Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays

Peter Müller, Giovanni Parmigiani, Christian Robert, and Judith Rousseau

We consider the choice of an optimal sample size for multiple-comparison problems. The motivating application is the choice of the number of microarray experiments to be carried out when learning about differential gene expression. However, the approach is valid in any application that involves multiple comparisons in a large number of hypothesis tests. We discuss two decision problems in the context of this setup: the sample size selection and the decision about the multiple comparisons. We adopt a decision-theoretic approach, using loss functions that combine the competing goals of discovering as many differentially expressed genes as possible, while keeping the number of false discoveries manageable. For consistency, we use the same loss function for both decisions. The decision rule that emerges for the multiple-comparison problem takes the form of the rules proposed in the recent literature to control the posterior expected false-discovery rate. For the sample size selection, we combine the expected utility argument with an additional sensitivity analysis, reporting the conditional expected utilities and conditioning on assumed levels of the true differential expression. We recognize the resulting diagnostic as a form of statistical power facilitating interpretation and communication. As a sampling model for observed gene expression densities across genes and arrays, we use a variation of a hierarchical gamma/gamma model. But the discussion of the decision problem is independent of the chosen probability model. The approach is valid for any model that includes positive prior probabilities for the null hypotheses in the multiple comparisons and that allows for efficient marginal and posterior simulation, possibly by dependent Markov chain Monte Carlo simulation.

KEY WORDS: False-discovery rate; Genomic data analysis; Multiple comparison.

1. INTRODUCTION

We consider the problem of sample size selection for experiments involving massive multiple comparisons. Approaching the sample size question as a decision problem, we argue that a solution needs to address the choice of sample size as part of a larger decision problem, involving both the sample size decision before carrying out the experiment and the later decision about the multiple comparisons once the data have been collected. We consider loss functions that combine the competing goals of controlling false-positive decisions and controlling false-negative decisions. For various reasonable loss functions, we show that the form of the terminal decision is the same: Reject all comparisons with marginal posterior probability beyond a certain threshold. We prove a formal result about the slow rate of change of the expected utility over a range of practically relevant sample sizes; this suggests that the preposterior expected utility alone may not suffice for a decisive sample size recommendation. Motivated by this, we conclude by recommending appropriate summaries of sensitivity of the expected utility with respect to the parameters of interest. The discussion includes specific algorithms to evaluate the proposed diagnostics. With a view toward the motivating application, we propose a non-parametric probability model that allows the use of pilot data to learn about the relevant sampling distribution for the sample size decision. Formally, this amounts to using the posterior predictive distribution from the pilot data as the prior model used in the sample size calculation.

Our discussion is motivated by the specific problem of choosing the number of replications in microarray experiments. Gene expression microarrays are technologies for simultaneously detecting differentially expressed genes in an organism (Schena, Shalon, Davis, and Brown 1995; Duggan, Bittner, Chen, Meltzer, and Trent 1999). (For a recent review of microarray technology and related statistical methods see, e.g., Kohane, Kho, and Butte 2002.) The range of applications is broad. Here we focus on controlled experiments that aim to search or screen for genes whose expressions are regulated by modifying the conditions of interest, either environmentally or genetically. A number of pressing biological questions can be addressed using this type of genomic screening. Because microarrays are costly, the design of the experiment and choice of sample size result in a difficult trade-off between the allocation of limited research resources and statistical learning. Our approach is applicable to the process of selecting the number of biological replicates (such as the number of mice to be assigned to each group), as well as the selection of the number of technological replicates (such as the number of separate aliquots of RNA to be extracted from two biological samples being compared). In our theoretical discussion, we use the term “replicate” to refer to either type. Each situation requires a different interpretation of the array-to-array variability, as well as different priors or different pilot samples. Our illustration specifically concerns biological replicates. General issues of experimental design in microarray experiments, beyond the sample size selection considered in this article, have been discussed by Kerr and Churchill (2001), Yang and Speed (2002), and Simon, Radmacher, and Dobbin (2002).

The general goal of the genomic screening is to discover as many of the genes that are differentially expressed across the experimental conditions as possible, while keeping the number of false discoveries at a manageable level. The consequences of a false discovery are often similar across genes. The decision-making process when selecting a sample size can benefit from explicitly acknowledging these experimental goals by following a formal decision-theoretic approach. Sample size selection using decision theoretics, including the multistage nature central to our discussion, was formalized within a Bayesian framework as early as 1961 through the work of Raiffa and Schlaifer.
Following this paradigm, we present a general decision-theoretic framework for the choice of sample size for genomic screening or for use in a similar selection problem. Central to our analysis is the concept of the false-discovery rate (FDR), introduced by Benjamini and Hochberg (1995). In controlled experiments, it is plausible to assume that genes can be divided into two groups, truly altered genes and truly unaltered genes. For a given approach to selecting a set of putatively altered genes, the FDR is the fraction of truly unaltered genes among the genes classified as differentially expressed. Common microarray software uses the FDR to guide gene selection (see, e.g., Tusher, Tibshirani, and Chu 2001). Applications of FDRs to microarray analysis have been discussed by Storey and Tibshirani (2003), and extensions have been given by Genovese and Wasserman (2003), who also introduced the definition of the posterior expected FDR as we use it here. We show that the decision-theoretic approach leads to a multiple-comparison decision of the form described by Genovese and Wasserman (2003). They focused on decision rules of the following kind. Assume that for each comparison, some univariate statistic $v_i$ is available. This could be, for example, a $p$ value or any other univariate statistic related to the comparison of interest. All comparisons with $v_i$ beyond a certain cutoff $t$ are considered discoveries. Central to their approach is the use of an upper bound on the FDR to calibrate that cutoff $t$.

In the context of microarray experiments, important initial progress toward the evaluation of sample sizes has been made by Pan, Lin, and Le (2002), who developed traditional power analyses for use in the context of microarray experiments. Their modeling is realistic in that it acknowledges heterogeneity in gene-specific noise and specifies a mixture model for regulated and unregulated genes. Further progress is needed, however. Pan et al. (2002) did not exploit heterogeneity in developing screening statistics, as is done by hierarchical models. This can potentially underestimate the power, especially in the critical range of experiments with very few replicates. Also, their power analysis considers a single effect size for all regulated genes. Finally, explicit consideration of properties of the entire selection, such as FDR, is preferable in the context of multiple testing. Zien, Fluck, Zimmer, and Lengauer (2002) proposed an alternative approach to an informed sample size choice. They considered receiver operating characteristic-type curves and showed achievable combinations of false-negative and false-positive rates. Mukherjee et al. (2003) discussed sample size considerations for classification of microarray data. They assumed a parametric learning curve for empirical error as a function of the sample size. Their approach is based on estimating the parameters in that learning curve. Lee and Whitmore (2002) considered an ANOVA setup including, among other parameters, interaction effects for the gene and biological conditions. Hypothesis testing for these interactions formalizes the desired inference about differential expression. They assumed approximate normality for an estimator of these interaction effects and proceeded with a conventional power analysis. Bickel (2003) proposed a framework for inference on differential gene expression that includes a loss function consisting of a payoff for correct discoveries and a cost for false discoveries. (See Sec. 2.1 for a definition of these events.) The net desirability function defined by Bickel (2003) is equivalent to one of the loss functions introduced in Section 2.1.

As in many traditional sample size problems, the practical use of the proposed approach will be as a decision support tool. We do not expect investigators to blindly trust the proposed solution. Rather, we envision that an investigator may be operating under budget and resource constraints that allow for a narrow range of sample size choices. The proposed methods can guide the choice within that range by informing the investigator about the likely payoffs and decision summaries.

In Section 2 we outline the decision problem and our approach to the solution in a general form, without referring to a specific probability model. In Section 3 we develop an efficient simulation approach for evaluating the required sample size selection criteria. We define a Monte Carlo simulation method that allows us to evaluate the expected false-negative rate (FNR) and power across the sample sizes. We demonstrate that because of their preposterior nature, the required simulations are easier and less computationally intensive than posterior simulation in the underlying probability model. In Section 4 we introduce a specific probability model that we then use in Section 5 to show results in an example. In Section 6 we conclude with a final discussion.

2. THE DECISION PROBLEMS

To highlight the general nature of the proposed approach, we first proceed without reference to a specific probability model or comparison of interest. We let $\omega$ and $y$ denote the model parameters and expression measurements, and let $z_i \in \{0, 1\}$ denote an indicator for the regulation of gene $i$. Regulation is broadly defined to include any of the typical questions of interest, such as differential expression across two conditions, time trends, sensitivity to at least one out of a panel of compounds, and so forth. We assume that the probability model includes indicators $z_i$ as parameters or easily imputed latent variables. We write $y_j$ when we want to highlight that the data $y$ are a function of the sample size $J$. We assume that the underlying probability model allows for efficient posterior simulation. Let $v_i = P(z_i = 1|y)$ denote the marginal posterior probability for the $i$th comparison. Computation of $v_i$ could involve some analytical approximations, such as empirical Bayes estimates for hyperparameters. In Section 4 we introduce the probability model used in our implementation and discuss posterior inference in that model.

