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

Peter Müller,[1] **Giovanni Parmigiani,**[2] **Christian Robert,**[3] and **Judith Rousseau**[4]

Abstract

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 comparison 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. The focus of the discussion is on the sample size selection. For the multiple comparison we assume an approach as in Genovese and Wasserman (2002), based on controlling posterior expected false discovery rate (FDR). For the sample size selection we adopt a decision theoretic solution, using expected false negative rate (FNR) as decision criterion, combined with a power analysis as sensitivity diagnostic. Posterior expected FDR and marginal FNR are computed with respect to an assumed parametric probability model. In our implementation we use a version of the model proposed in Newton et al. (2001). But the discussion 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 efficient marginal and posterior simulation. Posterior and marginal simulation can be done by dependent Markov chain Monte Carlo simulation.

Key words: Multiple comparison; False discovery rate.

[1]Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX.

[2]Department of Oncology, Biostatistics and Pathology, Johns Hopkins University, Baltimore, MD

[3]CEREMADE, Université Paris Dauphine, and CREST, INSEE, France

[4]Université Rene Descartes, Paris, and CREST, INSEE, France

The order of authors is strictly alphabethical.

1 Introduction

We consider the problem of choosing the sample size for experiments involving massive multiple comparisons. Our discussion is motivated by the specific problem of choosing the number of replications in microarray experiments. Gene expression microarrays are technologies for simultaneously quantifying the level of transcription of a large portion of the genes in an organism (Schena et al., 1995; Duggan et al., 1999). For a recent review of microarray technology and related statistical methods see, for example, Kohane et al. (2002). The range of applications is broad. Here we focus on controlled experiments whose goal is to search, or screen, for genes whose expression is regulated by modifying conditions of interest, either environmentally or genetically. There are a number of pressing biological questions that can be addressed using this type of genomic screens, and microarrays are costly. This results in difficult tradeoffs in the allocation of limited research resources. Careful choice of sample sizes is critical in addressing these tradeoffs. General issues of experimental design in microarray experiments are discussed by Kerr and Churchill (2001), Yang and Speed (2002) and Simon et al. (2002).

Although microarrays offer efficient technologies for exploring high-dimensional biological phenomena, they still involve a significant amount of noise. It is common practice to verify putative discoveries obtained from microarrays by means of more accurate, and more expensive, assays, such as RT-PCR (reverse transcription-polymerase chain reaction), as well as other additional experiments. This confirmation work generally requires a significant effort. Thus the typical result of a microarray data analysis is a list of putative regulated genes. This selection problem can also be framed as a massive multiple hypothesis test, with one or more hypotheses per gene. The goal of the genomic screen is, to a good degree of approximation, that of discovering as many of the genes that are differentially expressed across conditions as possible, while keeping the number of false discoveries manageable. The consequences of a false discovery are often similar across genes.

We take the viewpoint that choice of sample size can benefit from acknowledging these experimental goals explicitly. Decision theoretic approaches to sample size choice offer a systematic way to connect sample size determination to the goals of the inferences that will be made using the data collected. The problem is formalized as a two-stage decision: first sample size selection and then gene selection for the given sample size. This multistage nature of sample size determination has been formalized within a Bayesian framework as early as Raiffa and Schlaifer (1961).

Implementation requires a joint probability model for the unknowns under investigation (the parameters describing regulation of genes) and the observations, as well as a quantitative measure of the consequences of the gene selection. The latter, a utility or a loss, drives both the gene selection and the sample size determination. The sample size is decided before experimenting, requiring consideration of the marginal distribution of observations. In contrast inference about differential expression is made after the experiment is concluded, conditionally on the observed data. See also Lindley (1997) or Adcock (1997) and references therein for discussions of sample size determination.

Following this paradigm, we present in this paper a general decision theoretic framework for the choice of sample size in a genomic screen, or in a similar selection problem. Central to our analysis is the concept of false discovery rate, or 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 and truly unaltered genes. For a given approach to selecting a set of putatively altered genes, the FDR is the fraction of truly altered genes among the genes classified as differentially expressed. Commonly used microarray software uses FDR to guide gene selection; see for example Tusher et al. (2001). Applications of FDR's to microarray analysis are discussed by Storey and Tibshirani (2003). Extensions are discussed by Genovese and Wasserman (2002), who also introduce the definition of posterior expected FDR as we use it here.

Our specific selection approach is based on the work in Genovese and Wasserman (2002). They focus on decision rules of the following kind. Assume that for each comparison some univariate summary 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. Although Genovese and Wasserman (2002) make no explicit reference to an underlying probability model on the observable data, we develop a version based on a multilevel parametric model, with mixture components corresponding to regulated and unregulated genes. We then use as summary statistic the marginal posterior probabilities of differential expression.

Our hierarchical specification is similar to the Newton-Kendziorski model for cDNA arrays (Newton et al., 2001; Newton and Kendziorski, 2003). In microarray analyses, because gene expression measurements are subject to a number of common sources of noise, large gains in efficiency can result from considering the ensemble of gene expression measures at once, rather than considering each gene in isolation. This is in agreement with the Bayesian viewpoint and, in particular, Bayesian hierarchical models, such as the Newton-Kendziorski model and others (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Ibrahim et al., 2002; Parmigiani et al., 2002). Empirical Bayes models (Efron et al., 2001) also offer practical ways to achieve these gain in efficiency. Bayesian hierarchical models allow to realistically capture the variation in effect sizes within the set of truly regulated genes. This is critical in sample size calculations, as most existing methodologies examine situations in which all genes are regulated by the same amount across conditions.

An attractive aspect of Bayesian approaches to sample size determination is the ability to capture existing information about the likely sources of noise and likely magnitudes of the unknown gene-specific effects via prior distributions. The genomic and gene-specific noise structure varies significantly from laboratory to laboratory. Most laboratories repeatedly use the same type of control RNA on the same chip type, so that useful evidence about model parameters accumulates. This can be formally exploited into sample size choice via the prior distribution.

In the context of microarray experiments, important initial progress towards the evaluation of sample sizes has been made by Pan et al. (2002), who developed traditional power analyses for the microarray context. 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, however, is necessary: Pan et al. (2002) do not exploit heterogeneity in developing screening statistics, as done by hierarchical models. This can potentially underestimate 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 et al. (2002) propose an alternative approach to informed sample size choice. They consider ROC-type curves, showing achievable combinations of false negative and false positive rates.

In summary, gene selection in genomic screens challenges traditional sample size approaches along several dimensions: there is heterogeneity in the likely magnitude of effects across genes, and therefore in the alternative hypotheses one needs to consider; there is heterogeneity in the sources of noise across genes, which makes treatment of nuisance parameters challenging; and the goals of investigation are most appropriately stated in terms of global properties of gene selection procedures, rather than individual tests. Our modeling approach addresses all three of these needs explicitly; our computational approach ensures an efficient implementation.

In Section 2 we outline the decision problem and our approach to the solution in general form, without reference to a specific probability model. In Section 3 we outline an efficient simulation approach for evaluating the required sample size selection criteria. We define a Monte Carlo simulation method that allows to evaluate expected FNR and power across sample sizes. We demonstrate that, despite their preposterior nature, the required simulations are easier and less computation intensive than posterior simulation in the underlying probability model. A probability model is introduced in Section 4. Section 5 reports results in an example, We conclude with some discussion points in Section 6.

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 ω and y denote the model parameters and expression measurements, and $z_i \in \{0, 1\}$ denote an indicator for 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. In a two-condition comparison, data y are indicated by X_{ij}, Y_{ij}, for J experiments, $j = 1, \ldots, J$. We will write y_J when we want to highlight that y is function of the sample size J. We assume that the underlying probability model allows 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, like empirical Bayes estimates for hyperparameters, etc. Later, in Section 4 we will 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 the gene selection. This will be either a selection, (also referred to as discovery, or rejection, and denoted as $d_i = 1$ for comparison i) or not (also referred to as a negative and denoted as $d_i = 0$). 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 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 ω. The relevant probability model with respect to which we average is the prior probability on (z, ω) and the conditional sampling distribution on y given (z, ω). At the time of the decision about the multiple comparison the data y is known and the relevant probability model is the posterior distribution conditional on y. The choice of a specific loss function is complicated by the fact that the experiment involves (at least) two competing goals, minimizing the number of false discoveries (false positives) on one hand, and minimizing the number of false negatives on the other hand.

We discuss four alternative utility functions that combine the two goals. All four loss functions are chosen such that the implied solution for the multiple comparison problem is essentially the rule proposed in Genovese and Wasserman (2002). Details are discussed below, in Section 2.1. Using the same loss function we proceed in Section 2.2 to address the sample size selection. By using a single loss function for both decisions we ensure a coherent non-dominated procedure. While we consider this a desirable property, we are also concerned about sensitivity to prior specifications and frequentist properties of the derived decision rules. This concern and the relatively flat nature of the expected loss (as a function of sample size) motivate us to propose additional sensitivity analysis. In particular, we propose to look at the probability of a false negative as a function of the true level of differential expression. In the context of one of the proposed loss functions this can be interpreted as the contribution of one gene to the expected loss, conditioning on an assumed true level of differential expression, and marginalizing with respect to all other unknowns. The diagnostic takes the form of a power plot showing the probability of a false negative (or the complementary event of correct discovery) as a function of the assumed parameter value. Details are discussed below.

