

Chapter 3

A GENERAL FDR-BASED COMPUTATIONAL FRAMEWORK FOR SAMPLE SIZE PLANNING IN MICROARRAY STUDIES

Danh V. Nguyen^{1}, Hao Liu¹ and Damla Şentürk²*

¹Division of Biostatistics, University of California
Davis, California 95616, USA

and

²Department of Statistics, Pennsylvania State University
University Park, Pennsylvania 16802, USA

Abstract

The false discovery rate (FDR) is a measure of error in multiple testing widely used for DNA microarray data. FDR is a particularly useful measure of error especially in exploratory microarray studies where investigators aim to screen for differentially expressed genes. Various experimental designs and statistical methods have been proposed for detecting differential expression. However, it is important to determine the statistical power of a given design and method to detect differentially expressed genes at the planning stage. In this paper, we provide a general computational framework to determine the sample size and associated power using FDR as a measure of error. We illustrate the proposed sample size calculation framework with a model of gene expression that incorporates (1) heterogeneity of variance, (2) non-constant gene expression structure, and (3) measurement (technical) errors in the technology.

Key Words: Differential gene expression; DNA Microarray studies; False discovery rate (FDR); Measurement error; Multiple hypothesis testing; *P*-value; Simulation.

1. Introduction and Background

DNA microarray experiments are important tools in biomedical research. In particular, exploratory DNA microarray studies often aim to identify genes that are differentially expressed under different biological conditions. Follow-up validation studies, such as real

*E-mail address: ucdnguyen@ucdavis.edu, Phone: (530)754-6510; Correspondence: Danh V. Nguyen, Division of Biostatistics, MS1C, University of California, Davis, CA 95616, USA

time RT-PCR, of the set of identified genes are often carried out after DNA microarray experiments (Chuaqui et al., 2002; Davidson et al., 2004). Although DNA microarrays can be designed to be confirmatory tools, they are still largely used for exploratory experiments to monitor thousands of genes for differential expression. (We refer the reader to Nguyen et al. (2002) for a more thorough discussion of DNA microarray technologies.) Examples of such studies, include the identification of genes differentially expressed between breast cancer patients with mutation in the *BRCA1* versus *BRCA2* gene (Hedenfalk et al., 2001) and the the identification of differential gene expression among groups given different dietary fatty acids enrichment in a study of colon cancer in the rat (Davidson et al., 2004), among others.

In the context of exploratory microarray studies, the focus is on searching for genes that exhibit differential expression irrespective of any other genes. The search involves multiple statistical hypothesis testing since thousands of statistical tests, one for each gene (probe), are carried out to determine the set of differentially expressed genes. In this process, the ability to identify genes that exhibit true differential expression while controlling for errors can substantially reduce the cost of follow-up validation of microarray experiments. In addition, because of the exploratory nature and size of genome-wide microarray studies/applications, it is beneficial to use multiplicity (error) controlling procedures that minimize a rate of false positives. Procedures to control the false discovery rate (FDR), or the expected proportion of genes *declared expressed that are in fact actually not expressed*, have been useful in many microarray studies. FDR was developed by Benjamini and Hochberg (1995). Many have advocated the use of FDR as a measure of error in the context of searching for differential gene expression based on DNA microarray data (Tusher et al., 2001; Efron et al., 2001; Reiner et al., 2003; Storey, 2002, 2003; Storey and Tibshirani, 2003, among others).

Various experimental designs and statistical methods have been proposed for detecting differential expression (for examples, see Kerr and Churchill (2001), Kerr et al. (2001), Lee et al. (2002), and Tusher et al. (2001), among others). For a given statistical approach to detect differential gene expression in microarray experiments it is important, at the planning stage, to determine the statistical power of a given design and method to detect differentially expressed genes. This involves determining the sample size for the study to achieve a specified level of power. Numerous sample size planning procedures have been proposed for specific designs/methods and error control measures. For example, the seminal work of Lee and Whitmore (2002) is on sample size calculation to achieve a specified power and controlling the expected proportion of type I errors in microarray experiments, analogous to the traditional sample size calculation setting. More precisely, they adapted the following type I and II error definitions in the context of multiple testing, respectively: $E(V)/m_0$ and $E(T)/m_1$, where V is the number of truly unexpressed genes that are (falsely) declared expressed, T is the number of expressed genes that are (falsely) declared unexpressed, m_0 is the total number of unexpressed genes, and m_1 is the total number of expressed genes. Jung et al. (2005) proposed sample size calculation for the t-test based on a Bonferroni-type improved single-step method and controlling for the familywise error rate (the probability of at most one erroneous rejection). Wang and Chen (2004) proposed a method to determine the sample size needed to detect a fraction of truly expressed genes (out of m_1 total truly expressed genes) for the one- and two-sample t-tests. More recent approaches focus on

determining the sample size and controlling for the false discovery rate, defined as the expected proportion of false discoveries, $E(V/R)$ when $R > 0$ (and 0 otherwise), where R is the total number genes declared expressed (Yang et al., 2003; Dobbin and Simon, 2005; Hu et al., 2005; Pawitan et al., 2005; Jung, 2005). The FDR error measure will be discussed in more detail in Section 2..