An important aspect of the problem is that the earlier decision about the sample size needs to take into account the later decision about gene selection. This will be either a selection (also referred to as discovery, or rejection, and denoted by $d_i = 1$ for comparison $i$) or not (also referred to as a negative and denoted by $d_i = 0$). Decision-theoretic approaches to sample size selection assume that the investigator is a rational decision maker choosing an action that minimizes the loss of the possible consequences—averaging with respect to all of the relevant unknowns (Raiffa and Schlaifer 1961; DeGroot 1970). At the time of the sample size decision, the relevant unknowns are the data $y$, the indicators $z = (z_1, \ldots, z_n)$, and the model parameters $\omega$. The relevant probability model with respect to which we average is the prior probability on $(z, \omega)$ and the conditional
sampling distribution on \( y \) given \((z, \omega)\). At the time of the decision about multiple comparisons, the data \( y \) are known and the relevant probability model is the posterior distribution conditional on \( y \).

The solution proceeds in a traditional backward induction fashion by first considering the terminal multiple comparison decision of the gene selection. Knowing the optimal policy for the eventual gene selection, we can then approach the initial sample size problem. It is thus natural to first discuss inference about the multiple comparison decisions \( d_i, i = 1, \ldots, n \).

2.1 Terminal Decision

The choice of a decision rule for multiple comparisons is driven by the following considerations. First, the rule should have a coherent justification as the solution that minimizes the expected loss under a sensible loss function. Second, inference about the multiple comparison decision is nested within the sample size selection, making computational efficiency an important issue. In the type of experiments considered here, a relatively small number of genes are regulated, and the noise levels are relatively high. Finally, although our approach is based on joint probability models on data and parameters (i.e., in essence Bayesian), we are concerned about frequentist operating characteristics for the proposed rule. The use of frequentist properties to validate Bayesian inference is common practice in the context of medical decision making.

With these considerations in mind, we propose loss functions that have the following characteristics: they capture the typical goals of genomic screening; they are easy to evaluate; lead to simple decision rules; and can be interpreted as generalizations of frequentist error rates. We consider four alternative loss functions that amount to terminal decision rules of the same form. We start with a notation for various summaries that formalize the two competing goals of controlling the false-negative and false-positive decisions. Writing \( D = \sum d_i \) for the number of discoveries, we let

\[
FDR(d, z) = \frac{\sum d_i (1 - z_i)}{D + \epsilon} \quad \text{and} \quad FNR(d, z) = \frac{\sum (1 - d_i) z_i}{n - D + \epsilon}
\]

denote the realized FDR and FNR. FDR(\( \cdot \)) and FNR(\( \cdot \)) are the percentage of wrong decisions, relative to the number of discoveries and negatives, where the additional term \( \epsilon \) avoids a zero denominator. (See, e.g., Genovese and Wasserman 2003 for a discussion of FNR and FDR.) Conditioning on \( y \) and marginalizing with respect to \( z \), we obtain the posterior expected FDR and FNR,

\[
FDR(d, y) = \int FDR(d, z) \, dp(z | y)
= \frac{\sum d_i (1 - v_i)}{D + \epsilon}
\]

and

\[
FNR(d, y) = \int FNR(d, z) \, dp(z | y)
= \frac{\sum (1 - d_i) v_i}{n - D + \epsilon}.
\]

Let \( FD = \sum d_i (1 - v_i) \) and \( FN = \sum (1 - d_i) v_i \) denote the posterior expected count of false discoveries and false negatives. We consider four ways of combining the goals of minimizing false discoveries and false negatives. The first two specifications combine FNRs and FDRs and numbers, leading to the posterior expected losses

\[
L_N(d, y) = \epsilon FD + FN,
\]

and \( L_R(d, y) = \epsilon FDR + FNR \). The loss function \( L_N \) is a natural extension of \((0, 1, c)\) loss functions for traditional hypothesis testing problems (Lindley 1971). From this perspective, the combination of error rates in \( L_R \) seems less attractive. The loss for a false discovery and false negative depends on the total number of discoveries or negatives. Alternatively, we consider bivariate loss functions that explicitly acknowledge the two competing goals, leading to the posterior expected losses

\[
L_{2R}(d, y) = (FDR, FNR), \quad L_{2N}(d, y) = (FD, FN).
\]

Using posterior expectations, we marginalize with respect to the unknown parameters, leaving only \( d \) and \( y \) as the arguments of the loss function. The sample size is indirectly included in the dimension of the data vector \( y \). For the bivariate loss functions, we need an additional specification to define the combination of the bivariate functions. A traditional approach to select an action in multicriteria decision problems is to minimize one dimension of the loss function while enforcing a constraint on the other dimensions (Keeney, Raiffa, and Meyer 1976). We thus define the optimal decisions under \( L_{2N} \) as the minimization of \( FN \) subject to \( FD \leq \alpha_N \). Similarly, under \( L_{2R} \), we minimize \( FNR \) subject to \( FDR \leq \alpha_R \). Under all four loss functions, the optimal decision for the multiple comparisons takes the same form.

**Theorem 1.** Under all four loss functions, the optimal decision takes the form

\[ d_i = I(v_i \geq t^*). \]

The optimal cutoff \( t^* \) is \( t^*_N = c/(c + 1) \), \( t^*_R(y) = v_{(n-D^*)} \), \( t^*_N(y) = \min\{s: FD(s, y) \leq \alpha_N\} \), and \( t^*_R(y) = \min\{s: FDR(s, y) \leq \alpha_R\} \). In the expression for \( t^*_N \), \( v_{(i)} \) is the \( i \)th order statistic of \( \{v_1, \ldots, v_n\} \), and \( D^* \) is the optimal number of discoveries. See the proof in the Appendix for a constructive definition of \( D^* \).

The proof proceeds by straightforward algebra (see App. A for details). Under \( L_R, L_{2N}, \) and \( L_{2R} \), the optimal threshold \( t^* \) depends on the observed data. The nature of the terminal decision rule \( d_i \) is the same as that in the work of Genovese and Wasserman (2003), which is a more general rule allowing the decision to be determined by cutoffs on any univariate summary statistic \( v_i \). Using \( v_i = P(z_i = 1 | y) \) is a special case.

For simplicity, in what follows we focus on \( L = L_{2N} \) only, omitting the subscript \( 2N \) to simplify the notation. In Section 5.3 we revisit the other three loss functions. Also, by a slight abuse of the notation, we write \( d = t \) for the decision rule \( d_i = I(v_i \geq t) \). Finally, we note that not all loss functions lead to decisions \( d_i = I(v_i \geq t) \). For example, assuming a loss of a false negative that depends on the true level of differential ex-
pression would lead to different rules. One could argue that discovering a gene that shows a very small differential expression in a given experiment may not be as interesting as discovering one that shows a major change in its expression.

2.2 Sample Size

2.2.1 Marginal FN and FNR. In contrast to the terminal decision of the gene selection, which is made conditional on the observed data, the sample size is decided before conducting the experiment. Thus we now consider the marginal prior mean of the proposed loss functions, also known as the posterior expected loss (Raiffa and Schlaifer 1961), after substituting the optimal terminal decision for the multiple-comparison decision. The relevant loss function \( L^m(J) \) for the sample size selection is

\[
L^m(J) = E \left[ \min_d \{ L(d, y) \} \right] = E \left[ \min_d \{ \text{FN}(d, y_J) | FD \leq \alpha N \} \right] = E \left[ \text{FN}(t^*(y_J), y_J) \right].
\]

The conditioning bar in the nested optimization indicates that the minimization is subject to the bound on \( FD \). The sequence of alternating between the expectation and the optimization is characteristic for sequential decision problems. (See, e.g., DeGroot 1970 and Berger 1985 for a discussion of sequential decision problems in general.) The expectation is determined with respect to the prior probability model on the data \( y_J \) under a given sample size \( J \). The only argument left after determining the expectation and the minimization is the sample size \( J \). The nested minimization with respect to \( d \) is the solution of the multiple comparison problem. It reduces to

\[
d_i = I \{ v_i > t^*(y_J) \}.
\]

We denote the pre-posterior expected FN by \( \text{FN}_m(J) = E \left[ \text{FN}(t^*(y_J), y_J) \right] \), with the bound on \( FD \) being implicit in the definition of \( t^*(y_J) \). Thus we could alternatively write (2) as \( L^m(J) = \text{FN}_m(J) \). We use analogous definitions for \( \text{FNR}_m, \text{FDR}_m, \text{FD}_m \). The latter is equal to \( \alpha N \) by definition of \( t^*(y_J) \).

In Section 3 we introduce a simulation-based algorithm for a practical evaluation of the expectation and nested optimization in (2). Using this algorithm, one could evaluate and then plot the marginal expected utility (i.e., \( \text{FN}_m \)) against \( J \) to select a sample size. At this time one could add a (deterministic) sampling cost, if desired. But in practical application this would require the difficult choice of a relative weight for the sampling cost versus an inference loss. Alternatively, we take a goal-oriented perspective and use the plot of \( L^m(J) \) versus the sample size \( J \) to find a sample size for any set goal of \( L^m(J) \).