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

2.1 Terminal Decision

The choice of decision rule for the multiple comparison is driven by the following considerations. First, the rule should have a coherent justification as the solution that minimizes expected loss under a sensible loss function. Second, inference about the multiple comparison 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 is regulated, and the noise levels are relatively high. Finally, a good rule should allow for dependencies across genes. Although the typical prior probability model does not include dependence, the data might allow inference about dependence. 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 capture the typical goals of genomic screens, 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 all lead to terminal decision rules of the same form. Writing $D = \sum d_i$ for the number of discoveries we let

$$\mathrm{FDR}(d, y, , z) = \frac{\sum d_i(1 - z_i)}{D + \epsilon} \text{ and } \mathrm{FNR}(d, y, , z) = \frac{\sum (1 - d_i) z_i}{n - D + \epsilon} \tag{1}$$

denote the realized false discovery rate and false negative rate, respectively. $\mathrm{FDR}(\cdot)$ and $\mathrm{FNR}(\cdot)$ are the percentage of wrong decisions, relative to the the number of discoveries and negatives, respectively (the additional term ϵ avoids a zero denominator). See, for example, Genovese and Wasserman (2002) for a discussion of FNR and FDR. Conditioning on y and marginalizing with respect to z we obtain the posterior expected FDR and FNR

$$\overline{\mathrm{FDR}}(d, y) = \int \mathrm{FDR}(t, y, , z)\, \mathrm{d}p(z \mid y) = \sum d_i(1 - v_i)/(D + \epsilon)$$

and

$$\overline{\mathrm{FNR}}(d, y) = \int \mathrm{FNR}(t, y, z)\, \mathrm{d}p(z \mid y) = \sum (1 - d_i) v_i/(n - D + \epsilon).$$

Let $\overline{\mathrm{FD}} = \sum d_i(1 - v_i)$ and $\overline{\mathrm{FN}} = \sum (1 - d_i) v_i$ denote the posterior expected count of false discoveries and negatives. We consider the following four ways of combining the goals of minimizing false discoveries and false negatives. The first two specifications combine false negative and false discovery rates and numbers, leading to the following posterior expected losses:

$$L_N(d, y) = c\,\overline{\mathrm{FD}} + \overline{\mathrm{FN}},$$

and $L_R(d, y) = c\,\overline{\mathrm{FDR}} + \overline{\mathrm{FNR}}$. 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, respectively. In the upcoming discussion we will eventually build an argument against L_R. However, in many aspects inference under L_R is very similar to inference under L_N and we will therefore continue to include L_R in the discussion. Alternatively, we consider bivariate loss functions that explicitly acknowledge the two competing goals, leading to posterior expected losses:

$$L_{2R}(d, y) = (\overline{\text{FDR}}, \overline{\text{FNR}}), \quad L_{2N}(d, y) = (\overline{\text{FD}}, \overline{\text{FN}}).$$

Using posterior expectations we have marginalized 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.

We now show that for all four loss functions, L_R, L_N, L_{2R} and L_{2N}, the optimal terminal decision takes the same form: reject (declare a discovery) if the marginal posterior probability v_i is beyond a certain cutoff t,

$$d_i = I(v_i > t).$$

The decisions only differ in how this cutoff t is determined. Under L_R and L_N the derivation is straightforward. Start by considering $L_R(d, y)$, subject to a fixed total number of discoveries D. The expected loss simplifies to

$$L_R(d, y|D) = C_1(D) - C_2(D) \sum_{i=1}^{n} d_i v_i + C_3(D) \sum_{i=1}^{n} v_i \tag{2}$$

with $C_1(D) = cD/(D + \epsilon)$, $C_2(D) = c/(D + \epsilon) + 1/(n - D + \epsilon)$ and $C_3(D) = 1/(n - D + \epsilon)$. Only the second term includes d_i. For fixed D, a minimum is achieved by setting $d_i = 1$ for the D largest posterior probabilities v_i. Under the optimal decision the first sum in (2) reduces to the sum over the highest D order statistics $v_{(i)}$ of $(v_1, \ldots, v_n)$.

$$\min_d L_R(d, y|D) = C_1(D) - C_2(D) \sum_{i=n-D+1}^{n} v_{(i)} + C_3(D) \sum v_i. \tag{3}$$

The global optimum is found by minimizing (3) with respect to D to find the optimal $D = D^*$. Thus the optimal decision is $d_i = (v_i > t)$ and $t = t_R(y) \equiv v_{(n-D^*)}$. An analogous argument holds for L_N. We find that $\min L_N(d, y|D) = cD - (c + 1) \sum_{i=n-D+1}^{n} v_{(i)} + \sum v_i$ and the global minimum is achieved for $t_N = c/(c + 1)$.

Under L_{2N} and L_{2R} we need an additional argument. 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 et al., 1976). Using this strategy we now show that the optimal terminal decision again takes the form of a threshold on v_i. First, consider the loss L_{2R}. To minimize $\overline{\text{FNR}}$ subject to $\overline{\text{FDR}} \leq \alpha$ we write the Lagrangian function

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

Using Lagrangian relaxation (Fisher, 1985) we find a weight $\lambda^* \geq 0$ such that the minimization of $f_{\lambda^*}(d)$ provides an approximate solution to the original constrained optimization problem (The solution is only approximate because of the discrete nature of the decision space). But $f_{\lambda^*} = L_R$ with $c = \lambda^*$. Thus the solution must have the same form as described above. The only difference to before is that the implied coefficient c itself is a complicated function of the data. By a slight abuse of notation we write $\overline{\text{FDR}}(t, y)$ for $\overline{\text{FDR}}(d, y)$ when $d_i = I(v_i > t)$. Knowing the structure of the solution we can solve the decision problem by finding the cutoff $t_{2R}(y) = \min\{s : \overline{\text{FDR}}(s, y) \leq \alpha\}$. A similar argument holds for L_{2N}, with $t_{2N}(y) = \min\{s : \overline{\text{FD}}(s, y) \leq \alpha\}$.

In summary, all four loss functions lead to very similar decision rules. The only difference lies in the choice of the cutoff t. All but L_N lead to data dependent cutoffs. Under L_{2R} the threshold is chosen to fix $\overline{\text{FDR}}$, thus leading to exactly the rule proposed in Genovese and Wasserman (2002). In fact, Genovese and Wasserman (2002) discuss 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. The loss function L_{2N} leads to a very similar solution, replacing the constraint on $\overline{\text{FDR}}$ by a constraint on $\overline{\text{FD}}$. An important implication of the different strategies for choosing the cutoff is the nature of $\overline{\text{FDR}}$ as a function of sample size J. Under L_{2R} it remains, by design, constant over J. This has awkward implications. Imagine the asymptotic case with large sample size when the true z_i are practically known. To achieve the desired $\overline{\text{FDR}}$ 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 equally pathological asymptotics. In contrast, under L_N the cutoff t is fixed across sample size, leading to vanishing $\overline{\text{FDR}}$ in the limit as $J \to \infty$, due to posterior consistency. However, it is interesting to include losses L_{2R} and L_{2N} in the discussion. The problem might only be of asymptotic relevance and not of concern for moderate sample sizes. We will further discuss the issue in the light of the results shown in Section 5. Apart from these concerns, all four loss functions are very similar with regard to their properties, nature of the inference and implementation details. We will therefore continue to consider all four loss functions in the upcoming discussion

2.2 Sample Size

2.2.1 Marginal FN and FNR

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

$$L^m_R(J) = E[\min_d \{L_R(d, y_J)\}], \quad L^m_{2R}(J) = E[\min_d \{\overline{\text{FNR}}(d, y_J) \mid \overline{\text{FDR}}(d, y_J) \leq \alpha\}],$$
$$L^m_N(J) = E[\min_d \{L_N(d, y_J)\}], \quad L^m_{2N}(J) = E[\min_d \{\overline{\text{FN}}(d, y_J) \mid \overline{\text{FD}}(d, y_J) \leq \alpha\}] \tag{4}$$

The sequence of alternating expectation and optimization is characteristic for sequential decision problems. See, for example DeGroot (1970) and Berger (1985), for a discussion of sequential decision problems in general. The expectation is with respect to the prior probability model on the data y_J under a given sample size J. The only argument left after the expectation and minimization is the sample size J. The nested minimization with respect to d is the solution of the multiple comparison problem. With all four loss functions it reduces to $d_i = I\{v_i > t(y)\}$. We will denote preposterior expected FNR by $\overline{\text{FNR}}_m(J, L) = E[\overline{\text{FNR}}(t^*_L, y_J)]$ with L denoting the loss function with respect to which the threshold t^* for the terminal decision is selected in the nested minimization. We use analogous definitions for $\overline{\text{FN}}_m$, $\overline{\text{FDR}}_m$ and $\overline{\text{FD}}_m$. Thus we could alternatively write (4) as $L^m_{2R}(J) = \overline{\text{FNR}}_m(J, L_{2R})$, $L^m_{2N}(J) = \overline{\text{FN}}_m(J, L_{2N})$, $L^m_R(J) = c\overline{\text{FDR}}_m(J, L_R) + \overline{\text{FNR}}_m(J, L_R)$, and $L^m_N(J) = c\overline{\text{FD}}_m(J, L_N) + \overline{\text{FN}}_m(J, L_N)$.

In summary, the following procedure emerges. Given the complexity of the expectations in (4), analytical computation is out of reach and both expectation and minimization must be approximated by simulation. For a grid of J's compute by prior Monte Carlo simulation the expectation in (4). For each simulated data set y_J we first compute the marginal posterior probabilities $v_i = P(z_i = 1|y_J)$, then we find the optimal threshold $t^*(y_J)$, and finally we substitute $d_i = I(v_i > t^*(y)_J)$ to evaluate $\overline{\text{FNR}}(t^*, y_J)$ and $\overline{\text{FN}}(t^*, y_J)$. The Monte Carlo average of simulated $\overline{\text{FNR}}$, $\overline{\text{FN}}$, $\overline{\text{FDR}}$ and $\overline{\text{FD}}$ for a given sample size provides an estimate of $\overline{\text{FNR}}_m$, $\overline{\text{FN}}_m$, $\overline{\text{FD}}_m$ and $\overline{\text{FDR}}_m$. Implementation details, including important simplifications to improve computational efficiency are discussed in the next section. One could then plot marginal FNR (or FN) against J to select a sample size. At this time one could add a (deterministic) sampling cost, if desired. But this would require the practically difficult choice of a relative weight for sampling cost versus inference loss. Alternatively, we take a goal programming perspective in fixing the sample size and use the plot of $L^m(J)$ versus sample size J to find a sample size for any set goal of $L^m(J)$.

However, in doing so a practical complication arises. For practically 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 3 for an example. The slow rate of decrease is a general feature of $\overline{\text{FNR}}$ and $\overline{\text{FN}}$.

Theorem 1 *Under the three loss functions L_{2R}, L_{2N} and L_N, $\overline{FNR}$ and $\overline{FN}$ decrease asymptotically as*

$$\overline{FNR}(t^*, y_J) = O_P(\sqrt{\log J/J}),$$

where t^* generically indicates the optiomal cutoff under each of the three loss functions, and

$$\overline{FN}(t^*, y_J) = O_P(n\sqrt{\log J/J}).$$

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

2.2.2 Power

For a practical sample size selection we propose to consider sensitivity diagnostics in addition to marginal loss, i.e., FNR or FN. We assume that the probability model includes parameters ρ_i that represent 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 4.1 we would use $\rho = \log\theta_1/\theta_0$. Recall that $v_i(y) = P(z_i = 1 \mid y)$ denotes the marginal posterior probability. To explore sensitivity with respect to the assumed prior on ρ_i we propose to evaluate

$$\beta(\rho_i) = P\{v_i(y) > t(y) \mid \rho_i\} = \int I(v_i(y) > t(y))\, \mathrm{d}p(y|\rho_i)\,. \tag{5}$$

The expectation is with respect to the joint probability model on the data y. In particular the expectation appropriately adjusts for dependencies, uncertainties on other model parameters, and the entire process of constructing the threshold $t(y)$. 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 $\beta(\rho)$ has interesting interpretations. We propose it because of the link with the traditional notion of power. The definition of β is essentially the power for one hypothesis in the multiple comparison, although with the twist of marginalizing with respect to all other parameters. At the same time $1 - \beta(\rho)$ is the contribution of one gene to the marginal loss function $L_{2N}(J)$, when we condition on $\rho_i = \rho$. Figure 6 shows a typical example.