Most of the current sample size calculation methods which aims to control the FDR are designed for a specific statistic (e.g. t-test) and/or study design. Although such approaches are useful, their generalization to more complex models of gene expression can be limited. We propose a more general computational framework to determine the sample size needed for any given microarray study using FDR as a measure of multiple testing error. The proposed computational framework allows for the calculation of sample size and power for a broad spectrum of experimental designs and statistical methods.

The paper is organized as follows. In Section 2., we describe, in more details, the popular sequential FDR methods used in practice to adjust for multiple testing in microarray studies. We also describe recent works that incorporate a more precise estimator of $\pi_0 \equiv m_0/(m_0 + m_1)$, the proportion of truly unexpressed genes, into the FDR procedure. Such an approach has been shown to improve the power to detect truly expressed genes (Storey, 2002; Nguyen, 2004a). These recent FDR developments are used in the proposed simple, but more general, simulation-based procedure for sample size calculation (Section 3.). We illustrate the proposed computational framework with a model of gene expression that incorporates (1) heterogeneity of variance, (2) a more complex gene expression structure, and (3) measurement (technical) errors in microarray technology. We illustrate the use of the resulting power surface (as a function of π_0 and sample size), to help guide the selection of the required sample size in Section 4..

2. FDR and Multiple Testing in Microarray Studies

We now introduce the basics of the false discovery rate in the context of multiple testing in microarray studies. As mentioned earlier, a widely used measure of error in multiple testing based on microarray data is the FDR, the expected proportion of false discoveries among R declared discoveries or rejections, introduced by Benjamini and Hochberg (1995). More precisely, FDR is defined as

$$\text{FDR} = E\left(\frac{V}{R}I_{\{R>0\}}\right) = E\left(\frac{V}{R} \mid R > 0\right) \Pr(R > 0), \quad (1)$$

where V is the number of erroneous rejections (Type I errors), R is the total number of rejections and $I_{\{A\}}$ denotes the indicator function for event A . Please refer to Table 1, where the outcomes for testing m hypotheses are summarized. The total number of genes being tested is m , among which m_0 and m_1 genes are truly unexpressed and truly expressed, respectively. As we discuss below, using a more precise estimator for the proportion of genes truly unexpressed, namely $\pi_0 \equiv m_0/m$, will result in more powerful FDR procedures.

Note in Table 1 that only R , W , and m are observable and all other quantities in the table are not observable. In particular, the random variables V and S that count the number of false and true positives (i.e., false and true declaration of genes that are differentially

expressed) are not observable. Also note that the well-known familywise error rate, the probability of rejecting any null hypothesis erroneously, is $\Pr(V > 0)$. Thus, FDR provides a much less strict criterion to control than the FWER does in multiple testing. Hence, an obvious gain in power is expected when controlling FDR compared to controlling FWER (Benjamini and Hochberg, 1995).

Table 1. Notations for the possible outcomes of testing m hypotheses (Benjamini and Hochberg, 1995). The proportion of true null hypotheses is $\pi_0 \equiv m_0/m$ and $\text{FDR} = E(\frac{V}{R} I_{\{R>0\}})$.

	Accept (Declare unexpressed)	Reject (Declare expressed)	Total
Null true (Truly unexpressed)	U	V (error I)	m_0
Alternative true (Truly expressed)	T (error II)	S	m_1
Total	W	R	m

2.1. FDR Procedures: Sequential and Direct Approaches

The traditional *sequential* approach to FDR, introduced by Benjamini and Hochberg (1995) (herein BH) and widely applied in microarray multiple testing, requires fixing an FDR level of control, say α ($0 < \alpha < 1$). Denote the observed p -values by p_1, \dots, p_m and the corresponding ordered observed p -values as $p_{(1)}, \dots, p_{(m)}$. The Benjamini and Hochberg FDR (BH-FDR) controlling procedure is to find

$$\hat{k}_{\text{BH}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \alpha \right\} \quad (2)$$

and rejecting null hypotheses corresponding to $p_{(1)}, \dots, p_{(\hat{k}_{\text{BH}})}$, where $\alpha \in (0, 1)$ is the pre-specified target control level. BH proved that the procedure (2) results in $\text{FDR} \leq \pi_0 \alpha$ for $0 \leq m_0 \leq m$, where $\pi_0 = m_0/m$ is the proportion of true null hypotheses. Since $0 \leq \pi_0 \leq 1$, it follows that FDR is controlled at level α for all configuration of m_0 . It was subsequently shown by Finner and Roters (2001) that the BH-FDR procedure (2) gives precisely $\text{FDR} = \pi_0 \alpha$. Therefore, the level of FDR control is actually $\pi_0 \alpha$, which is less than or equal to α . For this reason, the BH-FDR controlling procedure (2) is increasingly conservative as π_0 approaches zero. Consequently, this leads to a loss in power to detect true alternative hypotheses.

