STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3543–3570 (DOI: 10.1002/sim.1335)

Power and sample size for DNA microarray studies

Mei-Ling Ting Lee[1,2,3,*,†] and G. A. Whitmore[4]

[1] *Department of Medicine, Brigham and Women's Hospital, Boston, U.S.A.*
[2] *Harvard Medical School, Boston, U.S.A.*
[3] *Biostatistics Department, Harvard School of Public Health, Boston, U.S.A.*
[4] *McGill University, Montreal, Canada*

SUMMARY

A microarray study aims at having a high probability of declaring genes to be differentially expressed if they are truly expressed, while keeping the probability of making false declarations of expression acceptably low. Thus, in formal terms, well-designed microarray studies will have high power while controlling type I error risk. Achieving this objective is the purpose of this paper. Here, we discuss conceptual issues and present computational methods for statistical power and sample size in microarray studies, taking account of the multiple testing that is generic to these studies. The discussion encompasses choices of experimental design and replication for a study. Practical examples are used to demonstrate the methods. The examples show forcefully that replication of a microarray experiment can yield large increases in statistical power. The paper refers to cDNA arrays in the discussion and illustrations but the proposed methodology is equally applicable to expression data from oligonucleotide arrays. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian inferences; false discovery rate; family type I error; microarray studies; multiple testing; power and sample size

1. INTRODUCTION

Microarray studies aim to discover genes in biological samples that are differentially expressed under different experimental conditions. Experimental designs for microarray studies vary widely and it is important to determine what statistical power a particular design may have to uncover a specified level of differential expression. In this paper we adopt an ANOVA model for microarray data. Selected interaction parameters in the ANOVA model measure differential expression of genes across experimental conditions. The paper discusses conceptual issues and presents computational methods for statistical power and sample size in microarray

[*] Correspondence to: Mei-Ling Ting Lee, Channing Laboratory, BWH/HMS, 181 Longwood Avenue, Boston, MA, 02115-5804, U.S.A.
[†] E-mail: stmei@channing.harvard.edu

Contract/grant sponsor: National Institute of Health; contract/grant numbers: HG02510-01, HL66795-02.
Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada.

Received December 2001
Accepted April 2002

studies. The methodology takes account of the multiple testing that is part of all such studies. The link to implementation algorithms for multiple testing is described. A Bayesian perspective on power and sample size determination is also presented. The discussion encompasses choices of experimental design and replication for a study. Practical examples and case studies are used to demonstrate the methods. Abbreviated sample size and power tables are included for several standard designs. The examples show forcefully that replication of a microarray experiment can yield large increases in statistical power. The paper refers to cDNA arrays in the discussion and illustrations but the proposed methodology is equally applicable to expression data from oligonucleotide arrays.

2. TEST HYPOTHESES IN MICROARRAY STUDIES

The key statistical quantity in a microarray study is the differential expression of a gene in a given experimental condition. Our study of statistical power centres on an analysis of variance (ANOVA) model that incorporates a set of interaction parameters reflecting differential gene expression across experimental conditions. For illustrations of ANOVA models in microarray studies refer to Kerr and Churchill [1], Kerr *et al.* [2], Lee *et al.* [3, 4] and Wolfinger *et al.* [5], among others.

We take the response variable for the ANOVA model as the logarithm (to base 2) of the machine reading of intensity and refer to it simply as the *log-intensity*. Thus, if W is the intensity measurement, the response variable in the ANOVA model is taken as $Y = \log_2(W)$. We assume that W is positive so the logarithm is defined. If the readings are background corrected then we assume that only corrected readings with positive values are used in the analysis. There is some question about whether background correction is advisable. We do not wish to address this dispute in our study here. It is a common practice to calculate the log-ratio of machine readings for the red and green dyes in some cDNA experiments. This practice corrects in a direct way for the varying amount of DNA deposited on the array across the spots. Our ANOVA model accommodates this kind of effect by the inclusion of appropriate main effects and interaction effects, as we will describe shortly. Our experience shows that the use of log-intensity with an appropriate selection of explanatory factors and their interactions provides a powerful modelling framework for a wide variety of microarray experimental designs.

2.1. The ANOVA model

A typical ANOVA model incorporates various factors and their interactions to take account of sources of variability in the microarray data. For example, one might include factors such as *gene G*, *specimen or experimental condition C*, *array slide S*, and *dye D* in the model, with each factor having several levels.

The generic structure of our ANOVA model is as follows:

$$Y_{\mathbf{b}} = \gamma_0 + \gamma_1(b_1) + \gamma_2(b_2) + \cdots + \gamma_L(b_L) + \sum_{l=1}^{L} \sum_{k>l}^{L} \gamma_{lk}(b_l, b_k) + \cdots + \varepsilon_{\mathbf{b}} \tag{1}$$

Here $l = 1, \ldots, L$ denotes a set of L experimental factors. Parameter γ_0 is a constant term. Parameter $\gamma_l(b_l)$ denotes a main effect for factor l when it has level b_l, for $l = 1, \ldots, L$, respectively. Similarly, parameters $\gamma_{lk}(b_l, b_k)$ denote pairwise interaction terms for factors l and k when they have their respective levels b_l and b_k, with $l, k = 1, \ldots, L$. For example, with $L = 3$ factors, the parameter $\gamma_{13}(b_1, b_3)$, where $(b_1, b_3) = (5, 4)$, signifies the interaction parameter for factors 1 and 3 when these two factors have their levels 5 and 4, respectively. The model can be expanded to include third- and higher-order interaction terms if needed, as indicated by the series of dots in (1). The error term is denoted by $\varepsilon_{\mathbf{b}}$. The index $\mathbf{b}$ is a vector of the form $(b_1, \ldots, b_L)$ where b_l denotes the level of factor l. Also, if required, an additional index component can be added to label replicated observations at any given factor level combination.

The factor *condition* refers to the biological specimen or experimental condition. Kerr and Churchill [1] use the agricultural word 'variety' for this term. The term *dye* refers to the dye colour of the intensity reading. When these and other factors are included in the ANOVA model as main effects, they serve the role of normalizing the gene expression data.

2.2. *Interaction effects*

Sets of interaction terms are typically needed in the ANOVA model to account for variability in gene expression. Although all pairwise sets of interaction effects may not be included, those involving gene-by-condition, gene-by-slide, and gene-by-dye interactions, that is, $G \times C$, $G \times S$ and $G \times D$, are usually needed. The first of these sets, the $G \times C$ interaction effects, are the quantities of scientific interest because, as we will discuss shortly, these reflect the differential expression of genes across the specimens or experimental conditions.

Returning to the comment about the common use of the log-ratio of red and green intensities in microarray analyses, we point out that the correlation of these intensities at the same spot which results from the varying amount of deposited DNA is accounted for in ANOVA model (1) by the inclusion of appropriate interaction terms. With a single array slide, for instance, inclusion of the gene-by-dye interaction term $G \times D$ suffices. Where there are multiple slides, the third-order interaction $G \times D \times S$ might be included to capture this source of variability. The microarray design must include a dye-colour reversal feature to allow these important interactions to be estimated.

2.3. *Parameter estimation*

The parameters of the ANOVA model may be estimated by various methods. We shall assume that ordinary least squares methods are used here but the methodology is readily modified for other estimation approaches, such as those based on the L_1 norm. As pointed out by Lee *et al.* [4], the main-effect and interaction parameters involving G will be as numerous as the genes themselves and, hence, typically, may number in the thousands. They propose a two-stage estimation procedure for the ANOVA model in which the parameter estimates involving genes are derived in a second-stage analysis where the estimation proceeds gene by gene.

As we have just noted, the set of gene-by-condition interaction effects, $G \times C$, is the principal set of interest in a microarray study. We shall denote these interaction effects by the symbols $\mathcal{I}_{gc}$, with their estimates denoted by $\hat{\mathcal{I}}_{gc}$. Here, indices g and c refer to gene g and condition c, with ranges $g = 1, \ldots, G$ and $c = 1, \ldots, C$, respectively. It is a standard requirement of ANOVA models that the parameters $\mathcal{I}_{gc}$ and their estimators $\hat{\mathcal{I}}_{gc}$ be subject to estimability constraints that hold simultaneously across conditions $c = 1, \ldots, C$ and genes $g = 1, \ldots, G$. We will adopt the constraint form where their (weighted) sums equal zero. We refer to this kind of constraint as an *interaction sum constraint*. The constraint implies that $\hat{\mathcal{I}}_{gc}$ is interpreted as the estimated differential expression intensity for gene g under condition c *relative* to the average for all genes and conditions in the study.

To illustrate the kind of quantity that $\hat{\mathcal{I}}_{gc}$ represents, suppose that $\hat{\mathcal{I}}_{gc}$ happens to equal 1.231. Then we know that the absolute expression intensity for gene g in experimental condition c is $2^{1.231} = 2.37$ times the (weighted geometric) mean expression levels for all genes and conditions in the study, other factors in the study being held constant. In other words, $\hat{\mathcal{I}}_{gc} = 1.231$ implies a 2.37-fold over-expression or up-regulation of gene g. As we shall show later, the pattern of the estimates $\hat{\mathcal{I}}_{gc}$ for all conditions c will form the basis of statistical inference about whether a gene g exhibits differential gene expression across the experimental conditions.

ANOVA model (1) assumes that the main effects and interaction effects of the model are fixed, not random. In some studies, however, it may be quite reasonable to treat some of these effects as random and, more specifically, to assume they are normally distributed. For example, the main effect for array slide S may very well be a random outcome from a normal population of array effects. A mixed model approach to the analysis of microarray data has been considered by Wolfinger *et al.* [5]. The mixed model would provide different parameter estimates and, hence, possibly different substantive results. The implications of random effects for power levels of designs remain to be investigated in depth.

