-1}}} = \delta_j \sqrt{N a_1 a_2} + Z_j + o_p(1),$$

where

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

and $o_p(1)$ converges to 0 in probability as $N \to \infty$. Here, $(Z_1, \ldots Z_m)$ is the limit of test statistics $(T_1, \ldots, T_m)$ under H_0 that will be calculated from the data of the future study. It is easy to show that $(Z_1, \ldots, Z_m)$ 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 = α, the critical value c_α satisfies

$$\alpha=\mathrm{P}(\max_{1\le j\le m}|Z_j|>c_\alpha) \tag{2}$$

from (1).

Suppose that there are m_1 (prognostic) genes with non-zero effect sizes and $m_0(= m-m_1)$ (non-prognostic) genes with 0 effect sizes. Let $\mathcal{M}_1$ 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 γ true rejections with probability $1-\beta_\gamma$ by controlling the FWER at α. Then, we need to solve

$$1-\beta_\gamma=\mathrm{P}\{\sum_{j\in\mathcal{M}_1}1(|\delta_j\sqrt{Na_1a_2}+Z_j|>c_\alpha)\ge\gamma\} \tag{3}$$

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

$$1-\beta_0=\mathrm{P}(\max_{1\le j\le m}|\delta_j\sqrt{Na_1a_2}+Z_j|>c_\alpha). \tag{4}$$

Note that $1-\beta_0$ denotes the probability of any rejections while $1 \bullet \beta_1$ denotes the probability of any true rejections. Wang and Chen (2004), Jung (2005), Tsai et al. (2005), and Shao and Tseng (2007) consider the probabilities of γ true rejections for different multiple methods.

In order to solve these equations, we need to approximate the probabilities (2)–(4) involving the high dimensional random vector $(Z_1, \ldots, Z_m)$. Lin (2005) proposes an efficient resampling method to approximate the null distribution of test statistics for multiple testing. We modify his method to approximate the alternative as well as null distributions of the test statistics that will be obtained from the future study using pilot data. If a pilot data set $\{(x_{ki1}, \ldots, x_{kim}), i = 1, \ldots, n_k, k = 1, 2\}$ is available and its size is reasonably large for a reliable estimation of the true correlation coefficients, then the distribution of $(Z_1, \ldots, Z_m)$ can be approximated using the historical or pilot data. That is, we approximate the distribution of $(Z_1, \ldots, Z_m)$ from the simulated data conditioning on pilot data, $(\tilde{Z}_1, \ldots, \tilde{Z}_m)$, where

$$\tilde{Z}_j=\frac{\tilde{x}_{1j}-\tilde{x}_{2j}}{\sqrt{v_j}}$$
$$\tilde{x}_{kj}=n_k^{-1}\sum_{i=1}^{n_k}(x_{kij}-\overline{x}_j)\varepsilon_{ki},$$

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

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.

Algorithm for Sample Size Calculation

I. Input variables:

- **i.** Pilot data $\{(x_{ki1}, \ldots, x_{kim}), i = 1, \ldots, n_k, k = 1, 2\}$.
- **ii.** Number of prognostic genes m_1, their identifiers $\mathcal{M}_1 = \{j_1, \ldots, j_{m1}\}$, and effect sizes $(\delta_{j1}, \ldots, \delta_{jm1})$. For the remaining $m_0 (= m - m_1)$ genes, effect sizes are 0.
- **iii.** FWER = α.
- **iv.** Number of minimum true rejections γ ($\le m_1$), and the probability of γ true rejections $1 - \beta_\gamma$.
- **v.** Proportion of subjects in each group, a_1 and a_2 ($a_1 + a_2 = 1$).

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

III. Given FWER = α, calculate c_α by the upper 100α percentile of $\tilde{u}_1, \ldots, \tilde{u}_B$, where $\tilde{u}_b = \max_{1 \le j \le m} |\tilde{z}_{bj}|$.

IV. Let

$$h(N) = B^{-1}\sum_{b=1}^{B} I\{\sum_{j \in \mathcal{M}_1} I(|\delta_j \sqrt{N a_1 a_2} + \tilde{z}_{bj}| > c_\alpha) \ge \gamma\}.$$

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

Note that above method can be used for sample size recalculation in the midst of a study. At the design stage of a study, we calculate an approximate sample size $\hat{N}$ based on pilot data using this algorithm or projected correlation coefficients as in Jung, Bang and Young (2005). Often, the first stage sample size n is chosen by half of $\hat{N}$. We collect the first stage data $\{(x_{ki1}, \ldots, x_{kim}), 1 \le i \le n_k, k = 1, 2\}$, 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, $\{(x_{ki1}, \ldots, x_{kim}), n_1 + 1 \le i \le N_k, k = 1, 2\}$, and conduct the multiple testing procedure using the cumulative data $\{(x_{ki1}, \ldots, x_{kim}), 1 \le i \le N_k, k = 1, 2\}$.