Therefore, incorporating a less conservative, hence, more precise estimate of π_0 into the FDR controlling procedure (2) can improve power, as was done in Benjamini, Krieger, and Yekutieli (2001) and Benjamini and Hochberg (2000). Similarly, Storey (2002, 2003) and Storey and Tibshirani (2003) also recognized that the estimation of π_0 is critical in the context of DNA microarray applications; however, they proposed a non-sequential, *direct*

approach to estimate the FDR for a fixed rejection region. Storey (2002) proposed a direct estimate of FDR for a fixed rejection region $[0, \gamma]$. More precisely, the proposed estimator of FDR is

$$\widehat{\text{FDR}}_\lambda(\gamma) = \hat{\pi}_0(\lambda) \frac{\gamma}{\widehat{\text{Pr}}(P \leq \gamma)}, \quad (3)$$

where $\widehat{\text{Pr}}(P \leq \gamma) = \#\{p_j \leq \gamma\}/m$ and $\hat{\pi}_0(\lambda)$ is a conservatively biased estimator of π_0 with parameter $\lambda \in (0, 1)$. We will elaborate on the estimator $\hat{\pi}_0(\lambda)$ in the following section. The estimator (3) is conservatively designed in the sense that $E[\widehat{\text{FDR}}_\lambda(\gamma)] \leq \text{FDR}(\gamma)$ for all γ and π_0 (Storey, 2002; Theorem 2).

2.2. More Powerful FDR Procedure Via Direct Estimation of π_0

For the original sequential FDR controlling procedure (2), we have that $\text{FDR} = \pi_0 \alpha$. Setting π_0 to its upper bound of one gives the desired level of FDR control, α . Thus, for the BH-FDR procedure, $\hat{\pi}_0(\text{BH}) \equiv 1$, because no information about π_0 was actually utilized from the distribution of observed p -values, $\{p_j\}_{j=1}^m$. Thus, we can express the original sequential FDR controlling procedure (2) as

$$\hat{k}_{\text{BH}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\hat{\pi}_0(\text{BH})} \right) \right\}. \quad (4)$$

Clearly, the choice of $\hat{\pi}_0(\text{BH}) = 1$ represents the *most* conservative choice. The extreme opposite to this most conservative choice would be to use π_0 itself in the BH-FDR procedure. Since the original BH-FDR provides $\text{FDR} = \pi_0 \alpha \leq \alpha$, it is conservative by a factor of $\pi_0 = m_0/m$. If π_0 (or equivalently m_0) is known, then the conservativeness can be corrected by applying the BH-FDR procedure at level $\alpha' = \alpha/\pi_0$, instead of α . This correction provides FDR control at level α , since $\text{FDR} = \pi_0 \alpha' = \alpha$. Thus, we define the *least conservative* FDR (LC-FDR) procedure as finding

$$\hat{k}_{\text{LC}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\pi_0} \right) \right\} \quad (5)$$

and then rejecting the hypotheses corresponding to the p -values $p_{(1)}, \dots, p_{(\hat{k}_{\text{LC}})}$. Although procedure (5) is ideal, it is not computable in practice since π_0 is unknown. However, we can approximate it by incorporating a conservative estimate of π_0 . Next, we describe a family of FDR procedures that are more powerful than the BH-FDR procedure (4) and they approximate the LC-FDR procedure (5).

As described earlier, because the direct approach to FDR incorporates a more precise estimator of π_0 , it is substantially more powerful than the BH-FDR procedure (Storey, 2002). Such an estimator, denoted by $\hat{\pi}_0(\lambda)$ and proposed by Storey (2002), is

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1 - \lambda)}, \quad (6)$$

for $\lambda \in (0, 1)$. The parameter λ balances between bias and variance. Note that as λ approaches 1, the bias is lower since $\#\{p_j > \lambda\}$ consists mostly of truly null p -values. The numerical properties of (6), with respect to power and FDR control, were studied in details

by Nguyen (2004a). The estimator $\hat{\pi}_0(\lambda)$ was designed to be conservatively biased (that is, $E[\hat{\pi}_0(\lambda)] \geq \pi_0$) in order to achieve FDR control in expectation: $E[\widehat{\text{FDR}}_\lambda(\gamma)] \leq \text{FDR}(\gamma)$ for all γ and π_0 .

However, the sequential FDR approach can be made as powerful as the direct estimation approach by incorporating the estimator (6) into the sequential FDR procedure (Nguyen, 2004a, 2004b, 2005). More precisely, using the estimator $\hat{\pi}_0(\lambda)$, the following family of sequential FDR algorithms (indexed by λ) can be used to better approximate the optimal (least conservative) FDR controlling procedure (5). This procedure involves finding

$$\hat{k}_\lambda = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\hat{\pi}_0(\lambda)} \right) \right\} \quad (7)$$

and then rejecting the hypotheses corresponding to the p -values $p_{(1)}, \dots, p_{(\hat{k}_\lambda)}$. The above FDR algorithm, proposed and studied in more details by Nguyen (2005), simply replaces the unknown π_0 in the optimal sequential FDR procedure with the conservatively designed estimator $\hat{\pi}_0(\lambda)$. Note that the FDR algorithm given by (7) is sequential and it utilizes the exact information (estimator) of π_0 as in the direct estimation approach; hence, the two approaches yield equivalent power. The interpretation and use are also straightforward. The user simply specifies the level of FDR control, α , and the calculation of (7) is easy. Thus, due to the improved power of the sequential procedure (7) to detect genes truly expressed and also due to the ease of interpretation, we utilize it for the sample size planning procedure in Section 3.. However, we first illustrate the simple calculations involved with the sequential FDR procedure (7) in the next section.