2.4. The null and alternative hypotheses

Let $\boldsymbol{\mathcal{I}}_g = (\mathcal{I}_{gc},\ c = 1,\ldots,C)'$ denote the column vector of interaction parameters for gene g, where the prime denotes transposition. With respect to differential gene expression, the null and alternative (research) hypotheses of interest for any given gene g can be stated in terms of $\boldsymbol{\mathcal{I}}_g$ as follows:

H_0 : $\boldsymbol{\mathcal{I}}_g = \mathbf{0}$, the zero vector, that is, gene g is not differentially expressed
H_1 : $\boldsymbol{\mathcal{I}}_g = \boldsymbol{\mathcal{I}}^d$, a specified non-zero vector, that is, gene g is differentially expressed

The non-zero vector $\boldsymbol{\mathcal{I}}^d$ in H_1 is a *target* vector of differential expression levels that it is desired to detect. For instance, a study may include four experimental conditions such that conditions $c = 1$ and $c = 2$ replicate a treatment condition and conditions $c = 3$ and $c = 4$ replicate a control condition. In this illustrative study, it may be desired to detect any gene g that has a differential expression pattern of form $\boldsymbol{\mathcal{I}}^d = (1.5, 1.5, -1.5, -1.5)'$. This pattern is equivalent to testing for an 8-fold up-regulation under treatment relative to control, that is, $2^{1.5-(-1.5)} = 2^3 = 8$.

A test of hypotheses exposes an investigator to two types of error. A principal aim of a microarray study is to have a high probability of declaring a gene to be differentially expressed if it is truly differentially expressed, while keeping the probability of making a false declaration of differential expression acceptably low. The achievement of this objective is the purpose of this paper.

In an actual microarray study, genes that are truly differentially expressed will generally do so to different degrees, some weakly some strongly. Therefore, the components $\mathcal{I}_{gc}$ of the interaction parameter vectors $\boldsymbol{\mathcal{I}}_g$ will have values that vary over a continuum as g varies. It is important for us to stress here, however, that this distribution of true expression levels does not directly enter into the power calculation. Instead, the alternative hypothesis H_1 refers only to a single non-zero vector $\mathcal{I}_g$, specifically, the target vector $\mathcal{I}^d$. It is this target vector that is to be used as the reference differential expression pattern for a power calculation. The assumption is that the target vector $\mathcal{I}^d$ (or any vector that lies an equivalent 'distance' from zero) represents a pattern of differential expression that the investigator wishes to detect with high probability (that is, with high power).

2.5. Distributional form of estimated differential expression

In many applications, it is reasonable to assume that estimate vector $\hat{\mathcal{I}}_g$, where $\hat{\mathcal{I}}_g = (\hat{\mathcal{I}}_{gc},\ c = 1,\ldots,C)'$, has an approximate multivariate normal distribution with a mean zero and covariance matrix $\mathbf{\Sigma}$ under the null hypothesis H_0. The claim to a normal approximation is especially strong where the microarray study involves repeated observations of gene expression across conditions so that the interaction estimates $\hat{\mathcal{I}}_{gc}$ are averages of independent log-intensity readings. An appeal to the central limit theorem then supports the assumption of approximate normality. Likewise, under the alternative hypothesis H_1, $\hat{\mathcal{I}}_g$ also has an approximate multivariate normal distribution with the same covariance matrix but now with non-zero mean $\mathcal{I}^d$. We note that the covariance matrix $\mathbf{\Sigma}$ will have rank $C-1$ because of the interaction sum constraint.

2.6. Summary measures of estimated differential expression

On the basis of the ANOVA modelling approach, different statistics may be used to summarize differential expression for single genes in microarray studies. We shall calculate power for some summary measure $V_g = h(\hat{\mathcal{I}}_g)$ of the estimated differential expression vector $\hat{\mathcal{I}}_g$ for gene g, where h is any function specified by the investigator that captures the particular differential expression features that are of scientific interest in the statistical test. The variable V_g is a random variable for gene g that will have some realization v_g in the microarray study. Under null hypothesis H_0, summary measure V_g has a probability density function (PDF) that we denote by $f_0(v)$. Similarly, under the alternative hypothesis H_1, summary measure V_g has a PDF that we denote by $f_1(v)$. We shall show that it is the statistical distance between these two density functions, in a precise sense, that defines the level of power for a microarray study.

3. ERROR CONTROL AND MULTIPLE TESTING IN MICROARRAY STUDIES

As microarray studies typically involve the simultaneous study of thousands of genes, the probabilities of producing incorrect test conclusions (false positives and false negatives) must be controlled for the whole gene set. In this development, we adapt the logic behind power and sample size calculations that are taking place at the planning stage of a study.

3.1. Multiple testing context

The following framework, adapted from Benjamini and Hochberg [8], is useful for understanding multiple testing and the control of inferential errors in microarray studies:

Multiple testing framework

True hypothesis	Test declaration: Unexpressed	Expressed	Number of genes
Unexpressed H_0	A_0	R_0	G_0
Expressed H_1	A_1	R_1	G_1
Total	A	R	G

This framework postulates that there are, in fact, only two possible situations for any gene. Either the gene is not differentially expressed (hypothesis H_0 true) or it is differentially expressed at the level described by the alternative hypothesis H_1. Thus, as discussed earlier, the hypothesis testing framework abstracts from the reality of genes having varying degrees of differential expression. The test declaration (decision) is either that the gene is differentially expressed (H_0 rejected) or that it is unexpressed (H_0 accepted). Thus, there are four possible test outcomes for each gene corresponding to the four combinations of true hypothesis and test declaration.

The total number of genes being tested is G with G_1 and G_0 being the numbers that are truly expressed and unexpressed, respectively. The counts of the four test outcomes are shown by the entries A_0, A_1, R_0 and R_1 in the multiple testing framework. These counts are random variables in advance of the analysis of the study data. The counts A_0 and A_1 are the numbers of true and false negatives (that is, true and false declarations that genes are not differentially expressed). The counts R_1 and R_0 are the numbers of true and false positives (that is, true and false declarations of genes being differentially expressed). The totals, A and R, are the numbers of genes that the study declares are unexpressed (H_0 accepted) and are differentially expressed (H_0 rejected), respectively.

The framework shows that proportions $p_0 = G_0/G$ and $p_1 = G_1/G = 1 - p_0$ of the genes are truly unexpressed and expressed, respectively. The counts G_0 and G_1 and, hence, the proportions p_0 and p_1, are generally unknown. As we show later, their values must be anticipated prior to the conduct of a study. Usually, G_0 will be much larger than G_1 and, indeed, in some studies it may be uncertain if any gene is actually differentially expressed (that is, it may be uncertain if $G_1 > 0$).

We index the genes for which H_0 and H_1 hold by the sets $\mathcal{G}_0$ and $\mathcal{G}_1$, respectively. We must remember, of course, that the memberships of these index sets are unknown because we do not know in advance if any given gene is differentially expressed or not. The test outcomes counted by R_0 are false positives reflecting type I errors. We use α_0 to denote the probability of a type I error for any single gene in the index set $\mathcal{G}_0$ under the selected decision rule. Thus,

$$\alpha_0 = \text{probability of type I error for any gene} = E(R_0)/G_0 \tag{2}$$

Likewise, the test outcomes counted by A_1 are false negatives reflecting type II errors. We use β_1 to denote the probability of a type II error for any single gene in the index set $\mathcal{G}_1$ under the decision rule, that is

$$\beta_1 = \text{probability of type II error for any gene} = E(A_1)/G_1 \tag{3}$$

The *power* of any hypothesis test is defined as the probability of concluding H_1 when, in fact, H_1 is true. In the context of multiple testing, power is defined as the expected proportion of truly expressed genes that are correctly declared as expressed, that is

$$\text{power} = \frac{\text{expected number declared expressed}}{\text{actual number truly expressed}} = \frac{E(R_1)}{G_1} = 1 - \beta_1 \tag{4}$$

Another related performance measure in multiple testing is the *false discovery rate* or FDR for short, proposed by Benjamini and Hochberg [8]. This measure refers to the expected proportion of falsely rejected null hypotheses in multiple tests. With reference to the notation in the multiple testing framework, the FDR is defined as the expected value $E(R_0/R)$, with the ratio R_0/R taken as 0 if $R = 0$.

$$\text{FDR} = E\left(\frac{R_0}{R}\right) \tag{5}$$

In the context of microarrays, FDR is the expected proportion of declared expressed genes that are actually unexpressed. The FDR will be considered in our discussion of a Bayesian perspective on power in Section 5.

3.2. Test outcome dependencies

The vector estimates $\hat{\boldsymbol{\mathcal{I}}}_g$ may be probabilistically dependent for different genes in the same microarray study. This implies that test outcomes for different genes may be probabilistically dependent. For example, a subarray of spots on a microarray slide may share an excess of fluorescence because of contamination of the slide. Careful modelling of such effects can reduce dependencies if they are anticipated. Delongchamp *et al.* [9], for instance, suggest segmentation of an array into subarrays to account for the effects of irregular areas of an array slide that they describe as ‘splotches’. We wish to emphasize in our discussion of dependence that we are not discussing biological dependencies of differential expression levels among genes (that is, co-regulation). These kinds of dependencies are certainly going to be present in every microarray study. For example, H_1 may be true for a group of genes because they are differentially expressed together under given experimental conditions. The focus of our concern is whether estimation errors in the components of $\hat{\boldsymbol{\mathcal{I}}}_g$, representing departures between observed and true values, are intercorrelated among genes. The summary measures V_g for the affected genes and, hence, their test outcomes (H_0 or H_1) will then reflect this dependence in estimation errors. We can envisage practical cases where dependence may be a major concern and others where it may be minor.

Given the potential for some dependence of the errors in the vector estimates $\hat{\boldsymbol{\mathcal{I}}}_g$, even after careful modelling of effects, we consider two different ways to proceed. If the dependency is judged to be substantial or we wish to be conservative in the control of false positives, we may adopt a Bonferroni approach, which we describe shortly. On the other hand, if the dependency is judged to be insignificant we may wish to calculate power or sample size on the assumption that the vector estimates are mutually independent.

3.3. Family type I error probability and the expected number of false positives

There are several ways of specifying the desired control over type I errors in the planning context of multiple testing. We consider two ways: (i) a specification of the *family type I* *error probability*; (ii) a specification of the *expected number of false positives*. The first specification refers to the probability of producing one or more false positives for genes in index set $\mathscr{G}_0$, which we denote by α_F. Thus, in the notation of the preceding multiple testing framework, we have

$$\alpha_F = \text{family type I error probability} = P(R_0 > 0) \tag{6}$$

The second specification refers to the expected number of genes in index set $\mathscr{G}_0$ for which H_0 is incorrectly rejected, that is, the quantity $E(R_0)$. In the following development, we show the connection between the type I error risk for an individual test, denoted earlier by α_0, and the multiple testing control quantities α_F and $E(R_0)$.

