5 The Role of Sample Size on Measures of Uncertainty and Power

Gary L. Gadbury, Qinfang Xiang, Jode W. Edwards, Grier P. Page, and David B. Allison

CONTENTS

5.1 INTRODUCTION

Since an initial focus on "fold change" (cf., Reference 1), researchers have recognized the need to quantify statistical significance of estimated differences in genetic expression between two or more treatment groups using microarrays [2]. In light of this, replication has been recognized as a critical element of the design of microarray studies [3]. Replication may imply spotting a single gene multiple times on one array [4] or multiple tissue samples that each have their own array (e.g., [5,6]). We consider the latter and use the term "sample size" to refer to number of arrays in a study. For another discussion of levels of replication see Simon and Dobbin [7].

Replicate arrays in an experiment provide for statistical tests of differential expression at the level of a specific gene [8–10]. Regardless of the test used, the result is often a measure of "certainty" (e.g., a *p*-value) in rejecting a null hypothesis of no differential expression. However, in microarray studies there are questions that remain

to be answered and new ones that emerge. Examples are as follows:

1. At what threshold is a p-value statistically significant?
2. Of those genes declared differentially expressed, what proportion are truly differentially expressed?
3. Of those not so declared, what proportion are *not* differentially expressed?
4. Of the genes that are truly differentially expressed, what proportion do we expect to detect in a particular study?
5. What role does sample size play in all of these questions?

Obtaining answers to these questions can be a challenging task due to small sample sizes that are typical in many microarray experiments. If testing for a difference in mean genetic expression at a specific gene using a parametric test, the validity of a p-value may be questioned when sample sizes are small. Nonparametric tests also have difficulties since randomization distributions or bootstrapped distributions can be course with small samples, and p-values cannot attain small enough values to be "statistically significant."

This chapter presents techniques that facilitate answers to questions 1 to 5. We consider quantities of interest in a microarray study shown in Table 5.1. Two of these quantities are the expected number of genes that are (1) differentially expressed and will be detected as "significant" at a particular threshold and (2) not differentially expressed and will not be detected as such, denoted D and A, respectively. The other two quantities are the expected number of genes that are differentially expressed but are not so declared (B) and are not differentially expressed but are so declared (C). Proportions based on these quantities, defined in Gadbury et al. [11] are:

$$\text{TP} = \frac{D}{C+D}, \quad \text{TN} = \frac{A}{A+B}, \quad \text{EDR} = \frac{D}{B+D} \tag{5.1}$$

where each is defined as zero if its denominator is zero. TP is true positive; TN, true negative; and EDR, the expected discovery rate, which is the expected proportion of genes that will be declared significant at a particular threshold among all genes that

TABLE 5.1
Quantities of Interest in Microarray Experiments

	Genes for which there is no real effect	Genes for which there is a real effect
Genes not declared significant at designated threshold	A	B
Genes declared significant at designated threshold	C	D

Note: $A + B + C + D$ = the number of genes analyzed in a microarray experiment.

are truly differentially expressed. EDR sounds like but is not identical to the notion of power. It is an "expected proportion" and there may be no gene — specific test with a power identical to the EDR. Moreover, TP and TN are expected proportions and there may be no specific gene that has a "true positive probability" (or true negative) equal to TP (or TN). The focus of this chapter is to discuss techniques to estimate TP, TN, and EDR from microarrray experiments, and to present methods that evaluate the role of sample size in bringing these proportions to desired levels. Traditional power calculations for planning future sample sizes to detect gene specific effects are notably problematic with microarray data since information about variances and meaningful effect sizes are typically absent [12].

5.2 TP, TN, AND EDR IN MICROARRAY EXPERIMENTS

In this chapter it is the interplay between threshold, sample size, and the proportions in Equation 5.1 that is of interest. An example is shown in Figure 5.1. The figure is based on an "experimental situation" (details discussed later) with two treatment groups of equal size, n = integers 2 to 10, 20, 40 and a threshold at which a gene is declared differentially expressed equal to 0.1, 0.05, 0.01, 0.001, 0.0001, and 0.00001 shown on a logarithm (base 10) scale. The figure shows that as sample size increases, so does EDR. However, EDR is smaller for smaller thresholds since the criteria for declaring a gene differentially expressed are more strict (this assumes that there are indeed genes that are truly differentially expressed and it is our ability to detect them

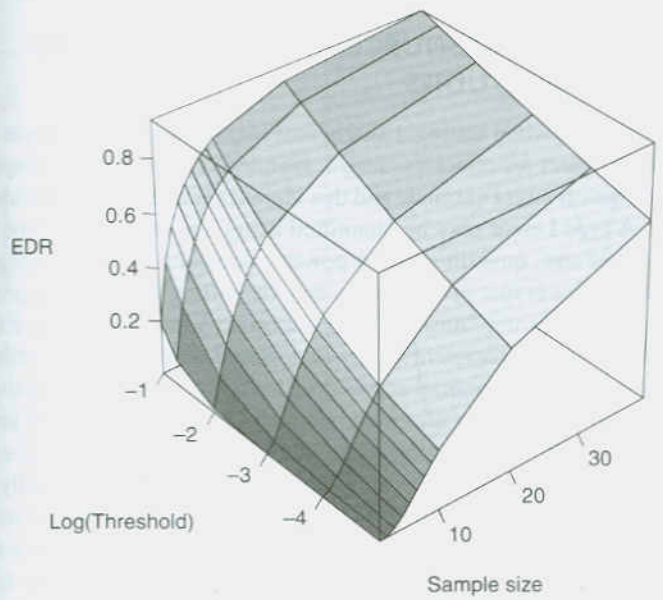

FIGURE 5.1 Three-dimensional plot showing an expected discovery rate (EDR) for varying sample sizes and threshold (on logarithm base 10 scale).

as such that is in question). Using larger thresholds, (and thus increasing EDR) comes at the cost of reducing TP (not shown in Figure 5.1).