3 SIMULATIONS

Extensive simulations were conducted to evaluate the accuracy of our sample size calculation method that is derived using large sample approximations. Suppose that there are $m = 1000$ candidate genes whose expression levels have a multivariate Gaussian distribution consisting of 100 independent blocks. Within each block, 10 genes have a compound symmetry correlation structure with a common correlation coefficient ρ(= 0.3 or 0.6). The marginal distributions have means $\mu_{1j} = 0.5$ for $1 \le j \le 10$, $\mu_{1j} = 0$ for $11 \le j \le 1000$ and μ_{2j}

= 0 for $1 \le j \le 1000$, and unit marginal variances for both groups $\sigma_j = 1$ for $j = 1, \ldots, m$. Note that the standardized effect sizes are $\delta_j = 0.5$ for $\mathcal{M}_1 = \{1 \le j \le 10\}$ and $\delta_j = 0$ for $\mathcal{M}_0 = \{11 \le j \le 1000\}$. We consider the sample size calculation based on a pilot data set as discussed above. At first, we generate pilot data of size n, $\{(x_{ki1}, \ldots, x_{kim}), 1 \le i \le n_k, k = 1, 2\}$, with equal allocation ($a_1 = a_2 = 1/2$). Using each pilot data set, we estimate the total sample size N for a power of $1 - \beta_\gamma = 0.8$ for $\gamma = 0, 1, 3, 5$ or 7 by controlling the FWER at 5%. For each pilot data set, $B = 5000$ sets of $(\varepsilon_{ki}, 1 \le i \le n_k, k = 1, 2)$ are generated for estimating N. We generate 1000 pilot data sets of size n, so that we will have 1000 estimated sample sizes N, under each simulation setting.

Figure 1 displays the empirical distribution of the estimated sample sizes N from 1000 pilot data sets of size $n = 20$, 50 or 100 for $\rho = 0.3$ and $\gamma = 5$. We observe that the variance of N decreases as n increases. This outcome is natural since, with a larger n, we have a better estimation of the true correlation structure of the gene expression data, and consequently we obtain a more accurate sample size estimate. Furthermore, the center of the distribution of estimated N's increases in n.

Figure 2 displays the quartiles of the estimated N's for $10 \le n \le 150$ and $\gamma = 1, 3, 5, 7$, and 9. As expected, N increases in γ. The required N under $\rho = 0.6$ is larger than that under $\rho = 0.3$ for $\gamma \le 5$, but the comparison switches to the opposite direction for $\gamma \ge 7$. The quartiles of N tend to increase in n for small n values, but become stable for large n, say for $n > 60$. The interquartile range of N decreases in n as observed in Figure 1 also, and is quite narrow over a wide range n values implying that the proposed sample size formula provides a stable sample size estimate with pilot data of a moderate size. In these simulations, a relatively small m(= 1, 000) was chosen to loosen the computation burden.

Now, we want to check if an estimated N really guarantees the intended power $1 - \beta_\gamma$. We consider the case with $n = 100$ and $\rho = 0.3$ among the previous simulation settings. Given γ, we choose N by one of the three quartiles from Figure 2, generate 1000 sets of microarray data with size N under the chosen simulation setting, and apply the permutation-based multiple testing procedure with $\alpha = 0.05$ to each simulation data set. For each simulation data, we estimate the critical value c_α from $B = 5000$ permutations. Empirical power, $1 - \hat{\beta}_\gamma$, is obtained as the proportion of simulation data sets with at least γ true rejections. Table 1 summarizes the simulation results. We observe that the estimated N is very stable (the interquartile range is at most 7 under the simulation settings). The estimated sample sizes within the interquartile range closely maintain the intended power, $1 - \beta_\gamma = 0.8$, except for $\gamma = 1$. Due to simulation error and the narrow interquartile range, a small increase in N between Q_2 and Q_3 does not necessarily lead to an increased empirical power.