Thus the following modification to the approach outlined in Section 2.2.1 emerges. The investigator fixes a level 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 6, together with FNR (or FN) in the marginal loss function allows the investigator an informed sample size choice. The FNR and FN plot adds the experimentwise dimension to the marginal summary 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.

A constructive description of the proposed sample size selection is as follows. We compute posterior probabilities of differential expression and posterior mean FNR for a grid of possible sample sizes and simulated data sets. Part of this simulation is the evaluation of $t^*(y_J)$ for each simulated data set y_J. This requires posterior simulation and evaluation of $\overline{\text{FDR}}$ for a grid of cutoffs. Appropriate Monte Carlo averages over simulated experiments y_J estimate $\overline{\text{FNR}}_m$ and $\overline{\text{FN}}_m$ under the four alternative loss functions. Ergodic averages over subsamples of genes stratified by (simulated) true differential expression estimate marginal power.

Plotting $\overline{\text{FNR}}_m$, $\overline{\text{FN}}_m$ and marginal power for an alternative of interest against sample size J allows then an informed sample size decision. Details of this simulation process are explicated in Section 3. We discuss some computational simplifications that greatly reduce the required simulation effort. We argue that finding J^* requires a computational effort comparable to posterior inference for one given data set.

3 Simulation

The described approach to sample size selection involves several calculations that are typically analytically intractable. Details depend on the specific probability model. Typically 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 we give a step-by-step algorithm we introduce notation and review the important steps in the algorithm in words. The discussion is still possible without reference to a particular probability model. We will introduce a specific probability model in Section 4.

For a given sample size J we simulate data $y_J \sim p(y_J)$. Simulating from the marginal $p(y_J) = \int p(y_J \mid \omega, z)\, \mathrm{d}p(\omega, z)$ is conveniently implemented by first generating "true" parameters (ω, z) from the prior, and then generating from the assumed sampling model $p(y_J \mid \omega, z)$ given the simulated parameter. Here (ω, z) are all unknown parameters in the model, including latent indicator variables z_i for differential expression of gene i. To distinguish this prior simulation from posterior MCMC simulation required later in the algorithm we will denote the realizations of this prior simulation by a ^{o} superindex as in ω^o, etc.

For each simulated data set y_J we compute the posterior mean process $\overline{\mathrm{FDR}}(t, y_J)$ and $\overline{\mathrm{FD}}(t, y_J)$ as functions of t, and find the appropriate cutoff. For example, for L_{2R} we find the cutoff $t^*_{2R}(y_J) = \min\{t : \overline{\mathrm{FDR}}(t, y_J) \leq \alpha\}$ that achieves a pre-determined posterior mean FDR level α. And similarly for the other loss functions. The only posterior summary required for evaluation of $\overline{\mathrm{FDR}}$ are the marginal posterior probabilities $v_i = p(z_i = 1 \mid y_J)$. We compute these by a posterior MCMC simulation.

Plugging in the threshold $t^*(y_J)$ we next compute posterior mean $\overline{\mathrm{FNR}}\,\{t^*(y_J), y_J\}$. and $\overline{\mathrm{FN}}\,\{t^*(y_J), y_J\}$. Averaging over y_J by (independent) Monte Carlo simulation $y_J \sim p(y_J)$ we compute

$$L^m_{2R}(J) = \overline{\mathrm{FNR}}_m(J, L_{2R}) = E_{y_J}\left\{\overline{\mathrm{FNR}}(t^*(y_J), y_J)\right\} \tag{6}$$

and similarly $L^m_N(J), L^m_R(J)$ and $L^m_{2N}(J)$. Plugging in $t^*(y_J)$ hinders interpretation of (6) as one joint integral with respect to the joint distribution $p(\omega, y_J)$ on parameters and data. Instead we need to proceed with two nested steps, as described above. Finally, evaluating (6) across J we find the sample size J^* to achieve a desired marginal expected FNR or FN.

The information in the FNR and FN curves is supplemented by power curves $\beta(\rho)$. The curve for $\beta(\rho)$ can be evaluated within the same simulation used for FNR and FN. For each simulated experiment we record (J, ρ^o_i, d_i), $i = 1, \ldots, n$. Here ρ^o_i is the true simulated level of differential expression. The recorded simulations are then arranged by J and ρ (possibly on a grid) and summarized in a plot like Figure 6.

Implementation is facilitated by several simplifications that increase computational efficiency. First, we will 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 $\overline{\mathrm{FNR}}_m(J, L)$ and $\overline{\mathrm{FN}}_m(J, L)$ 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 posterior expected $\overline{\mathrm{FDR}}(t, y_J)$ and $\overline{\mathrm{FNR}}(t, y_J)$. Both require posterior simulation $\omega \sim$

$p(\omega|y_J)$ by MCMC. In the context of the preposterior simulation we can start the MCMC at the true parameter values ω^o used to simulate the data y_J. Details are explained in the step by step algorithm below.

Finally, when computing $\widehat{L}(J)$ 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 $\widehat{L}^m(J)$ through all points of the scatterplot. This formalizes borrowing strength across different sample sizes J, exploiting the fact that $L^m(J)$ is smooth across J. In fact, we recommend to enforce the smooth fit $\widehat{L}^m$ to be monotone decreasing and to follow the $(\log J/J)$ asymptotics. We used a least squares fit of a linear regression of the observed $\overline{\text{FNR}}(t^*, y_J)$ values on $\sqrt{\log J/J}$. For comparison we fit a smoothing spline without any such constraints. The spline fit is practically undistinguishable from the simple regression, validating the use of the asymptotic law for the curve fitting. See Section 5.

Algorithm 1: Sample Size Determination

1. *Simulation:* Loop over repeated simulations $y_{J_1} \sim p(y_{J_1})$. Repeat the following steps for $m = 1, \ldots, M$ (To simplify notation we do not include an additional $_m$ subindex).

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

    1.2. *Data simulation:* $y_{J_1} \sim p(y_{J_1} \mid \omega^o, z^o)$.
    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 $y_J \subset y_{J_1}$ denote the size J subset of the maximal data set.

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

    a. Initialize MCMC posterior simulation at the true parameters, $(\omega, z) = (\omega^o, z^o)$.

    b. Simulate S transitions of the posterior MCMC.

    1.3.3. *Posterior probabilities:*
    Compute $v_i = P(z_i = 1|y_J)$ as the appropriate ergodic average and evaluate

$$\overline{\text{FD}}(t, y_J) = \sum (v_i > t)\,(1 - v_i) \text{ and } \overline{\text{FDR}}(t, y_J) = \frac{\overline{\text{FD}}(t, y_J)}{D + \epsilon}$$

    for $t \in \{v_1, \ldots, v_J\}$.

1.3.4. *Thresholds:* Compute the cutoffs t^*:

$$t^*_{2R}(y_J) = \min\{t : \overline{\mathrm{FDR}}(t, y_J) \le \alpha\},$$
$$t^*_{2N}(y_J) = \min\{t : \overline{\mathrm{FD}}(t, y_J) \le \alpha\},$$

$t^*_N(y_J) = c/(1+c)$ and $t^*_R(y_J) = v_{(n-D^*)}$, where D^* is given in (3).

1.3.5. $\overline{FN}(t^*_{2N}, y_J)$ *and* $\overline{FNR}(t^*_{2N}, y_J)$*:*
Using the the cutoff t^* evaluate $\overline{\mathrm{FN}} = \sum(v_i \le t^*(y_J))\, v_i$ and $\overline{\mathrm{FNR}} = \overline{\mathrm{FN}}/(n - D^* + \epsilon)$. Record the pairs $(J, \overline{\mathrm{FNR}})$ and $(J, \overline{\mathrm{FN}})$. separately for each loss (using the appropriate cutoff t^* in the definition of $\overline{\mathrm{FN}}$ and $\overline{\mathrm{FNR}}$).

1.3.6. *Power:* Let $d_i = I(v_i > t^*)$ and record the triples (J, ρ^o_i, d_i).

2. *Curve Fitting of Monte Carlo Experiments:*

   2.1. Expected loss $L^m(J)$*,* $\overline{FN}_m$ *and* $\overline{FNR}_m$*:* For each loss function fit a curve through the M observed pairs $(J, \overline{\mathrm{FNR}})$ and $(J, \overline{\mathrm{FN}})$ to estimate $\overline{\mathrm{FN}}_m(J, L)$ and $\overline{\mathrm{FNR}}_m(J, L)$, respectively. Use the asymptotic expressions reported in Theorem 1 to guide the curve fitting. Report estimated loss $\widehat{L}^m(J)$ as the appropriate combination of $\overline{\mathrm{FN}}_m$ and $\overline{\mathrm{FNR}}_m$ (see equation (4)).

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

3. *Optimal sample size:*
   Find J^* to achieve a desired $\widehat{L}^m(J)$. Alternatively, use $\widehat{L}^m(J)$ and power curves as in Figure 6 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. So far we described the approach without reference to a specific probability model. Any probability model that allows the simulations required to fix $t^*(y_J)$, and efficient marginal simulation $y_J \sim p(y_J)$ can be used. Evaluation of $t^*(y_J)$ essentially reduces to evaluation of marginal posterior probabilities $v_i = P(z_i = 1 \mid y)$. Thus the probability model needs to include, explicitly or as easily imputed latent data, indicators z_i for differential expression.

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 be easy to include prior data when available.

For the purpose of the design argument the model need not include all details of the data cleaning process, including spatial dependence of measurement errors across the microarray, correction for misalignments, etc. While such detail is critical for the analysis of 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. So in the following discussion we assume that the data are appropriately standardized and normalized and that the noise distribution implicitly include consideration of those. See, for example, Tseng et al. (2001), Baggerly et al. (2001), or Yang et al. (2002) for a discussion of the process of normalization.