We first define an acceptance interval $\mathscr{A}$ for the summary statistic V_g that gives the desired α_0 risk for a test on a single gene. Specifically, we wish to use the following decision rule to judge whether gene g is differentially expressed or not:

$$\text{If } v_g \in \mathscr{A} \text{ then conclude } H_0, \text{ otherwise conclude } H_1 \tag{7}$$

Under the null hypothesis H_0, summary measure V_g for any single gene g will fall in acceptance interval $\mathscr{A}$ with the following probability.

$$P(V_g \in \mathscr{A}) = \int_{\mathscr{A}} f_0(v)\,\mathrm{d}v = 1 - \alpha_0 \quad \text{for each gene } g \in \mathscr{G}_0 \tag{8}$$

As we demonstrate later, we can use (8) to calculate $\mathscr{A}$ from knowledge of the form of the null PDF $f_0(v)$. Interval $\mathscr{A}$ is chosen to be the shortest among those intervals satisfying (8).

We now describe two testing approaches, depending on whether the estimation errors in $\hat{\boldsymbol{\mathcal{I}}}_g$ are independent or not. We refer to these as the Sidak and Bonferroni approaches, respectively.

3.3.1. Independent estimation errors: the Sidak approach. Under the assumption of independence, the family type I error probability α_F and the type I error probability for an individual test α_0 are connected as follows for the gene index set $\mathscr{G}_0$:

$$P(R_0 = 0) = (1 - \alpha_0)^{G_0} = 1 - \alpha_F \tag{9}$$

In most microarray studies, G_0 is large and, hence, even a small specification for α_0 will translate into a large value for the family type I error probability α_F. In addition, in most studies it is uncertain what number of genes are unexpressed. In this situation, an investigator may wish to assume that all genes are unexpressed (so $G_0 = G$) and change the exponent in (9) from G_0 to G.

With independence, the random variable R_0 follows a binomial distribution with parameters G_0 and α_0. Thus, the expectation $E(R_0)$ equals $G_0\alpha_0$. When G_0 is reasonably large and α_0 is small, the *number of false positives* R_0 that will arise under the assumption of independence will follow an approximate Poisson distribution with mean parameter

$$E(R_0) = G_0\alpha_0 \approx -\ln(1 - \alpha_F) \tag{10}$$

For example, if the family type I error α_F is 0.20 and G_0 is large, the Poisson mean is $E(R_0) = -\ln(0.80) = 0.223$. In this case, the probability of experiencing no false positive is $\exp(-0.223) = 0.80$. The probability of exactly one false positive is $0.223\exp(-0.223) =$

$0.223(0.80)=0.18$. The probability of experiencing two or more false positives is therefore 0.02. Because of the direct connection between α_F and the mean $E(R_0)$ in this case, either value may be used to specify the desired control over the family type I error risk.

As another example, if an investigator feels that expecting 2.5 false positives is tolerable then this specification implies that $E(R_0)=-\ln(1-\alpha_F)=2.5$ and, hence, a family type I error probability of $\alpha_F=1-\exp(-2.5)=0.918$. This α_F value may appear very high. The illustration reminds us, however, that a large value of α_F may be reasonable in microarray studies where a few false positives among thousands of genes must be tolerated in order to avoid missing many truly expressed genes (that is, to avoid false negatives). The design of a microarray study involves a careful balancing of costs of false positives and false negatives. The connection between α_F and α_0 in this last example is

$$\alpha_0=\frac{E(R_0)}{G_0}\approx\frac{-\ln(1-\alpha_F)}{G_0} \tag{11}$$

For instance, if G_0 happens to equal 5000 then, $\alpha_0=2.5/5000=-[\ln(1-0.918)]/5000=0.00050$.

Under the independence approach represented by rule (9), we may wish to focus more directly on the number of false positives by using the following property of order statistics for simple random samples: the k_1th lowest and k_2th highest order statistics of the summary measures v_g for genes $g\in\mathscr{G}_0$ span an expected combined tail area of $k/(G_0+1)$ where $k=k_1+k_2$. This property may be used to set the acceptance interval $\mathscr{A}$ based on the anticipated values of extreme order statistics under the null PDF $f_0(v)$. Specifically, the acceptance interval in (8) may be defined by the following specification:

$$\alpha_0=\frac{k}{G_0+1} \tag{12}$$

Substitution of (12) into (9) gives the following implied value for the family type I error probability for this rule:

$$\alpha_F=1-\left(1-\frac{k}{G_0+1}\right)^{G_0}\approx[1-\exp(-k)] \quad \text{for large } G_0 \tag{13}$$

The mean number of false positives for this rule is approximately $k=E(R_0)=-\ln(1-\alpha_F)$. Although the form of (12) is motivated by the theory of order statistics in which k is a whole number, (12) and (13) can be used with fractional values of k.

3.3.2. Dependent estimation errors: the Bonferroni procedure. The Bonferroni procedure is widely used in statistics for error control where simultaneous inferences are being made. The procedure makes use of the Bonferroni probability inequality to control the family type I error probability. The inequality holds whatever may be the extent of statistical dependence among the estimated differential expression vectors $\hat{\mathcal{I}}_g$ of the gene set.

For this procedure, the acceptance interval in (8) may be defined by the following specification:

$$\alpha_0=\frac{\alpha_F}{G_0} \tag{14}$$

This definition of the acceptance interval $\mathscr{A}$ guarantees that the following inequality holds for the family type I error probability:

$$P\left[\bigcap_{g\in\mathscr{G}_0}(V_g\in\mathscr{A})\right]\geqslant 1-\alpha_{\mathrm{F}} \tag{15}$$

Thus, the inequality in (15) assures us that the Bonferroni procedure keeps the family type I error probability at level α_{F} or lower. In subsequent discussion in the Bonferroni context, we refer to α_{F} as the family type I error probability although the inequality (15) implies that the true error probability may be somewhat lower. We note that, for given G_0 and α_{F}, the Bonferroni rule (14) will always choose a wider acceptance interval $\mathscr{A}$ than the rule based on the independence assumption in (9).

With respect to the expected number of false positives, using the Bonferroni procedure in (14) provides the following result:

$$E(R_0)=G_0\alpha_0=\alpha_{\mathrm{F}} \tag{16}$$

It can be seen that the expected number of false positives equals the family type I error probability in this case. Thus, necessarily, the expected number $E(R_0)$ cannot exceed one (although the actual number R_0 is not so constrained).

Unlike the independence approach discussed in the preceding section, there is no direct link between the probability distribution for the number of false positives R_0 and the family type I error probability α_{F} under the Bonferroni approach. The Bonferroni procedure controls the chance of incurring one or more false positives but provides no probability statement about how many false positives may be present if some do occur (that is, the approximate Poisson distribution does not apply).

3.4. Family power level and the expected number of true positives

As with type I errors, we can quantify type II error control in several ways in the context of multiple testing. We focus on two ways: (i) the *family type II error probability* (or, equivalently, one minus the *family power level*); (ii) the *expected number of true positives*. The first measure refers to the probability of producing one or more false negatives for genes in index set $\mathscr{G}_1$, which we denote by β_{F}. Thus, in the notation of the preceding multiple testing framework, we have

$$\beta_{\mathrm{F}}=\text{family type II error probability}=P(A_1>0) \tag{17}$$

The corresponding family power level is then $1-\beta_{\mathrm{F}}$. The second measure refers to the expected number of genes in index set $\mathscr{G}_1$ that are correctly declared as differentially expressed, that is, the quantity $E(R_1)$. In the following development, we show the connection between the type II error risk for an individual test, denoted earlier by β_1, and the values of the multiple testing quantities β_{F} and $E(R_1)$.

The power of the test for any *single* gene that is differentially expressed at the level defined in H_1 equals $1-\beta_1$. This declaration is equivalent to having the summary measure $V_g=h(\hat{\boldsymbol{\mathcal{I}}}_g)$ for the gene in question fall outside the acceptance interval $\mathscr{A}$. We denote this rejection interval by the complement $\mathscr{A}^c$. The power for a single differentially expressed gene is therefore given by

$$P(V_g \in \mathscr{A}^c) = \int_{\mathscr{A}^c} f_1(v)\,\mathrm{d}v = 1 - \beta_1 \quad \text{for any gene } g \in \mathscr{G}_1 \tag{18}$$

In essence, therefore, $1 - \beta_1$ is fixed by the rejection interval which, in turn, is fixed by the specified control on the family type I error risk and the specification for the alternative hypothesis H_1. Our use of PDF $f_1(v)$ for this power calculation means that we are examining the power for any and all differential gene expression target vectors $\boldsymbol{\mathcal{I}}^d$ whose estimates map into the same random variable $V_g = h(\hat{\boldsymbol{\mathcal{I}}}_g)$ having the PDF $f_1(v)$.

As with type I errors, we encounter the Sidak and Bonferroni formulae for power, depending on whether estimation errors in $\hat{\boldsymbol{\mathcal{I}}}_g$ are independent or not. We can now abbreviate the presentation because the underlying logic is clear from the earlier development.

3.4.1. Independent estimation errors: the Sidak approach. The anticipated count G_1 when taken together with the power level $1 - \beta_1$ for an individual test can be used to calculate either measure of family type II error control. Under the assumption of independence, the family type II error probability β_{F} and the type II error probability for an individual test β_1 are connected as follows for the gene index set $\mathscr{G}_1$:

$$P(A_1 = 0) = (1 - \beta_1)^{G_1} = 1 - \beta_{\mathrm{F}} \tag{19}$$

Also, under independence, the random variable R_1 follows a binomial distribution with parameters G_1 and $1 - \beta_1$. Thus, the expected number of true positives is given by

$$E(R_1) = G_1(1 - \beta_1) \tag{20}$$

In many studies, it is uncertain what number of genes G_1 will be differentially expressed, if any. In this situation, an investigator may wish to consider power only for the case of an isolated gene that is differentially expressed (so $G_1 = 1$). In this case, $1 - \beta_{\mathrm{F}} = 1 - \beta_1 = E(R_1)$.

To illustrate these power calculations numerically, consider a microarray study for which G_1 is anticipated to be 50 genes and for which $1 - \beta_1 = 0.99$. In this case, $1 - \beta_{\mathrm{F}} = (0.99)^{50} = 0.605$. Observe how high the power level must be for a single gene in index set $\mathscr{G}_1$, namely 0.99, in order to have even a moderate probability of discovering all 50 differentially expressed genes (0.605). In this same situation, the expected number of true positives among the 50 differentially expressed genes would be $G_1(1 - \beta_1) = 50(0.99) = 49.5$. In other words, 99 per cent of the truly expressed genes are expected to be declared as such.