---

**Figure 1.** \( L_{2N}^N \): Expected Loss and Other Relevant Summaries. See Section 5 for a description of the Example. (a) Expected loss function \( L^m = \text{FN}_m (\text{FN in the plot}) \). (b)–(e) Expected false-negative rate \( \text{FNR}_m \) (FNR in the plot), false-discovery count \( \text{FD}_m \) (labeled FD), and rate \( \text{FDR}_m \) (labeled FDR), the number of discoveries \( D \) and the threshold \( t^* \) (TSTAR). In each panel, the dots show the values recorded in each of the simulations [in (d), some dots fall outside the range of the figure]. The false-discovery count \( \text{FD}_m \) is fixed by design, leading to an increasing number of discoveries \( D_m \). The dashed curves for \( \text{FNR}_m \) and \( \text{FN}_m \) (almost indistinguishable from the solid line) show an alternative curve fit. See the text for an explanation of the curve fit.
However, doing this entails a practical complication. For relevant sample sizes of \( J \leq 20 \), the decrease in \( L^m(J) \) is too flat to allow a conclusive choice of sample size; see Figure 1(a) for an example. The slow rate of decrease is a general feature of \( \text{FNR} \) and \( \text{FN} \).

**Theorem 2.** Consider the three loss functions \( L = L_{2N}, L_N, \) and \( L_{2R} \). The false-negative rate and counts of \( \text{FNR} \) and \( \text{FN} \) decrease asymptotically as

\[
\text{FNR}(t^*, y_J) = O_P(\sqrt{\log J/J}),
\]

where \( t^* \) generically indicates the optimal cutoff under each of the three loss functions, and

\[
\text{FN}(t^*, y_J) = O_P(n\sqrt{\log J/J}).
\]

For both results, we must assume that the genes are “randomly chosen,” that is, a fraction \( p, 0 < p < 1 \), of the genes are truly differentially expressed. In other words, we assume that the level of differential expression is neither equal to 0 (or very small) nor always different from 0. A formal argument is given in Appendix B. The argument starts with a Laplace approximation for \( v_i = P(z_i = 1|y_J) \). Based on this approximation, we show that only genes with a low or zero differential expression, are included in the negative set, that is, the set of genes with \( d_i = 0 \). We then approximate the average in \( \text{FNR} \) (or \( \text{FN} \)) by an integral, exploiting the fact that these are genes with small differential expressions. Finally, we recognize that the integral expression is on the order of \( \sqrt{\log J/J} \).

### 2.2.2 Conditional Preposterior Expected Utility.

The relatively flat nature of the expected utility \( L^m(J) \) does not allow for a conclusive sample size recommendation. A natural next step is to investigate the expected utility as a function of an assumed true value for some key parameters of the probability model. Specifically, we assume that the probability model includes a parameter \( \rho_i \) that represents the level of differential expression for gene \( i \), with \( \rho_i = 0 \) if \( z_i = 0 \) and \( \rho_i \neq 0 \) when \( z_i = 1 \). For example, in the probability model discussed in Section 4, we would use \( \rho = \log \theta_1/\theta_0 \). We thus proceed by considering the expected utility, conditional on an assumed true level of \( \rho_i \). Recall the definition of \( \text{FN} \) as the false-negative count. Conditioning on \( \rho_i \) only changes the term that is related to gene \( i \). For a large \( n \), conditioning \( \rho_i \) for one gene leads to only negligible changes in inference for other genes (a posteriori, as well as a priori). Finally, note that for \( \rho_i \neq 0 \) gene \( i \) contributes only to \( \text{FN} \), not to \( \text{FD} \). Thus we can characterize the conditional expected utility as a function of \( \rho_i \) by considering

\[
\beta(J, \rho_i) = P[v_i(y_J) > t^*|\rho_i] = \int I(v_i(y_J) > t^*) \, dp(y_J|\rho_i),
\]

writing \( v_i(y_J) \) to highlight the nature of the marginal posterior probability \( v_i \) as a function of the data. The expectation is determined with respect to the joint probability model on data \( y \). In particular, the expected utility appropriately adjusts for dependencies, uncertainties on other model parameters, and the entire process of finding and applying \( t^*(y_J) \). Assuming that the genes are a priori exchangeable, the marginal expectation is the same for all \( i \), allowing us to drop the \( i \) subindex.

The diagnostic \( \beta(J, \rho) \) has interesting interpretations. We define it as the term in the conditional expected utility that varies as a function of \( \rho_i \). Our main reason to propose it is due to its link with the traditional notion of power. The definition of \( \beta \) is essentially the power to test one hypothesis in the multiple-comparison decision, although with an added twist of marginalizing it with respect to all other parameters. To simplify the terminology, we refer to \( \beta(J, \rho) \) as “power,” with the understanding that the definition includes the mentioned marginalizations. Figure 2 shows a typical example.

Thus the following modification to the approach outlined in Section 2.2.1 emerges. The investigator fixes a minimum level

\[
\beta = E_{\theta_J}(P(d = 1|\rho, y_J)) \quad \text{by } J
\]

\[
\beta \quad \text{against } J \quad \text{by } \rho
\]

**Figure 2**. Power \( \beta \) Against True Effect \( \rho \) and Against Sample Size \( J \). See Section 5 for a description of the Example. In (b), power curves for two-fold (\( \rho = \log 2 \)) and four-fold over expression (\( \rho = \log 4 \)) are highlighted in bold. Power \( \beta(J, \rho) \) is defined in (3) as the average posterior probability of discovery, conditional on the true level of differential expression \( \rho_i = \log(\theta_1/\theta_0) \).
of differential expression that is of interest in the given experiment and the desired probability of discovering a gene that is differentially expressed at that level. Inspection of a power plot like Figure 2, together with $\text{FNR}_m$ and $\text{FD}_m$ in the marginal loss function, allows the investigator to obtain an informed sample size choice. The $\text{FD}_m$ and $\text{FNR}_m$ plots, along with plots of related summaries $\text{FDR}_m$ and $\text{FNR}_m$, add the experiment-wise dimension to the marginal, comparison-wise summary that is provided by the power plot. It tells the investigator how many false-negatives might be missed, averaging over the range of likely differential expression levels and summing over all genes. Computation of $\beta(J, \rho)$ is achieved within the same pre-posterior simulation that is used to evaluate $\text{FNR}_m$ and $\text{FD}_m$.

3. SIMULATION

The approach to sample size selection that we describe involves several calculations that are typically analytically intractable. Details depend on the specific probability model. Often the posterior mean probabilities $v_i$, the threshold $t^*(y_J)$, and the expected FNR are not available in closed form; however, all can be computed by Monte Carlo simulation. In this section we describe how such Monte Carlo simulation is implemented. Before giving a step-by-step algorithm, we introduce the notations and review the important steps in the algorithm in words. The discussion is still possible without reference to a particular probability model.

For a given sample size $J$, we simulate data $y_J \sim p(y_J)$. Simulating from the marginal $p(y_J) = \int p(y_J|\omega, z) dP(\omega, z)$ is conveniently implemented by first generating “true” parameters $\omega, z$ from the prior, and then generating from the assumed sampling model $p(y_J|\omega, z)$ given the simulated parameter. To distinguish this prior simulation from a posterior Markov chain Monte Carlo (MCMC) simulation that is required later in the algorithm, we mark the realizations of this prior simulation by a superindex as in $\omega^p$, and so on.

For each simulated dataset $y_J$, we compute the marginal posterior probabilities $v_i = p(z_i = 1|y_J)$ and evaluate $\text{FDR}(t, y_J)$ and $\text{FD}(t, y_J)$ on a grid over $t$ to find the optimal cutoff $t^*(y_J)$. Plugging in the optimal cutoff $t^*$ in $d_i = I(v_i > t)$, we evaluate the posterior means $\text{FNR}(t^*, y_J)$ and $\text{FNR}(t^*, y_J)$. Averaging over $y_J$ by (independent) Monte Carlo simulation, with repeated simulation of $y_J \sim p(y_J)$, we compute $L^m(J) = E_{y_J} \{ \text{FNR}(t^*, y_J) \}$. (4)

The nonlinear thresholding $d_i = I(v_i > t^*)$ implicit in the definition of $\text{FNR}$ hinders the interpretation of (4) as one joint integral with respect to the joint distribution $p(\omega, z)$ on parameters and data. Instead, we need to proceed with two nested steps, as described earlier. Finally, evaluating (4) across $J$, we find the sample size $J^*$, which allows us to achieve a desired marginal expected loss.

The information in the marginal loss $L^m(J)$ is supplemented by power curves $\beta(J, \rho)$. Power $\beta(J, \rho)$ as defined in (3) is a summary of the pre-posterior expected utility. It is evaluated as part of the same simulation described earlier to find $L^m$. For each simulated experiment, we record $(J, p^o_i, v_i, d_i)$, $i = 1, \ldots, n$. Here $p_i^o$ is the true simulated level of the differential expression. The recorded simulations are then arranged by $J$ to compute $\text{FNR}_m$ as described earlier. Arranging the same simulations by $J$ and $\rho$ (possibly on a grid), we estimate $\beta(J, \rho)$, which can be summarized in plots like those given in Figure 2.

Implementation is facilitated by several simplifications that increase the computational efficiency. First, we use common random numbers across $J$, in the following sense. We consider sample sizes on the interval $J_0 \leq J \leq J_1$. We start by generating one large sample $y_{J_1}$, and use appropriate subsamples $y_J \subset y_{J_1}$ to compute $\text{FNR}_m(J)$, $\text{FD}_m(J)$, $\text{FNR}_m(J)$, and $\text{FDR}_m(J)$ for $J$ over a grid $J_0 \leq J \leq J_1$. Using the common underlying data reduces sampling variation across $J$.