The five questions in the introduction and the quantities in Equation 5.1 have interested others and have led to a body of literature on this topic. Much of the literature in this area names three quantities of interest that are related to those in Equation 5.1. First is the false discovery rate (FDR) that is analogous to 1 − TP, a false negative rate (i.e., 1 − TN), and what some have called "power," analogous to EDR (e.g., [13]), though some have called "power" to be one minus the probability of a false negative, interpreting "false negative" as a type II error [14]. Motivating the need for these results can be illustrated by considering Question 1 from the introduction. The answer may seem straightforward enough were it not for the many thousands of simultaneous tests that are conducted in a microarray experiment. Some of the traditional corrections for multiple comparisons such as the Bonferroni technique were not developed for this context and are far too conservative, particularly when results from an initial study might be used to plan follow-up studies, that is, too small a threshold may "miss" many interesting genes worthy of further attention. Less conservative methods have been developed that, rather than controlling for an experiment-wise error, controls instead the expected proportion of falsely rejected null hypotheses [15–17].

The approach to estimating the above described quantities from microarray data has generally fallen into two (at least) areas: permutation based methods, and model based methods involving Bayesian posterior probabilities. All have generally recognized the importance of sample size in bringing these quantities to desired levels.

5.3 SAMPLE SIZE AND SOURCES OF UNCERTAINTY IN MICROARRAY STUDIES

It is well known that when testing a single null hypothesis, small sample sizes have lower power to detect an effect vs. larger samples. Test statistics computed using small samples have a larger variance and this variance makes it more difficult to "see" a true effect. A type I error may be quantified using a p-value and a type II error, at a particular effect size, quantified using power calculations. An emerging paradigm in microarray studies is that investigators may be willing to tolerate a proportion of type I errors in favor of not "missing" any important genes that do have differential expression. In an ideal experiment all genes declared differentially expressed would be ones that have a true differential expression due to the treatment condition, TP = 1 or FDR = 0. Those genes that are ruled out as differentially expressed are also correctly determined, TN = 1. Finally, when planning and conducting an experiment, one would hope to "expect to discover" all of the important differentially expressed genes (high EDR). In practice, these quantities depend on variance of test statistics and some have used estimates of TP and TN to reflect uncertainty in a microarray experiment (e.g., low TP implies more uncertainty). The sample size has a direct effect on measures of TP, TN, and EDR in a microarray experiment. Figure 5.2 shows a diagram depicting a hypothetical two-dye microarray experiment. Only two arrays are shown (due to space restrictions) in each of two treatment groups. Issues related to

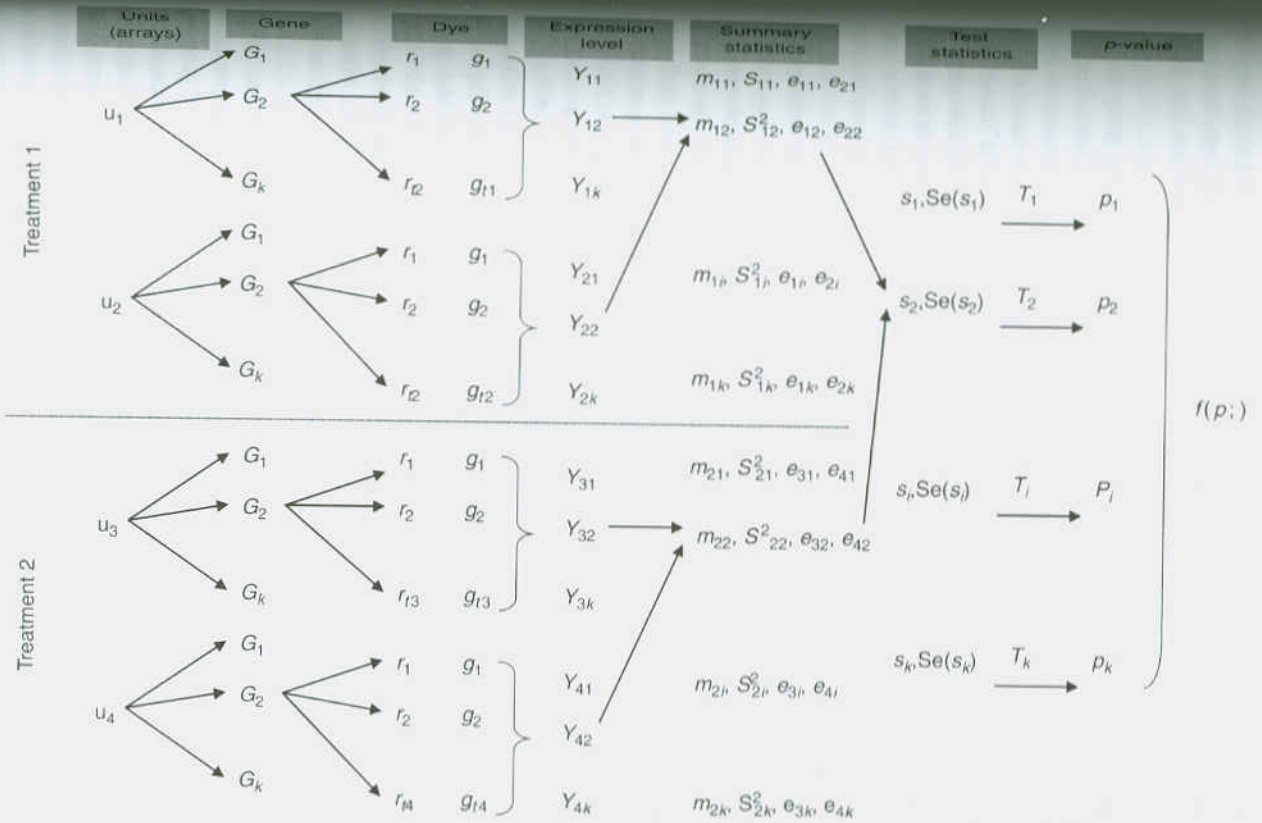

FIGURE 5.2 Diagram of a hypothetical 2-dye microarray experiment with four arrays divided into two treatment groups of two arrays each.

background correction and normalization are not considered. On each array, k genes are spotted. Figure 5.2 expands on the second gene, G_2, on each array. The r_i and g_i denote pixel intensities at each spot on the red and green channels, respectively. Pixel level data can be used to evaluate measurement (technical) errors in an experiment [18] but this level of experimental variance is not considered here. We assume that the pixel intensities are summarized into an expression level for the ith experimental unit (array) for G_2, denoted Y_{i2}.