2.3. An Example Illustrating the Sequential FDR Calculations

To illustrate the simple calculations involved with the sequential FDR procedure (7), we use the well known hereditary breast cancer microarray data from Hendenfalk et al. (2001), which is available publicly at <http://research.nhgri.nih.gov/microarray/NEJM.Supplement/>. A central interest in this microarray study is to identify genes differentially expressed in breast cancer patients with mutations in the *BRCA1* gene relative to those with *BRCA2* mutations. Following Storey and Tibshirani (2003), we used $m = 3,170$ genes after pre-processing. Thus, for each of the $m = 3,170$ genes of interest in the study, a p -value was assigned to each gene based on a chosen comparison statistic. For illustration, we use permutation p -values corresponding to the two sample t-statistics to obtain the set of p -values: p_1, \dots, p_{3170} . For demonstration of the calculation, we choose $\lambda = 1/2$. Then we simply count the $\#\{p_j > 1/2\} = 1074$ and so the estimated proportion of unexpressed genes is $\hat{\pi}_0(\lambda) = 1074/(3170 \times 0.5) = 0.678$, as specified in equation (6). Next, to determine \hat{k}_λ in the FDR procedure (7), we order the p -values: $p_{(1)}, \dots, p_{(3170)}$. For an FDR level of $\alpha = 0.10$, for instance, we compute $\hat{k}_\lambda = \max\{j : p_{(j)} \leq (j/3170) \times (0.1/0.678)\} = 281$. Thus, 281 genes were identified as differentially expressed, corresponding to $p_{(1)}, \dots, p_{(281)}$. For illustrating the simple calculations involved with (7), we selected $\lambda = 1/2$. An optimal selection of λ based on minimizing the mean square error, defined as $\hat{\pi}_0(\text{OPT}) \equiv \lim_{\lambda \rightarrow 1} \hat{\pi}_0(\lambda)$, is available (Storey and Tibshirani, 2003). For this example, $\hat{\pi}_0(\text{OPT}) = 0.688$, which is quite close to the estimate of $\hat{\pi}_0(0.5) = 0.678$.

3. Sample Size Calculation Using FDR as a Measure of Error in Multiple Testing

3.1. A General FDR-based Algorithm for Sample Size Determination

Generally, sample size planning depends on various parameters determined by the investigators. These include (1) the variability structure/model of the population, (2) the magnitude of the expression changes among experimental conditions under a specified gene expression model, (3) the power to detect the expression changes, and (4) the level of error control α . In the traditional sample size calculation setting involving a single hypothesis test of the null hypothesis \mathcal{H}^0 versus the alternative hypothesis \mathcal{H}^A , (4) is simply the significance level α , which is the specified Type I error = $Pr(\text{reject } \mathcal{H}^0 \text{ when } \mathcal{H}^0 \text{ is true})$. In biomedical research, the typical values of the Type I error used is $\alpha = 0.05$ or 0.01 . In the context of testing m hypothesis (one for each gene) based on microarray gene expression data considered here, (4) is the false discovery rate level $\alpha \in (0, 1)$. Typically the FDR level of control used in practice is less than 10% ($\alpha < 0.10$); however, the investigator may specify the needed level of FDR error control, depending on the specific goals of the experiment. The magnitude of gene expression changes in item (2) should be biologically meaningful or at levels that the investigator wishes to detect generally. Items (1)-(3) are typically based on a combination of prior data (preliminary/ pilot or existing data in the published literature) and explicit assumptions. Incorporating similar prior data to quantify the information in items (1)-(3) for sample size planning of a new microarray study is feasible due to the availability of data from previous microarray experiments. Examples include the Gene Expression Omnibus database (<http://www.ncbi.nlm.nih.gov/geo/>), the Stanford MicroArray Database <http://genome-www5.stanford.edu/>, and other published microarray results/data.

Additionally, utilizing FDR procedure (7) for FDR error control necessarily requires an estimate of the proportion of unexpressed genes $\pi_0 = m_0/m$, namely $\hat{\pi}_0(\lambda)$. This estimate can be obtained from similar existing data or preliminary data. However, at the study planning stage, it is informative to consider the power corresponding to a sequence of sample size n for a range of π_0 values. Such an approach can be particularly informative when the estimates of π_0 are difficult to ascertain at the planning stage. However, for many microarray experiments, π_0 is much less than one and most likely above 0.50. Therefore, for the power and sample size analysis, we advocate examining power and sample size for a *range* of possible π_0 values. Given the power curve (surface), as a function of π_0 (at a fixed FDR level), the investigator can decide on the appropriate sample size. A detailed example of this is provided in Section 4..

First, we outline the general algorithm to generate the power surface as a function of the sample size n and π_0 , for a specified FDR level α in this section. We then illustrate the proposed algorithm with a specific model for microarray gene expression data, which incorporates (1) heterogeneity of variance, (2) more complex gene expression structures, and (3) measurement (technical) errors in the technology in Section 3.2..

Procedure to generate the power surface

1. **FDR control level.** Specify the desired FDR control level $\alpha \in (0, 1)$.