4 EXAMPLE

Huang *et al.*(2003) 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 12625 probe sets, called genes in this section. Expression values were calculated using the Robust Multichip Average (RMA) method (Irizarry *et al.*, 2003). RMA estimates are based upon a robust average of background corrected PM intensities. Normalization was done using quantile normalization (Bolstad *et al.*, 2003). We removed all "AFFX" genes and filtered out the genes for which there were less than 8 present calls among the 37 present/marginal/absent calls. The filtering yielded $m = 6599$ genes which were then used in the subsequent analyses.

The estimated standardized effect sizes of top 20 genes are distributed between 1.177 and 1.869 in absolute value. 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.*(2003) as pilot data. We specify the set of prognostic genes $\mathcal{M}_1$ by the top 20 genes. 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.

Figure 3 displays the estimated N for γ (∞ [0, 20]) true rejections by $\alpha = 0.05$ 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 $m = 6599$ genes that 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 3, the required sample size monotonically increases from $N = 37$ for $\gamma = 1$ to $N = 192$ for $\gamma = 20$. Note that $n = 37$ of the Huang *et al.*(2003) data is about the right size for at least one true rejection, i.e. $\gamma = 1$. The expected number of false rejections, defined as $\Sigma_{j\in \mathcal{M}_0} P(|T_j| > c_\alpha)$ and estimated from the $B = 5000$ simulations, is only about 0.1. As a referee suggested, we also checked the empirical FWER under H_0, and found that it was 0.053, which is close to the nominal 5%, with $n = 37$.

It is difficult to make direct comparisons of our sample size with those by other methods since different methods are based on different kinds of type I error control and different types of power. Pan, Lin and Le (2002) calculate sample size of Bonferroni test by specifying the effect size of a prognostic gene and the marginal power. For a prognostic gene with effect size δ_0, the marginal power given N is expressed as

$$p_N = \overline{\Phi}(z_{\alpha/(2m)} - \delta_0 \sqrt{N a_1 a_2}),$$

where $\overline{\Phi}(z) = \int_z^\infty \varphi(z)dt$ and $\varphi(\cdot)$ is the probability density function of the standard normal distribution. In order to expand the formula by Pan, Lin and Le (2002), suppose that all m_1 prognostic genes have the same effect size δ_0. Then, ignoring the possible dependency among genes, we can calculate the probability of at least γ true rejections by

$$1 - \beta_\gamma = \sum_{y=\gamma}^{m_1} \binom{m_1}{y} p_N^y (1 - p_N)^{m_1 - y}. \quad (5)$$

Given $(m_1, \delta, \gamma, 1 - \beta_\gamma, a_1)$, we obtain the sample size by solving this equation with respect to N using a numerical method, such as the bisection method. Let's consider above breast cancer data example assuming that the top $m_1 = 20$ genes have equal an effect size $\delta_0 = 1$ and the remaining $m_0 = 6579$ genes have 0 effect sizes. For $\gamma = 15$ and $1 - \beta_\gamma = 0.9$, the required sample size for Bonferroni method is obtained as $N = 118$ from (5), while our formula reflecting the dependency observed from the Huang *et al.*(2003) data gives $N = 102$ under the same parameter setting.

5 DISCUSSIONS

Our sample size calculation method is based on large sample approximations. So, an accurate sample size estimation requires pilot data with a reasonable sample size. Even though a pilot data set may not be 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. From extensive simulations, we observe that a small pilot data set tends to give an underestimated sample size *N*. If *n* is smaller than 50% of the calculated *N*, we recommend to increase the final *N* by 5% to 10% based on our experience from the simulation studies.

When no pilot data are available, we propose a two-stage microarray study design for accurate sample size calculation reflecting the dependency among different genes and an approximation to the number of real effects. Two-stage design methods for gene discovery have been proposed by many researchers, e.g. Satagopan *et al.*(2002), Zehetmayer, Bauer, Posch (2005), Lin (2006), Moerkerke and Goetghebeur (2008). These methods screen a small number of promising genes based on the significance observed from the first stage data, and the final testing is conducted at the end of the second stage accounting for the two-stage testing procedure. In our two-stage design, however, a definitive testing is done only at the second stage, so that we do not need to adjust the type I error for the two-stage design at the second stage.