The expression levels for the second gene within each treatment group are then summarized by some statistic, often using a mean and variance. Then a "test" for differential expression at each gene can be constructed using some measure of differential expression and a standard error, $\delta, \text{Se}_{(\delta)}$, allowing computation of a test statistic, T, that may be a contrast in an ANOVA model. A "p-value" results from the test statistic being compared against a null reference distribution. This reference distribution may be obtained via permutation or bootstrap procedures or an assumed model form for T. Finally all tests produce a distribution of p-values, $p = p_1, \ldots, p_k$, that can be modeled by some density function, $f(p;\theta)$, a concept used in recent papers [19–21].

Figure 5.2 helps in the following discussion that reviews various approaches to computing or estimating the proportions (or analogous quantities) in Equation 5.1. Pepe et al. [22] used rank based statistics to order genes by degree of observed differential expression. Their measure of variability was the probability that a gene, g, is ranked in the top c genes, denoted by $P_g(c) = P[\text{Rank}(g) \leq c]$. This probability may be thought of as analogous to TP in Equation 5.1, that is, it is a measure of "certainty" that a gene will reappear as important (i.e., the top c genes, where c is selected by the researcher) in follow-up studies. They used a bootstrap routine to estimate this probability where the resampling unit was at the level of the tissue (array). Their example data set included two groups of 30 and 23 arrays making the bootstrap a useful technique to compute quantities associated with sampling variability. A simulation technique was proposed for sample size calculations (e.g., power computations) where the original data set served as a population model for a bootstrap routine. The technique has limitations with small sample sizes where the bootstrap distribution would be too coarse to be useful for probability estimates. An advantage of the approach is that the correlation structure among genes is preserved since the entire array is the sampling unit.

Tusher et al. [23] ranked genes by order of differential expression using a modified t-statistic where the denominator was inflated by a small constant to compensate for genes with very small variance. A selected threshold, Δ, then determined genes that were differentially expressed. Uncertainty was measured by an estimated false discovery rate (FDR). This was computed by permuting arrays across treatment conditions, computing the "t-statistic," ordering the genes by these statistics, and counting how many genes appeared above (or below) the threshold. Since the permutation across treatment conditions mimics the situation of "no treatment effect," this number of genes was an estimate of a number falsely declared. This number was recorded for all permutations and then averaged across permutations. This average divided by the number originally declared differentially expressed is an estimated FDR, analogous to 1 − TP in Equation 5.1.

Kerr et al. [24] used an ANOVA model that may include array, dye, treatment, and gene effects to model the gene expression values, for example, Y_{ij} in Figure 5.2. They used a residual bootstrap technique [25] to simulate a reference distribution for an F-statistic and to obtain standard errors of contrasts, but they did not directly deal with the proportions in Equation 5.1. Wolfinger et al. [14] employed mixed models and some parametric assumptions to model gene expression measurements, and they did suggest a method for power analysis. The method involved specification of an exemplary data set, variance components, fitting the model to the exemplary data holding variance components at their specified values, computing standard errors of specified contrasts, and computing power using a noncentral t-distribution and a specified false positive rate.

Efron and Tibshirani [26] modeled the test statistics arising from a Wilcoxon test. The model was of the form, $f_T(t) = p_0 f_0(t) + p_1 f_1(t)$ where $f_i(t)$ is the distribution of the test statistics under the null hypothesis ($i = 0$ meaning no differential expression), or under the alternative ($i = 1$ meaning there is differential expression). The $p_i, i = 0, 1$, are prior probabilities of no differential expression or differential expression, respectively. A use of Bayes theorem resulted in posterior probabilities such as the probability that a gene is differentially expressed given the test statistic, analogous to TP in Equation 5.1. They fitted a model using empirical Bayes methods. They also provided comparisons between their method and that of Benjamini and Hochberg's false discovery rate [15].

Lee and Whitmore [13] considered a table like Table 5.1, and investigated sample size requirements on types I and II error probabilities. "Power" was equal to $1 - P$ (type II error), which is analogous to our quantity EDR in Equation 5.1. They defined FDR as an expected proportion of falsely rejected null hypothesis, analogous to 1 – TP. Their defined type I error was $C/(A + C)$, where A and C are cell entries in Table 5.1. They also presented a Bayesian perspective on power and sample size using mixture models fitted to summary statistics of differential expression, where the prior probability, p_1, was an "anticipated" proportion of truly differentially expressed genes. Their focus, however, was on evaluating required sample size and power for linear summaries of differential expression involving computation of a null variance, effect size, and specification of an expected number of false positives. They noted, in particular, that specification of a null variance is problematic since it requires knowledge of the inherent variability of the data in the planned study. They also extended their results to situations where there may be more than two treatment groups where interest is in determining differential expression among several treatment groups.

Pan et al. [6] used a t-type statistic to quantify differential expression. However, the threshold to declare significance was obtained by creating a reference distribution that was a mixture of normal distributions. This model, when fitted to a "pilot" data set, could then be used to assess the number of replicates required to achieve desired power at a given significance level. The fitted model was considered fixed, a type I error was specified, and power computed for any specified effect size, for example, standardized difference in mean expression levels between two groups.

Zein et al. [5] considered sample size effects on pairwise comparisons of different groups and discussed the role of both technical and biological variability. Actual data

sets were used to develop parameter specifications for simulated data sets. They used the term sensitivity as analogous to EDR in Equation 5.1, and specificity that is analogous to TP. They evaluated the effect of varying sample size on these two quantities for various simulated data sets and using different types of statistical tests for differential expression, for example, t-tests and a rank-based test.

Many other contributions have been made where quantities related to those in Equation 5.1 were computed to reflect uncertainty in conclusions, or to assess power and the role of sample size in ensuring power rises to acceptable levels. A slightly different approach was taken by Pavlidis et al. [27] who used several real data sets to assess the effect of replication on microarray experiments. Mukherjee et al. [28] estimated sample size requirements for a classification methodology. Van der Laan and Bryan [29] developed a technique that incorporates sampling variability into a cluster type analysis and provide results for sensitivity, proportion of false positives, and a sample size formula.