Another simplification arises in the setup of the posterior simulations required to evaluate the posterior expected $\text{FNR}(t, y_J)$ and $\text{FD}(t, y_J)$. Both require posterior simulation $\omega \sim p(\omega|y_J)$ by MCMC. In the context of the pre-posterior simulation, we can start the MCMC at the true parameter values $\omega^o$ used to simulate the data $y_J$. Details are explained in the step-by-step algorithm given later in this section.

Finally, when computing $L^m(J)$, we borrow strength across different sample sizes. Instead of averaging separately for each $J$ the computed values $L(t^*, y_J)$ for that $J$, we proceed as follows. Consider a scatterplot of all pairs $(J, L(t^*, y_J))$. We fit a smooth curve $\tilde{L}^m(J)$ through all points of the scatterplot. This formalizes the borrowing strength across different sample sizes $J$, exploiting the fact that $L^m(J)$ is smooth across $J$. In fact, we recommend enforcing the smooth fit $\tilde{L}^m$ to be monotone and decreasing and to follow the $(\log J/J)$ asymptotics. We used a least squares fit of a linear regression of the observed $\text{FNR}(t^*, y_J)$ values on $\sqrt{\log J/J}$. For comparison, we fit a smoothing spline without any such constraints. The spline fit is practically indistinguishable from the simple regression, validating use of the asymptotic law for the curve fitting (see Sec. 5).

Algorithm 1: Sample Size Determination.

1. Simulation: Loop over repeated simulations $y_{J_i} \sim p(y_{J_i})$.

1.1 Prior simulation: $(\omega^o, z^o) \sim p(\omega, z)$.

1.2 Data simulation: $y_{J_i} \sim p(y_{J_i}|\omega^o, z^o)$.

We simulate data for the largest sample size $J_1$ considered in the design.

1.3 Loop over $J$: loop over a grid of sample sizes $J = J_1, \ldots, J_0$.

Let $y_J \subset y_{J_i}$ denote the size $J$ subset of the maximal dataset.

1.3.1 Posterior simulation $\omega \sim p(\omega|y_J)$.

a. Initialize MCMC posterior simulation with $(\omega, z) = (\omega^o, z^o)$.

b. Simulate $S$ transitions of the posterior MCMC.

1.3.2 Posterior probabilities: Compute $u_i = P(z_i = 1|y_J)$ as the appropriate ergodic average and evaluate

$$\text{FD}(t, y_J) = \sum (v_i > t)(1 - v_i)$$

and

$$\text{FDR}(t, y_J) = \text{FD}(t, y_J)/(D + \epsilon)$$
for $t \in \{v_1, \ldots, v_J\}$ and find the optimal cutoff $t^*(y_j)$.

Record $(J, \mathcal{F}_D(t^*, y_j), \mathcal{F}_N(t^*, y_j), F_{DR}(t^*, y_j), F_{NR}(t^*, y_j))$.

1.3.3 Power: Let $d_i = I(v_i > t^*)$ and record the triples $(J, \rho_i^*, d_i)$.

2. Curve fitting of Monte Carlo experiments:

2.1 Preposterior expectations $L^m(J, \mathcal{F}_D, \mathcal{F}_N, \mathcal{F}_{DR}, \mathcal{F}_{NR}$, and $\mathcal{F}_{NR}$m$: For each of the last four quantities fit a curve through the observed pairs $(J, \mathcal{F}_N)$, and so on. Use the asymptotically shaped reported in Theorem 1 to guide the curve fitting.

2.2 Power vi: Use the triples $(J, \rho_i^*, d_i)$ computed in step 1 to estimate $\beta(J, \rho)$.

3. Optimal sample size: Use $\tilde{L}^m(J)$ and power curves as in Figure 2 to make an informed sample size choice.

4. THE PROBABILITY MODEL

Our approach to sample size selection assumes an encompassing probability model that specifies a joint distribution across comparisons and across repeated experiments. In general, the model should be sufficiently structured and detailed to reflect the prior expected levels of noise, a reasonable subjective judgment about the likely numbers of differentially expressed genes, and some assumption about dependencies, if relevant. It should also easily incorporate prior data when available.

The design argument can be developed with a simplified model, ignoring all details of the data-cleaning process, including the spatial dependence of measurement errors across the microarray, correction for misalignments, and so on. Although such detail is critical for the analysis of the observed microarray data, it is an unnecessary burden for the design stage. The variability resulting from preprocessing and normalization can be subsumed as an aggregate in the prior description of the expected noise. In what follows we thus assume that the data are appropriately standardized and normalized and that the noise distribution implicitly includes the consideration of those processes (see, e.g., Tseng, Oh, Rohlin, Liao, and Wong 2001; Baggerly et al. 2001; Yang et al. 2002, for a discussion of the process of normalization).

For the implementation in the example we choose a variation of the model introduced by Newton, Kendziorski, Richmond, Blattner, and Tsui (2001) and Newton and Kendziorski (2003). We focus on the comparison of two conditions and assume that data will be available as arrays of appropriately normalized intensity measurements $X_{ij}$ and $Y_{ij}$ for gene $i$, $i = 1, \ldots, n$, and experiment $j$, $j = 1, \ldots, J$, with $X$ and $Y$ denoting the intensities in the two conditions.

Newton et al. (2001) proposed a hierarchical gamma/gamma model. The model starts by assuming that the observed intensities are sampled from gamma distributions, with a conjugate gamma prior on the scale parameters. The model includes a positive prior probability mass for matching the means of the gamma distribution for the same gene under the two conditions of interest. For the purpose of the sample size design, we extend the model to multiple experiments, $j = 1, \ldots, J$. We assume a gamma sampling distribution for the observed intensities $X_{ij}$, $Y_{ij}$ for gene $i$ in sample $j$,

$$X_{ij} \sim \text{Gamma}(a, \theta_{ij}) \quad \text{and} \quad Y_{ij} \sim \text{Gamma}(a, \theta_{ij}).$$

The scale parameters are gene-specific random effects $(\theta_{ij}, \theta_{ij})$, with positive prior probability for a tie,

$$\text{Pr}(\theta_{ij} = \theta_{ij}) = \text{Pr}(z_i = 0) = p.$$

Conditional on latent indicators $z_i$ for differential gene expression, $z_i = I(\theta_{ij} \neq \theta_{ij})$, we assume conjugate gamma random-effects distributions,

$$\theta_{ij} \sim \text{Gamma}(a_0, \nu), \quad (\theta_{ij} | z_i = 1) \sim \text{Gamma}(a_0, \nu), \quad \text{and} \quad (\theta_{ij} | z_i = 0) \sim \text{Gamma}(\theta_{ij}).$$

The model is completed with a prior $p(\eta)$ for the parameters $\eta = (a, a_0, \nu, p)$. In the implementation for the example in Section 5, we fix $\nu$. We assume a priori independence and use marginal gamma priors for $a_0$ and $a$ and a conjugate beta prior for $p$. As in the approach of Newton et al. (2001), the foregoing model leads to closed-form marginal likelihoods $p(X_i, Y_i | z_i = 0, \eta)$, $p(X_i, Y_i | z_i = 1, \eta)$, and $p(X_i, Y_i | \eta)$ after integrating out $\theta_{ij}, \theta_{ij}$ but still conditional on $\eta = (p, a, a_0)$. This greatly simplifies the posterior simulation.

We add two more generalizations to the model. First, we want to modify the model to allow the use of a pilot dataset to learn about the sampling distribution of the observed gene expressions across genes and repeated samples. We envision a system wherein the investigator collects some pilot data (on control tissue) before going through the sample size argument. These pilot data can then be used to learn about the important features of the sampling distribution. If the observed pilot data can be adequately fit by the marginal model $p(X_i | z_i = 0)$ under the gamma/gamma hierarchical model, then the sample size design should proceed as before. But if the pilot data show evidence against the gamma/gamma model, then the system should estimate a model extension and proceed with the extended model.

A convenient way to achieve the desired extension is a scale mixture extension of the basic model (5). In particular, we assume that

$$X_{ij} \sim \int \text{Ga}(a, \theta_{ij} r_{ij}) \, d\rho(r_{ij} | w, m)$$

and

$$Y_{ij} \sim \int \text{Ga}(a, \theta_{ij} s_{ij}) \, d\rho(s_{ij} | w, m),$$

where $p(r | w, m)$ is a discrete mixing measure with $P(r = m_k) = w_k$ ($k = 1, \ldots, K$). Locations $m = (m_1, \ldots, m_K)$ and weights $w = (w_1, \ldots, w_K)$ parameterize the mixture. To center the mixture model at the basic model, we fix $m_1 = 1.0$ and assume a high prior probability for a large weight $w_1$. We use the same mixture for $s_{ij}$, $P(s_{ij} = m_k) = w_k$. The model is completed with $m_k \sim \text{Ga}(b, b)$, $k > 1$ and a Dirichlet prior $w \sim \text{Dir}_K(M \cdot w, \ldots, M \cdot w)$. Selecting a large factor $M$ in the Dirichlet priors assigns a high prior probability for a large $w_1$. By assuming a dominating term with $m_1 = 1.0$ and $E(m_2) = \cdots = E(m_K) = 1$, we allocate a large prior probability for the basic model and maintain the interpretation of $\rho_i = \theta_{ij} / \theta_{ij}$ as the level of differential expression.