We generate IID random numbers, (ε_{ki}, $1 \le i \le n_k$, $k = 1, 2$), from $N(0, 1)$ distribution. In fact, they can be generated from any distribution with mean 0 and variance 1. However, the normal distribution was shown to provide the best small sample approximation (simulation results not reported). Note that the approximate normality of the test statistics results from the asymptotic theory for large number of arrays rather than the normality of gene expression data.

The proposed sample size calculation method is based on the single-step multiple testing procedure (SSP). If one wants a sample size for a given global power $1 - \beta_0$, then the required sample size for SSP will be identical to that for a step-down procedure (SDP), since the two types of procedures have the same global power. Theoretically, SDP has a slightly larger true rejection probability than SSP, especially with a small *m*. With a large *m* and a relatively small m_1 as in most microarray data, however, the true rejection probability is almost identical between the two types of multiple testing procedures, so that the proposed sample size method can be used for SDP also. Furthermore, the proposed sample size estimation method can be easily modified for SDP when the true rejection probability, $1 - \beta$ γ with $\gamma > 0$, is specified.

The proposed sample size estimation method is useful for any studies collecting high dimensional data such as SNP or proteomic studies. It can be easily modified for a FDR-based multiple testing procedure too.

Acknowledgments

This research was supported by NIH Grant 1 UL1 RR024128-01.

References

Black MA, Doerge RW. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics. 2002; 18:1609–1616. [PubMed: 12490445]

Bolstad BM, Irizarry RA, Astrand M, Speed TPA. Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics. 2003; 19:185–193. [PubMed: 12538238]

Cui, X.; Churchill, GA. Methods of Microarray Data Analysis II. Norwell, MA: Kluwer Academic Publishers; 2003. How many mice and how many arrays? Replication in mouse cDNA microarray experiments; p. 139-154.

Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statistical Science. 2003; 18:71–103.

Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica. 2000; 12:111–139.

Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test. 2003; 12:1–44.

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1998; 75:800–802.

Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statististics. 1979; 6:65–70.

Huang E, Cheng S, Dressman H, Pittman J, Tsou M, Horng C, Bild A, Iversen E, Liao M, Chen C. Gene expression predictors of breast cancer outcomes. Lancet. 2003; 361:1590–1596. [PubMed: 12747878]

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. [PubMed: 12925520]

Jung SH, Bang H, Young SS. Sample size calculation for multiple testing in microarray data analysis. Biostatistics. 2005; 6:157–169. [PubMed: 15618534]

Jung SH. Sample size for FDR-control in microarray data analysis. Bioinformatics. 2005; 21:3097–3104. [PubMed: 15845654]

Lee MLT, Whitmore GA. Power and sample size for DNA microarray studies. Statistics in Medicine. 2002; 22:3543–3570. [PubMed: 12436455]

Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics. 2005; 21:781–787. [PubMed: 15454414]

Lin DY. Evaluating statistical significance in two-stage genomewide association studies. American Journal of Human Genetics. 2006; 78:505–509. [PubMed: 16408254]

Moerkerke B, Goetghebeur E. Optimal screening for promising genes in 2-stage designs. Biostatistics. 2008; 9:700–714. [PubMed: 18349035]

Pan W, Lin J, Le CT. How many replicated of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biology. 2002; 3:1–10.

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. Two-stage designs for gene-disease association studies. Biometrics. 2002; 58:163–170. [PubMed: 11890312]

Shao Y, Tseng CH. Sample size calculation with dependence adjustment for FDR-control in microarray studies. Statistics in Medicine. 2007; 26:4219–4237. [PubMed: 17328091]

Tibshirani R. A simple method for assessing sample sizes in microarray experiments. BMC Bioinformatics. 2006; 7:1–6. [PubMed: 16393334]

Tsai CA, Wang SJ, Chen DT, Chen JJ. Sample size for gene expression microarray experiments. Bioinformatics. 2006; 21:1502–1508. [PubMed: 15564298]

Wang SJ, Chen JJ. Sample size for identifying differentially expressed genes in microarray experiments. Journal of Computational Biology. 2004; 11:714–726. [PubMed: 15579240]

Westfall PH, Wolfinger RD. Multiple tests with discrete distributions. American Statistician. 1997; 51:3–8.

Westfall PH, Young SS. P-value adjustments for multiple tests in multivariate binomial models. Journal of the American Statistical Association. 1989; 84:780–786.

Westfall, PH.; Young, SS. Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. New York: Wiley; 1993.