Not discussed thus far regarding Figure 5.2 are some results based on the distribution of p-values resulting from statistical tests on all genes. Results based on this distribution are valid assuming that a valid test was used to produce a p-value. A distribution of p-values can be modeled and this distribution can provide estimates of the proportions in Equation 5.1 as well as shed light on the answers to questions 1 to 5, posed in the introduction. We now expand on this idea and indicate that further details are available in Allison et al. [19], Gadbury et al. [20], and Gadbury et al. [11]. First, however, we highlight some history regarding the use of p-values as random variables.

5.4 ON THE DISTRIBUTION OF p-VALUES

An often overlooked characteristic of a p-value is that, since it is computed from the sample, it too is a random variable [30]. The earliest work on the stochastic properties of a p-value may have been by Dempster and Shatzoff [31]. Other work has subsequently appeared in Schervish [32] and Donahue [33]. A key result related to our work here is the well-known probability integral transform that states that a cumulative distribution function evaluated at a random variable is a uniform random variable. Applied to p-values, this states that a test statistic, under a null hypothesis of no differential expression will produce a uniformly distributed p-value on the interval (0,1) as long as the distribution of the test statistic is known. We will refer to this latter condition as using a test that produces a "valid p-value."

Schweder and Spjøtvoll [34] may have been the first to consider this in the context of multiple testing. They produced the "p-value plot" as a means to (visually) quantify the number (proportion) of false null hypotheses. This used the idea that if several null hypotheses were not true, then there should be a larger number of "small" p-values than would have been expected if all null hypotheses were true. Hung et al. [35] derived the exact distribution of p-values under the alternative hypothesis and under various distributional assumptions for the data. They showed that, for their specific cases, these distributions depended on the effect size (under the alternative) and sample size. Parker and Rothenberg [36] suggested modeling a distribution of

p-values using a mixture of a uniform distribution and one or more beta distributions, the beta distribution being chosen for its flexibility in modeling shapes on the unit interval.

Allison et al. [19] adopted this idea from Parker and Rothenberg [36] and developed a method for modeling the distribution of p-values from microarray experiments. An idea later echoed in Pounds and Morris [21], they used the model to estimate proportions in Equation 5.1. Yang et al. [37] also noted how a distribution (histogram) of p-values will have a "peak" near zero when many null hypotheses are not true, but they did not directly model this distribution. Next, we review the method from Allison et al. [19] and discuss a procedure by which sample size effects can be assessed based on the method. Details of the latter are in Gadbury et al. [11].

5.5 A MIXTURE MODEL FOR THE DISTRIBUTION OF p-VALUES

The following example is used to illustrate the mixture model method of Allison et al. [19]. Human rheumatoid arthritis synovial fibroblast cell line samples were stimulated with tumor necrosis factor-α where one group ($n = 3$) had the Nf-κB pathway taken out by a dominant negative transiently transfected vector and the other group ($n = 3$) had a control vector added.

Figure 5.3 shows a histogram of p-values obtained from two sample t-tests of a null hypothesis $H_o : \mu_{1j} = \mu_{2j}$ vs. a two tailed alternative where μ_{ij} is a population mean expression for gene j in treatment group i and where $j = 1, \ldots, k = 12{,}625$ genes. A p-value from any valid statistical test can be used such as a test of a contrast from an ANOVA model.

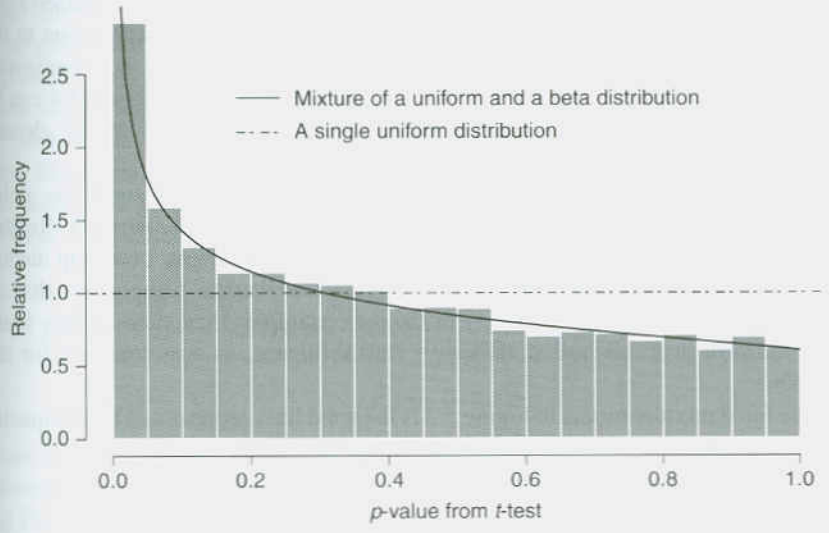

FIGURE 5.3 Distribution of continuous p-values obtained from two tailed t-tests on 12,625 genes from the rheumatoid arthritis data set, and the fitted mixture model is shown by the curve.

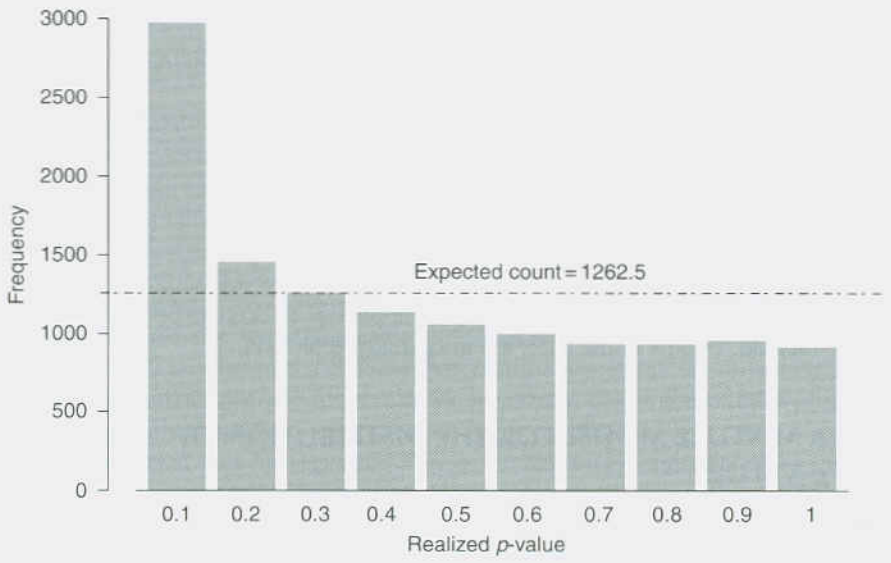

FIGURE 5.4 Distribution of discrete p-values obtained from two tailed randomization tests on 12,625 genes from the rheumatoid arthritis data set.