2. **Simulation model.** Specify gene expression model and parameters. Based on the model, H (e.g. $H = 500$) Monte Carlo data sets of size $n \times m$ are generated for each sample size n and the proportion of true null hypotheses π_0 .
3. **Hypotheses, test-statistics, and p -values.** Specify the null and alternative hypotheses, \mathcal{H}_j^0 and \mathcal{H}_j^A , for genes $j = 1, \dots, m$. Compute the associated test-statistics, denoted by t_1, \dots, t_m and their corresponding p -values by p_1, \dots, p_m . Obtain the sorted p -values, $p_{(1)} \leq \dots \leq p_{(m)}$.
4. **Sequential FDR procedure.** Estimate the proportion of the null hypotheses given in (6): $\hat{\pi}_0(\lambda) = \#\{p_j > \lambda\}/[m(1 - \lambda)]$ and compute the FDR procedure given by (7): $\hat{k}_\lambda = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\hat{\pi}_0(\lambda)} \right) \right\}$. From the set of null hypotheses $\{\mathcal{H}_j^0; j = 1, \dots, m\}$, reject the null hypotheses corresponding to the p -values $p_{(1)}, \dots, p_{(\hat{k}_\lambda)}$. (The method of Storey and Tibshirani (2003) can be used to choose λ , although for choose $\lambda = 0.5$ to illustrate the concepts here.)
5. **Generate power surface.** Note that the computations in (3)-(4) are repeated for each generated data set $h = 1, \dots, H$. For data set h , the key quantities summarized in Table 1 are tracked. In particular, denote v_h to be the observed number of false rejections, r_h to be the total number of rejections, and s_h to be the observed number of (correct) rejections, when the null hypotheses are false. The power corresponding to each sample size n and π_0 , denoted $P(n, \pi_0)$, is obtained as $H^{-1} \sum_{h=1}^H (s_h/m_1)$. It is the average proportion of the true alternative hypotheses correctly identified (discovered), averaged over the H data sets (simulations).
6. **Check FDR control level.** Due to the complexity of the model, possibly with general dependence structure, it is important to check the FDR control from the simulation. For each sample size n and proportion of true null hypotheses π_0 , we can obtain the average proportion of false rejections as $H^{-1} \sum_{h=1}^H v_h / \max\{r_h, 1\}$.
7. **Determine sample size needed.** The sample size needed for a specified level of power over a range of π_0 values can be directly obtained from the generated power surface $P(n, \pi_0)$ in item 5 described above.

3.2. A Specific Example: Incorporating Measurement Error, Heterogeneity of Variance, and More Complex Gene Expression Models

To illustrate the proposed FDR-based sample size algorithm, we consider a model for gene expression data that incorporates biological variability and technical or measurement errors in the microarray technologies. First, we describe the technical or measurement errors for microarray gene expression. A model for gene expression with both additive and multiplicative measurement errors is

$$x_k = \mu_k e^\eta + \epsilon, \quad k = 1, \dots, K, \quad (8)$$

where x_k is the observed (gene expression) intensity measurement and μ_k is the true (unknown) gene expression (Rocke and Durbin, 2001) in group k . (For example, $K = 2$ for

two groups when comparing a control and an experimental group.) This gene expression measurement error model has been widely adopted and provides a reasonable approximation to empirical data (see, for example, Rocke and Durbin, 2001; Zien et al. 2003; Huber et al., 2002 and references therein). In the above error model, the additive measurement error is $\epsilon \sim N(0, \sigma_\epsilon^2)$ and represents the error associated with genes that are not expressed or expressed at low levels. The multiplicative (proportional) measurement error is $\eta \sim N(0, \sigma_\eta^2)$ and it represents the proportional error for genes expressed at high levels. Model (8) is a two-component error model which approximates a constant standard deviation for low expression levels and a constant coefficient of variation for higher expression levels.

To illustrate the proposed procedure for FDR-based sample size calculation, we consider the following lognormal model of gene expression for μ_k (from Zien et al. (2003)),

$$\mu_k = \mu_k^* e^\beta, \quad k = 1, \dots, K, \quad (9)$$

where $\beta \sim N(0, \sigma^2)$, μ_k^* is the mean gene expression level in group k , and the parameter σ represents the standard deviation of the biological variability. For illustration, we take $K = 2$ for a two-group comparison study. The family of lognormal distributions have been used as a model for gene expression (Nguyen and Rocke, 2004; Zien et al., 2003; Nguyen, 2004a, 2005b; Konishi, 2004, among others). For an introduction to the use of the log-normal distribution in the sciences, see Limpert et al. (2001).

We based the choice of the specific model parameters on the previous works of Zien et al. (2003) and Rocke and Durbin (2002), where the parameters were estimated based on replicated microarray experiments of Bartosiewicz et al. (2000), Stuart et al. (2001), and Lemon et al. (2002). The standard deviations of additive and multiplicative measurement errors are $\sigma_\epsilon = 100$ and $\sigma_\eta = 0.12$, respectively. Rather than fixing the biological variability parameter for all genes, we incorporate heterogeneity of biological variability by allowing σ^2 to take on values from the set $\{0.01, 0.04, 0.09, 0.16, 0.25\}$. Also, varying levels of gene expression are incorporated into the gene expression model (9). To do this, note that the fold ratio of expression between group 1 and 2 for gene j is $\theta_j \equiv \mu_{j1}^*/\mu_{j2}^*$. Also, the signal to noise ratio, averaged over groups 1 and 2, is $\delta_j \equiv \sqrt{\mu_{j1}^*\mu_{j2}^*}/\sigma_\epsilon$. The gene expression fold change between groups 1 and 2, namely θ_j , is allowed to vary from the set $\{7, 5, 4, 3, 2, .5, .25, .1\}$. Thus, both over- and under-expressed genes are represented. We also allowed the expression signal to noise ratio (δ) to vary, ranging from about 2.7 to 0.32. From a total of $m = 2000$ genes, m_1 are differentially expressed and the remaining $m_0 = m - m_1$ are unexpressed. We simulate 500 data sets of size $n \times m$, based on the above model, for $\pi_0 = m_0/m$ ranging from 0.10 to 0.95 and $n_k = 6, 8, \dots, 28$ ($k = 1, 2$).