4.1 A Hierarchical Gamma/Gamma Model

For the implementation in the example we choose a variation of the model introduced in (Newton et al., 2001; 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) propose probabilistic modeling for the observed gene frequencies in a single-slide experiment. For each gene i we record a pair (X_i, Y_i) of intensities corresponding to transcript abundance of a gene in the two samples. The true unknown mean expression levels are denoted by θ_{0i} and θ_{1i}. Other parameters, like scale or shape parameters, are denoted a. The aim of the experiment is to derive inference about the ratio θ_{1i}/θ_{0i}. Uncertainty about θ_{0i} and θ_{1i} is described by hyperparameters (a_0, ν, p). A latent variable $z_i \in \{0, 1\}$ is an indicator for equal mean values for gene i, i.e., equal expression. We use $z_i = 0$ to indicate equal expression, and $z_i = 1$ to indicate differential expression. Figure 1 summarizes the structure of the probability model. Conditional on the observed fluorescence intensities, the posterior distribution on z_i contains all information about differential expression of gene i. Let $r_i = X_i/Y_i$ denote the observed relative expression for gene i, and let $\eta = (a_0, \nu, p, a)$. Newton et al. (2001) gives the marginal likelihood likelihood $p(r_i \mid z_i, \eta)$, marginalized with respect to θ_{1i} and θ_{0i}. They proceed by maximizing the marginal likelihood for η by an implementation of the EM algorithm.

We will use a simple hierarchical extension of the described model by assuming repeated

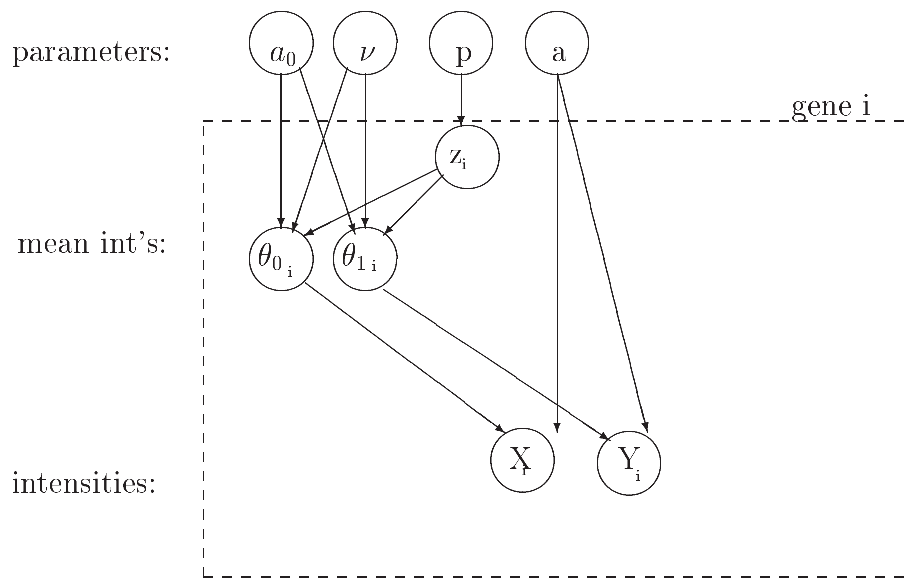


Figure 1: A model for differential gene expression from Newton et al. (2001) for fluorescence intensity measurements in a single slide experiment. For each gene, i.e., each spot on the chip, we record a pair (X, Y) of intensities corresponding to transcript abundance of a gene in both samples. The true unknown mean expression values are characterized by θ_{0i} and θ_{1i}. The aim of the experiment is to derive inference about equality of θ_{1i} and θ_{0k}. Uncertainty about θ_{0i} and θ_{1i} is described by parameters (a_0, ν). The variable z_i is an binomial indicator for equal mean values, i.e., equal expression, associated with a probability p. All information about differential expression of gene i is contained in the posterior distribution for z_i.

measurements X_{ij} and Y_{ij}, $j = 1, \ldots, J$, to be conditionally independent given the model parameters. We assume a Gamma sampling distribution for the observed intensities X_{ij}, Y_{ij} for gene i in sample j,

$$X_{ij} \sim Gamma(a, \theta_{0i}) \text{ and } Y_{ij} \sim Gamma(a, \theta_{1i}).$$

The scale parameters are gene specific random effects $(\theta_{0i}, \theta_{1i})$. The model includes an *a priori* positive probability for lack of differential expression

$$Pr(\theta_{0i} = \theta_{1i}) = Pr(z_i = 0) = p.$$

Conditional on latent indicators z_i for differential gene expression, $z_i = I(\theta_{0i} \neq \theta_{1i})$, we assume conjugate gamma random effects distributions

$$\begin{aligned} \theta_{0i} &\sim Gamma(a_0, \nu) \\ (\theta_{1i}|z_i = 1) \sim Gamma(a_0, \nu) &\text{ and } (\theta_{1i}|z_i = 0) \sim \mathbb{I}_{\theta_{0i}}(\theta_{1i}). \end{aligned} \tag{7}$$

The model is completed with a prior for the parameters $(a, a_0, \nu, p) \sim \pi(a, a_0, \nu, p)$ In the implementation for the example in Section 5 we fix ν. We assume *a priori* indpendence and use marginal gamma priors for a_0 and a, and a conjugate beta prior for p.

As in Newton et al. (2001) , the above model leads to a closed form marginal likelihood after integrating out θ_{1i}, θ_{0i}, but still conditional on $\eta = (p, a, a_0)$. Let $X_i = (X_{ij},\ j = 1, \ldots, J)$ and $Y_i = (Y_{ij},\ j = 1, \ldots, J)$. We find

$$p(X_i, Y_i | z_i = 0, \eta) = \left\{ \frac{\Gamma(2Ja + a_0)}{\Gamma(a)^{2J}\Gamma(a_0)} \right\} \frac{(\nu)^{a_0} \left(\prod_j X_{ij} \prod_j Y_{ij}\right)^{a-1}}{[(\sum_j X_i + \sum_j Y_i + \nu)]^{2a+a_0}}$$

and

$$p(X_i, Y_i | z_i = 1, \eta) = \left\{ \frac{\Gamma(aJ + a_0)}{\Gamma(a)^J \, \Gamma(a_0)} \right\}^2 \frac{(\nu\nu)^{a_0} \left(\prod_j X_{ij} \ \prod_j Y_{ij}\right)^{a-1}}{[(\sum_j X_{ij} + \nu)(\sum_j Y_{ij} + \nu)]^{a+a_0}},$$

and thus the marginal distribution is

$$p(X_i, Y_i | \eta) = p \, p(X_i, Y_i \mid z_i = 0, \eta) + (1 - p) \, p(X_i, Y_i \mid z_i = 1, \eta) \tag{8}$$

Availability of the closed form expression for the marginal likelihood greatly simplifies posterior simulation. Marginalizing with respect to the random effects reduces the model to the 3-dimensional marginal posterior $p(\eta \mid y) \propto p(\eta) \prod_i p(X_i, Y_i|\eta)$. Conditional on currently imputed values for η we can at any time augment the parameter vector by generating $z_i \sim p(z_i \mid \eta, X_i, Y_i)$ as simple independent Bernoulli draws, if desired. This greatly simplifies posterior simulation.

4.2 A Mixture Model Extension: Sample Size Choice with Pilot Data

One limitation of a parametric model like the hierarchical Gamma/Gamma model discussed before is the need to fix specific model assumptions. The investigator has to select hyper-parameters that reflect the relevant experimental conditions. Also, the investigator has to assume that the sampling distribution for observed gene expressions can adequately be approximated by the assumed model. To mitigate problems related with these requirements we consider a model extension that still maintains the computational simplicity of the basic model, but allows for additional flexibility. In particular, we want a model that allows the use of a pilot data set to learn about the sampling distribution of observed gene expressions across genes and repeated samples. We envision a system where the investigator collects some pilot data (on control tissue) before going through the sample size argument. These pilot data could then be used to learn about important features of the sampling distribution. If the observed pilot data can be adequately fit by the marginal model $p_0(X_i|z_i = 0)$ under the Gamma/Gamma hierarchical model then the sample size design should proceed as before. If, however, 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 inference is a scale mixture extension of the basic model. In particular, we will replace the Gamma distributions for $p(X_{ij}|\theta_{0i})$ and $p(Y_{ij}|\theta_{1i})$ by scale mixtures of Gamma distributions

$$X_{ij} \sim \int Ga(a, \theta_{0i}\, r_{ij})\, \mathrm{d}p(r_{ij}|w, m) \text{ and } Y_{ij} \sim \int Ga(a, \theta_{1i}\, s_{ij})\, \mathrm{d}p(s_{ij}|w, m) \tag{9}$$

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 high prior probability for large weight w_1. We use the same mixture for s_{jk}, $P(s_{jk} = m_h) = w_h$. The model is completed with $m_k \sim Ga(b, b)$, $k > 1$ and a Dirichlet prior $w \sim Dir_K(M \cdot W, W, \ldots, W)$. Selecting a large factor M in the Dirichlet prior assigns high prior probability for large w_1, as desired. By assuming a dominating term with $m_1 = 1.0$ and $E(m_k) = 1$, $k > 1$, we allocate large prior probability for the basic model and maintain the interpretation of θ_{0i}/θ_{1i} as level of differential expression.

Model (9) thus replaces the Gamma sampling distribution in the Gamma/Gamma hierarchical model (7) with a scale mixture of Gamma distributions. This is important in the context of microarray data experiments, where technical details in the data collection process typically introduce noise beyond simple sampling variability due to the biological process. A concern related to microarray data experiments prompts us to introduce a further generalization to allow for occasional slides that are outliers compared to the other arrays in the experiment. This happens for reasons unrelated to the biologic effect of interest but needs to be accounted for in the modeling. We achieve this by adding a second mixture to (9)

$$(X_{ij}|r_{ij}, g_j) \sim Ga(a, \theta_{0i}\, g_j\, r_{ij}) \text{ and } (Y_{ij}|s_{ij}, g_j) \sim Ga(a, \theta_{1i}\, g_j\, s_{ij}), \tag{10}$$

with an additional slide specific scale factor g_j. Paralleling the definition of $p(r_{ij}|w, m)$ we use a finite discrete mixture $P(g_j = m_{gk}) = w_{gk},\ k = 1, \ldots, L$ with a Dirichlet prior $(w_{g1}, \ldots m_{gL}) \sim Dir_L(M_g \cdot W_g, W_g, \ldots, W_g)$, $m_{gk} \sim Ga(b_g, b_g)$ for $k > 1$ and $m_{g1} \equiv 1$.