Gadbury et al. [20] described a procedure based on randomization tests for a difference in expression at each gene. The result is a discrete distribution of p-values with only 10 distinct values. The distribution shown in Figure 5.4 shows a similar shape as the continuous distribution if Figure 5.3. The number of p-values equal to 0.1 was 2975 vs. an expected 1262.5 if there were no differences in genetic expression for any genes. As a comparison, the number of genes with p-values obtained from the t-tests that were less than or equal to 0.1 was 2797, in close agreement to the randomization test results. Thus the randomization test can served as a "sensitivity check" for other testing procedures. The discrete distribution in Figure 5.4 can be modeled using a multinomial distribution but the information that can be gleaned from it is less rich than that available in a continuous model.

Returning to Figure 5.3, many more p-values than expected under the global null hypothesis cluster near zero. A mixture of a uniform plus one beta distribution captures this shape. Allison et al. [19] describe a parametric bootstrap method that can be used to estimate the number of beta distribution components that are required to model the shape. In many of the data examples they studied, they found that a uniform and one beta distribution was sufficient, as was the case for this example.

The fitted mixture model in Figure 5.3 is the solid line, represented by an equation of the form

$$f^*(p) = \prod_{i=1}^{k} [\lambda_0 + \lambda_1 \beta(p_i; r, s)], \quad p_i \in (0,1), \ i = 1, \ldots, k = 12{,}625 \qquad (5.2)$$

where $\beta(p; r, s)$ is a beta distribution with shape parameters r and s, and $\lambda_1 = 1-\lambda_0$. Parameters of the model were estimated using maximum likelihood. For these data, λ_1 is estimated as 0.395, suggesting that 39.5% of the genes are differentially expressed — an unusually strong signal and not representative of the many microarray data sets we have analyzed. The estimates for r and s are 0.539 and 1.844, respectively. The p-values were obtained from t-tests using pooled degrees of freedom. Using a Welsh correction on the t-test, for comparison, the estimated λ_1, r, s are 0.384, 0.686, and 1.944, respectively. The value of the maximum log-likelihood was 1484 vs. zero, which would be the value for a strictly uniform distribution (i.e., a distribution with no signal). The log-likelihood thus serves as a measure of relative model fit.

Calling the expression in Equation 5.2 a likelihood presupposes independence of p-values, but Allison et al. [19] also used simulations to examine the extent to which the log-likelihood would be affected by moderate dependence among genes. Gadbury et al. [20] did the same for the resulting distribution of discrete p-values obtained from randomization tests. It is challenging to model dependence among gene expression levels due to the small sample sizes and very large number of genes. Allison et al. [19] and Gadbury et al. [20] used a multivariate normal distribution with a mean and variance structure similar to the data and the correlation structure implemented through a block diagonal equicorrelation matrix with a parameter ρ occupying the off-diagonal entries of the blocks. The expressions levels were assumed independent across different blocks. The correlation ρ varied from 0.0 to 0.8. Moderate correlation was considered to be values of ρ around 0.4 with stronger dependence at 0.8. Negative correlation was not feasible with this approach since the correlation matrix was not positive definite for large negative values of ρ. In simulations where no genes were differentially expressed, the variance of the sampling distribution of the maximum log-likelihood increased with ρ. This suggests that for data fitted with the mixture model where the value of the log-likelihood is not large, some caution must be exercised since the value could be attributed to some genes being differentially expressed, or no genes differentially expressed but correlated instead. The effects of correlated expression levels on results from statistical methods for microarray data have been given limited attention in the literature and is a subject of continuing investigation.

Immediately available from the fitted mixture model, Equation 5.2, are maximum likelihood estimates of TP, TN, and EDR. Suppose a threshold τ is selected that determines genes that are declared differentially expressed (p-value $\leq \tau$) or not (p-value $> \tau$). Then at this threshold,

$$\widehat{\text{TP}} = \frac{\hat{\lambda}_1 B(\tau; \hat{r}, \hat{s})}{\hat{\lambda}_0 \tau + \hat{\lambda}_1 B(\tau; \hat{r}, \hat{s})}, \quad \widehat{\text{TN}} = \frac{\hat{\lambda}_0 (1-\tau)}{\hat{\lambda}_0 (1-\tau) + \hat{\lambda}_1 [1 - B(\tau; \hat{r}, \hat{s})]}, \quad \widehat{\text{EDR}} = B(\tau; \hat{r}, \hat{s}) \tag{5.3}$$

where $B(\tau; \hat{r}, \hat{s})$ is the cumulative distribution function of a beta distribution with estimated shape parameters, evaluated at τ. Choosing $\tau = 0.05$ resulted in $\widehat{\text{TP}} = 0.790$ and $\widehat{\text{TN}} = 0.671$, and $\widehat{\text{EDR}} = 0.287$. The high estimate of TP results from the strong signal present in this cell line dataset. The low estimate of EDR is attributed to small sample size (3 per group). The bootstrap can be used to estimate standard errors and

confidence intervals for TP, TN, and EDR. Sampling variability, in this context, is variability associated with a realization of k p-values from a model of the form given by Equation 5.2. The p-values from the t-tests were resampled 1000 times, each time fitting a mixture model to the bootstrap sample. An approximate 95% confidence interval for TP is (0.768, 0.812), for TN it is (0.608, 0.734), and for EDR it is (0.254, 0.320). As a comparison, the estimate of TP using the method of Tusher et al. [23] on these data was 0.77 using a threshold of $\Delta = 1.2$. In the next section we review a method from Gadbury et al. [11] that evaluates, for a given model fitted to actual data, the effect of threshold selection and sample size on estimated TP, TN, and EDR.