3.4.2. Dependent estimation errors: the Bonferroni procedure. Under an assumption of dependence for the estimated differential expression vectors, recourse to the Bonferroni inequality gives the following specification for the family power level $1 - \beta_{\mathrm{F}}$ as a function of the level of β_1 for a single test:

$$1 - \beta_{\mathrm{F}} \geqslant \max(0, 1 - G_1\beta_1) \tag{21}$$

As $1 - \beta_{\mathrm{F}}$ must be non-negative, a minimum of zero is imposed in (21). Thus, the Bonferroni inequality gives us a lower bound on the family power level.

The expected number of true positives under the Bonferroni approach is given by

$$E(R_1) = G_1(1 - \beta_1) \tag{22}$$

For the previous numerical example, where $G_1 = 50$ and $1 - \beta_1 = 0.99$, the lower bound on the family power level is $1 - 50(0.01) = 0.50$. The expected number of true positives is $E(R_1) = 50(0.99) = 49.5$. Thus, again, 99 per cent of the truly expressed genes are expected to be declared as such. As is the case with false positives, there is no direct link between the family type II error probability β_F and the probability distribution for the number of true positives R_1 under dependence. The Bonferroni procedure controls the chance of incurring one or more false negatives but provides no probability statement about how many false negatives may be present if some do occur.

3.5. Relation of error control in the planning stage to multiple testing for observed data

We now discuss the relation between specifications for type I and type II error controls at the planning stage of a microarray study before the microarray experiments are conducted and the implementation algorithms used at the actual testing stage (for example, step-down p-values) after the gene expression data have been collected.

For planning purposes, our methodology posits the index sets $\mathscr{G}_0$ and $\mathscr{G}_1$ for unexpressed and differentially expressed genes, respectively. Although the planning does not identify the members of each set, it does specify the cardinality of each. In this statistical setting, test implementation algorithms seek to maximize the power of detecting which genes are truly in the index set $\mathscr{G}_1$ while still controlling either the family type I error probability or a related quantity, such as the false discovery rate (discussed later in Section 5).

Many approaches have been proposed for actual test implementation once the microarray data are in hand. For example, step-down p-value algorithms and methods for controlling the false discovery rate have been widely adopted for error control in microarray studies, see, for instance, Dudoit *et al.* [6] and Efron *et al.* [7]. Observed p-values for the G hypothesis tests in a microarray study will be derived from the null PDF $f_0(v)$ or its estimate, evaluated at the respective realizations v_g, $g = 1, \ldots, G$. The observed p-values, say $p_1, \ldots, p_G$, will vary from gene to gene because of inherent sampling variability and also because the null hypothesis may hold for some genes but not for others. The information content of the observed p-values is used in these testing procedures to assign genes to either the index set $\mathscr{G}_0$ or the index set $\mathscr{G}_1$ without knowing the size of either set. These approaches exploit information in the data themselves (specifically, the observed p-values) and, hence, are data-dependent. In contrast, in planning for power and sample size, we must anticipate the sizes of these two index sets, and control the two types of errors accordingly, without the benefit of having the observed p-values themselves. The p-values derived from the actual observed microarray data not only allow classification of individual genes as differentially expressed or not but also provide a report card on the study plan and whether its specifications were reasonable or not.

4. POWER CALCULATIONS FOR DIFFERENT SUMMARY MEASURES

We now present power calculations for the two classes of functions $V_g = h(\hat{\boldsymbol{\mathcal{I}}}_g)$ mentioned earlier, both of which are important in microarray studies.

4.1. Designs with linear summary of differential expression

Consider a situation where the summary measure V_g is a linear combination of differential expression estimates $\hat{\mathcal{I}}_{gc}$ of the following form:

$$V_g = h(\hat{\boldsymbol{\mathcal{I}}}_g) = \boldsymbol{\lambda}'\hat{\boldsymbol{\mathcal{I}}}_g = \sum_{c \in C} \lambda_c \hat{\mathcal{I}}_{gc} \tag{23}$$

where $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_C)$ is a vector of specified coefficients. Examples of such linear combinations include any single differential expression estimate, say $\hat{\mathcal{I}}_{g1}$, or any difference of such estimates, say $\hat{\mathcal{I}}_{g1} - \hat{\mathcal{I}}_{g2}$. Frequently the linear combination of interest will be a contrast of interaction estimates that reflects, for example, the difference between treatment and control conditions.

As discussed earlier, we may assume that the vector $\hat{\boldsymbol{\mathcal{I}}}_g$ has an approximate multivariate normal distribution with mean zero under the null hypothesis and covariance matrix $\boldsymbol{\Sigma}$. This assumption is reasonable, first, because of the application of the central limit theorem in deriving individual estimates $\hat{\mathcal{I}}_{gc}$ from repeated readings and, second, from a further application of the central limit theorem where the linear combination in (23) involves further averaging of the individual estimates. It then follows from this normality assumption that the null PDF $f_0(v)$ is an approximate normal distribution with mean zero and null variance

$$\sigma_0^2 = \mathrm{var}(V_g | H_0) = \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda} \tag{24}$$

Under the alternative hypothesis H_1, we assume that the $\hat{\mathcal{I}}_{gc}$ have the same multivariate normal distribution but with mean $\boldsymbol{\mathcal{I}}^d$. In other words, that the null distribution is simply translated to a new mean position. In this case, the summary measure V_g has an approximate normal PDF $f_1(v)$ with the same variance σ_0^2 and mean parameter

$$\mu_1 = E(V_g | H_1) = \boldsymbol{\lambda}'\boldsymbol{\mathcal{I}}^d \tag{25}$$

We consider only linear combinations for which μ_1 is non-zero.

Here are the steps for computing power:

1. Compute the null variance σ_0^2 in (24) from specifications for the vector $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\Sigma}$.
2. Compute μ_1 in (25) from specifications for the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mathcal{I}}^d$.
3. Specify the family type I error risk α_{F} or, under independence, the equivalent mean number of false positives $E(R_0) \approx -\ln(1 - \alpha_{\mathrm{F}})$.

The first step is the most difficult because it requires some knowledge of the inherent variability of the data in the planned microarray study. As we discuss later, this inherent variability is intimately connected with the experimental error in the scientific process, the experimental design and the number of replicates of the design used in the study.

We now present a brief numerical example of a power calculation based on the methodology for a linear function of differential expression.

4.1.1. Numerical example for linear summary of differential expression. Consider a microarray study in which interest lies in the difference between two experimental conditions

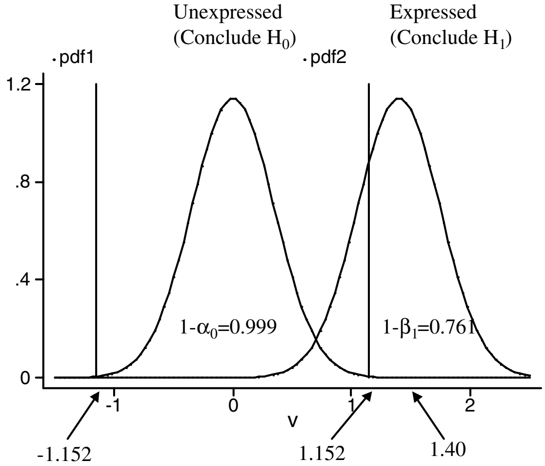

Figure 1. Illustration of a power calculation in the non-central normal case.

representing, say, two tissue types. Thus, differences in differential gene expression of the form $\hat{\mathcal{I}}_{g1} - \hat{\mathcal{I}}_{g2}$ are being considered. The null standard deviation for such differences is expected to be similar to that found in a previous study, namely, $\sigma_0 = 0.35$ on a log-scale with base 2. We suppose that a difference of $\mu_1 = 1.40$ (on a log-2 scale) is the target difference under the alternative hypothesis. Observe that this difference represents a $2^{1.40} = 2.64$-fold difference in gene expression and is four times the null standard deviation (that is, $\mu_1/\sigma_0 = 1.40/0.35 = 4.0$). The study involves a gene set of $G = 2100$ genes. It is anticipated that $G_0 = 2000$ genes will show no differential expression, while the remaining $G_1 = 100$ will be differentially expressed at the target level μ_1. We assume statistical independence among the estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ of the gene set. The order statistic rule (12) with $E(R_0) = k = 2$ will be used for setting the acceptance interval $\mathscr{A}$. It then follows that $\mathscr{A}$ is defined by $\pm z\sigma_0$ where z denotes the standard normal percentile $z(2000/2001) = z(0.9995) = 3.2905$. The resulting interval is $(-1.152, 1.152)$ on a log-2 scale. In making this determination of $\mathscr{A}$, we have used the interval centred on zero as it is the shortest interval. The situation is illustrated in Figure 1. Observe that the area spanned by the acceptance interval under the null PDF in this illustration corresponds to $1999/2001 = 0.999$ so $\alpha_F = 1 - (0.999)^{2000} = 0.8648$ and $E(R_0) = -\ln(1 - 0.8648) = 2$, as required. Finally, the power for detecting a single differentially-expressed gene is given by the area under the alternative PDF in Figure 1, labelled $1 - \beta_1$. Reference to the standard normal distribution gives $1 - \beta_1 = 0.761$. Thus, any single gene with a 2.64-fold difference in expression between the two tissue types has probability 0.761 of being classified as differentially expressed in this study (that is, of leading to conclusion H_1). This probability is the same whether the difference refers to an up- or down-regulated gene. This same power value implies that about 76 per cent of the anticipated $G_1 = 100$ differentially-expressed genes in the array will be correctly declared as differentially expressed. The probability of detecting all 100 of these genes is the family power level $1 - \beta_F$ given by (19). Here that family power level is 0.761^{100}, a vanishingly small value.