Witte JS, Elston RC, Cardon LR. On the relative sample size required for multiple comparisons. Statistics in Medicine. 2000; 19:369–372. [PubMed: 10649302]

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology. 2001; 8:625–637. [PubMed: 11747616]

Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. Bioinformatics. 2005; 21:3771–3777. [PubMed: 16091414]

Appendix

We want to show that the unconditional distribution of $(Z_1, \ldots, Z_m)$ can be approximated by the conditional distribution of $(\tilde{Z}_1, \ldots, \tilde{Z}_m)$ given the pilot data $\mathcal{D} = \{(x_{ki1}, \ldots, x_{kim}), i = 1, \ldots, n_k, k = 1, 2\}$ with a large n. In the following proof, we assume a nearby alternative hypothesis with respect to $n(< N)$, so that we use approximations, $S_j^2/\sigma_j^2 \approx 1$ and $v_j/\sigma_j^2(n_1^{-1}+n_2^{-1}) \approx 1$. Under $\max_{1\le j\le m}\sigma_j^2 \le M(<\infty)$, by the central limit theorem for large N, $(Z_1, \ldots, Z_m)$ is approximately normal with marginal means 0 and variances 1, and correlation coefficients $\rho_{jj'}$.

Now, we derive the conditional distribution of $(\tilde{Z}_1, \ldots, \tilde{Z}_m)$ given the pilot data for large n. Suppose that $(\varepsilon_{ki}, 1 \le i \le n_k, k = 1, 2)$ are IID $N(0, 1)$ random variables independent of $\mathcal{D}$. Since $\tilde{Z}_j$ are linear combinations of $(\varepsilon_{ki}, 1 \le i \le n_k, k = 1, 2)$, $(\tilde{Z}_1, \ldots, \tilde{Z}_m)$ has a normal distribution conditioning on $\mathcal{D}$. Hence, it suffices to show that, conditioning on $\mathcal{D}$, $(\tilde{Z}_1, \ldots, \tilde{Z}_m)$ has marginal means 0 and variances 1, and correlation coefficient matrix $\Sigma = (\rho_{jj'})_{1\le j,j' \le m}$. Since $\mathrm{E}(\tilde{x}_{kj}|\mathcal{D}) = n_k^{-1}\sum_{i=1}^{n_k}(x_{kij} - \bar{x}_j)\mathrm{E}(\varepsilon_{ki}) = 0$, we have

$$\mathrm{E}(\tilde{Z}_j|\mathcal{D}) = \frac{\mathrm{E}(\tilde{x}_{1j} - \tilde{x}_{2j}|\mathcal{D})}{\sqrt{v_j}} = 0.$$

Also,

$$\mathrm{var}(\tilde{Z}_j|\mathcal{D}) = \frac{\mathrm{var}(\tilde{x}_{1j}) + \mathrm{var}(\tilde{x}_{2j}|\mathcal{D})}{v_j} = 1$$

since $\mathrm{var}(\tilde{x}_{kj}|\mathcal{D}) = n_k^{-1}\sum_{i=1}^{n_k}(x_{kij} - \bar{x}_j)^2\mathrm{var}(\varepsilon_{ki}) = s_{kj}^2/n_k$ and $v_j = s_{1j}^2/n_1 + s_{2j}^2/n_2$. For $1 \le j \ne j' \le m$,

$$\mathrm{cov}(\tilde{Z}_j, \tilde{Z}_{j'}|\mathcal{D}) = \frac{\mathrm{cov}(\tilde{x}_{1j}, \tilde{x}_{1j'}|\mathcal{D}) + \mathrm{cov}(\tilde{x}_{2j}, \tilde{x}_{2j'}|\mathcal{D})}{\sqrt{v_j v_{j'}}}.$$

Here, $\sqrt{v_j v_{j'}} = \sigma_j\sigma_{j'}(n_1^{-1}+n_2^{-1}) + o_p(n^{-1})$ and

$$\begin{aligned}\mathrm{cov}(\tilde{x}_{kj}, \tilde{x}_{kj'}|\mathcal{D}) &= n_k^{-2}\sum_{i=1}^{n_k}(x_{kij} - \bar{x}_j)(x_{kij'} - \bar{x}_{j'})\mathrm{var}(\varepsilon_{ki}) \\ &= n_k^{-2}\sum_{i=1}^{n_k}(x_{kij} - \bar{x}_j)(x_{kij'} - \bar{x}_{j'}),\end{aligned}$$

which equals $n_k^{-1}\sigma_j\sigma_{j'}\rho_{jj'} + o_p(n^{-1})$. This completes the proof.