A concern related to microarray data experiments prompts us to introduce a further generalization to allow for the occa-
sional presence of slides that are outliers compared with the other arrays in the experiment. This occurs for reasons unrelated to the biologic effect of interest, but needs to be accounted for in the modeling nevertheless. We achieve this by adding a second mixture to (7),

$$X_{ij} \sim Ga(a, \theta_{ij} g_{ij}) \quad \text{and} \quad Y_{ij} \sim Ga(a, \theta_{ij} g_{sj} s_{ij}),$$

with an additional slide-specific scale factor $g_{ij}$. Paralleling the definition of $p(r_{ij}|w, m)$, we use a finite discrete mixture

$$P(g_j = m_k = w_{kg} = 1, \ldots, L) \quad \text{with a Dirichlet prior (wg1, \ldots, m_{gL})} \sim \text{Dir}_L(M_g = W_g, W_g, \ldots, W_g), \quad m_g = 1, \quad \text{and} \quad m_{gK} \sim Ga(b_g, b_g) \quad \text{for } k = 2, \ldots, L.$$  

An important feature of the proposed mixture model is its computational simplicity. We proceed in two stages. First, we use the pilot data to fit the mixture model. Let $X_{ij}^o$, $j = 1, \ldots, J^o$, denote the pilot data. We use posterior MCMC simulation to estimate the posterior mean model. We do this once, before starting the optimal design. Posterior simulation in mixture models like (8) is a standard problem. We include reversible-jump moves to allow for random size mixtures (Green 1995).

We then fix the mixture model at the posterior modes $\hat{K}$ and $\hat{L}$, and the posterior means $(\hat{w}, \hat{m}, \hat{w}_g, \hat{m}_g) = E(w, m, w_g, m_g | X^o, \hat{K}, \hat{L})$. We proceed with the optimal sample size approach, using model (8) with the fixed mixtures. We conduct the procedure, including all posterior and marginal simulation, exactly as before, with only one modification. We add a step to impute $r_{ij}, s_{ij}$ and $g_j$. Conditional on $(r_{ij}, s_{ij}, g_j)$, we replace $X_{ij}$ by $X_{ij}/(r_{ij} g_j)$ and $Y_{ij}$ by $Y_{ij}/(s_{ij} g_j)$. Everything else remains unchanged. Updating the mixture variables $r_{ij}, s_{ij}$, and $g_j$ is straightforward. The following algorithm summarizes the proposed approach with the pilot data.

**Algorithm 2: Sample Size Determination With Pilot Data.**

1. Pilot data: Assume that pilot data $X^o = \{X_{ij}^o, i = 1, \ldots, n, j = 1, \ldots, J^o\}$, from control tissue is available.
2. Mixture model: Estimate the mixture model and report the posterior modes $(\hat{K}, \hat{L})$, and the conditional posterior means $(\hat{w}, \hat{m}, \hat{w}_g, \hat{m}_g) = E(w, m, w_g, m_g | X^o, \hat{K}, \hat{L})$. Both are computed by posterior MCMC simulation for the mixture model (8).
3. Optimal sample size: Proceed as in Algorithm 1, replacing $X_{ij}$ with $X_{ij}/(r_{ij} g_j)$ and $Y_{ij}$ by $Y_{ij}/(s_{ij} g_j)$, and adding an additional step in the posterior MCMC to update the mixture indicators $r_{ij}$ and $s_{ij}$ (step 1.2 in Algorithm 1).

The indicators are initialized with the (true) values from the data simulation. The mixture model parameters remain fixed at $(\hat{w}, \hat{m}, \hat{w}_g, \hat{m}_g, \hat{K}, \hat{L})$.

Rescaling with the iteratively updated latent scale factors $r_{ij} g_j$ and $s_{ij} g_j$ formalizes the use of the pilot data to inform the sample size selection by changing the prior model (as in Algorithm 1) to the preposterior simulation, conditional on the pilot data.

**5. EXAMPLE**

We analyze the data reported by Richmond, Glasner, Mau, Jin, and Blattner (1999). These data were also used as an illustration by Newton et al. (2001). We use the control data to plan for a hypothetical future experiment.

**5.1 Implementation**

We proceed as proposed in Algorithm 2. First, we estimate the mixture model (8), using the available control data as a pilot dataset $X^o$. Estimation of (8) is implemented as a MCMC posterior simulation with reversible-jump moves. We use split-merge moves (Richardson and Green 1997) for both mixtures defined in (8). Recall that the mixtures are defined with respect to the discrete mixing measures $p(r_{ij}|w, m)$ and $p(g_j|w_g, m_g)$. The third mixture, with respect to $s_{ij}$, does not appear in the model because the pilot data include only the control data. We find the posterior mode for the size of the mixture models at $\hat{K} = 3$ and $\hat{L} = 2$.

To define the probability model for the design calculations, we fix $K = 3$ and $L = 2$ and set the mixture model parameters $(m, w, m_g, w_g)$ at their posterior means (conditional on the fixed size of the mixture). Maintaining the randomness of the mixture parameters in the design model would not significantly complicate the procedure, but it also would not contribute much to the final inference. Implementing Algorithm 1, we compute the expected losses and power $\beta(J, \rho)$ across a grid of sample sizes $J$.

Algorithm 1 is implemented as three nested loops. The outer loop is simply a repeated simulation from model (8), with fixed priors for the scale factors $r_{ij}, s_{ij}$, and $g_j$. We start by generating the hyperparameters $\eta = (a, a_0, p)$ for the prior model, given in (5)–(8) (step 1.1 of Algorithm 1). Let $Ga(a, \beta)$ denote a gamma distribution with a shape parameter $a$ and a mean of $a/\beta$. We use a $Ga(2, 2)$ prior for $a$, a $Ga(12, 1)$ prior for $a_0$, and a $Be(1, 10)$ beta prior for $p$. We include the additional prior constraints $a < 1$, $a_0 < 1$, and $.01 < p < .15$. Next, we generate indicators $z_i$ and random effects $(\theta_{ij}, \theta_{ij})$, $i = 1, \ldots, n$. We conclude simulation for the outer loop by simulating hypothetical data $(X_{ij}, Y_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, J_1$ (step 1.2).

We then proceed with the second loop, nested within the first, by iterating over $J = J_1, \ldots, J_0$ (step 1.3). For each $J$, we generate a posterior Monte Carlo sample from $p(o|y, J)$. This is achieved by the third nested loop (step 1.3.1), which implements a posterior MCMC simulation. Posterior simulation is initialized with the known true parameter values (saved from step 1.1). In each iteration of the MCMC simulation, we update $r_{ij}, s_{ij}, g_j$, and the hyperparameters $(a, a_0, p)$. The first three steps are draws from the multinomial complete conditional posterior for the respective indicators. The last three steps are implemented as random-walk Metropolis–Hastings steps to update the hyperparameters. The random-walk proposals are generated from a truncated univariate normal centered at the current values of the respective parameter, with normal standard deviations of .05, .1, and .05 for $a, a_0$, and $p$. Implementation of the MCMC is greatly simplified by noting that $p(o|y, J)$ can be analytically marginalized with respect to the random effects $\theta_{ij}, \theta_{ij}$, and $z_i$. (See Newton et al. 2001 for a statement of the marginal likelihood.) At the end of each sweep of the posterior MCMC, we compute the posterior probabilities of the differential expression $p(z_i = 1|\eta, r, g, s, y_J)$. On completion of the innermost loop, we use ergodic averages of the conditional probabilities $p(z_i = 1|\eta, r, g, s, y_J)$ to approximate $v_i = P(z_i = 1|y_J)$ (step 1.3.2). Using the marginal posterior probabilities $v_i$, we then evaluate the posterior
false-discovery and false-negative counts and rates and cor-
responding decisions $d_i$ and record them for later use. We also
record the triples $(J, \rho_i^t, d_i)$ (step 1.3.3).

On completion of the outer loop, we summarize the observed
FDR, FD, FNR, FN, and the triples $(J, \rho_i^t, d_i)$. For example,
the sample average over the simulated values of FDR under a
given sample size $J$ provides a Monte Carlo estimate of the
preposterior expected FDR$_m(J)$. The fraction of $d_i = 1$ under
a given sample size and the true effect $\rho_i^t = \rho$ provides a Monte
Carlo estimate for $\beta(J, \rho)$. To ensure sufficient Monte Carlo
sample size, the latter is done for a grid on $\rho$.

5.2 Results

Recall that FNR$_m(J)$ and FNR$_m(J)$ denote the preposterior
expectations of FN and FNR. We use analogous notations
D$_m(J)$ and $\overline{D}_m(J)$ to denote the preposterior expectations
of the number of discoveries $D$ and the threshold $t^*$, computed
under the loss $L$ and sample size $J$, defined analogously to
FNR$_m$ and FNR$_m$. We computed all inference in one run of Al-
gorithm 1, collecting the necessary Monte Carlo averages for
all summaries.