An important feature of the proposed mixture model is computational simplicity. We will proceed in two stages. In a first stage the pilot data is used to fit the mixture model. Let X^o_{ij}, $j = 1, \ldots, J^o$, denote the pilot data. We will use posterior MCMC simulation to estimate the posterior mean model. This is done once, before starting the optimal design. Posterior simulation in mixture models like (9) is a standard problem. We include reversible jump moves to allow for random size mixtures. Our reversible jump implementation includes a merge move to combine two terms in the current mixture, a matching split move, a birth move and a matching death move. Details are similar to Richardson and Green (1997), with the mixture of gammas replacing the mixture of normals. Inference is based on a geometric prior on the number of terms K and L in both mixtures.

We then fix the mixture model at the posterior modes $\widehat{K}$ and $\widehat{L}$, and the posterior means $(\bar{w}, \bar{m}, \bar{w}_g, \bar{m}_g) = E(w, m, w_g, m_g \mid X^o, \widehat{K}, \widehat{L})$. We proceed with the optimal sample size approach, using model (9) with the fixed mixtures. The procedure, including all posterior and marginal simulation, is exactly as before, with only one modification. We add an additional 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 pilot data.

Algorithm 2: Sample Size Determination with Pilot Data

1. *Pilot data:* Assume pilot data $X^o = \{X^o_{ij}, i = 1, \ldots, n, j = 1, \ldots, J^o\}$, from control tissue is available.

2. *Mixture model:* Estimate the mixture model and report the posterior modes $(\widehat{K}, \widehat{L})$, and

the conditional posterior means $(\bar{w}, \bar{m}, \bar{w}_g, \bar{m}_g) = E(w, m, w_g, m_g \mid X^o, \widehat{K}, \widehat{L})$. Both are computed by posterior MCMC simulation for the mixture model (9).

3. *Optimal Sample Size:* Proceed as in Algorithm 1, replacing X_{ij} with $X_{ij}/(r_{ij}g_j)$ and Y_{ij} by $Y_{ij}/(s_{ij}g_j)$.

   The indicators are initialized with the (true) values from the data simulation (Step 1.2. in Algorithm 1). Updating s_{ij}, r_{ij} and g_j adds an additional step in the posterior simulation (Step 1.3.1.b of Algorithm 1). The mixture model parameters remain fixed at $(\bar{w}, \bar{m}, \bar{w}_g, \bar{m}_g, \widehat{K}, \widehat{L})$.

5 Implementation

Consider the data reported in Richmond et al. (1999). The data are also used as illustration in Newton et al. (2001). We use the control data to plan for a hypothetical future experiment. We proceed as proposed in Section 4.2. First we estimate the mixture model (9), using the available control data as a pilot data set. Estimation of (10) is implemented as a straightforward Markov chain Monte Carlo posterior simulation with reversible jump moves, as described in Section 4.2. The posterior mode for the size of the mixture models was found at $K = 3$ and $L = 2$.

To define the probability model for the design calculations we fixed $K = 3$ and $L = 2$ and set the mixture model parameters (μ, w, λ, u) at their posterior means (conditional on the fixed size of the mixture). Maintaining the mixture parameters random also in the design model would not significantly complicate the procedure, but would not contribute much to the final inference. Implementing Algorithm 2 we compute expected losses, and power $\beta(\rho)$ across a grid of sample sizes J. Summaries are reported in Figures 2 through 6. Recall that $\overline{\text{FN}}_m(J, L)$ and $\overline{\text{FNR}}_m(J, L)$ denote the preposterior expectations of $\overline{\text{FN}}$ and $\overline{\text{FNR}}$. We will use analogous notation $\overline{\text{D}}_m(J, L)$ and $\overline{t^*}_m(J, L)$ to denote preposterior expectations of the number of discoveries D and the threshold t^*, computed under loss L and sample size J, defined analoguously to $\overline{\text{FNR}}_m$ and $\overline{\text{FN}}_m$. All inference was computed in one run of the program, collecting the necessary Monte Carlo averages for all four alternative loss functions.

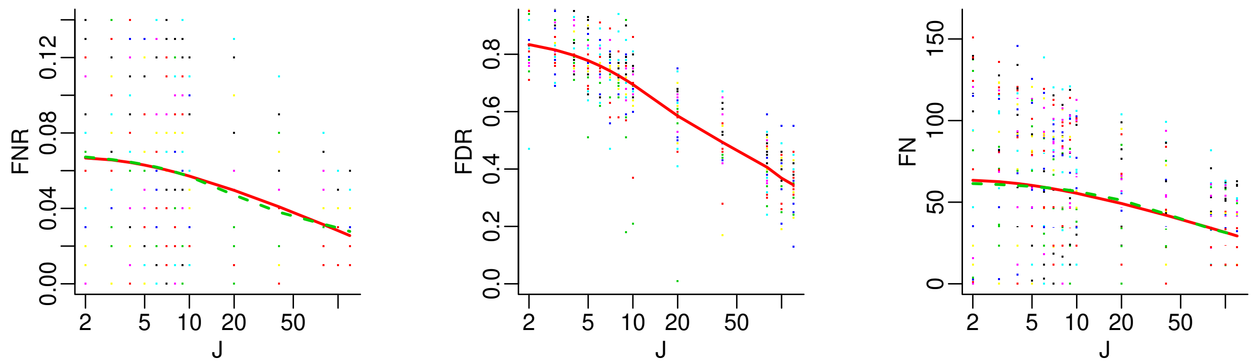
Figure 2 summarizes inference under L_N. Let $\bar{p} = E(p)$ denote the prior mean for the probability of differential expression. The trade-off parameter c is set to $c = 2\cdot\bar{p}/(1-\bar{p}) = 0.2$. Consider L_N under the two extreme choices $t^* = 1.0$ and $t^* = 0$. Choosing $c = \bar{p}/(1-\bar{p})$ would exactly match L_N under both, equally undesirable extremes. The additional factor 2 was fixed after some trial and error. Under L_N, the threshold t^* is fixed at $t^* = c/(1-c)$. The fixed threshold leads to decreasing FNR and FDR as a function of sample size. The estimated curves $\overline{\text{FNR}}_m(J, L)$ and $\overline{\text{FN}}_m(J, L)$ were derived by fitting a linear model with predictor $\sqrt{\log(J+1)/(J+1)}$ to the observed pairs $(J, \overline{\text{FNR}}(t_N^*, y_J))$. This is motivated by the asympototic results of Theorem 1 (with the offset +1 to avoid a singularity at $J = 0$). For comparison we estimated the same curve using a cubic smoothing spline, using the `smoothing.spline` function in R with default parameters. The corresponding curves for $\overline{\text{FNR}}_m$ and $\overline{\text{FN}}_m$ are shown as dashed lines. For $\overline{\text{FDR}}_m$, $\overline{\text{FD}}_m$, $\overline{\text{D}}_m$ and $\overline{t^*}_m$ we used cubic smoothing splines to estimate the mean value as a function of J.

Figure 3 shows inference for the loss function L_{2R}. The threshold on $\overline{\text{FDR}}$ was set to $\alpha = 0.40$. The posterior mean $\overline{\text{FDR}}$ is fixed by design of the optimal decision approach under L_{2R}. To maintain the fixed $\overline{\text{FDR}}$ the procedure has to eventually start lowering the threshold t^* to reject comparisons with increasingly lower posterior probability of differential expression.

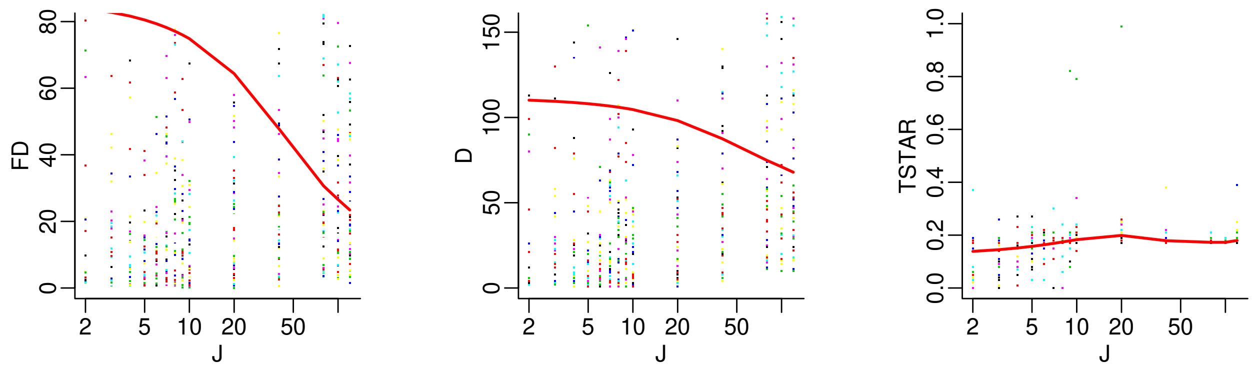
Figure 4 shows inference for the loss function L_{2N}. We set the threshold for $\overline{\text{FD}}$ as $\alpha_N = 7.1$. The value is computed as $\alpha_N = 0.1\ n\bar{p}\alpha/(1-\alpha)$, chosen to match a bound $\overline{\text{FDR}} \leq \alpha$ under the assumption that 10% of the true differential expressions are discovered. Overall, performance is more balanced than L_{2R}. Fixing the count $\overline{\text{FD}}$ instead of the rate $\overline{\text{FDR}}$ slows the awkward decrease in the threshold that was required to maintain the fixed false discovery rate 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}.

Finally, Figure 5 repeats similar plots for inference under L_R. The tradeoff factor c is fixed as in L_N. At an intermediate sample size the threshold t^* swiftly moves from an initial value $t^* \approx 0$ to the other extreme $t^* \approx 1$. In contrast to the other three loss functions the optimal decision under L_R does not constrain an error, error rate or cutoff. This allows the sudden jump in $\overline{\text{FDR}}$ and t^* that can be observed in Figure 5.