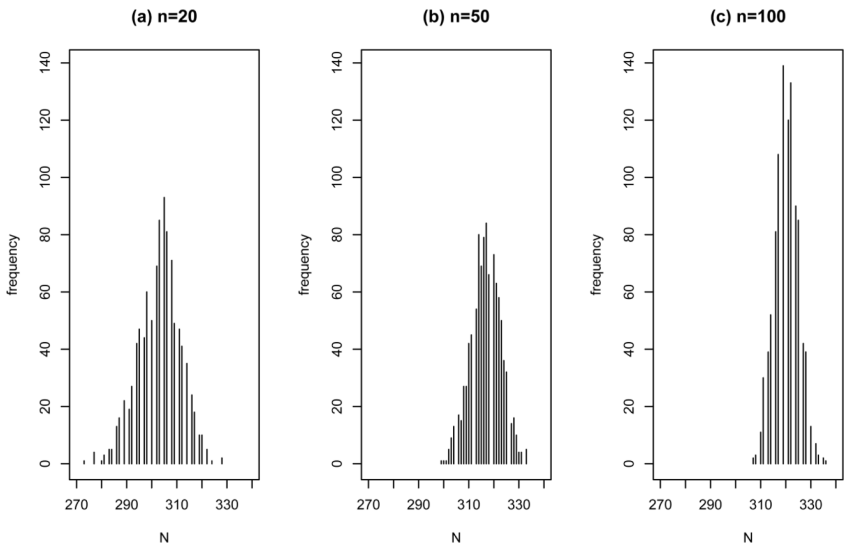

Figure 1.
Distribution of the estimated sample size N from 1000 pilot data sets of size n = 20, 50 or 100 under ρ = 0.3 and γ = 5

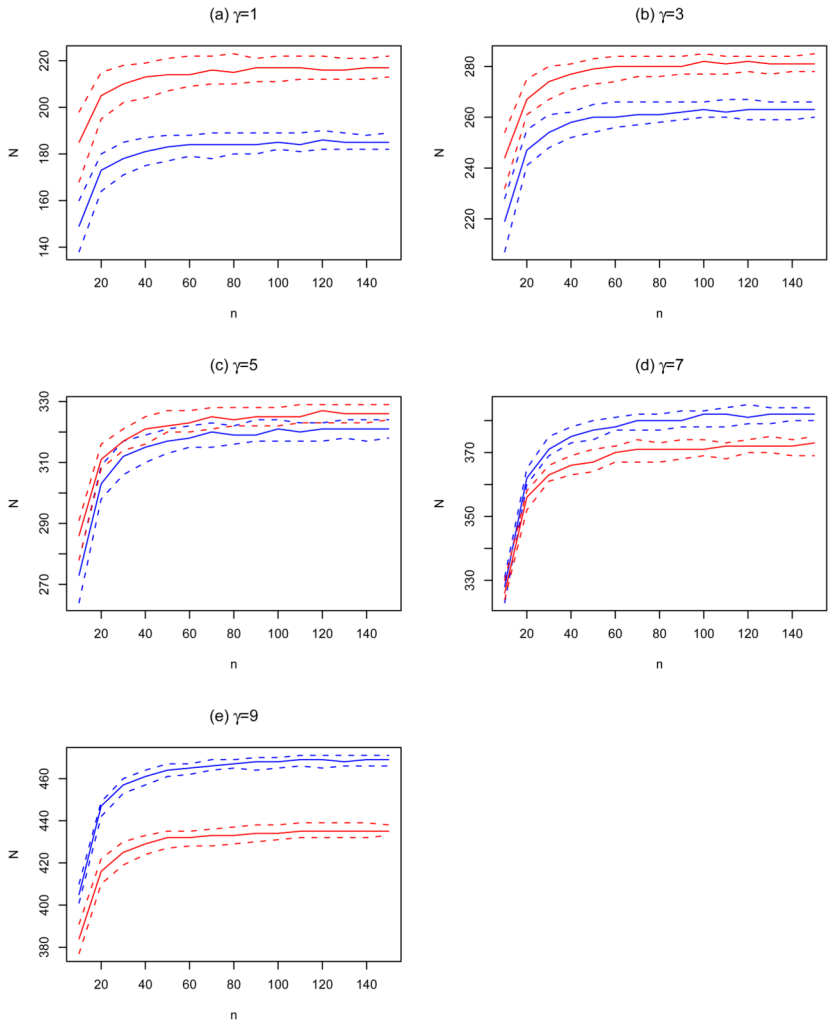

Figure 2.
Median (solid line) and the first and third quartiles (broken line) of the estimated sample sizes N from 1000 pilot data sets of size n with $\rho = 0.3$ in blue and $\rho = 0.6$ in red