4.2. Designs with quadratic summary of differential expression

We now consider a situation where the summary measure V_g is a quadratic form. To represent this quadratic form symbolically, we restrict vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ and $\boldsymbol{\mathcal{I}}^d$ to their first $C-1$ components and restrict matrix $\boldsymbol{\Sigma}$ to the principal submatrix defined by the first $C-1$ interaction parameters. We denote these restricted forms by $\hat{\boldsymbol{\mathcal{I}}}_{g|\mathrm{R}}$, $\boldsymbol{\mathcal{I}}^d_{\mathrm{R}}$ and $\boldsymbol{\Sigma}_{\mathrm{R}}$, respectively. These restricted forms are required by the interaction sum constraint which makes one component of each vector redundant. With this restricted notation, the quadratic form of interest is expressed as follows:

$$V_g = \hat{\boldsymbol{\mathcal{I}}}_{g|\mathrm{R}}' \boldsymbol{\Sigma}_{\mathrm{R}}^{-1} \hat{\boldsymbol{\mathcal{I}}}_{g|\mathrm{R}} \tag{26}$$

We see that this measure implicitly takes account of all differential expression estimates and, hence, is responding to differential expression in any of the C experimental conditions in the study. Statistic V_g in (26) is larger whenever one of the interaction estimates in $\hat{\boldsymbol{\mathcal{I}}}_g$ is larger. It is therefore a comprehensive measure of differential gene expression. Measure V_g in (26) can be interpreted as the squared statistical distance between the restricted interaction estimate vector $\hat{\boldsymbol{\mathcal{I}}}_{g|\mathrm{R}}$ and the zero vector $\mathbf{0}$ specified in the null hypothesis. It is also intimately connected to the sum of squares for the set of interaction effects $G \times C$ in the ANOVA model.

The quadratic measure in (26) is suitable for microarray studies which examine an assortment of experimental conditions with the simple aim of discovering genes that are differentially expressed *in any pattern* among the conditions. For example, a microarray study may examine tissues from C different tumours with the aim of seeing if there are genetic differences among the tumours. As another example, the experimental conditions may represent a biological system at C different time points and interest may lie in the time course of genetic change in the system, if any. Thus, measure (26) is suited to uncovering differential gene expression in a general set of experimental conditions where theory may provide no guidance about where among the conditions the differential expression is likely to arise.

To apply measure (26), we assume, as before, that the estimate vector $\hat{\boldsymbol{\mathcal{I}}}_g$ is approximately multivariate normal with covariance matrix $\boldsymbol{\Sigma}$ and mean zero under the null hypothesis. The theory of quadratic forms then states that V_g follows an approximate chi-square distribution with $C-1$ degrees of freedom. One degree of freedom is lost because of the interaction sum constraint. Under the alternative hypothesis, V_g has a non-central chi-square distribution with non-centrality parameter θ_1 where

$$\theta_1 = \boldsymbol{\mathcal{I}}^{d\prime}_{\mathrm{R}} \boldsymbol{\Sigma}_{\mathrm{R}}^{-1} \boldsymbol{\mathcal{I}}^d_{\mathrm{R}} \tag{27}$$

We caution that the assumption of chi-square and non-central chi-square distributions for quadratic measure V_g in (26) is a little more sensitive to the assumed normality of the vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ than is the case with the linear summary measure (23). The reason is that the quadratic measure does not have the benefit of a secondary application of the central limit theorem from taking a linear combination of estimates.

The non-central chi-square PDF can be used in (18) to calculate the power of the microarray study. The steps for computing power are as follows:

1. Compute the non-centrality parameter θ_1 in (27) from specifications for vector $\boldsymbol{\mathcal{I}}^d$ and covariance matrix $\boldsymbol{\Sigma}$.

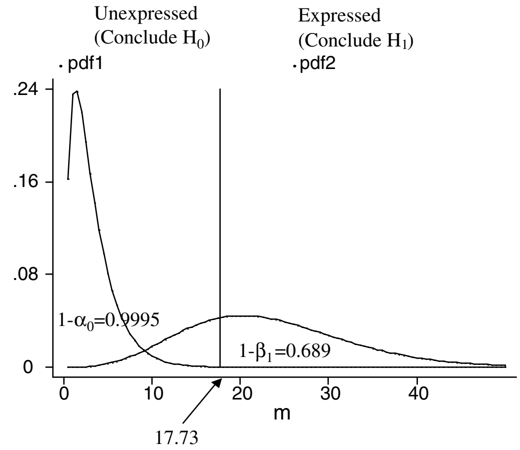

Figure 2. Illustration of a power calculation in the non-central chi square case.

2. Specify the family type I error risk α_F or, under independence, the equivalent mean number of false positives $E(R_0) \approx -\ln(1-\alpha_F)$.

As before, the first step is the most difficult because it requires some knowledge of the inherent variability of the data in the planned microarray study, which depends on the experimental error in the scientific process, the experimental design and the number of replicates of the design used in the study.

4.2.1. Numerical example for quadratic summary of differential expression. As a brief numerical example of a power calculation for the quadratic summary measure, assume that the gene set contains $G=2100$ genes and that the study includes $C=4$ experimental conditions. Furthermore, assume statistical independence among the estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ of the gene set. We suppose that the non-centrality parameter θ_1 under the alternative hypothesis is calculated from (27) and equals 20.0. It is anticipated that $G_0=2000$ of the 2100 genes will not be differentially expressed and $G_1=100$ genes will be differentially expressed at the target level θ_1. We will use the order statistic rule (12) with $E(R_0)=k=1$ to set the acceptance interval $\mathscr{A}$. It then follows that $\mathscr{A}$ is defined by $\chi_3^2(2000/2001)=\chi_3^2(0.9995)=17.73$. The resulting acceptance interval $\mathscr{A}$ under the null PDF $f_0(v)$ is $(0, 17.73)$ as shown in Figure 2. Finally, the power of this microarray study to detect a single differentially-expressed gene is given by the area under the alternative PDF $f_1(v)$, labelled $1-\beta_1$ in Figure 2. Reference to the relevant non-central chi-square distribution gives this power value as $1-\beta_1=0.689$. Thus, 69 per cent of the 100 differentially expressed genes in the index set $\mathscr{G}_1$ are expected to be detected by this study.

4.3. Relation to methods based on t- and F-statistics

It can be seen that the preceding test methodology and power calculations use neither the observed mean square error (MSE) nor t- and F-statistics at the level of the individual gene. (Recall that MSE enters the denominator of both t- and F-statistics.) Our experience with microarray data sets has made us cautious about assuming a normal error term for the ANOVA model. We have found that data anomalies and non-normal features of the error term, which are frequently encountered in microarray data, make MSE values more susceptible to distortion than the ANOVA interaction estimates $\hat{\mathcal{I}}_g$ upon which our approach relies. Moreover, the problem is aggravated by the fact that many microarray experimental designs provide few degrees of freedom for the error term at the individual gene level. Other investigators have noted similar problems and chosen alternative strategies for dealing with them. Some, such as Dudoit *et al.* [6], are successful in using tests based on t- and F-statistics. Some have used permutation tests to avoid assumptions about the error distribution, although this approach does depend on having reasonably large degrees of freedom at the individual gene level. Another strategy is adopted in Efron *et al.* [7] where a variance-offset is used to improve reliability. They compute expression scores of the form $\bar{D}_i/(a_0+S_i)$ for each gene i. Here $\bar{D}_i$ and S_i are the mean and standard deviation of expression differences between treatment and control and a_0 is a fixed quantity (the 90th percentile of all S_i values in this instance). The constant a_0 helps to stabilize the scores. They note that setting $a_0=0$ (that is, omitting the constant) is a 'disasterous choice' in their application (Reference [7] p. 1156).

5. A BAYESIAN PERSPECTIVE ON POWER AND SAMPLE SIZE

Lee *et al.* [3, 4] and Efron *et al.* [7] describe a mixture model for differential gene expression that provides a Bayesian posterior probability for the event that a given gene is differentially expressed. This mixture model has a useful interpretation in terms of our study of power and sample size.

In the multiple testing framework presented earlier, we defined p_1 and its complement $p_0=1-p_1$ as the respective probabilities that a randomly selected gene would be differentially expressed (H_1) or not (H_0). We now take these probabilities as prior probabilities in a Bayesian model for the summary measure of gene expression V_g for gene g. The marginal PDF for the summary statistic V_g under this model is

$$f(v)=p_0f_0(v)+p_1f_1(v) \tag{28}$$

This model simplifies reality in two respects. First, it assumes that the prior probabilities are the same for all genes, although this assumption can be relaxed easily. Second, it assumes that if a gene is differentially expressed then it is expressed at the target level specified in H_1.

From Bayes theorem, the posterior probabilities for any gene g having summary statistic $V_g=v_g$ can be calculated from the components of the mixture model (28) as follows:

$$P(H_0|v_g)=\frac{p_0f_0(v_g)}{f(v_g)},\quad P(H_1|v_g)=\frac{p_1f_1(v_g)}{f(v_g)} \tag{29}$$

The posterior probabilities $P(H_1|v_g)$ and $P(H_0|v_g)$ are the respective probabilities that gene g is truly differentially expressed or not, given its summary measure has outcome v_g.

5.1. Connection to local true and false discovery rates

For some classification cut-off value v_*, defined by an appropriate balancing of misclassification costs, each gene g can be *declared* as differentially expressed or not, depending on whether $v_g > v_*$ or not. Probability $P(H_1|v_g)$ is then the probability of a correct declaration of differential expression when an expression reading of $v_g > v_*$ is presented.

Efron *et al.* [7] interpret the posterior probability $P(H_0|v_g)$ in (29) as the false discovery rate for all genes sharing summary measure v_g, for given $v_g > v_*$. As $P(H_0|v_g)$ describes the FDR in the locality of v_g, Efron *et al.* [7] refer to it as the *local false discovery rate* or local FDR for short. By analogy, the complementary posterior probability $P(H_1|v_g)$ might be interpreted as a *local true discovery rate* or local TDR. The local TDR is the proportion of truly expressed genes among those genes sharing summary measure $v_g > v_*$.

5.2. Representative local true discovery rate

A representative value of the local TDR can be chosen to summarize the ability of a microarray study to correctly classify genes declared to be differentially expressed. As a representative value of the posterior probability $P(H_1|v_g)$, we suggest replacing v_g by the corresponding parameter $h(\mathcal{I}^d)$ under the alternative hypothesis H_1. The resulting probability is

$$P[H_1|h(\mathcal{I}^d)] = \frac{p_1 f_1[h(\mathcal{I}^d)]}{f[h(\mathcal{I}^d)]} \tag{30}$$

When h is the linear function in (25), parameter $h(\mathcal{I}^d)$ is μ_1. When h is the quadratic function in (27), parameter $h(\mathcal{I}^d)$ corresponds to θ_1.