For this example, the null hypothesis for gene j is that there is no differential expression between groups 1 and 2, namely $\mathcal{H}_j^0 : \mu_{j1}^* = \mu_{j2}^*$, for $j = 1, \dots, m$. As we expect that there are both over- and under-expressed genes, the natural alternatives are $\mathcal{H}_j^A : \mu_{j1}^* \neq \mu_{j2}^*$, $j = 1, \dots, m$. We examine the following comparison t-statistic, with different group variances, for the j th gene: $t_j = (\bar{x}_{j2} - \bar{x}_{j1})/s_j^*$. The denominator is $s_j^* = \{(s_{j1}^2/n_1) + (s_{j2}^2/n_2)\}^{1/2}$, where s_{j1}^2 and s_{j2}^2 are the sample variances of group 1 and 2, respectively. The assumptions of the standard t-statistics are not tenable under the current model, which has non-constant variance, varying mean structure, non-identically distributed observations, and measurement errors. To avoid distributional assumptions,

we consider permutation-based p -values. For the b th permutation of the $n = n_1 + n_2$ group labels, denote the corresponding comparison statistics based on the permuted data by $\{t_j^{0b}\}_{j=1}^m$ ($b = 1, \dots, B$). The permutation-based p -value for gene j can be computed as $p_j = \sum_{b=1}^B \#\{k : |t_k^{0b}| \geq |t_j|\} / (mB)$. The FDR procedure (7) is then applied to the ordered p -values $p_{(1)}, \dots, p_{(m)}$.

4. Power and FDR Control Surfaces for Study Design

4.1. Power and Sample Size

To obtain the power surface, we generate $H = 500$ Monte Carlo data sets from the model (8)-(9) for each π_0 and sample size n combination. More precisely, we generate 500 data sets for each combination of $\pi_0 = .1, .2, \dots, .9, .95$ and $n = 6, 8, \dots, 28$, for a specified FDR control level $0 < \alpha < 1$. (For illustration, we used an FDR level of $\alpha = 0.05$.) In the current multiple testing setting considered here, power is calculated as the proportion of true alternative hypotheses correctly rejected (averaged over the H simulations): $H^{-1} \sum_{h=1}^H (s_h/m_1) \equiv P(n, \pi_0)$, where s_h is the observed number of correct rejections (discoveries). To guide in the FDR-based sample size selection, we examine the power surface $P(n, \pi_0)$, illustrated in Figure 1. Because the proportion of true null hypotheses $\pi_0 = m_0/m$ is not known precisely *a priori*, examining $P(n, \pi_0)$ to guide the selection of n for a range of π_0 values is useful. Thus, $P(n, \pi_0)$ given in Figure 1 is over a range of π_0 between 0 and 1. For example, using Figure 1, a sample size of $n_i = 22$ yields 89%-80% power corresponding to $0.3 \leq \pi_0 \leq 0.6$. Such direct use and interpretation of $P(n, \pi_0)$ may be more useful in designing gene expression microarray experiments since the uncertainty in π_0 is quite large in practice, especially for novel experiments.

4.2. Expected FDR Control

Figure 2 displays the estimate of $\text{FDR} = E[(V/R)I_{\{R>0\}}]$, the expected proportion of false rejections, based on the $H = 500$ Monte Carlo data sets for each n and π_0 . More precisely, we computed $H^{-1} \sum_{h=1}^H v_h / \max\{r_h, 1\}$, where v_h is the observed number of false rejections (i.e. number of truly unexpressed genes falsely declared to be expressed) and r_h is the total number of rejections. In this example, the desired/specified level of FDR control is $\alpha = 0.05$. The FDR surface, corresponding to sequential FDR procedure (7), displayed in Figure 2 indicates that the FDR control level $\alpha = 0.05$ was achieved.

5. Discussion

In this work, we have proposed an FDR-based computational approach to determine the sample size and associated power for gene expression microarray studies. Although we have chosen to illustrate the algorithm with an explicit model for a two-group comparison study, which incorporates (1) heterogeneity of variance, (2) a more complex gene expression structure, and (3) measurement (technical) errors in microarray technology, the procedure is applicable to any assumed model of gene expression and the corresponding statistical method of analysis. For example, extensions to more than two comparisons and