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 $\overline{\text{FDR}}$, $\overline{\text{FD}}$, and t^*, respectively. Under L_R the tradeoff is explicitly included as a coefficient in the loss function. The constraint on $\overline{\text{FDR}}$ under L_{2R} has the awkward implication that with increasing sample size we have to 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 $\overline{\text{FDR}}$ and t^*. This leads us to favor L_{2N} and L_N. Both lead to very similar inference, with L_{2N} having

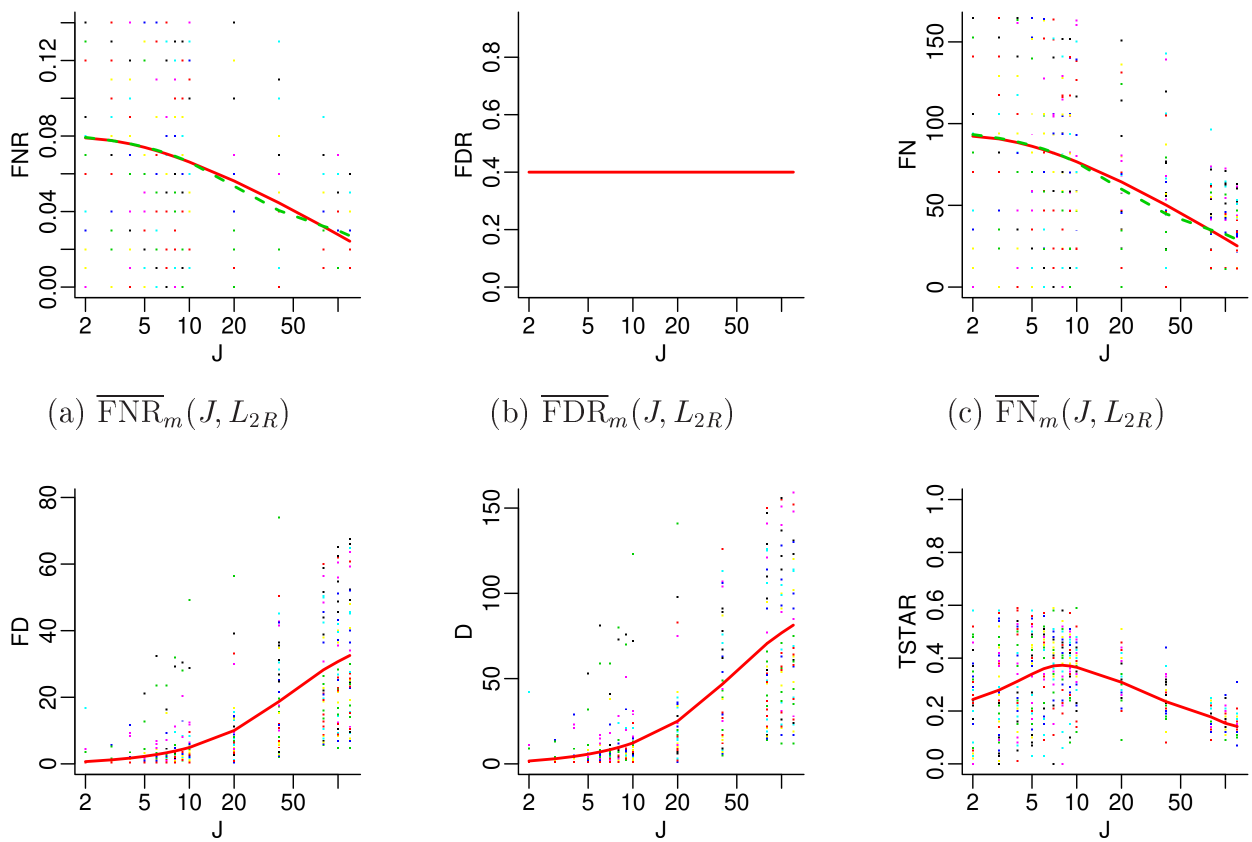


(a) $\overline{\mathrm{FNR}}_m(J, L_N)$ (b) $\overline{\mathrm{FDR}}_m(J, L_N)$ (c) $\overline{\mathrm{FN}}(J, L_N)$



(d) $\overline{\mathrm{FD}}_m(J, L_N)$ (e) $\overline{\mathrm{D}}_m(J, L_N)$ (f) $\overline{t^*}_m(J, L_N)$

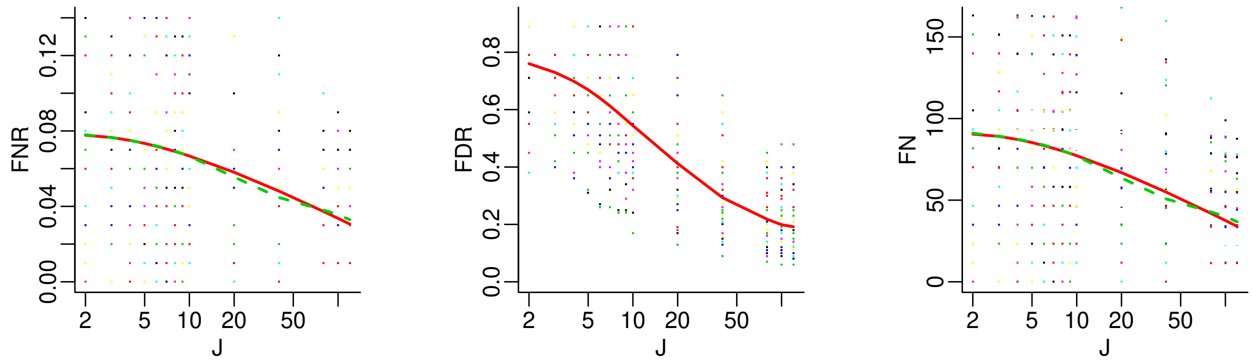
Figure 2: L_N: Expected loss and other relevant summaries. Expectations indicated with $\overline{X}_m(J, L_N)$ are with respect to the marginal distribution on the data, i.e., $\overline{X}_m(\cdot) \equiv \int X \, \mathrm{d}p(y_J)$. Panel (a) shows the expected loss function $\overline{\mathrm{FNR}}_m$. Panels (b) through (e) plot the expected $\overline{\mathrm{FDR}}_m$, the expected counts $\overline{\mathrm{FN}}_m$ and $\overline{\mathrm{FD}}_m$, the expected number of discoveries $\overline{\mathrm{D}}_m$, and the average cutoff $\overline{t^*}_m$. In each panel, the dots show the values recorded in each of the simulations, and the lines show the estimated marginal expectations. To allow comparison across the different loss functions the vertical limits are matched across Figures 2 through 5. This leads to some of the simulation points being outside the range in some panels, like panel (d) in this figure. The cutoff t^* is fixed by design, leading to decreasing $\overline{\mathrm{FDR}}_m$ and $\overline{\mathrm{FNR}}_m$. The dashed curves for $\overline{\mathrm{FNR}}_m$ and $\overline{\mathrm{FN}}_m$ (almost undistinguishable from the solid line) show an alternative curve fit. See the text for an explaination of the curve fit.



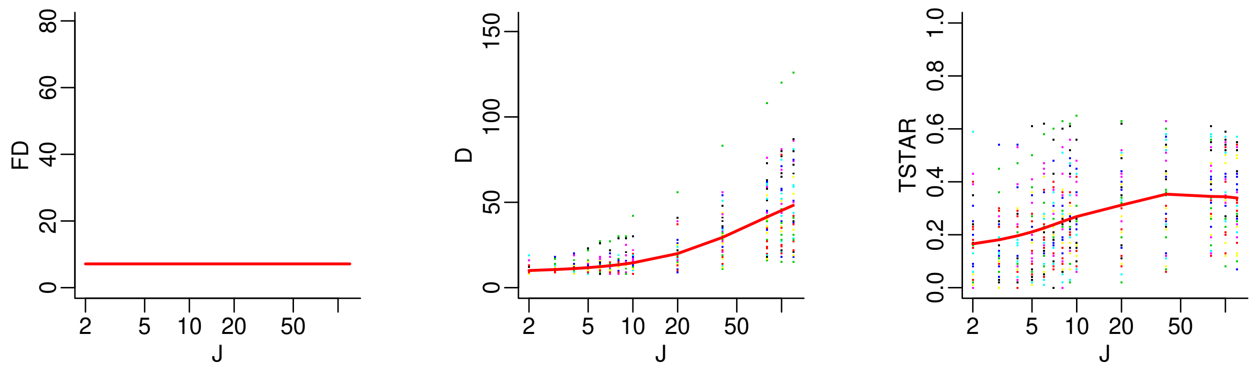
(a) $\overline{\mathrm{FNR}}_m(J, L_{2R})$ (b) $\overline{\mathrm{FDR}}_m(J, L_{2R})$ (c) $\overline{\mathrm{FN}}_m(J, L_{2R})$

(d) $\overline{\mathrm{FD}}_m(J, L)$ (e) $\overline{\mathrm{D}}_m(J, L_{2R})$ (f) $\overline{t^*}_m(J, L_{2R})$

Figure 3: L_{2R}: Same as in Figure 2. Inference under L_{2R} fixes $\overline{\mathrm{FDR}}_m$ by design. For large J the constraint on $\overline{\mathrm{FDR}}_m$ requires the procedure to include an increasing number of genes with $v_i \approx 0$, leading to the steep decline in t^* (notice the log scale on J) and the matching increase in $\overline{\mathrm{FD}}_m$ and $\overline{\mathrm{D}}_m$.

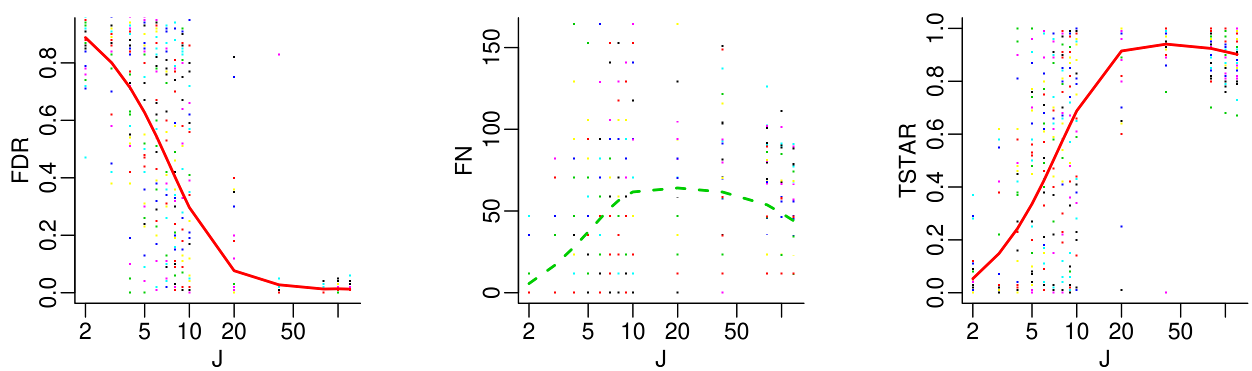


(a) $\overline{\mathrm{FNR}}_m(J, L_{2N})$ (b) $\overline{\mathrm{FDR}}_m(J, L_{2N})$ (c) $\overline{\mathrm{FN}}_m(J, L_{2N})$



(d) $\overline{\mathrm{FD}}_m(J, L_{2n})$ (e) $\overline{\mathrm{D}}_m(J, L_{2n})$ (f) $\overline{t^*}_m(J, L_{2N})$

Figure 4: L_{2N}: same as Figure 2. Under L_{2N}, the false discovery count $\overline{\mathrm{FD}}$, and thus the preposterior expectation $\overline{\mathrm{FD}}_m$ are fixed. To achieve the fixed number of false discoveries the number of discoveries has to increase with J, leading to a decreasing $\overline{\mathrm{FDR}}_m$.