Figure 1 shows the expected loss $L''(J) = \text{FNR}_m(J)$ and
other summaries under the loss function $L = L_{2R}$. We set
the threshold for FD as $\alpha R = 7.1$. We chose this to match a
bound $\text{FDR} \leq \alpha R$ for $\alpha R = 40\%$. We compute the value as
$\alpha N = .1n \bar{p}_R/(1 - \alpha R)$. Under the assumption that 10\%
of the true differential expressions are discovered. Under $L_{2N}$,
the false-discovery count FD, and thus also the preposterior
expectation FD$_m$, is fixed by definition at FD = $\alpha N$. To main-
tain the fixed FD, the procedure must eventually start lowering
the threshold $t^*$ to reject comparisons with increasingly
lower posterior probabilities of differential expression. The esti-

mated curves FNR$_m(J)$ and FNR$_m(J)$ are derived by fitting a lin-
ear model with predictor $\sqrt{\log(J + 1)/(J + 1)}$ to the observed
pairs $(J, \text{FNR}(\rho_i^t, y))$. This is motivated by the asymptotic re-

tsults of Theorem 1 (with the offset +1 to avoid a singularity
at $J = 0$). For comparison, we estimate the same curve with
a cubic smoothing spline, using the smoothing.spline function
in R with default parameters. The corresponding curves for
FNR$_m$ and FNR$_m$ are shown as dashed lines. For FDR$_m$, FD$_m$,
D$_m$, and $\overline{D}_m$, we use cubic smoothing splines to estimate the
mean value as a function of $J$.

The relatively flat nature of FNR$_m$ and FD$_m$ does not enable
a conclusive sample size choice. We propose considering addi-
tional power curves, as defined in (3). Figure 2 shows $\beta(J, \rho)$
as a function of $\rho$ and $J$. Figure 2(a) plots the power against
the assumed true level of differential expression $\rho$, with a sepa-
rate curve for each sample size $J$. Figure 2(b) plots the same
summary against the sample size $J$, arranged by the level of
differential expression $\rho$. In practice, a sample size argument
would then proceed as follows. First, the investigator deter-
mines a minimum level of differential expression that would be
considered biologically meaningful, say two-fold expression at
$\rho = \log 2$. Using a pilot dataset, we proceed with Algorithm 2 to
compute the expected FNR, FN, and power across the sample
sizes. Inspection of the power plot for the level $\rho$ of interest,
together with the FNR and FN plots, informs the investigator
about the minimum sample size needed to achieve the desired
power and/or error rates.

5.3 Alternative Loss Functions

Although the general nature of the loss function as trading off
false-positives and false-negatives is natural, the specific form
of combining them is less clear. A strength of the proposed ap-
proach is that it allows us to evaluate the alternative loss func-
tions that combine the basic summaries FN, FD, FDR, and FNR
in different ways with minimal computational effort. We dis-
cuss the results for three alternative loss functions, $L_{2R}$, $L_N$,
and $L_R$ (introduced in Sec. 2.1).

We have already established (in Thm. 1) that there is a com-
mon optimal terminal decision rule under all four loss func-
tions. This allows us to easily adapt Algorithm 1 for all four loss
functions. The only required change is in step 1.3.2. For $L_{2R}$,
$L_N$, and $L_R$ different definitions of $t^*$ are required. The rest of
the algorithm proceeds unchanged. It is possible to use a single
implementation of Algorithm 1, recording $J, t^*, \text{FD}, \text{FD}$,
and FDR for all four loss functions. Appropriate summaries of
the saved Monte Carlo samples allow us to produce summaries
such as those shown in Figure 1 for all four loss functions, based
on a single run of Algorithm 1.

An important implication of the different strategies for
choosing the cutoff is the nature of FDR as a function of
the sample size $J$. Under $L_{2R}$ it remains, by design, constant
over $J$. This has awkward implications. Imagine the asymptotic
case with a large sample size when the true $z_i$ are practically
known. To achieve the desired FDR, we have to knowingly
flag some genes as differentially expressed even when $v_i \approx 0$.
By the same argument, the loss $L_{2N}$ leads to similar asympto-
tics. However, fixing the count FD instead of the rate FDR
slows the awkward decrease in the threshold required to main-
tain the fixed FDR under $L_{2R}$. The number of discoveries is
still increasing, but starts at a higher level and avoids the steep
increase seen under $L_{2R}$. In contrast, under $L_N$ the cutoff $t$
is fixed across the sample size, leading to a vanishing FDR in the
limit as $J \to \infty$, due to posterior consistency. But these prob-
lems might be of only asymptotic relevance and of no concern
for moderate sample sizes. Apart from these concerns, all three
loss functions $L_{2R}$, $L_N$, and $L_{2N}$ are very similar in terms of
their properties, nature of the inference, and implementation
details.

Inference under $L_R$ leads to different behaviors among the
various summaries. In contrast with the other three loss func-
tions, the optimal decision under $L_R$ does not constrain an er-
or, the error rate, or the cutoff. Considering plots similar to
those in Figure 1, we find that at an intermediate sample size,
the threshold $t^*$ moves swiftly from an initial value of $t^* \approx 0$
to the other extreme of $t^* \approx 1$. This unintuitive behavior con-

firms our initial reservations against $L_R$ for including penalties
for a false discovery and false negative that depend on the total
number of discoveries or negatives.

In summary, inference under the four loss functions differs in
how the competing goals of reducing false positives and false
negatives are balanced. The loss functions $L_{2R}$, $L_{2N}$, and $L_N$
define the trade-off implicitly by fixing FDR, FD, and $t^*$. Un-
der $L_R$, the trade-off is explicitly included as a coefficient in the
loss function. The constraint on FDR under $L_{2R}$ has the awk-
ward implication that with an increasing sample size, we must
knowingly include an increasing number of false positives in
the rejection region to maintain the set false-positive rate. The
loss function $L_R$ induces counterintuitive jumps in FDR and $t^*$, with leads us to favor $L_{2N}$ and $L_N$. Both lead to very similar inference, with $L_{2N}$ having the advantage of a constraint on the practically more important FD, rather than $t^*$, as in $L_N$.

6. CONCLUSION

The design of microarray experiments for measuring gene expression is a critical aspect of genomic analyses in biology and medicine. Microarrays are costly, and difficult trade-offs must be evaluated when allocating resources to alternative investigations. Even in the simplest two-sample comparison setting, microray analyses pose difficult challenges to traditional sample size approaches. First, in terms of hypothesis testing, they present a multitude of heterogeneous alternatives; second, they are generally performed with goals that are best captured by properties of the ensemble of the choices made; and third, they mandate the incorporation of existing knowledge, because signal-to-noise ratios vary significantly with the specific technology, the source of RNA, and the overall experience of the laboratory’s personnel.

Our goal of this article has been to develop a formal decision-theoretic framework to address these challenges. This gives investigators the opportunity to quantify both the a priori uncertainty about the likely expression levels and the implications of sample size choices on the performance of inference about differential expression. The consequences of decisions are captured by loss functions related to genome-wise error rates. We argue for using posterior expected error rates for the terminal decision about the multiple comparisons, and marginal expected error rates for the design decision about the sample size, consistently with a Bayesian sequential approach. Similar issues recur in other high-dimensional multiple-comparison problems and in the detection of faint signal levels in noisy data; the methods that we propose are applicable more generally to these problems as well.

In situations requiring complex decision making, decision models such as ours are best thought of as decision support tools. As is common in simpler settings, we envision investigators exploring various scenarios rather than simply eliciting input and blindly trusting the emerging sample size recommendation. A reasonable situation is also one in which an investigator has in mind a certain sample size that is feasible within given resource constraints. The proposed method informs the investigator about the effect sizes that she or he can realistically expect to discover with the proposed sample size, and also about the ensuing error rates.

An interesting application of the proposed method is in a sequential framework. An investigator could proceed in steps, starting with an initial batch of experiments and stopping when a satisfactory balance of classification error rates is achieved. This could be implemented without preposterior calculations. Because genome-wise error rates refer to the ensemble of genes, an investigator could not sample to a foregone conclusion about any individual genes by using this stopping rule.

In our model we assume that genes are from a discrete mixture in which some genes are altered across the two samples and others are completely unaltered. This assumption is realistic in tightly controlled experiments, but less so in the comparison of RNA samples across organs or across organisms. These broader comparisons are often made to produce exploratory analyses, such as clusters. The choice of sample sizes in these circumstances is different from that used in controlled experiments. Some insight into this issue has been offered by Simon et al. (2002) and Bryan and van der Laan (2001).

An important practical indication for microarray design arises from the illustration described in Section 5. In particular, for a realistic set of parameters and pilot data, we have shown that the improvement in the genome-wide error rate appears to be nonconcave, with a small initial plateau at very small sample size. In some cases the payoff of increasing the sample size from, say, two to three appears to be negligible. This has implications for the common practice of planning experiments with only two or three replicates. We suggest that an analysis of the kind presented in Figure 1 can provide valuable information to investigators entertaining experiments with a very small number of replicates.

APPENDIX A: OPTIMAL TERMINAL DECISION

We prove Theorem 1. We start by considering $L_N$, subject to a fixed total number of discoveries $D$. We find that $L_N(d,y|D) = cD - (c + 1) \sum \bar{d}_i + \sum v_i$. The last term does not involve the decision. For fixed $D$, the rest is minimized by setting $d_i = 1$ for the $D$ largest $v_i$. In other words, for any $D$, the optimal rule is of the type $d_i = \mathbb{1}(v_i > t)$, where $t$ is simply the $(n - D)$th order statistic of $\{v_1, \ldots, v_n\}$. Thus we conclude that the global minimum must be of the same form, and it remains only to find the globally optimal $t$. Straightforward algebra shows that the global minimum is achieved for $t^* = c/(c + 1)$.