5.6 PLANNING FUTURE EXPERIMENTS: THE ROLE OF SAMPLE SIZE ON TP, TN, AND EDR

Gadbury et al. [11] used a computational procedure to consider the effect of threshold and sample size on TP, TN, and EDR. The procedure assumes that an experiment has been conducted with $N = 2n$ units divided into two groups of equal size and a mixture model $f^*(p)$ (Equation 5.2) fitted to the distribution of p-values obtained from a t-test of differential expression on each gene. A p-value from any valid test could be used as long as it can be back-transformed to the test statistic that produced it. Equal sample sizes in each group is convenient but is not required. The model is fitted using maximum likelihood and the estimated parameters are now considered fixed and equal to the true values, conceptually similar to Pan et al. [6].

A random sample $p^* = p_1^*, \ldots, p_k^*$ is generated from the mixture model $f^*(p)$, with the parameters estimated from the preliminary sample. The outcome of a Bernoulli trial first determines whether a p_i^* is generated from the uniform component with probability λ_0, or the beta distribution component with probability $\lambda_1 = 1 - \lambda_0$. From this sample of p-values, a set of adjusted p-values, $p^{**} = p_1^{**}, \ldots, p_k^{**}$, is created by transforming the p_i^* that were generated from the beta distribution to the corresponding t-statistic t_i^* and computing a new p-value, p_i^{**}, using a new sample size, n^*. The p_i^* generated from the uniform distribution are left unchanged.

From the new p^{**}, estimates of TP, TN, and EDR can be computed. To illustrate this, let $Z = \{1, 2, \ldots, k\}$ be a set of indices corresponding to the genes in the study, and let T be a subset of Z representing the set of genes that have a true differential expression across two experimental groups, that is, $T \subseteq Z$. Let

$$I_{[T]}(i) = \begin{cases} 1, & i \in T \\ 0, & i \notin T \end{cases} \quad \text{for } i = 1, \ldots, k$$

then $\sum_{i=1}^{k} I_{[T]}(i)$ represents the number of genes under study that are truly differentially expressed, unknown in practice but known and calculable in computer simulations.

A gene is declared to be differentially expressed if the p-value (calculated on observed data) from a statistical test falls below a predetermined threshold (τ). The

resulting decision function, when equal to 1, declares a gene differentially expressed:

$$\psi_i(x_i) = \begin{cases} 1, & p_i \leq \tau \\ 0, & p_i > \tau \end{cases}$$

where x_i is a vector of length N representing the data for the ith gene, $i = 1, \ldots, k$, hereafter abbreviated as ψ_i.

Estimates for the values in Table 5.1 that can be calculated in computer simulation experiments are given by

$$\begin{aligned} \hat{A} &= \sum_{i=1}^{k} (1 - \psi_i)[1 - I_{\{T\}}(i)], \quad & \hat{B} &= \sum_{i=1}^{k} (1 - \psi_i) I_{\{T\}}(i) \\ \hat{C} &= \sum_{i=1}^{k} \psi_i [1 - I_{\{T\}}(i)], & \hat{D} &= \sum_{i=1}^{k} \psi_i I_{\{T\}}(i) \end{aligned} \tag{5.4}$$

The values A, B, C, and D in Table 5.1 are defined using the expectations of the estimates in Equations 5.4, which are taken with respect to the fitted mixture model. These are,

$$E(\hat{A}) = A = k\lambda_0(1 - \tau), \quad E(\hat{B}) = B = k\lambda_1(1 - \mathrm{B}(\tau; r, s))$$
$$E(\hat{C}) = C = k\lambda_0\tau, E(\hat{D}) = D = k\lambda_1 B(\tau; r, s)$$

It can be seen that

$$\mathrm{TP} = \frac{D}{C + D}, \quad TN = \frac{A}{A + B}, \quad \mathrm{EDR} = \frac{D}{B + D}$$

defined in Equation 5.1, have the same form as $\widehat{\mathrm{TP}}, \widehat{\mathrm{TN}}$, and $\widehat{\mathrm{EDR}}$ in Equation 5.3, if estimated model parameters are taken as fixed. In simulations, Gadbury et al. [11] proposed estimating TP, TN, and EDR using

$$\widehat{\mathrm{TP}} = \frac{\hat{D}}{\hat{C} + \hat{D}}, \quad \widehat{\mathrm{TN}} = \frac{\hat{A}}{\hat{A} + \hat{B}}, \quad \widehat{\mathrm{EDR}} = \frac{\hat{D}}{\hat{B} + \hat{D}} \tag{5.5}$$

where the quantities in Equation 5.4 are readily available in the simulations.

The process is repeated M times thus obtaining M values of $\widehat{\mathrm{TP}}, \widehat{\mathrm{TN}}$, and $\widehat{\mathrm{EDR}}$ given in Equation 5.5. The value of M is chosen sufficiently large so that Monte Carlo estimates of $E[\widehat{\mathrm{TP}}], E[\widehat{\mathrm{TN}}]$, and $E[\widehat{\mathrm{EDR}}]$ can be accurately estimated using the average over the M values of $\widehat{\mathrm{TP}}, \widehat{\mathrm{TN}}$, and $\widehat{\mathrm{EDR}}$. This expectation is with respect to the simulation process. Since the model has been fixed, sampling variability in the estimates in Equation 5.5 is due to simulation uncertainty rather than model uncertainty. This is analogous to traditional power calculations where a desired effect size is fixed, a sample size is fixed, and then power computed at that effect size and sample size using some statistical model or distribution. Thus, standard errors in

the averages of the M values of $\widehat{\text{TP}}$, $\widehat{\text{TN}}$, and $\widehat{\text{EDR}}$ are generally small. The above described process is repeated for different values of n^* and τ.