The representative local TDR that we have just defined is not directly comparable to the power level of a test although it does convey closely related information about the ability of a microarray study to correctly identify genes that are truly differentially expressed. As defined at the outset of the paper, classical power refers to the conditional probability of declaring a gene as differentially expressed when, in fact, that is true. In this Bayesian context, local TDR refers to the conditional probability that a gene is truly differentially expressed when, in fact, it has been declared as expressed by the test procedure. The conditioning events of these two probabilities are reversed in the classical and Bayesian contexts.

5.3. Numerical example

To give a numerical example of local TDR and FDR, we consider the demonstration depicted in Section 4.2 and Figure 2. The gene set contains $G = 2100$ genes. Four experimental conditions are under study, so $C = 4$. It is anticipated that $G_0 = 2000$ of the 2100 genes will not be differentially expressed and $G_1 = 100$ genes will be differentially expressed. These counts correspond to prior probabilities of $p_0 = 0.952$ and $p_1 = 0.048$. The non-centrality parameter θ_1 has been specified as 20 under H_1. Setting v_g equal to $\theta_1 = 20$, the probability densities $f_0(20)$ and $f_1(20)$ for central and non-central chi-square distributions with $C - 1 = 3$ degrees of freedom are calculated as 0.00008096 and 0.004457, respectively. The marginal probability density $f(20)$ is then calculated from (28) as 0.0002893. Finally, the desired posterior probabilities in (30) are calculated as 0.266 and 0.734, respectively. Thus, for example, the probability that H_1 is true rises from a prior level of 0.048 to a posterior level of 0.734 if the gene has an observed differential expression level v_g equal to $\theta_1 = 20$. Assuming $v_g = 20$ is above the classification cut-off v_*, these respective probabilities are the local FDR and TDR. In other words, among genes having observed differential expression at level $v_g = 20$, about 73 per cent will be truly differentially expressed and 27 per cent will not.

6. APPLICATIONS TO SOME STANDARD MICROARRAY DESIGNS

We now show applications of the power methodology to some standard microarray designs and present representative sample size and power tables for these designs.

6.1. Matched-pairs design

Consider a microarray study with n matched pairs of treatment and control conditions. For example, in a study of liposarcoma, each treatment–control pair may be liposarcoma tissue and normal fat tissue taken from a matched pair of patients. Thus, there are $C=2n$ experimental conditions in total. To be explicit, we assume that indices $c=1,\ldots,n$ denote the treatment conditions and $c=n+1,\ldots,2n=C$ denote the matching control conditions. The assumption is made that a given gene g either has no difference in log-expression between the treatment and control conditions (null hypothesis H_0) or has a difference in log-expression equal to some non-zero value Δ_1 (alternative hypothesis H_1). This assumption implies that the interaction parameters $\mathcal{I}_{gc}$ have the following values under the alternative hypothesis:

$$\mathcal{I}_{gc}=\begin{cases}\Delta_1/2 & \text{for } c=1,\ldots,n \quad \text{treatment conditions}\\ -\Delta_1/2 & \text{for } c=n+1,\ldots,2n \quad \text{control conditions}\end{cases} \tag{31}$$

Observe that these parameter values sum to zero as required by the interaction sum constraint.

We illustrate the methodology for a linear function of differential gene expression. The linear combination of interest in this context involves a contrast of gene expression under treatment and control conditions. We choose the convenient definition

$$\boldsymbol{\lambda}'=(1/n,\ldots,1/n,-1/n,\ldots,-1/n)$$

where there are n coefficients of each sign. Thus, from (25) and (24), we have

$$\mu_1=E(V_g|H_1)=\Delta_1 \tag{32}$$

$$\sigma_0^2=\text{var}(V_g|H_0)=\sigma_D^2/n \tag{33}$$

Here σ_D^2 signifies the variance of the difference in log-expression between treatment and control conditions in a matched pair.

6.1.1. Sample size table for matched-pairs design. Table I(a) gives the number of matched treatment–control pairs n required to achieve a specified individual power level $1-\beta_1$ for the experimental design we have just described. The calculations for the table assume that the estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ are mutually independent across genes. The table is entered based on the specified mean number of false positives $E(R_0)$, ratio $|\Delta_1|/\sigma_D$, anticipated number of unexpressed genes G_0 and desired individual power level $1-\beta_1$. If G_0 is expected to be similar to the total gene count G, the table could be entered using G without introducing great error. To conserve space, only two individual power levels are offered in the table, 0.90 and 0.99. The sample size shown in the table is the smallest whole number that will yield the specified power. The total number of experimental conditions C is double the entry in the table, that is, $C=2n$. Observe that the ratio $|\Delta_1|/\sigma_D$ can be interpreted as

Table I. Sample size for matched-pairs designs and completely randomized designs with a linear summary of differential expression. The number listed in a cell is the sample size (n) required in the treatment and control groups to yield the specified individual power level $1 - \beta_1$, which is the expected proportion of truly expressed genes that will be correctly declared as expressed by the tests. The requisite total sample size is $C = 2n$. Gene number G_0 denotes the anticipated number of unexpressed genes involved in the experiment.

(a) *Sidak approach*: estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ are assumed to be mutually independent across genes. $E(R_0)$ denotes the mean number of false positives. The family power level $1 - \beta_\mathrm{F}$ and expected number of true positives $E(R_1)$ can be calculated from $1 - \beta_1$ using (19) and (20).

	Mean number of false positives											
	$E(R_0) = 1$				$E(R_0) = 2$				$E(R_0) = 3$			
	Distance $\lvert\Delta_1\rvert/\sigma_\mathrm{D}$				Distance $\lvert\Delta_1\rvert/\sigma_\mathrm{D}$				Distance $\lvert\Delta_1\rvert/\sigma_\mathrm{D}$			
	1.0	1.5	2.0	2.5	1.0	1.5	2.0	2.5	1.0	1.5	2.0	2.5
Genes G_0	*Power = proportion correctly declared as expressed = 0.90*											
500	20	9	5	4	18	8	5	3	17	8	5	3
1000	21	10	6	4	20	9	5	4	19	9	5	3
2000	23	11	6	4	21	10	6	4	20	9	5	4
8000	27	12	7	5	25	11	7	4	24	11	6	4
Genes G_0	*Power = proportion correctly declared as expressed = 0.99*											
500	30	14	8	5	28	13	7	5	26	12	7	5
1000	32	15	8	6	30	14	8	5	29	13	8	5
2000	34	15	9	6	32	15	8	6	31	14	8	5
8000	38	17	10	7	36	16	9	6	35	16	9	6

(b) *Bonferroni approach*: estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ may be dependent across genes. Value α_F denotes the family type I error probability for the gene set. A lower bound on the family power level $1 - \beta_\mathrm{F}$ and the expected number of true positives $E(R_1)$ can be calculated from $1 - \beta_1$ using (21) and (22).

	Family type I error probability											
	$\alpha_\mathrm{F} = 0.01$				$\alpha_\mathrm{F} = 0.10$				$\alpha_\mathrm{F} = 0.50$			
	Distance $\lvert\Delta_1\rvert/\sigma_\mathrm{D}$				Distance $\lvert\Delta_1\rvert/\sigma_\mathrm{D}$				Distance $\lvert\Delta_1\rvert/\sigma_\mathrm{D}$			
	1.0	1.5	2.0	2.5	1.0	1.5	2.0	2.5	1.0	1.5	2.0	2.5
Genes G_0	*Power = proportion correctly declared as expressed = 0.90*											
500	31	14	8	5	26	12	7	5	21	10	6	4
1000	33	15	9	6	27	12	7	5	23	11	6	4
2000	35	16	9	6	29	13	8	5	25	11	7	4
8000	38	17	10	7	32	15	8	6	28	13	7	5
Genes G_0	*Power = proportion correctly declared as expressed = 0.99*											
500	44	20	11	7	37	17	10	6	32	15	8	6
1000	46	21	12	8	39	18	10	7	34	15	9	6
2000	48	22	12	8	41	19	11	7	36	16	9	6
8000	52	23	13	9	45	20	12	8	41	18	11	7

the statistical distance (that is, the number of standard deviations) between the treatment and control log-expression levels under the alternative hypothesis. An examination of Table I(a) shows that the required sample size is most sensitive to the ratio $|\Delta_1|/\sigma_{\mathrm{D}}$ and the required power level and least sensitive to the mean number of false positives $E(R_0)$. The required sample size is also moderately sensitive to the number of unexpressed genes G_0 because of the effect of controlling for simultaneous inferences. The practical lesson to be drawn from this last observation is that the gene set $\mathscr{G}_0$ should be kept as small as possible, consistent with the scientific objective of the microarray study. Inclusion of superfluous genes in the analysis, possibly for reasons of data exploration or data mining, will have a cost in terms of power loss. Of course, housekeeping genes and genes included on the arrays as positive controls may be used for diagnostic and quality-control checks but do not enter the main analysis. Such monitoring genes should not be counted in the number G_0 used in power calculations.

As one example of a reference to Table I(a), consider a study for which $G_0 = 2000$ unexpressed genes. The investigator wishes to control the mean number of false positives at $E(R_0) = 1.0$ and to detect a twofold difference between treatment and control conditions with an individual power level of 0.90. Previous studies by the investigator may suggest that the standard deviation of gene expression differences in matched pairs will be about $\sigma_{\mathrm{D}} = 0.5$ on a log-2 scale. The twofold difference represents a value of $\log_2(2) = 1.00$ for $|\Delta_1|$ on a log-2 scale. Thus, the ratio $|\Delta_1|/\sigma_{\mathrm{D}}$ equals $1.00/0.5 = 2.0$. Reference to Table I(a) for these specifications shows that $n = 6$. Thus, six pairs of treatment and control conditions are required in the study. The specified individual power level of 0.90 indicates that 90 per cent of the differentially expressed genes are expected to be discovered.