A similar argument holds under $L_R(d, y)$. We find that

$$L_R(d, y|D) = C_1(D) - C_2(D) \sum_{i=n-D+1}^n v_i + C_3(D) \sum v_i,$$

(A.1)

with $C_1(D) = cD/(D + \epsilon)$, $C_2(D) = c/(D + \epsilon) + 1/(n - D + \epsilon)$, $C_3(D) = 1/(n - D + \epsilon)$, and $v_i$ the $i$th order statistic of $v_i$. The global optimum is found by minimizing $L_R(d, y|D)$ with respect to $D$ to find the optimal $D = D^*$. Thus the optimal decision is $d_i = \mathbb{1}(v_i > t)$ and $t^* = c/(n-D^*)$.

Under $L_{2N}$ and $L_{2R}$, we need an additional argument. To minimize FNR subject to FDR $\leq \alpha$, we write the Lagrangian function

$$f_2(d) = \text{FNR} - \lambda(\alpha - \text{FDR}).$$

Using Lagrangian relaxation (Fisher 1985), we find a weight $\lambda^*$ such that the minimization of $f_2^*(d)$ provides an approximate solution to the original constrained optimization problem. (The solution is only approximate because of the discrete nature of the decision space.)

$$f_2^*(d) = L_R \text{ with } c = \lambda^*.$$ Thus the solution must have the same form as described earlier. The only difference is that the implied coefficient $c$ itself is a complicated function of the data. Knowing the structure of the solution, we can solve the decision problem by finding the cutoff $t_2R(y) = \min[s : \text{FDR}(s, y) \leq \alpha]$. A similar argument holds for $L_{2N}$, with $t_2N(y) = \min[s : \text{FDR}(s, y) \leq \alpha]$. Note that the optimal cutoff $t^*$ in all three new loss functions is now a function of the data. We write $t_2N^*(y)$ for the optimal cutoff under loss $L$ given data $y$.

APPENDIX B: ASYMPTOTIC FALSE–NEGATIVE RATE

We now prove Theorem 2, assuming a model with the same structure as in Section 4. The specific distributional assumptions, including the gamma sampling distribution for $(X_{ij}, Y_{ij})$ and the gamma prior for $(\theta_0, \theta_1)$, are not critical. We start the argument by establishing an asymptotic approximation for $P(\tau_j = 1 | y_j)$. We then use this result to argue that for a large $J$, the rejection region has to necessarily include...
some genes with zero or a small level of true differential expression. This is true under all three loss functions, \( L_{2R}, L_2N, \) and \( L_N. \) Thus the nonrejection region includes only small levels of true differential expression. We exploit this fact to approximate \( \text{FNR} \) by an integral that can be recognized as an expression of the order of \( \sqrt{\log J}. \) The integral approximation is valid only if a large number of genes are in the nonrejection region, allowing us to approximate the sum in the definition of \( \text{FNR} \) by an integral. We conclude the argument by showing that this is the case under all three loss functions, for a sufficiently large \( J. \)

We start with an asymptotic result for the posterior probability of differential expression. Let \( \eta = (a, a_0, p) \) denote the hyperparameters, and let \( y_i = \{X_{ij}, Y_{ij}, \ j = 1, \ldots, J\} \) denote the data for gene \( i. \) Because the number \( n \) of genes is very large, we have, for each gene,

\[
P(\zeta_i = 1|y_i) = \int P(\zeta_i = 1|y_i, \eta) d\eta(y_i|\eta) = P(\zeta_i = 1|y_i, \hat{\eta})(1 + \text{Op}(n^{-1/2})) = P(\zeta_i = 1|y_i, \eta)(1 + \text{Op}(n^{-1/2})),
\]

where \( \hat{\eta} \) is the maximum likelihood estimator and \( \eta \) are the true hyperparameters. Here \( X_n = \text{Op}(n^k) \), for a sequence of random variables \( X_n, \) is defined as

\[
\lim_{M \to \infty} \left\{ \limsup_{n} P[X_n/n^k > M] \right\} = 0.
\]

Moreover, for each gene \( i, \) the posterior probability of differential expression given \( \eta \) is

\[
P(\zeta_i = 1|y_i, \eta) = \frac{p_p(y_i|\zeta_i = 1, \eta)p(y_i|\zeta_i = 0, \eta)}{p_p(y_i|\zeta_i = 1, \eta) + (1 - p)p(y_i|\zeta_i = 0, \eta)}.
\]

Classical Laplace expansions imply that

\[
P(\zeta_i = 1|y_i, \eta) = \frac{1}{1 + c_i \exp[-J(\hat{\theta}_{ij} - \hat{\theta}_{ij})^2\tau_i/2]\sqrt{J}}.
\]

Finally, to show the same for \( \text{FDR}, \) we use (B.2) and the fact that all negatives have small \( p(z_i = 1|y_i, \eta) \) is very close to 1 (the error being essentially of the order \( n^{-1} \)).

We now use (B.2) to study the asymptotic behavior of the terminal decision. In particular, we consider \( \text{FDR}, \) \( \text{FD}, \) \( \text{FNR}, \) and \( \text{FN}. \) Let \( v_i \leq \cdots \leq v_N \) be the ordered posterior probabilities \( v_i = P(\zeta_i = 1|y) \) and recall that \( \text{FDR}(t, y) = \sum_i (1 - v_i)/D, \) where \( D = \sum_i I(t_i \geq t) \) is the number of discoveries. We use \( N = \sum_i I(y_i < t), \) \( p_p = \sum_i I(y_i > t)/\sum_i I(y_i = 0), \) and \( n_s = -n_t - n_1 \) to denote the number of negatives, false-positives, and differentially expressed genes. We use \( A_N, A_{FP}, A_1, \) and \( A_0 \) to denote the corresponding sets of genes. The foregoing expansions show that the ordering of \( v_i \) is asymptotically linked to the ordering of \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}|, \) with \( v_i \) monotone, increasing in \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| \) with, asymptotically,

\[
v_i \approx -c_i \sqrt{T} \exp[-J(\hat{\theta}_{ij} - \hat{\theta}_{ij})^2\tau_i/2].
\]

The false-discovery rate \( \text{FDR}(t, y) \) as a function of \( t \) is a step function taking values in \( \{1 - v(0), \ldots, 1 - (v(k)/\cdots + v(n))/n(n - k + 1), \ldots, 1 - (v(1) + \cdots + v(n))/n\}. \) Similarly, \( \text{FD}(t, y) \) is a step function with values \( \{1 - v(0), \ldots, 1 - v(1) + \cdots + 1 - v(n)\}. \) Both are monotone, decreasing in \( t. \) For a large \( J, \) the previous discussion shows that any gene with

\[
|\hat{\theta}_{ij} - \hat{\theta}_{ij}| > C \sqrt{\log J}/\sqrt{J}
\]

has a posterior probability of \( v_i \geq 1 - 1/\sqrt{J} \) when \( C \) is sufficiently large, uniformly in \( \theta_{ij} \) and \( \theta_{ij} \) belonging to some compact set, and with a large probability. We let

\[
S = \{i : |\hat{\theta}_{ij} - \hat{\theta}_{ij}| < C \sqrt{\log J}/\sqrt{J}\}
\]

denote the set of genes with small \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| \) that violate (B.3).

We now show that under all three losses, only genes with small \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| \) are classified as nondifferentially expressed, that is, \( A_N \subseteq S. \) Under \( L_N, \) the argument is straightforward. For all genes satisfying (B.3), the posterior probability \( p(z_i = 1) = 1 - 1/\sqrt{J} \) is beyond \( t_N = c/(1 - c) \) for sufficiently large \( J. \) Thus all genes in \( A_N \) satisfy \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| < C \sqrt{\log J}/\sqrt{J}. \)

To prove the claim under \( L_{2R}, \) we show that the opposite would violate the constraint on \( \text{FDR}. \) Assume that (B.3) holds for all \( i \in A_D. \) Then

\[
\text{FDR} = 1 - (v(n) - d + 1) + \cdots + v(n)/D \leq 1/\sqrt{J}.
\]

which is not enough to reach the set level or bound required for \( L_{2R}. \) Thus the rejection region \( A_D \) necessarily must include some genes that violate (B.3). Together with the monotonicity of \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| \) as a function of \( v_i, \) this proves the claim.

Finally, to show the same for \( L_{2N}, \) consider (B.3) with an even larger \( C. \) If \( C^2 > 1/\log (p - \alpha/2), \) then \( 1 - v(i) \leq \alpha/(2n) \) for all genes that satisfy (B.3) with such \( C. \) If only such genes are considered in the rejection region, then

\[
1 - v(k) + \cdots + 1 - v(n) \leq \alpha/2,
\]

which is not enough to reach the desired bound \( \text{FNR} = \alpha \) under \( L_{2N}. \)

We now use (B.2) and the fact that all negatives have small \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| \) to establish a bound on \( \text{FNR}, \)

\[
\text{FNR}(t, y) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{1 + c_j \sqrt{T} \exp[-J(\hat{\theta}_{ij} - \hat{\theta}_{ij})^2\tau_i/2](1 + \text{Op}(n^{-1/2}))},
\]

where \( c_j = \sqrt{\tau_j} / \sqrt{2\pi} \) with \( \tau_j = v(i_0) \ell(i_0) \ell(i_1)/(\ell(i_0) + \ell(i_1)) \) and \( \ell(i) \) is the Fisher information associated with the conditional model of \( X_i | (Y_i \eta) \) given \( \eta, \) when \( \eta \) is fixed.