5.7 SAMPLE SIZE AND THRESHOLD SELECTION: ILLUSTRATING THE PROCEDURE

Effects of sample size and threshold selection are evaluated using the above procedure on the rheumatoid arthritis cell line data set described earlier. Results are shown in Figure 5.5. The graph (a) in Figure 5.5 shows the minimum and maximum number (from $M = 100$ simulations) of 12,625 genes that were determined to be differentially expressed at three chosen thresholds for different sample sizes. The graph labeled (b) plots the average of 100 $\widehat{\text{TP}}$ values for the three thresholds at each sample size. Graphs (c) and (d) show the average of the 100 $\widehat{\text{TN}}$ and 100 $\widehat{\text{EDR}}$ values, respectively.

The number declared significant (graph a) was plotted since it reveals key information about TP. At very small sample sizes and thresholds, very few (and sometimes zero) genes are declared significant. This quantity estimates $C + D$ in Table 5.1, the denominator of TP. TP is defined to be zero when $C + D$ is zero; estimates, $\widehat{\text{TP}}$, are not expected to be very accurate when $\widehat{\text{C}} + \widehat{\text{D}}$ is a small positive number. This effect is seen in the plots for TP at small sample sizes. The TP plot also shows that the lines representing different thresholds cross over each other. Values of TP will be higher at lower thresholds as long as the sample size is large enough to detect differentially expressed genes.

Estimates of the quantities $A+B$ and $B+D$ (i.e., the denominators of TN and EDR, respectively) are more accurate at small sample sizes and small thresholds because A and B are usually large. However, estimates of EDR are small at these n and τ because D is small. So lines do not cross over in plots for EDR because a smaller threshold makes it more difficult to detect differentially expressed genes regardless of sample size. In the actual data set, $n = 3$, and one can see that the estimated EDR is quite small. One can also see from the procedure and resulting graph that EDR values rise to more acceptable levels as sample sizes approach around 20 arrays per treatment group.

5.8 DISCUSSION

The introduction posed five questions that are of interest in high-dimensional studies such as using microarrays. This chapter reviewed several approaches taken by others and provided some details of a technique based on mixture models and the use of simulated experiments from a given model. Answers to questions 2 to 4 are available using this technique. As far as questions 1 and 5 regarding threshold and sample size, the answer may very well be "it depends." We saw that there are trade-offs for choice of threshold. A very small threshold may miss many important genes (low EDR) but of the genes that are declared differentially expressed, one can be fairly certain that they are real (high TP). The described procedure relies on an initial study where a mixture model was fit to data and this model then becomes a standard for evaluating sample size effects on hypothetical future studies.

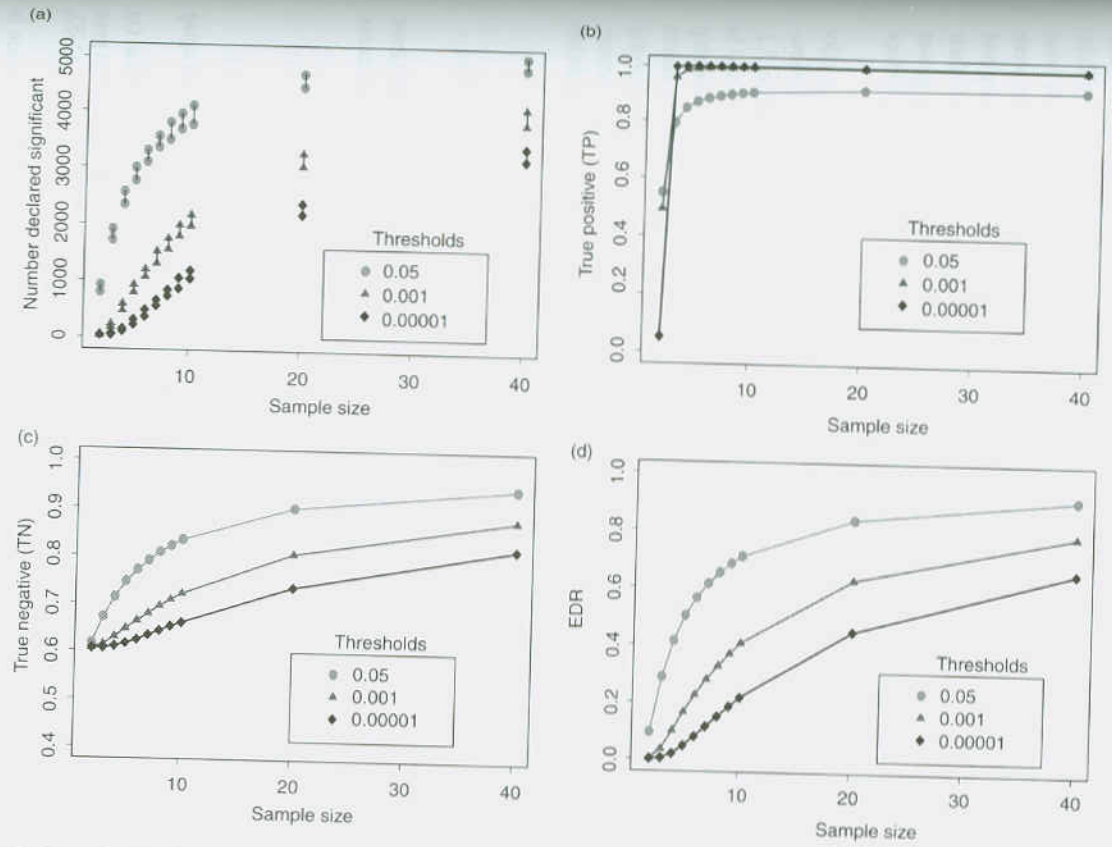

FIGURE 5.5 Effect of sample size (n = number of arrays per treatment group) on number declared significant (a), true positive (b), true negative (c), and EDR (d) for three selected thresholds, $\tau = 0.05, 0.001$, and 0.00001.