Should an investigator wish to avoid the assumption that the estimated differential expression vectors $\hat{\mathcal{I}}_g$ are mutually independent across genes and use the Bonferroni approach, then the required sample sizes are those shown in Table I(b). As shown in (16), we have $E(R_0) = G_0\alpha_0 = \alpha_{\mathrm{F}}$ in the Bonferroni approach. Thus, the expected number of false positives is necessarily smaller than 1 and, hence, cannot be controlled at an arbitrary level. As a consequence, Table I(b) is entered with reference to the desired level of the family type I error probability α_{F}. Three values of α_{F} are displayed in the abbreviated table, namely, 0.01, 0.10 and 0.50. As α_{F} approaches 1, the sample sizes approach those shown for $E(R_0) = 1$ in Table I(a). To illustrate Table I(b), consider the situation where α_{F} is set at 0.10, $G_0 = 2000$, $|\Delta_1|/\sigma_{\mathrm{D}} = 2.0$ and an individual power level of 0.90 is desired. In this case, Table I(b) indicates that eight pairs of treatment and control conditions are required in the study.

6.2. *Completely randomized design*

Tables I(a) and I(b) can also be used for a completely randomized design in which there are equal numbers of treatment and control conditions but they are not matched pairs. In this case, the variance of the difference in log-expression between treatment and control is given by $\sigma_{\mathrm{D}}^2 = 2\sigma^2$, where σ^2 is the experimental error variance of gene log-expression. By replacing ratio $|\Delta_1|/\sigma_{\mathrm{D}}$ by $|\Delta_1|/\sqrt{2}\sigma$, the sample sizes in Tables I(a) and I(b) can be used for a completely randomized design. The benefit of matching pairs of treatment and control conditions is indicated by the extend to which σ_{D}^2 is smaller than $2\sigma^2$.

To illustrate, suppose σ is anticipated to be 0.40 in a completely randomized design. Furthermore, suppose that $\Delta_1 = 1.00$, $E(R_0) = 1.0$, $G_0 = 2000$ and the desired individual power level is 0.90 as specified before. Then, reference is made to the ratio $|\Delta_1|/\sqrt{2}\sigma = 1.00/\sqrt{2}(0.40) = 1.77$ in the table. From Table I(a), the required sample size can be seen to be somewhere between 6 and 11. An exact calculation gives $n = 8$ (calculations not shown).

6.3. Isolated-effect design

Many microarray studies anticipate the presence of differential gene expression somewhere among the C experimental conditions but the investigator does not know in advance where the differential expression will appear among the C conditions. The science underpinning these studies is often at a formative stage so they are essentially exploratory in nature. The quadratic summary measure is useful in this kind of case.

Consider a microarray study in which one experimental condition, which we refer to as the *distinguished condition*, exhibits differential expression for a gene g relative to all other $C-1$ conditions under study. The latter $C-1$ conditions are assumed to be uniform in their gene expression. Without loss of generality, we take this distinguished condition as $c = 1$ and assume that the target difference in expression between condition $c = 1$ and all other conditions is Δ_1 on the log-intensity scale. This assumption implies that the interaction parameters $\mathcal{I}_{gc}$ have the following values under the alternative hypothesis H_1:

$$\mathcal{I}_{gc} = \begin{cases} \Delta_1(C-1)/C & \text{for } c = 1 \quad \text{distinguished condition} \\ -\Delta_1/C & \text{for } c = 2, \ldots, C \quad \text{all other conditions} \end{cases} \tag{34}$$

Observe that these parameter values sum to zero as required by the interaction sum constraint. The assumption that the differential gene expression occurs only in one isolated condition is conservative in the sense that it poses the most challenging situation for detection by the investigator. The differential expression in question may be either an up- or down-regulation, depending on the sign of the difference Δ_1. Finally, we assume that the microarray study is replicated r times. Hence, with this design, there are rC readings on each gene.

We now apply the quadratic summary measure of differential gene expression to this isolated-effect design. If the error variance of the ANOVA model is denoted by σ^2 then the non-centrality parameter (27) for the quadratic summary measure has the following form for this design (derivation not shown)

$$\theta_1 = \boldsymbol{\mathcal{I}}_{\mathrm{R}}^{d'} \boldsymbol{\Sigma}_{\mathrm{R}}^{-1} \boldsymbol{\mathcal{I}}_{\mathrm{R}}^{d} = \frac{r(C-1)}{C} \left(\frac{\Delta_1}{\sigma} \right)^2 \tag{35}$$

We note in (35) that the non-centrality parameter depends strongly on the number of replicates r and the statistical distance between the log-expression levels for the distinguished condition and all other conditions, as measured by the ratio $|\Delta_1|/\sigma$. The effect of the number of conditions C is less pronounced as the ratio $(C-1)/C$ approaches 1 as C increases.

6.3.1. Power table for isolated-effect design. Table II shows individual power levels $1-\beta_1$ for this design, expressed as percentages. Quantity $E(R_0)$ denotes the mean number of false positives. Parameter θ_1 is the non-centrality parameter for this design given in (35) and C denotes the number of experimental conditions. Gene number G_0 is the anticipated number

Table II. Power table for isolated-effect design with quadratic summary of differential expression. The power calculation is based on the quadratic summary function for the isolated-effect design. The number listed in each cell is the individual power level $1-\beta_\text{I}$ (in per cent); this value is the expected percentage of truly expressed genes that will be correctly declared as expressed by the tests. The family power level $1-\beta_\text{F}$ and expected number of true positives $E(R_1)$ can be calculated from $1-\beta_\text{I}$ using (19) and (20).

	Mean number of false positives											
	$E(R_0)=1$				$E(R_0)=2$				$E(R_0)=3$			
	Non-centrality θ_I				Non-centrality θ_I				Non-centrality θ_I			
	20	25	30	35	20	25	30	35	20	25	30	35
Genes G_0	*Number of conditions $C=5$*											
500	76	89	95	98	82	92	97	99	85	94	98	99
1000	70	85	93	97	76	89	95	98	80	91	96	99
2000	63	80	90	96	70	85	93	97	74	87	95	98
8000	50	69	83	92	57	75	87	94	60	78	89	95
Genes G_0	*Number of conditions $C=10$*											
500	58	76	87	94	66	81	91	96	70	85	93	97
1000	51	69	83	91	58	76	87	94	63	79	89	95
2000	44	63	78	88	51	69	83	91	55	73	85	93
8000	31	50	67	80	37	56	72	84	41	60	76	87
Genes G_0	*Number of conditions $C=20$*											
500	38	55	71	82	45	63	77	87	51	68	81	89
1000	31	48	64	77	38	55	71	82	42	60	74	85
2000	24	40	57	71	31	48	64	77	35	52	68	80
8000	15	28	44	59	19	34	50	65	22	38	54	69

Quantity $E(R_0)$ denotes the mean number of false positives.

Parameter θ_I is the non-centrality parameter for this design given in (35) which depends strongly on the number of replicates r and the ratio $|\Delta_1|/\sigma$.

Number C denotes the number of specimens or experimental conditions and, thus, rC is the number of readings on each gene.

Gene number G_0 denotes the number of unexpressed genes involved in the experiment.

Estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ are assumed to be mutually independent across genes.

of unexpressed genes involved in the experiment. If G_0 is expected to be similar to the total gene count G, the table could be entered using G without introducing great error. Estimated differential expression vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ are assumed to be mutually independent across genes.

As one example of a reference to Table II, consider a study involving $C=5$ experimental conditions and $G_0=2000$ unexpressed genes. Assume that the investigator wishes to control the mean number of false positives at $E(R_0)=1$ and to detect an isolated effect that amounts to a twofold difference between the distinguished condition and all others. The experimental error standard deviation is anticipated to be about $\sigma=0.40$ on a log-2 scale. The twofold difference represents a value of $\log_2(2)=1.00$ for $|\Delta_1|$ on a log-2 scale. Thus, the ratio $|\Delta_1|/\sigma$ equals $1.00/0.40=2.5$. Six replications are to be used ($r=6$). For these specifications, the non-centrality parameter (35) equals

$$\theta_1 = 6\,\frac{(5-1)}{5}\,(2.5)^2 = 30$$

Thus, reference to the cell corresponding to $E(R_0)=1,\ \theta_1=30,\ C=5$ and $G_0=2000$, shows an individual power level of $1-\beta_1=0.90$ or 90 per cent. Thus, 90 per cent of differentially expressed genes are expected to be discovered with this study design. The table can be used iteratively to explore the effect on power of specific design changes. For example, if $r=7$ replications were to be used in lieu of $r=6$ then recalculation of the non-centrality parameter gives $\theta_1=35$ and the individual power level is seen to rise to 96 per cent.

7. RELATION BETWEEN POWER, REPLICATION AND STUDY DESIGN

With given specifications for the family type I error probability $\alpha_{\rm F}$ and the vector $\boldsymbol{\mathcal{I}}^d$ in the alternative hypothesis, the power is determined by the properties of the distribution of $\hat{\boldsymbol{\mathcal{I}}}_g$ and, in particular, its covariance matrix $\boldsymbol{\Sigma}$. In Section 4 it was shown how the covariance matrix affects the variance of the null PDF $f_0(v)$ in (24) and the non-centrality parameter in (27). Both the sample size and experimental design influence this covariance matrix.

7.1. *Effects of replication*

If a given microarray design is repeated so that there are r independent repetitions of the design, then $\boldsymbol{\Sigma}$ for a single replicate is reduced by a multiple of $1/r$. Thus, σ_0^2 in (24), for example, is reduced by the factor $1/r$ and the non-centrality parameter θ_1 in (27) is multiplied by r. These reductions are illustrated in the previous standard designs we used to demonstrate the power and sample size methodology. For instance, referring to (33) for the matched-pairs design, we can see that $\sigma_{\rm D}^2$ is the variance of a single matched pair and n plays the role of the number of replicates (the number of matched pairs in this instance). Note that $\sigma_{\rm D}^2$ is reduced by the multiplier $1/n$ with n replications. Similarly, with the isolated-effect design, the number of replications r appears as a multiplier in the formula for the non-centrality parameter in (35).

The replication discussed here refers to the simple repetition of a basic experiment and, hence, we are considering a pure statistical effect that is captured by the number of repetitions r. This parameter does not reveal if the replicated design is a sound one in terms of the scientific question of interest. We trust that it is sound but do not explore this issue in the paper.