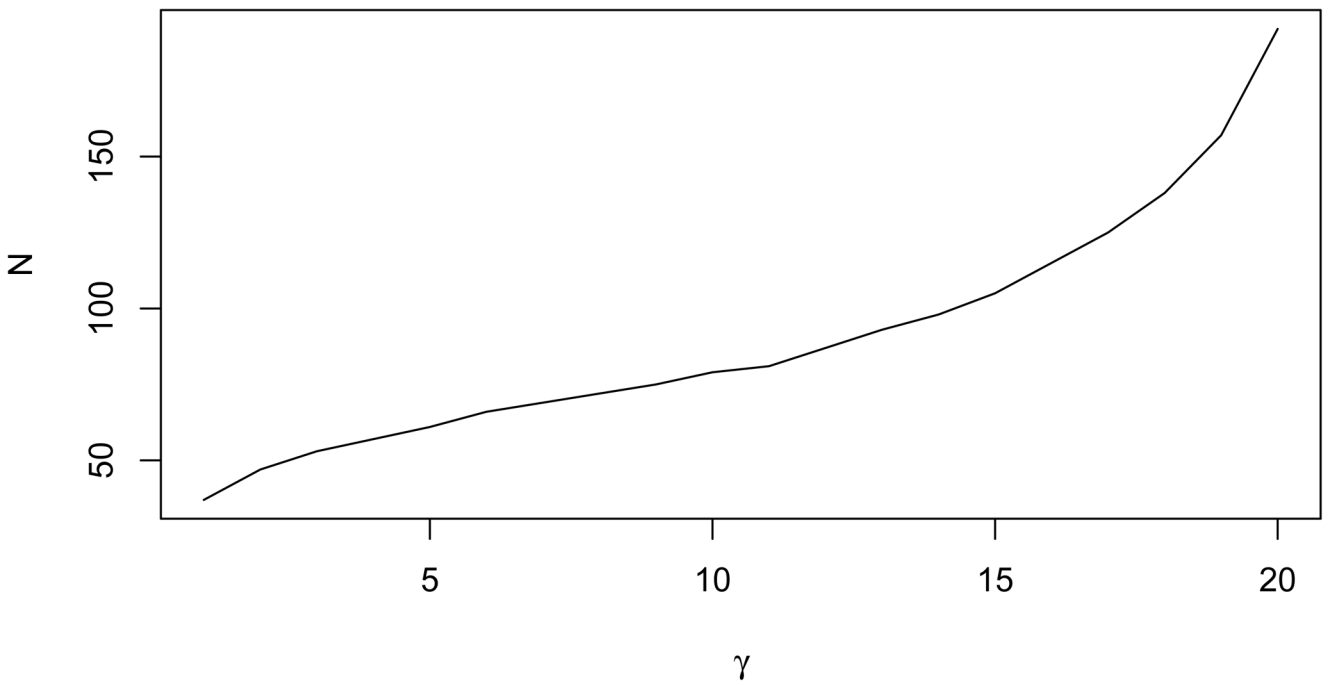

Figure 3.
Sample size required for a breast cancer study estimated using the Huang *et al.* (2003) data as pilot data

Table 1

Empirical power $1-\hat{\beta}_\gamma$: The three N's for each γ value are the first, second and third quartiles of estimated sample sizes from 1000 pilot data sets with n = 100 under $(\alpha, 1-\beta_\gamma) = (0.05, 0.8)$, $a_1 = a_2 = 1/2$, and $\rho = 0.3$

$\gamma = 1$		$\gamma = 3$		$\gamma = 5$		$\gamma = 7$	
N	$1-\hat{\beta}_\gamma$	N	$1-\hat{\beta}_\gamma$	N	$1-\hat{\beta}_\gamma$	N	$1-\hat{\beta}_\gamma$
182	0.775	260	0.774	317	0.778	378	0.757
185	0.780	263	0.794	321	0.799	382	0.808
189	0.807	266	0.790	324	0.782	383	0.780