An interesting twist to this idea, mentioned earlier, was given by Pavlidis et al. [27]. They identified several publicly available microarray data sets with varying sample sizes (6 arrays per group to 50 per group). Using the larger studies, they could sample subsets of arrays thus simulating a smaller experiment and evaluating the stability of results using small samples. They concluded that results become somewhat unstable with 5 or fewer replicates but that 10 to 15 replicates often provides reasonable stability, though the numbers are data dependent. In the simulations of Zein et al. [5], they noted that it was not possible to simultaneously constrain both false positive and false negative rates to reasonably low values when sample sizes were only 8 per group. Pepe et al. [22], as indicated earlier, had a rather large study and they assessed "power" by subsampling from the arrays in the actual study, but their smallest sample size was 15 arrays per group. Wolfinger et al. [14] noted that power can be very low even with replication but that the appropriate design can reduce variance of estimates and increase power substantially.

Several of the methods discussed throughout indicate the usefulness of a pilot data set to obtain effect sizes, variance estimates, and model estimates to plan follow-up studies by calculating estimates of quantities similar to those in Equation 5.1. As more microarray experiments are completed and as future funds are allocated to allow larger experiments, more knowledge will become available on the relationship between sample size and key criteria attesting to the importance of results as measured by quantities such as TP and TN. Traditional notions of significance level and power may be less than optimal for high dimensional studies. The approach outlined herein, as well as approaches proffered by others, provide alternative tools to the researcher to help plan follow-on investigations using microarrays.

ACKNOWLEDGMENTS

This research supported in part by NSF Grants 0090286 and 0217651, and NIH Grant U54CA100949-01. The authors are grateful to Professor John D. Mountz for access to the data used in this chapter.

REFERENCES

1. C. Lee, R.G. Klopp, R. Weindruch, and T.A. Prolla. Gene expression profile of aging and its retardation by caloric restriction. *Science* 285: 1390–1393, 1999.
2. M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 8: 37–52, 2001.
3. M.L.T. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences, USA* 97: 9834–9839, 2000.
4. M.A. Black, and R.W. Doerge. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* 18: 1609–1616, 2002.

5. A. Zien, J. Fluck, R. Zimmer, and T. Lengauer. Microarrays: how many do you need? *Journal of Computational Biology* 10: 653–667, 2003.
6. W. Pan, J. Lin, and C.T. Le. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* 3: 1–10, 2002.
7. R.M. Simon, and K. Dobbin. Experimental design of DNA microarray experiments. *Bio Techniques* 34: 16–21, 2003.
8. M.K. Kerr, C.A. Afshari, L. Bennett, P. Bushel, J. Martinez, N.J. Walker, and G.A. Churchill. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* 12: 203–217, 2002.
9. T. Ideker, V. Thorsson, A.F. Siegel, and L.E. Hood. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* 7: 805–817, 2000.
10. S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12: 111–139, 2002.
11. G.L. Gadbury, G.P. Page, J. Edwards, T. Kayo, T.A. Prolla, R. Weindruch, P.A. Permana, J. Mountz, and D.B. Allison. Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research* 13: 325–338, 2004.
12. Y.H. Yang, and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews in Genetics* 3: 579–588, 2002.
13. M.L.T. Lee, and G.A. Whitmore. Power and sample size for DNA microarray studies. *Statistics in Medicine* 21: 3543–3570, 2002.
14. R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 8: 625–637, 2001.
15. Y. Benjamini, and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society B* 57: 289–300, 1995.
16. H.J. Keselman, R. Cribble, and B. Holland. Controlling the rate of type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology* 55: 27–39, 2002.
17. J.D. Storey. A direct approach to false discovery rates. *Journal of Royal Statistical Society B* 64: 479–498, 2002.
18. J.P. Brody, B.A. Williams, B.J. Wold, and S.R. Quake. Significance and statistical errors in the analysis of DNA microarray data. *Proceedings of the National Academy of Sciences, USA* 99: 12975–12978, 2002.
19. D.B. Allison, G.L. Gadbury, M. Heo, J.R. Fernanfez, C. Lee, T.A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* 39: 1–20, 2002.
20. G.L. Gadbury, G.P. Page, M. Heo, J.D. Mountz, and D.B. Allison. Randomization tests for small samples: an application for genetic expression data. *Journal of Royal Statistical Society C* 52: 365–376, 2003.
21. S. Pounds, and S.W. Morris. Estimating the occurrence of false positive and false negative in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics* 19: 1236–1242, 2003.
22. M.S. Pepe, G. Longton, G.L. Anderson, and M. Schummer. Selecting differentially expressed genes from microarray experiments. *Biometrics* 59: 133–142, 2003.

23. V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA* 98: 5116–5121, 2001.
24. M.K. Kerr, M. Martin, and G.A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 7: 819–837, 2000.
25. B. Efron, and R.J. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.
26. B. Efron, and R.J. Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 23: 70–86, 2002.
27. P. Pavlidis, Q. Li, and W.S. Noble. The effect of replication on gene expression microarray experiments. *Bioinformatics* 19: 1620–1627, 2003.
28. S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub, and J.P. Mesirov. Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* 10: 119–142, 2003.
29. M.J. Van der Laan, and J. Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics* 2: 445–461, 2001.
30. H. Sackrowitz, and E. Samuel-Cahn. *P* values as random variables — expected *p* values. *The American Statistician* 53: 326–331, 1999.
31. A.P. Dempster, and M. Schatzoff. Expected significance level as a sensibility index for test statistics. *Journal of American Statistical Association* 60: 420–436, 1965.
32. M.J. Schervish. *P* values: what they are and what they are not. *The American Statistician* 50: 203–206, 1996.
33. R.M.J. Donahue. A note on information seldom reported via *p* values. *The American Statistician* 53: 303–306, 1999.
34. T. Schweder and E. Spjøtvoll. Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* 69: 493–502, 1982.
35. H.M. Hung, R.T. O'Neill, P. Bauser, and K. Köhne. The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics* 53: 11–22, 1997.
36. R.A. Paker, and R.B. Rothenberg. Identifying important results from multiple statistical tests. *Statistics in Medicine* 17: 1031–1043, 1988.
37. Y. Yang, J. Hoh, C. Broger, M. Neeb, J. Edington, K. Lindpaintner, and J. Ott. Statistical methods for analyzing microarray feature data with replications. *Journal of Computational Biology* 10: 157–169, 2003.