(a) $\overline{\mathrm{FDR}}_m(J, L_R)$ (b) $\overline{\mathrm{FN}}_m(J, L_R)$ (c) $\overline{t^*}_m(J, L_R)$

Figure 5: L_R: same as the corresponding panels in Figure 3. The undesirable jump in $\overline{\mathrm{FDR}}_m$ (and $\overline{t^*}_m$) make L_R impractical to use.

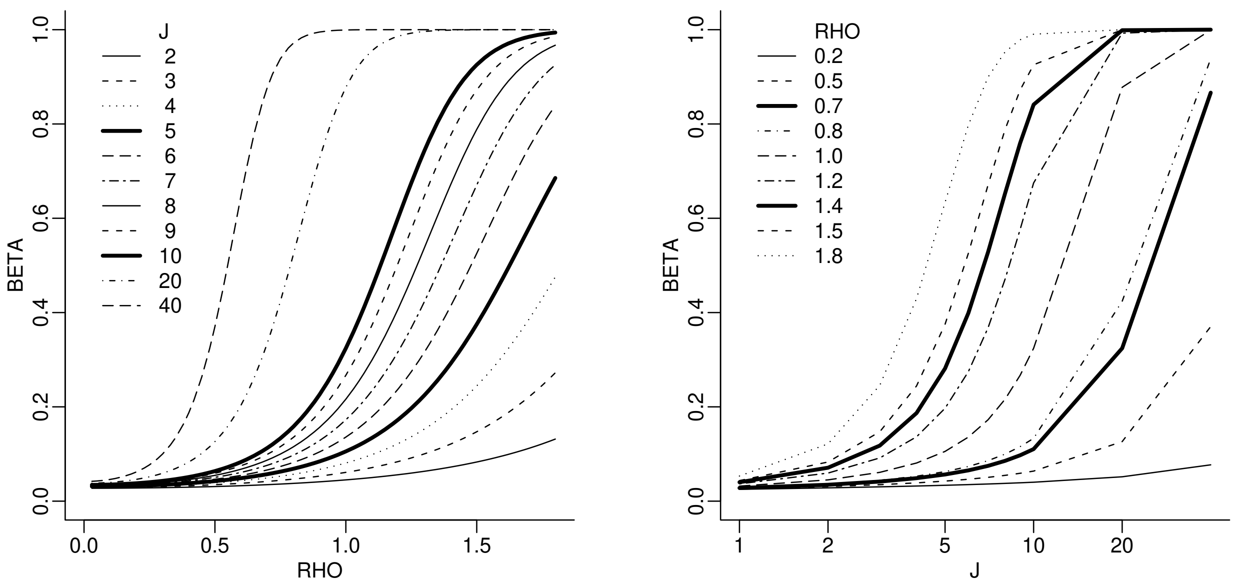
the advantage that the constraint is on the practically more important $\overline{\mathrm{FD}}$, rather than t^*, as in L_N.

For any loss function we advise to consider additional sensitivity diagnostic and power curves to help guide a sample size choice. Figure 6a shows power curves $\beta(\rho)$ as proposed in Section 2.2.2. Panel (a) plots power against the assumed true level of differential expression ρ, with a separate curve for each sample size J. Figure 6b plots the same summary against sample size J, arranged by level of differential expression ρ. In practice a sample size argument would then proceed as follows. First the investigator determines a level of differential expression that would be considered biologically meaningful, say two fold expression at $\rho = \log 2$. Using a pilot data set we proceed with Algorithm 2 to compute expected FNR, FN, and power across sample sizes. Inspection of the power plot for the level ρ of interest, together with the FNR and FN plots informs the investigator about the minimum sample size needed to achieve desired power and/or error rates.

6 Conclusion

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 need to be evaluated in the allocation of resources to alternative investigations. Even in the simplest two-sample comparison setting, micorarray analyses pose difficult challenges to traditional sample size approaches: first, in hypothesis testing terms, they present with 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; third, they mandate incorporation of existing knowledge, as signal-to-noise ratios vary significantly with the specific technology, the source of RNA, and a laboratory's overall experience.

Our goal in this article has been to develop a formal decision theoretic framework to address these challenges. This provides investigators the opportunity to quantify both the *a priori* uncertainty about likely expression levels and the implications of sample size choices on the performance of inference about differential expression. 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 in noisy data: the methods we



(a) $\beta = E_{y_J}\{P(d=1|\rho, y_J)\}$ against ρ (by J)

(b) β against J (by ρ)

Figure 6: Power against true effect ρ and against sample size J. In panel (b) power curves for two fold ($\rho = \log 2$) and four fold over expression ($\rho = \log 4$) are highlighted in bold. Here, power $\beta(\rho)$ is defined as the average posterior probability of discovery, conditional on the true level of differential expression $\rho_i = \log(\theta_{0i}/\theta_{1i})$. The probability is marginalized with respect to all other model parameters, the data y_J under sample size J, and the process of fixing the threshold t^*. Compare (5).

proposed would be applicable more generally to those problems as well.

In complex decision making situations, decision models such as ours are best thought of as a decision support tools. As is common in simpler settings, we envision investigators to explore 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 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, while 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 comparison are often made to produce exploratory analyses, such as clusters. The choice of sample sizes in these circumstances is different from controlled experiments. Some insight is offered by Simon et al. (2002) and Bryan and van der Laan (2001).

An important practical indication for microarray design arises from the illustration of Section 5. In particular, for a realistic set of parameters and pilot data, we show that the improvement in genome-wide error rate appears to be non-concave, 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 negligible. This has implications for the common practice of planning of experiments with only two or three replicates. We suggest that an analysis of the kind presented in Figure 2 would be valuable information for investigators entertaining experiments with a very small number of replicates.

Appendix 1: Asymptotic FNR

We discuss some asymptotic results, characterizing posterior probability of differential expression as J goes to infinity. The results help interpretation of the plots shown in Section 5. In the following discussion we consider a model with the same structure as in Section 4.1. The specific distributional assumptions, including the Gamma sampling distribution for (X_{ij}, Y_{ij}) and the Gamma prior for $(\theta_{0i}, \theta_{1i})$, are not critical. We start the argument by establishing an asymptotic approximation for $P(z_j = 1|y_J)$. We will then use this result to argue that for large J the rejection region has to necessarily include some genes with small or zero true differential expression. This is true under all three loss function, L_{2R}, L_{2N} and L_N. Thus the non-rejection region includes only small true differential expression. We exploit this fact to approximate $\overline{\text{FNR}}$ by an integral that can be recognized as an expression of order $\sqrt{\log J/J}$. The integral approximation is valid only if a large number of genes are in the non-rejection region, allowing us to approximate the sum in the definition of $\overline{\text{FNR}}$ by an integral. We conclude the argument by showing that this is the case under all three loss functions, for sufficiently large J.

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

$$\begin{aligned} P(z_i = 1|y) = \int P(z_i = 1|y_i, \eta)\mathrm{d}p(\eta|y_i) &= P(z_i = 1|y_i, \hat{\eta})(1 + O_P(n^{-1/2})) \\ &= P(z_i = 1|y_i, \eta)(1 + O_P(n^{-1/2})), \quad (11) \end{aligned}$$

where $\hat{\eta}$ is the maximum likelihood estimator, and η are the true hyperparameters. Here $X_n = O_P(n^k)$ 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 η is

$$P(z_i = 1|y_i, \eta) = \frac{p\ p(y_i|z_i = 1, \eta)}{p\ p(y_i|z_i = 1, \eta) + (1-p)\ p(y_i|z_i = 0, \eta)}$$

Classical Laplace expansions imply that

$$P(z_i = 1|y_i, \eta) = \frac{1}{1 + c_i e^{-J(\hat{\theta}_{0i} - \hat{\theta}_{1i})^2 \tau_i/2} \sqrt{J}} \quad (12)$$

$c_i, \tau_i = O_P(1)$ as J goes to infinity. The constant c_i includes the ratio $(1-p)/p$. Under suitable regularity conditions this result is uniform in $(\theta_{0i}, \theta_{1i}, \eta)$ over compact sets. In the non compact case, some conditions on the tails of the priors need to be added. See, for example, Guihenneuc and Rousseau (2002). Therefore, when $|\theta_{0i} - \theta_{1i}|$ is large $p(z_i = 1|y_i, \eta)$ goes to 1 at an exponential rate and thus $P(z_i = 1|y_i)$ is very close to 1 (the error being essentially of order n^{-1}).

We now use (12) to study asymptotic behavior of the terminal decision. In particular we consider $\overline{\text{FDR}}$, $\overline{\text{FD}}$, $\overline{\text{FNR}}$ and $\overline{\text{FN}}$. Let $v_{(1)} \leq \cdots \leq v_{(N)}$ be the ordered posterior probabilities $v_i = P(z_i = 1|y)$ and recall $\overline{\text{FDR}}(t, y) = \sum_i (1 - v_i)\, I(v_i \geq t)/D$, where $D = \sum_i I(v_i \geq t)$ is the number of discoveries. We will use $N = \sum I(v_i < t)$, $\text{FP} = \sum I(v_i > t) I(z_i = 0)$, $n_1 = \sum I(z_i = 1)$ and $n_0 = n - n_1$ to denote the number of negatives, false positives, differentially expressed and non-differentially expressed genes, respectively. We will use A_N, A_{FP}, A_1, and A_0 to denote the corresponding sets of genes. The above expansions show that the ordering of v_i is asymptotically linked to the ordering of $|\hat{\theta}_{0i} - \hat{\theta}_{1i}|$, with v_i monotone increasing in $|\hat{\theta}_{0i} - \hat{\theta}_{1i}|$, with asymptotically

$$v_i \approx 1 - c_i \sqrt{J} \exp[-J(\hat{\theta}_{0i} - \hat{\theta}_{1i})^2 \tau_i/2].$$

False discovery rate $\overline{\text{FDR}}(t, y)$ as a function of t is a step function taking values in $\{1 - v_{(n)},\ ...,\ 1 - (v_{(k)} + ... + v_{(n)})/(n - k + 1),\ ...,\ 1 - (v_{(1)} + ... + v_{(n)})/n\}$. Similarly, $\overline{\text{FD}}(t, y)$ is a step function with values $\{1 - v_{(n)},\ ...,\ 1 - v_{(1)} + ... + 1 - v_{(n)})\}$. Both are monotone decreasing in t. For large J, the earlier discussion shows that any gene with

$$|\hat{\theta}_{0i} - \hat{\theta}_{1i}| > C\sqrt{\log J}/\sqrt{J} \tag{13}$$

has posterior probability $v_i \geq 1 - 1/\sqrt{J}$, when C is large enough, uniformly in θ_{0i}, θ_{1i} belonging to some compact set, and with large probability. We denote with

$$S = \{i : \ |\hat{\theta}_{0i} - \hat{\theta}_{1i}| < C\sqrt{\log J}/\sqrt{J}\}$$

the set of genes with small $|\hat{\theta}_{0i} - \hat{\theta}_{1i}|$ that violate (13).