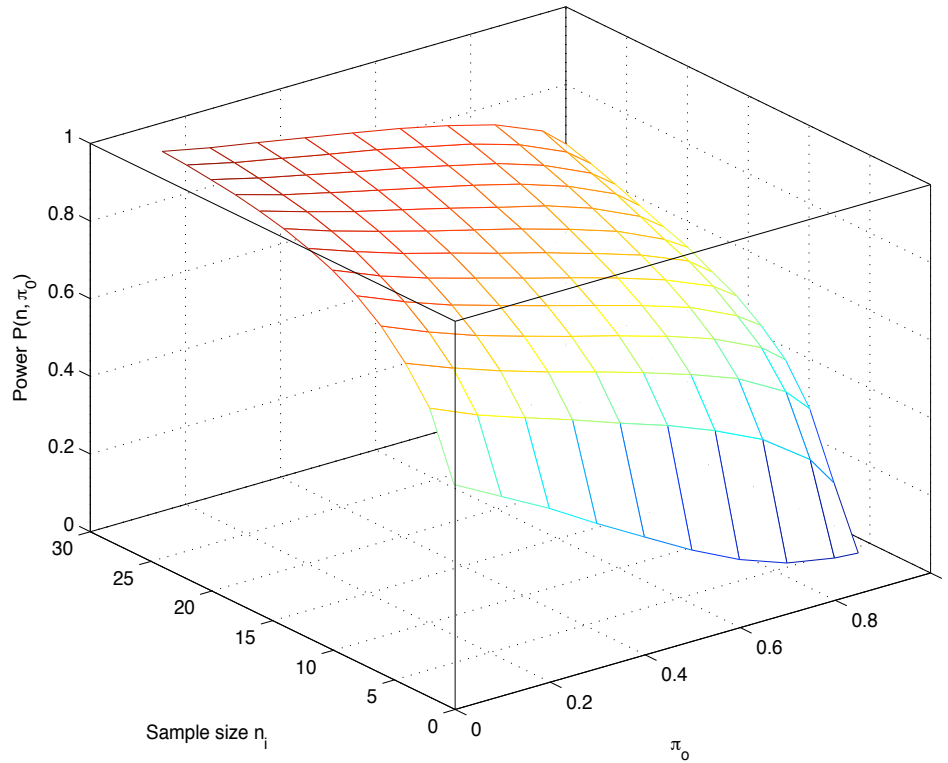


Figure 1. Power. The power surface, $P(n, \pi_0)$ as a function of sample size n and π_0 , the proportion of truly unexpressed genes.

other more complex gene expression models are feasible via the proposed computational approach. Also, under the current framework, sample size and power analysis can be implemented for (simulated) dependent gene expression data. Although it is difficult to simulate dependence network (structure) of thousands of genes simultaneously, data with some simple dependence structures can be simulated to understand the impact of dependence on sample size at the planning stage (Nguyen, 2004). The FDR-based approach used, namely (7) in the sample size algorithm, can be applied to other high dimensional data where m is large, including proteomics and metabolomics data among others.

We also note that in the proposed algorithm, we did not specify the sequence of sample sizes needed to achieve a level of power, typically 80%-95% power in practice. However, this can be achieved by using existing approximate sample size formulas, such as those proposed in Yang et al. (2003), Dobbin and Simon (2005) and Jung (2005) among others. Alternatively, an adaptive procedure can easily be implemented where the power, for a given sample size n , can be monitored and the sample size sequence is then adjusted (increased or decreased) automatically according to the desired level of power.

Finally, we note that for complex models and gene dependence structures, especially when the sample size is small, the assumptions needed for FDR theory may not be met. Thus, we emphasize the need to check the (average) FDR control level (item 7, Section

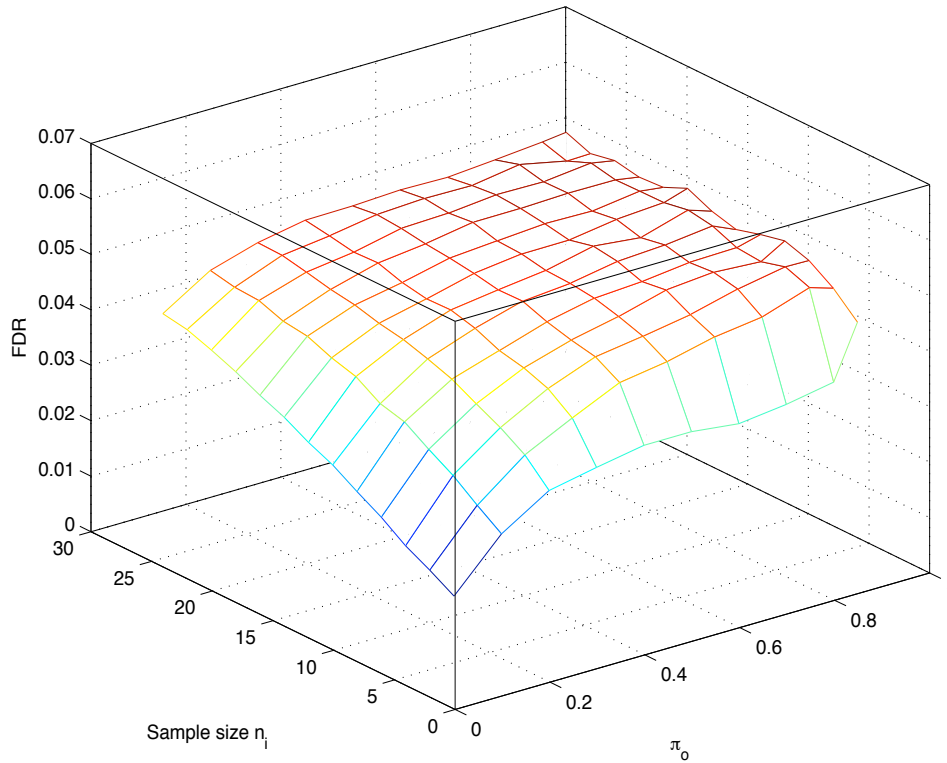


Figure 2. Expected FDR control. The FDR surface, given as function of sample size n and π_0 , the proportion of truly unexpressed genes. The level of FDR control is at $\alpha = 0.05$.