We do want to point out, however, that the nature of replication is an important issue in terms of the overall study plan. Contrast the following two situations. First, imagine an experiment in which a single small tissue fragment is cut from a tumour core and then used to prepare six arrays. Next, imagine an alternative experiment in which six small tissue fragments are cut from six different regions of the same tumour in a spatially randomized fashion and then used to prepare six arrays, one array being prepared from each of the tissue fragments. Both imaginary designs yield six arrays of data. The first design allows inferences to be made only about the single core tissue fragment based on a sample of size six. The ANOVA model for this design describes the population of all arrays that could be constructed from this single tissue fragment. The second design allows inferences about the whole tumour based on a sample of size six. The ANOVA model for this design describes the population of all arrays that could be prepared from the whole tumour. The respective sets of inferences clearly relate to different biological populations (the single tumour core fragment and the whole tumour, respectively). Both designs involve six replications ($r=6$) but the replications have different elemental designs. The variance structures of the two designs will differ, of course, making them essentially incommensurate. The method of calculating power, however, based on the two designs follows the logic we explained in the preceding paragraph. The choice of design here depends on the target population, is it the single core tissue fragment or the whole tumour that is of scientific interest to the investigator.

7.2. Controlling sources of variability

The choice of experimental design, as opposed to simple replication, has a more complicated influence on power. A good design will be one that takes account of important sources of variability in the microarray study and reduces the experimental error variance of the expression data. Kerr and Churchill [1], for example, discuss a number of alternative experimental designs for microarray studies that aim to be more efficient. Schuchhardt *et al.* [10] describe some of the many sources of variability in microarray studies including, among others, the probe, target and array preparation, hybridization process, background and overshining effects, and effects of image processing. Experience with different designs will give some indication of the correlation structure and the magnitudes of variance parameters that can be expected in covariance matrix $\mathbf{\Sigma}$. These expectations, in turn, can be used to compute the anticipated power.

To give a concrete illustration, suppose that an ANOVA model is modified by adding a main effect for the subarray in which each spot is located, the aim being to account for regional variability on the surface of the slide for the microarray. Furthermore, suppose that incorporation of this main effect would reduce the error variance by 15 per cent, other factors remaining unchanged. Then, the covariance levels in $\mathbf{\Sigma}$ are reduced by a multiple of 0.85. The direct effect of this refinement on the numerical example in Section 4.2, for instance, is to increase the non-centrality parameter θ_1 by a factor of $1/0.85=1.1765$ from 20 to 23.53. This change increases individual power level $1-\beta_1$ from 0.689 to 0.806, a worthwhile improvement.

8. ASSESSING POWER FROM MICROARRAY PILOT STUDIES

The purpose of a power calculation is to assess the ability of a proposed study design to uncover a differential expression pattern having the target specification $\mathcal{I}^d$. Thus, our methodology should find its main application at the *planning stage* of microarray studies. As part of this planning process, investigators sometimes wish to calculate the power of a pilot study in order to decide how the pilot study should be expanded to a full study or to decide on the appropriate scale for a new and related study. Power calculation for a pilot study involves an application of the same methodology but with the benefit of having estimates of relevant parameters needed for the calculation from the pilot study data. For instance, power calculations need estimates of inherent variability. The pilot study data can provide those estimates. As

Table III. Microarray design for a study of juvenile cystic kidney disease (PKD) in mice.

Array	Colour channel	
	1. Green	2. Red
1	1. Mutant	1. Mutant
2	1. Mutant	2. Wild type
3	2. Wild type	1. Mutant
4	2. Wild type	2. Wild type

illustrations of power calculations from pilot studies and as demonstrations of real applications of our methodology, we now consider two microarray case studies involving mice.

8.1. Case example 1: juvenile cystic kidney disease (PKD)

Lee *et al.* [4] considered an experiment where mice with the juvenile cystic kidney mutation PKD were used. Litter mates, 33 days old, were genotyped. Homozygous (mutant) and wild type mice were identified. Two pairs of kidneys from homozygous and wild type mice were isolated and pooled separately. Total RNA was isolated and four comparative array hybridization pairs were set up as illustrated in Table III. The table shows how the tissue types (mutant or wild type) were assigned to the four arrays and two colour channels of each array. A total of $G = 1728$ genes were under investigation.

The scientists in this study were interested in differential gene expression for the two tissue types, mutant (type 1) and wild type (type 2). Thus, the difference $\hat{\mathcal{I}}_{g1} - \hat{\mathcal{I}}_{g2}$ was the summary measure of interest for gene g. The alternative hypothesis for which power was to be calculated was $H_1{:}\mu_1 = 1.00$. This specification corresponds to a target 2.72-fold difference between mutant and wild type tissues on the natural log scale. The study data gave an estimate of $\hat{\sigma} = 0.2315$ for the standard deviation of the summary measure on the same log-scale. Estimation errors in vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ were assumed to be independent. The expected number of false positives was to be controlled at $E(R_0) = 2$. We let the total gene count G stand in for G_0. Using the methodology presented in Section 4.1, the individual power level for the study was calculated to be $1 - \beta_1 = 0.858$, which suggests that 86 per cent of truly differentially expressed genes are expected to be discovered.

8.2. Case example 2: opioid dependence

In another experiment designed to investigate how morphine dependence in mice alters gene expression, Lee *et al.* [11] considered a study involving two treatments (morphine, placebo) and four time points corresponding to consecutive states of opioid dependence, classified as *tolerance*, *withdrawal*, *early abstinence* and *late abstinence*. In the experiment, mice received either morphine (treatment) or placebo (control). Treatment mice were sacrificed at four time points corresponding to the tolerance, withdrawal, early abstinence and late abstinence states. Control mice were sacrificed at the same time points, with the exception of the withdrawal state which was omitted on the assumption that the tolerance and withdrawal states are identical with placebo. The microarray data resulted from hybridization of mouse spinal cord samples to a custom-designed array of 1728 cDNA sequences. At each time point (that is, at each state),

Table IV. Microarray design for a study of opioid dependence in mice.

Treatment	Dependence time stage			
	1. Tolerance	2. Withdrawal	3. Early abstinence	4. Late abstinence
1. Placebo	array 1	*	array 2	array 3
2. Morphine	array 4	array 5	array 6	array 7

* Omitted, no array.

in both the treatment and control groups, three mice were sacrificed, for a total of 21 mice. The paucity of spinal column mRNA in any single mouse required that the mRNA of the three mice sacrificed together be combined and blended into a single sample. The treatment and control samples were labelled with red dye. Other control samples, derived from mouse brain tissue, were labelled with green dye. The green readings were not used in the analysis reported here.

The experimental design is shown in Table IV. The natural logarithm of the raw red intensity reading, without background correction, was used as the response variable. As noted above, no array was created for the placebo-withdrawal combination (marked * in Table IV). The original intention was to place the spinal column sample on two spots of the same slide, yielding a replicated expression reading. This attempt was not entirely successful. The replicate was missed in the morphine-tolerance combination because of administrative error. Also, the array for the morphine-late abstinence combination had a large number of defective spots. Finally, several dozen spots in other arrays were faulty. The final microarray data set contains readings for only $G = 1722$ genes out of the original set of 1728. Six genes were dropped because they had defective readings for the morphine-tolerance combination, the unreplicated treatment combination in the design in Table IV. The fact that the replicated spots were nested within the same array was not taken into account in the analysis.

The aim of the study was to identify genes that characterize the tolerance, withdrawal and two abstinence states and to describe how gene expression is altered as a mouse moves from one state to the next. As this aim is somewhat broad it was decided to evaluate power on the assumption that a differential expression would appear in only one treatment combination, with all other combinations having a uniform expression level. This assumption is exactly what characterizes the isolated-effect design and, hence, the quadratic summary measure is of interest. We shall use the isolated-effect design as a template for the power calculation, recognizing that this power value will slightly overstate the power achieved in this actual study because of the failure to replicate one of the seven treatment combinations and the nesting of duplicate spots within the same arrays.

The alternative hypothesis H_1, for which power was to be calculated, had the target differential expression pattern given in (34) with $|\Delta_1| = 0.693$, which corresponds to a twofold differential expression on the natural log-scale. Thus, the target specification calls for a single treatment combination to exhibit a twofold up- or down-regulation relative to all other treatment combinations. We assume there are $r = 2$ replicates for each of the $C = 7$ treatment combinations. The study data gave an estimate of $\hat{\sigma} = 0.1513$ for the standard deviation of the ANOVA error variance. Estimation errors in vectors $\hat{\boldsymbol{\mathcal{I}}}_g$ were assumed to be independent. The expected number of false positives was controlled at $E(R_0) = 1$. The non-centrality parameter, calculated from (35), was $\theta_1 = 36.00$. We let the total gene count G stand in for G_0. Now, using the methodology presented in Section 4.2, the individual power level was calculated to be $1 - \beta_1 = 0.944$, which implies that 94 per cent of truly differentially expressed genes are expected to be discovered. Approximately the same power value is found in Table II for $E(R_0) = 1$, $C = 10$, $\theta_1 = 35$ and $G_0 = 2000$, with further refinement being provided by interpolation.

ACKNOWLEDGEMENTS

The authors thank the reviewers for constructive suggestions that greatly improved the manuscript. They also acknowledge with thanks the financial support provided for this research by grants from the National Institute of Health (Lee, HG02510-01 and HL66795-02) and the Natural Sciences and Engineering Research Council of Canada (Whitmore).

REFERENCES

1. Kerr MK, Churchill GA. Experimental design issues for gene expression microarrays. *Biostatistics* 2001; **2**: 183–201.
2. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 2001; **7**:819–837.
3. Lee M-LT, Kuo FC, Whitmore GA, Sklar JL. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences* 2000; **97**:9834–9839.
4. Lee M-LT, Lu W, Whitmore GA, Beier D. Models for microarray gene expression data. *Journal of Biopharmaceutical Statistics* 2002; **12**:1–19.
5. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. 2001, see http://brooks.statgen.ncsu.edu/ggibson/Pubs.htm.
6. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report 578, 2000, Department of Biochemistry, Stanford University School of Medicine, Stanford, California. See http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html.
7. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**:1151–1160.
8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:289–300.
9. Delongchamp RR, Velasco C, Evans R, Harris A, Casciano D. Adjusting cDNA array data for nuisance effects. Division of Biometry and Risk Assessment, HFT-20, 2001, National Center for Toxicological Research, Jefferson, Arkansas.
10. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. Normalization strategies for cDNA microarrays. *Nucleic Acids Research* 2000; **28**(10):e47.
11. Lee M-LT, Whitmore GA, Yukhananov RY. Analysis of unbalanced microarray data. *Journal of Data Science* 2002; (in press).