We now show that under all three losses, only genes with small $|\hat{\theta}_{0i} - \hat{\theta}_{1i}|$ are classified as non-differentially expressed, i.e., $A_N \subseteq S$.

Under L_N the argument is straightforward. For all genes satisfying (13) the posterior probability $v_i \approx 1 - 1/\sqrt{J}$ is beyond $t_N = c/(1-c)$ for sufficiently large J. Thus all genes in A_N satisfy $|\hat{\theta}_{0i} - \hat{\theta}_{1i}| < C\sqrt{\log J/J}$.

To prove the claim under L_{2R} we show that the opposite would violate the constraint on $\overline{\text{FDR}}$. Assume that (13) holds for all $i \in A_D$. Then

$$\overline{\text{FDR}} = 1 - (v_{(n-D+1)} + ... + v_{(n)})/D \leq 1/\sqrt{J}.$$

which is not enough to reach the set level α bound required for L_{2R}. Thus the rejection region A_D has to necessarily include some genes that violate (13). Together with the monotonicity of $|\hat{\theta}_{0i} - \hat{\theta}_{1i}|$ as a function of v_i this proves the claim.

Finally, to show the same for L_{2N} consider (13) with an even larger C. If $C^2 > 1/\tau_i[\log n - \log(\alpha/2)]$, then $1 - v_{(i)} \leq \alpha/(2n)$ for all genes that satisfy (13) with such C. If only such genes are considered in the rejection region then

$$1 - v_{(k)} + ... + 1 - v_{(N)} \leq \alpha/2,$$

which is not enough to reach the desired bound $\overline{\text{FN}} = \alpha$ under L_{2N}.

We now use (12) and the fact that all negatives have small $|\hat{\theta}_{0i} - \hat{\theta}_{1i}|$ to establish a bound on $\overline{\text{FNR}}$.

$$\overline{\text{FNR}}(t_*, y) = \frac{1}{N} \sum_{j=1}^{N} v_{(j)} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{1 + c_j \sqrt{J} e^{-J(\hat{\theta}_{0j} - \hat{\theta}_{1j})^2 \tau_j/2} (1 + O_P(n^{-1/2})},$$

where $c_j = \sqrt{\tau_j}/\sqrt{2\pi}$ with $\tau_j = i(\theta_{0j})i(\theta_{1j})/(i(\theta_{0j}) + i(\theta_{1j}))$ and $i(\theta)$ being the Fisher information associated with the conditional model of X_i (or Y_i) given θ, η, when η is fixed.

If N is large the sum can be approximated by an integral. The integral is with respect to the distribution of $v_{(j)}$ or, equivalently, the distribution of $(\hat{\theta}_{0j}, \hat{\theta}_{1j})$. We split the integral into two parts. With probability w_0 we have $\theta_{0i} = \theta_{1i} \equiv \theta_i$ and with probability w_1 we have $\theta_{0i} \neq \theta_{1i}$. Based on the earlier observation that we only fail to reject for small estimated differences, we can condition the latter term on $|\hat{\theta}_{0i} - \hat{\theta}_{1i}| < C\sqrt{\log J}/\sqrt{J}$. Let $\sqrt{J}(\hat{\theta}_{0i} - \hat{\theta}_{1i})\sqrt{\tau_j} = \sqrt{J}(\theta_{j0} - \theta_{j1})\sqrt{\tau_j} + \xi_j$, where ξ_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,

$$\begin{aligned}
\overline{\text{FNR}}(y, t_{2R}) &\approx w_0 \int_\xi \int_\theta \frac{1}{1 + \sqrt{i(\theta)}/\sqrt{2\pi}(1-p)/pe^{-\xi^2/2}\sqrt{J}} \mathrm{d}p(\xi)\mathrm{d}p(\theta) \\
&+ \quad w_1 \int_\xi \int_{\Theta_S} \frac{1}{1 + c(\theta_0, \theta_1)\sqrt{J}\, e^{-\xi^2/2} e^{-J(\theta_1 - \theta_0)^2 \tau(\theta_0, \theta_1)/2}} \mathrm{d}p(\xi)\mathrm{d}p(\theta_0, \theta_1).
\end{aligned}$$

Simple calculations imply that the above quantities are of order $\sqrt{\log J/J}$, when N is large.

Moreover, $\overline{\text{FN}} = N\,\overline{\text{FNR}}$. We now prove that $n/N = O_P(1)$ with high probability under all three losses.

We start with the argument for L_{2R}. Under the assumed sampling model $n_0 \approx p \cdot n$ genes satisfy $\theta_{0j} = \theta_{1j}$. If $N/n \to 0$, then a large proportion of genes satisfying $\theta_{0j} = \theta_{1j}$ would have posterior probabilities $v_j > t_{2R}$. Recall that A_{FP} is the set of false positives. This would imply that

$$\begin{aligned}
\overline{\text{FDR}} &\geq \frac{1}{n} \sum_{i \in A_{FP}} (1 - v_i) \\
&= \frac{1}{n} \sum_{i \in A_{FP}} \frac{c_j \sqrt{J} e^{-J(\hat{\theta}_{0i} - \hat{\theta}_{1i})^2 \tau_j / 2}}{1 + c_j \sqrt{J} e^{-J(\hat{\theta}_{0i} - \hat{\theta}_{1i})^2 \tau_j / 2}} (1 + O_P(n^{-1/2})) \\
&\geq \frac{p}{2} \int_\theta \int_\xi \frac{\sqrt{J} c(\theta) e^{-\xi^2/2}}{1 + \sqrt{J} c(\theta) e^{-\xi^2/2}} \mathrm{d}\xi \mathrm{d}p(\theta) (1 + O_P(n^{-1/2}),
\end{aligned}$$

when n is large enough, with high probability. The last inequality is true since under the assumption $N/n \to 0$ eventually more than $N/2 \approx np/2$ genes would be in A_{FP}. As J goes to infinity, the above term goes to $p/2$. This is a contradiction if $\alpha < p/2$, and we thus conclude $n/N = O_P(1)$.

Under L_{2N} we use an analogous argument for $\overline{\text{FD}}$. The right hand side in the first two (in-)equalities above remains unchanged, except for removing the leading $1/n$ factor. We conclude that $\overline{\text{FD}} \geq np/2$ and thus have a contradiction for $\alpha < np/2$.

Finally, under L_N, $t_N = c/(c+1)$, so $v_j \leq t_N \Leftrightarrow$

$$1 \leq c\, c_j e^{-J(\hat{\theta}_{0i} - \hat{\theta}_{1i})^2 \tau_j / 2} \sqrt{J} (1 + O_P(n^{-1/2}).$$

The number of genes $v_j \leq t_N$ is large with high probability. Indeed, if $\theta_{0i} = \theta_{1i}$,

$$P\left[1 > c\, c_j e^{-J(\hat{\theta}_{0i} - \hat{\theta}_{1i})^2 \tau_j / 2} \sqrt{J}\right] = O(J^{-1/2})$$

by Chebychev's inequality. Recall that FP is the number of genes satisfying $\theta_{0i} = \theta_{1i}$ and $v_i > t_N$. Then, FP is a binomial random variable $\mathrm{B}in(n_0, p_J)$, with $p_J = O_P(J^{-1/2})$ and where n_0 is the number of genes with $\theta_{0i} = \theta_{1i}$. Thus with probability $1 - e^{-c_1\sqrt{J}}$, for some positive constant c_1, $n_1 \leq c_2 n$, with $c_2 < 1$.

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 W., Z. (2001), "Identifying differentially expressed genes in cDNA microarray experiments," *Journal Computational Biology*, 8, 639–659.

Baldi, P. and Long, A. D. (2001), "A Bayesian framework for the analysis of microarray expression data: Regularized t–test and statistical inferences of gene changes," *Bioinformatics*, 17(6), 509–519.

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.

Bryan, J. and van der Laan, M. (2001), "Gene Expression Analysis with the Parametric Bootstrap," *Biostatistics*, 2(4), 445–461.

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

Duggan, D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. (1999), "Expression profiling using cDNA microarrays," *Nature Genetics*, 21, 10–14.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Fisher, M. (1985), "An Applications Oriented Guide to Lagrangian Relaxation," *Interfaces*, 15.

Genovese, C. and Wasserman, L. (2002), *Bayesian and Frequentist Multiple Testing*, Oxford: Oxford University Press, p. to appear.

Guihenneuc, C. and Rousseau, J. (2002), "Laplace expansions in MCMC algorithms for latent variable models," Technical report, CREST.

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002), “Bayesian Models for Gene Expression with DNA Microarray Data,” *Journal of the American Statistical Association*, 97, 88–99.

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

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.

Lindley, D. (1997), “The choice of sample size,” *The Statistician*, 46, 129–138.

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

Lönnstedt, I. and Speed, T. (2002), “Replicated Microarray Data,” *Statistica Sinica*, 12, 31–46.

Newton, M. A. and Kendziorski, C. M. (2003), “Parametric Empirical Bayes Methods for Micorarrays,” in *The analysis of gene expression data: methods and software*, New York: Springer.

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(5), research0022.1–0022.10.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002), “A statistical framework for expression-based molecular classification in cancer,” *Journal of the Royal Statistical Society, Series B*, 64, 717–736.

Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Boston: Harvard University Press, 1st ed.

Richardson, S. and Green, P. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society, series B*, 59, 731–792.

Richmond, C. S., Glasner, 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), "SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays," in *The analysis of gene expression data: methods and software*, New York: Springer.

Tseng, G. C., Oh, M. K., Rohlin, 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, H. and Speed, T. P. (2002), "Design issues for cDNA microarray experiments," *Nature Genetics Reviews*, 3, 579–588.

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.

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