3.1.). The required calculations needed for this task is a by-product of the FDR procedure (7); hence the FDR control level can be easily checked.

Acknowledgement

DVN was partially supported by an American Cancer Society IRG program grant, through the UC Davis Cancer Center.

References

- Bartosiewicz, M., Trounstein, M., Barker, D., Johnson, R., and Buckpitt, A. (2000). Development of a toxicological gene array and quantitative assessment of this technology. *Archives of Biochemistry and Biophysics*, **376**, 66–73.
- Benjamini, Y., and 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**, 289–300.

- Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.
- Benjamini, Y., Krieger, A., and Yekutieli, D. (2001). *Two staged linear step up FDR controlling procedure*. Technical report, Department of Statistics and Operation Research, Tel-Aviv University, and Department of Statistics, Wharton School, University of Pennsylvania.
- Chuaqui, R.F., Bonner, R.F., Best, C.J., Gillespie, J.W., Flaig, M.J., Hewitt, S.M., Phillips, J.L., Krizman, D.B., Tangrea, M.A., Ahram, M., Linehan, W.M., Knezevic, V., and Emmert-Buck, M.R. (2002). Post-analysis follow-up and validation of microarray experiments. *Nature Genetics*, **32** Supplement, 509–514.
- Davidson, L.A., Nguyen, D.V., Hokanson, R.M., Callaway, E.S., Isett, R.B., Turner, N.D., Dougherty, E.R., Lupton, J.R., Carroll, R.J., and Chapkin, R.S. (2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Research*, **64**, 6797–6804.
- Dobbin, K., and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.
- Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.
- Finner, H., and Rotter, M. (2000). On the false discovery rate and expected type I errors. *Biometrical Journal*, **43**, 985–1005.
- Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E. R., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A., and Trent J. (2001). Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, **344**, 539–548.
- Hu, J., Zou, F., and Wright, F.A. (2005). Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics*, **21**, 3264–3272.
- Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96-S104.
- Jung, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*, **21**, 3097–3104.
- Jung, S.-H., Bang, H., and Young, S. (2005) Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, **6**, 157–169.
- Kerr, M.K., and Churchill, G.A. (2001). Experimental design issues for gene expression microarrays. *Biostatistics*, **2**, 183–201.

- Kerr, M.K., Martin, M., and Churchill, G.A. (2001). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819–837.
- Konishi, T. (2004). Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics*, **5**, 5.
- Lee, M-LT., and Whitmore, G.A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, **21**, 3543–3570.
- Lee, M-LT., Lu, W., Whitmore, G.A., and Beier, D. (2002) Models for microarray gene expression data. *Journal of Biopharmaceutical Statistics*, **21**, 1–19.
- Lemon, W.J., Palatini, J.J., Krahe, R., and Wright, F.A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, **18**, 1470–1476.
- Limpert, E., Stahel, W.A., and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, **5**, 341–352.
- Nguyen, D.V. (2004a). On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray studies. *Computational Statistics and Data Analysis*, **47**, 611–637.
- Nguyen, D.V. (2004b). A comparison of direct and sequential false discovery rate algorithms: computational experiments for exploratory DNA microarray studies. *Computing Science Statistics*, **36**, in-press.
- Nguyen, D.V. (2005). A unified computational framework to compare direct and sequential false discovery rate algorithms for exploratory DNA microarray studies. *Journal of Data Science*, **3**, in-press.
- Nguyen, D.V. (2005b). Partial least squares dimension reduction for microarray gene expression data with a censored response. *Mathematical Biosciences*, **193**, 119–137.
- Nguyen, D.V., Arpat, A.B., Wang, N., and Carroll, R.J. (2002). DNA Microarray experiments: biological and technological aspects. *Biometrics*, **58**, 701–717.
- Nguyen, D.V. and Rocke, D.M. (2004). On partial least squares dimension reduction for microarray-based classification: a simulation study. *Computational Statistics and Data Analysis*, **46**, 407–425.
- Pawitan, Y., Michiels, S., Koscielny, S., Gusnato, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.
- Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.
- Rocke, D.M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, **8**, 557–569.

-
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, **64**, 479–498.
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q -value. *Annals of Statistics*, **31**, 2013–2031.
- Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.
- Stuart, O., Bush, T., and Nigam, K. (2001). Changes in global gene expression patterns during development of and maturation of rat kidney. *Proceedings of the National Academy of Sciences*, **98**, 5649–5654.
- Tusher, V.G., Tibshirani, R., and Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.
- Yang, M.C.K., Yang, J.J., McIndoe, R.A., and She, J.X. (2003). Microarray experimental design: power and sample size considerations. *Physiol Genomics*, **16**, 24–28.
- Wang S.-J., and Chen, J.J. (2004). Sample size for identifying differentially expressed genes in microarray experiments. *Journal of Computational Biology*, **11**, 714–726.
- Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2003). Microarrays: how many do you need? *Journal of Computational Biology*, **10**, 653–667.