If \( N \) is large, then the sum can be approximated by an integral with respect to the distribution of \( v(i) \) or, equivalently, the distribution of \( (\hat{\theta}_{ij}, \hat{\theta}_{ij}). \) We split the integral into two parts. With probability \( v_0, \) we have \( \hat{\theta}_{ij} \equiv i \) and with probability \( 1 - v_0, \) we have \( \hat{\theta}_{ij} \equiv i. \) Based on the earlier observation that we fail to reject the comparison only in the case of small estimated differences, we can condition the latter term on \( |\hat{\theta}_{ij} - \hat{\theta}_{ij}| < C \sqrt{\log J}/\sqrt{J}. \) Let \( \sqrt{T} \theta_{ij} - \tau_{ij} = \sqrt{T} (\theta_{ij} - \theta_{ij}) + \tau_{ij} + \xi_j, \) where \( \xi_j \) is a standard Gaussian random variable. Let \( \Theta_s = \{\theta_1, \theta_0) : |\theta_1 - \theta_0| < C \sqrt{\log J}/\sqrt{J}. \) Then

\[
\text{FNR}(y, t_{2R}) \approx w_0 \int_{\xi_j} \frac{1}{1 + \sqrt{(\theta)/\sqrt{2\pi}(1 - p)/pe^{-\xi_j^2/2}\sqrt{J}}} dp(\xi) d\theta + w_1 \int_{\xi_j} \frac{1}{1 + c(\theta, \theta_1) \sqrt{e^{-\xi_j^2/2}e^{-J(\theta_0 - \theta_1)^2/(\theta_0 - \theta_1)^2/2}}} \times dp(\xi) d\theta(\theta_0, \theta_1).
\]
Simple calculations imply that the foregoing quantities are of the order \( \sqrt{\log JT} \), when \( N \) is large.

Moreover, \( FNR = N \cdot FPR \). We now prove that \( n/N = OP (1) \) with a high probability under all three losses.

We start with the argument for \( L_{2R} \). Under the assumed sampling model, \( n_0 \approx n \cdot \theta_0 \) genes satisfy \( \theta_0 = \theta_j \). If \( N/n \to 0 \), then a large proportion of genes satisfying \( \theta_0 = \theta_j \) would have posterior probabilities \( v_j > t_2 \). Recall that \( FDR \) is the set of false positives. This would imply that

\[
FDR \geq \frac{1}{n} \sum_{i \in \text{App}} (1 - v_i) \geq \frac{1}{n} \sum_{i \in \text{App}} e_{ij} \sqrt{J} \exp[-J(\hat{\theta}_0 - \hat{\theta}_i)^2 \tau_j/2] \left(1 + O_P(n^{-1/2})\right)
\]

\[
\geq \frac{p}{2} \int_0^1 \int \sqrt{f_2(\tau) \exp^{-\xi^2/2}} \, d\tau \, d\xi \, d\xi (1 + O_P(n^{-1/2})),
\]

when \( n \) is sufficiently large, with a high probability. The last inequality is true because under the assumption that \( N/n \to 0 \), eventually more than \( N/2 \approx np/2 \) genes would be in \( \text{App} \). As \( J \) goes to infinity, the foregoing term goes to \( p/2 \). This is a contradiction if \( \alpha < p/2 \), and we thus conclude that \( n/N = OP (1) \).

Under \( L_{2N} \), we use an analogous argument for \( FDR \). The right side in the first two foregoing (in)equalities remains unchanged, except for removing the leading \( 1/n \) factor. We conclude that \( FDR \geq OP (n^{-1/2}) \), and thus have a contradiction for \( \alpha < p/2 \).

Finally, under \( L_{N} \), \( n_2 = c/(c+1) \), so \( v_j \leq n_2 \) if \( 1 \leq c \cdot \sqrt{N} \exp[-J(\hat{\theta}_0 - \hat{\theta}_i)^2 \tau_j/2] \), then

\[
P\left[1 > \frac{1}{c} \cdot \sqrt{N} \exp[-J(\hat{\theta}_0 - \hat{\theta}_i)^2 \tau_j/2] \right] = \mathcal{O}(1/\sqrt{N}),
\]

by Chebychev’s inequality. Recall that \( \text{FP} \) is the number of genes satisfying \( \theta_0 = \theta_j \) and \( v_j > N_2 \). Then \( \text{FP} \) is a binomial random variable \( \text{Bin}(n_0, p_1) \), with \( p_1 = OP(J^{-1/2}) \) and where \( n_0 \) is the number of genes with \( \theta_0 = \theta_j \). Thus with a probability of \( 1 - e^{-c_1 \sqrt{N}} \), for some positive constant \( c_1 \), \( n_1 \leq c_2 n \), with \( c_2 < 1 \).

\[
\text{REFERENCES}
\]

Adcock, C. J. (1997), “Sample Size Determination: A Review,” The Statistician, 46, 261–283.

Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V., and Zhang, W. (2001), “Identifying Differentially Expressed Genes in cDNA Microarray Experiments,” Journal of Computational Biology, 8, 639–659.

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.

Berger, J. (1985), Statistical Decision Theory and Bayesian Analysis, New York: Springer-Verlag.

Bickel, D. R. (2003), “Selecting an Optimal Rejection Region for Multiple Testing: A Decision-Theoretic Alternative to FDR Control, With an Application to Microarrays,” technical report, Medical College of Georgia.

Bryan, J., and van der Laan, M. (2001), “Gene Expression Analysis With the Parametric Bootstrap,” Biostatistics, 2, 445–461.

DeGroot, M. H. (1970), Optimal Statistical Decisions, New York: McGraw-Hill.

Duggan, D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. (1999), “Expression Profiling Using cDNA Microarrays,” Nature Genetics, 21, 10–14.

Fisher, M. (1985), “An Applications-Oriented Guide to Lagrangian Relaxation,” Interfaces, 15, 10–21.

Genovese, C., and Wasserman, L. (2003), “Bayesian and Frequentist Multiple Testing,” in Bayesian Statistics 7, eds. J. M. Bernardo, M. J. Bayarri, A. P. Dawid, J. O. Berger, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press, pp. 145–162.

Green, P. J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” Biometrika, 82, 711–732.

Guillen, C., and Rousseau, J. (2002), “Laplace Expansions in MCMC Algorithms for Latent Variable Models,” technical report, CREST.

Keeney, R. L., Raiffa, H. A., and Meyer, R. F. C. (1976), Decisions With Multiple Objectives: Preferences and Value Tradeoffs, New York: Wiley.

Kerr, M. K., and Churchill, G. A. (2001), “Experimental Design In Gene Expression Microarrays,” Biostatistics, 2, 183–201.

Kohane, I. S., Kho, A., and Butte, A. J. (2002), Microarrays for an Integrative Genomics, Cambridge, MA: MIT Press.

Lee, M.-L., and Whitmore, G. (2002), “Power and Sample Size for Microarray Studies,” in The Statistician, 11, 3543–3570.

Lindley, D. V. (1971), Making Decisions (2nd ed.), New York: Wiley.

(1997), “The Choice of Sample Size,” The Statistician, 46, 129–138.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T., and Mesirov, J. (2003), “Estimating Dataset Size Requirements for Classifying DNA Microarray Data,” Journal of Computational Biology, 10, 119–142.

Newton, M. A., and Kendziorski, C. M. (2003), “Parametric Empirical Bayes Methods for Microarrays,” in The Analysis of Gene Expression Data: Methods and Software, eds. G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, New York: Springer-Verlag. pp. 254–271.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), “On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data,” Journal of Computational Biology, 8, 37–52.

Pan, W., Lin, J., and Le, C. T. (2002), “How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach,” Genome Biology, 3, research 0022.1–0022.10.

Raiffa, H., and Schlaifer, R. (1961), Applied Statistical Decision Theory, Boston, MA: Harvard University Press.

Richardson, S., and Green, P. (1997), On Bayesian Analysis of Mixtures With an Unknown Number of Components,” Journal of the Royal Statistical Society, Ser. B, 59, 731–792.

Richmond, C. S., Glasser, J. D., Mau, R., Jin, H., and Blattner, F. (1999), “Genome-Wide Expression Profiling in Escherichia coli K-12,” Nucleic Acid Research, 27, 3821–3835.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995), “Quantitative Monitoring of Gene Expression Patterns With a Complementary DNA Microarray,” Science, 270, 467–470.

Simon, R., Radmacher, M. D., and Dobbin, K. (2002), “Design of Studies Using DNA Microarrays,” Genetic Epidemiology, 23, 21–36.

Storey, J. S., and Tibshirani, R. (2003), “SAGE Thresholding and False-Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays,” in The Analysis of Gene Expression Data: Methods and Software, eds. G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, New York: Springer-Verlag, pp. 272–290.

Tseng, G. C., Oh, M. K., Rohlfin, L., Liao, J., and Wong, W. (2001), “Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variations and Assessment of Gene Effects,” Nucleic Acids Research, 29, 2549–2557.

Tusher, V., Tibshirani, R., and Chu, G. (2001), “Significance Analysis of Microarrays Applied to the Ionizing Radiation Response,” Proceedings of the National Academy of Science USA, 98, 5116–5121.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. (2002), “Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation,” Nucleic Acids Research, 30, e15.

Yang, Y. H., and Speed, T. P. (2002), “Design Issues for cDNA Microarray Experiments,” Nature Genetics Reviews, 3, 579–588.

Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2002), “Microarrays: How Many Do You Need?” technical report, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany.