Statistical Methods in Medical Research 2004; **13**: 325–338

Power and sample size estimation in high dimensional biology

Gary L Gadbury Department of Mathematics and Statistics, University of Missouri – Rolla, MO, USA, **Grier P Page, Jode Edwards** USDA ARS, Department of Agronomy, Iowa State University, Ames, IA, USA, **Tsuyoshi Kayo** Wisconsin Regional Primate Research Center, Madison, WI, USA, **Tomas A Prolla** Department of Genetics and Medical Genetics, University of Wisconsin, Madison, WI, USA, **Richard Weindruch** Department of Medicine, University of Wisconsin and The Geriatric Research, Education, and Clinical Center, William S Middleton VA Hospital, Madison, WI, USA, **Paska A Permana** Phoenix Epidemiology and Clinical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Phoenix, AZ, USA, **John D Mountz** The Birmingham Veterans Administration Medical Center, University of Alabama at Birmingham, Birmingham, AL, USA and **David B Allison** Department of Biostatistics, Section on Statistical Genetics, and Clinical Nutrition Research Center, University of Alabama at Birmingham, Birmingham, AL, USA

Genomic scientists often test thousands of hypotheses in a single experiment. One example is a microarray experiment that seeks to determine differential gene expression among experimental groups. Planning such experiments involves a determination of sample size that will allow meaningful interpretations. Traditional power analysis methods may not be well suited to this task when thousands of hypotheses are tested in a discovery oriented basic research. We introduce the concept of expected discovery rate (EDR) and an approach that combines parametric mixture modelling with parametric bootstrapping to estimate the sample size needed for a desired accuracy of results. While the examples included are derived from microarray studies, the methods, herein, are 'extraparadigmatic' in the approach to study design and are applicable to most high dimensional biological situations. Pilot data from three different microarray experiments are used to extrapolate EDR as well as the related false discovery rate at different sample sizes and thresholds.

1 Introduction

Although our age has been termed the 'postgenomic era', a more accurate label may be the 'genomic era'.[1] Draft sequences of several genomes coupled with new technologies allow study of entire genomes rather than isolated single genes. This opens a new realm of high dimensional biology (HDB), where questions involve multiplicity at unprecedented scales. HDB can involve thousands of genetic polymorphisms, gene expression levels, protein measurements, genetic sequences or any combination of these and their interactions. Such situations demand creative approaches to the inferential process of research. Although bench scientists intuitively grasp the need for flexibility in the

Address for correspondence: David B Allison, Department of Biostatistics, 1665 University Avenue, Ryals Public Health Building, Suite 327, University of Alabama at Birmingham, Birmingham, AL 35294, USA. E-mail: dallison@ms.soph.uab.edu

 10.1191/0962280204sm369ra

inferential process, elaboration of formal statistical frameworks supporting this are just beginning. Here, we use microarray experiments to illustrate a novel approach to sample size estimation in HDB.

Microarray experiments commonly aim to identify differences in gene expression among groups of experimental units differing in genotype, age, diet and so on. Although recent, the technology is rapidly advancing and texts are now available describing biological foundations and associated statistical methodologies.[2,3] Fold changes were initially used to highlight genes thought to be differentially expressed but did not quantify precision or statistical significance of estimated differences.[4,5]

Statistical tests quantify evidence, against the null hypothesis, that a particular gene is not differentially expressed across groups.[6–10] *P*-values quantify the level of type I error that would be committed were the null hypothesis true. Owing to many tests conducted, numerous false positives may occur in microarray experiments,[5,11] prompting a need to control inferential error rates. Traditionally, control has been sought over family-wise error rate (the probability of committing at least one type I error over the entire collection of tests).[12] However, setting a threshold for this rate may be inconsistent with the *zeitgeist* of HDB. Genomicists are beginning to embrace false discovery rate (FDR) control as an alternative. FDR is the expected proportion of rejected null hypotheses, for which the null hypothesis is actually true.[13,14] It is a number that can be set by researchers depending on how willing they are to further investigate a gene whose expression level may not actually differ between groups.

The extraordinarily useful classical frequentist testing paradigm may not be optimal for basic scientists testing thousands of null hypotheses simultaneously, who will follow-up promising leads with subsequent research.[15] Such scientists conducting microarray experiments have interest in proportions related to two quantities: the expected number of genes that are 1) differentially expressed and will be detected as significant at a particular threshold and 2) not differentially expressed and will not be detected as such. These two numbers are denoted by D and A, respectively, in Table 1 along with the expected number of genes that are differentially expressed but are not so declared (B), and are not differentially expressed but are so declared (C). Although we refer to numbers and proportions of genes tested for differential expression in a microarray study, the concepts and formulae presented apply to any situation in which many null hypotheses are tested.

This paper focuses on three proportions:

$$\mathrm{TP} = \frac{D}{C+D}, \quad \mathrm{TN} = \frac{A}{A+B}, \quad \mathrm{EDR} = \frac{D}{B+D} \tag{1}$$

Table 1 Quantities of interest in microarray experiments

	Genes for which there is not a real effect	Genes for which there is a real effect
Genes not declared significant at designated threshold	*A*	*B*
Genes declared significant at designated threshold	*C*	*D*

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

Each proportion is defined as 0 if its denominator is 0. TP is true positive; TN is true negative and EDR is the expected discovery rate, which is the expected proportion of genes that will be declared significant at a particular threshold among all genes that are truly differentially expressed. EDR is akin but not identical to the notion of power. Ideal studies would have TP, TN and EDR close to 1.0. In practice, the closeness depends on sample size. Recently, Lee and Whitmore[16] studied sample size effects on FDR, which is 1 − TP, and a quantity they call power, which is analogous to EDR in Equation (1). However, power is generally defined as the probability of rejecting a null hypothesis given that it is false, whereas EDR does not denote this probability for any particular hypothesis. EDR is the average power across all null hypotheses tested, and there may be no specific hypothesis for which power equals the EDR. We extend these concepts, adapting a method reported in Allison *et al.*,[17] who made use of finite mixture distributions. Finite mixture distributions have found use in other high dimensional biological studies,[18] which also include microarray data analysis.[19] See, for example, Titterington *et al.*[20] for general background information on mixture distributions, and Everitt[21] for some descriptions of other applications in medicine/biology.

Allison *et al.*[17] used a mixture of uniform probability density function (PDF) and one or more beta PDFs to model a distribution of *P*-values obtained from a statistical test for differential expression for each gene in a microarray experiment. Starting with the results of such mixture modelling, the approach, herein, allows investigators to quantitatively assess TP, TN and EDR, and plan future experiments by quantifying the role of sample size in increasing these proportions to desired levels.

2 A mixture model approach

2.1 Description

Consider a two-group experiment with $N = 2n$ microarray chips, n chips per group (assaying k genes). For each gene, a null hypothesis of no difference in mRNA level between groups, H_{0i}: $\delta_i = 0$, $i = 1, \ldots, k$ (δ_i = true population mean difference in mRNA level between experimental conditions for the ith gene), is tested with a valid test statistic, generating k *P*-values. The k hypothesis tests can be used simultaneously to test a global null hypothesis that no differences in mRNA levels exist for any of the k genes, H_0: $\delta_i = 0$, $i = 1, \ldots, k$, versus an alternative hypothesis that mRNA levels differ between groups for a subset of m genes ($0 < m \leq k$).

Allison *et al.*[17] modelled the *P*-value distribution as a mixture of $v + 1$ beta distributions on the interval [0, 1].[22] Their technique applicable to any test producing valid *P*-values, may be considered more general than the use of mixtures of normal distributions.[19] The PDF of a random variable X that follows the beta distribution with parameters r and s is given by

$$\beta(x|r, s) = I_{(0,1)}(x)\frac{x^{r-1}(1-x)^{s-1}}{B(r,s)}, \quad \text{where}$$

$$B(r, s) = \int_0^1 u^{r-1}(1-u)^{s-1}\mathrm{d}u \text{ and } I_{(0,1)}(x) = 1 \text{ if } x \in (0,1)$$

and is otherwise equal to 0. Their mixture model can be expressed as

$$f(\underline{p}) = \prod_{i=1}^{k} \sum_{j=0}^{\nu} \lambda_j \beta(p_i | r_j, s_j) \tag{2}$$

where p_i is the *P*-value from a test on the *i*th gene, $\underline{p}$ is the vector of k *P*-values and λ_j is the probability that a randomly sampled *P*-value belongs to the *j*th component beta distribution, $\sum_{j=0}^{\nu} \lambda_j = 1$ and $r_0 \equiv s_0 \equiv 1$, thus indicating a uniform distribution for the initial component of the mixture model. If no genes are differentially expressed, the sampling distribution of *P*-values will follow a uniform distribution on the interval [0, 1]. If some genes are differentially expressed, additional beta distribution components will be required to model a set of *P*-values clustering near 0.

The number of components in the mixture model and estimates of parameters $\lambda_j, r_j, s_j, j = 0, \ldots, \nu$ can be obtained via maximum likelihood combined with a parametric bootstrap procedure.[23] The need for such a procedure (versus a usual likelihood ratio test) stems from the fact that regularity conditions do not hold for the asymptotic distribution of the likelihood ratio test statistic to be chi-squared.[24] More details of the implementation of this parametric bootstrap procedure are in Allison *et al.*[17] The global null hypothesis is H_0: $\lambda_1 = 0$, which questions whether differences in expression are evident for any of the k genes. For most data sets we have analysed using this method, a uniform distribution plus one beta distribution was sufficient for modelling *P*-values when H_0: $\lambda_1 = 0$ was rejected. Allison *et al.*[17] showed that simulation tests can evaluate the effect of correlated expression levels among some genes on estimated parameters in the mixture model. Maximum likelihood estimates (MLEs) of parameters in Equation (2) can be obtained using numerical methods. Evaluating the logarithm of Equation (2) at these estimates is the log-likelihood that helps to quantitatively distinguish a strong signal (i.e., evidence that $\lambda_1 > 0$) from a weaker one. The larger this value, the more certainty that there is a signal in the distribution of *P*-values.

We will adopt the convention of referring to the quantity calculated in Equation (2) as 'likelihood', although this terminology is only strictly correct if all the *P*-values are independent. Although the model in Equation (2) suggests independence of the *P*-values from k tests, the value of the likelihood remains a valid indicator of relative model fit if *P*-values are dependent. The global null hypothesis, H_0: $\lambda_1 = 0$, can be tested by incorporating potential dependency into a parametric bootstrap and a simulation procedure.

A key result from the fitted mixture model is a posterior probability curve, representing the probability that members of a set of genes are truly differentially expressed given that all members of the set have observed *P*-values $\leq p_i$. We will call this a TP for a specific gene set, denoted $\mathrm{TP}_i, i = 1, \ldots, k$. Suppose that the fitted model includes a uniform plus one beta component. Then

$$\mathrm{TP}_i = \frac{\lambda_1 B(p_i; r, s)}{\lambda_0 p_i + \lambda_1 B(p_i; r, s)} \tag{3}$$

where $B(p_i; r, s)$ is the cumulative distribution function of a beta distribution with parameters r and s, evaluated at p_i. A plot of TP_i versus p_i produces the TP posterior

probability plot. Similarly, producing a TN posterior probability plot requires computing

$$\mathrm{TN}_i = \frac{\lambda_0(1-p_i)}{\lambda_0(1-p_i)+\lambda_1[1-B(p_i;r,s)]} \tag{4}$$

The TP and TN posterior probability curves are estimated curves because the mixture model's parameters are estimated from data. As MLEs can be computed for λ_1, λ_0, r and s, an MLE can also be computed for TP and TN at any specified value of P. Moreover, standard errors can be estimated reflecting the uncertainty in the MLEs due to the distribution of P-values and the fitted model. Allison *et al.*[17] used a bootstrap procedure to estimate standard errors in a data example, and they assessed the behaviour of these estimates for varying distributional assumptions in a simulation study utilizing a nested bootstrap procedure.

Allison *et al.*[17] noted that the mixture modelling method as described works well for cases when the TP posterior probability plot is monotonically decreasing, that is, for $p_i < p_j, \mathrm{TP}_i \geq \mathrm{TP}_j$. This implies that the beta distribution is being used to model smaller P-values in the distribution. A TN posterior probability plot should be monotonically increasing.

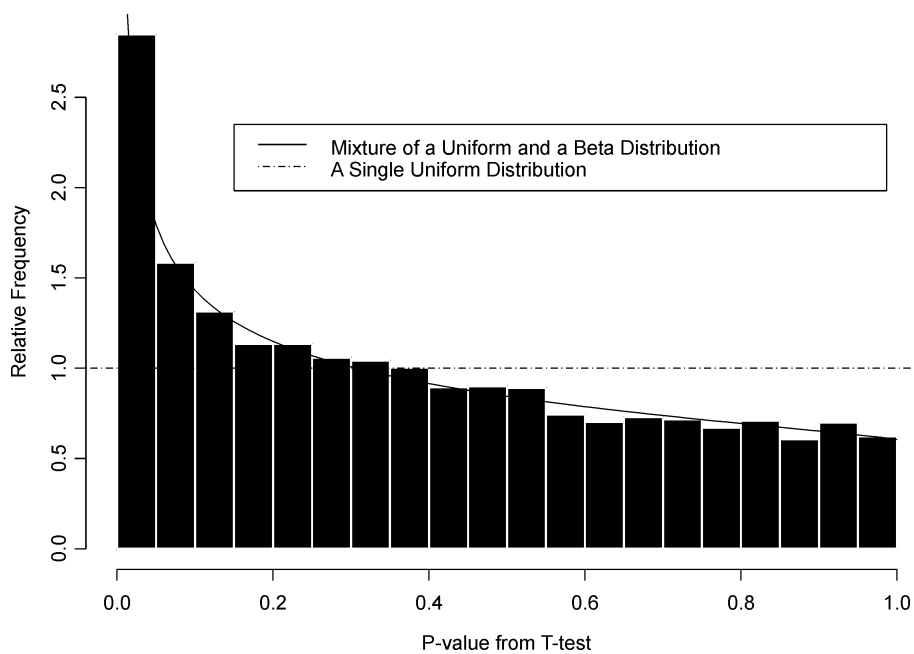

Figure 1 Fitted mixture model to 12 625 P-values. Dashed line represents a model for which the global null hypothesis would be true. The mixture model with an added beta distribution component captures the cluster of small P-values.

2.2 An example

In the first of three examples, human rheumatoid arthritis synovial fibroblast cell line samples were stimulated with tumour necrosis factor-α, where one group ($n = 3$) had the NF-κB pathway taken out by a dominant negative transiently transfected vector and the other group ($n = 3$) had a control vector added. Figure 1 shows a histogram of *P*-values obtained from two sample *t*-tests on $k = 12\,625$ genes. Many more *P*-values than expected under the global null hypothesis cluster near 0. A mixture of a uniform plus one beta distribution captures this shape. The mixture model is the solid line, represented by an equation of the form

$$f^*(\underline{p}) = \prod_{i=1}^{k} [\lambda_0 + \lambda_1 \beta(p_i; r, s)], \quad p_i \in (0, 1),\ i = 1, \ldots, k \tag{5}$$

For these data, λ_1 is estimated as 0.395, suggesting that 39.5% of the genes are differentially expressed – an unusually strong signal and not representative of the many microarray data sets we have analysed. The estimates for r and s are 0.539 and 1.844, respectively. The log-likelihood is equal to 1484 versus 0, which would be the value for a strictly uniform distribution (i.e., a distribution with no signal). This value of 1484 represented a marked departure from the null hypothesis that no genes are differentially expressed. Simulation studies suggest that this conclusion would hold even when high correlation of expression levels was present among some genes.[17]

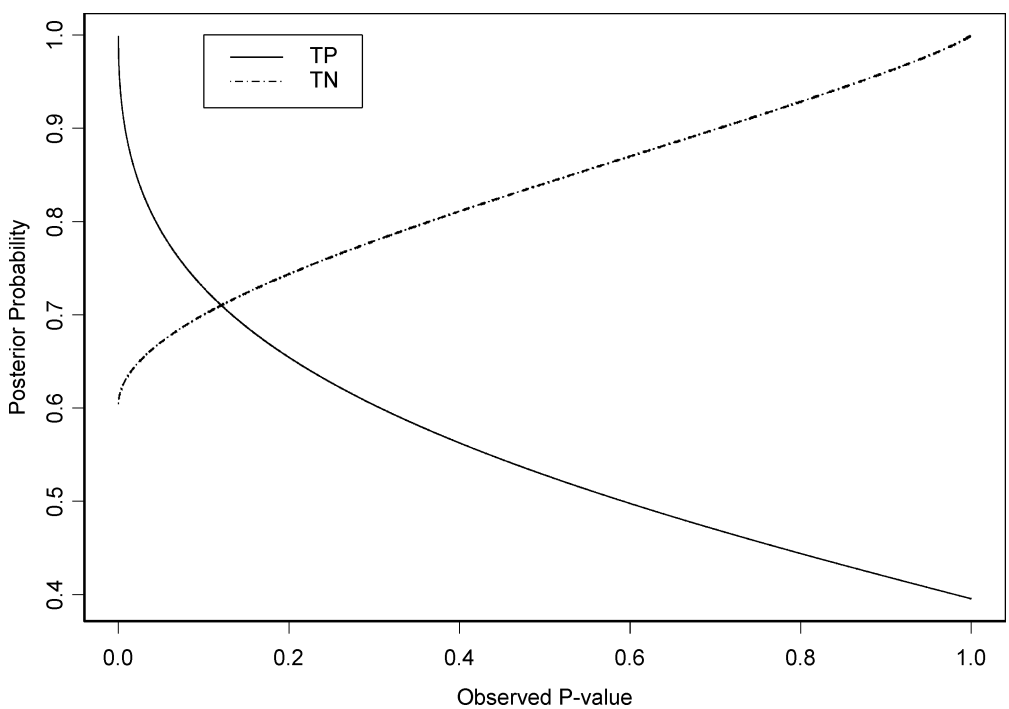

Figure 2 Posterior probability plots. TP and TN probabilities versus *P*-value for 12 625 genes from Example 1 data (arthritis).

More useful to the scientist are measures of TP and TN for the *P*-value thresholds corresponding to the *P*-values observed for each gene. Figure 2 shows posterior probability plots of TP and TN, computed using Equations (3) and (4), respectively. The plot illustrates the monotonicity of the curves. For example, if genes with *P*-values ≤ 0.1 are differentially expressed in this data set, $\approx 80\%$ of those conclusions will be true positives. If genes with *P*-values > 0.1 are not differentially expressed in this data set, $\approx 70\%$ of those conclusions will be true negatives. A one to one correspondence between any particular *P*-value and a value of TP or TN only holds within (and not across) data sets. More specifically, if the threshold to declare a gene differentially expressed is set to 0.1, then $\widehat{\text{TP}}$ is equal to 0.73 with a bootstrap estimated standard error of 0.014. $\widehat{\text{TN}}$ is equal to 0.70 with estimated standard error of 0.032. If the threshold is set to 0.001, then $\widehat{\text{TP}}$ is equal to 0.96 with a bootstrap estimated standard error of 0.003, and $\widehat{\text{TN}}$ is equal to 0.61 with estimated standard error of 0.028. Standard errors reported here reflect uncertainty in the fitted model to the distribution of *P*-values. Low values of the actual estimate, $\widehat{\text{TP}}$, reflect uncertainty due to smaller sample sizes. Estimates of TP will tend to become larger with increasing sample sizes.

3 Using the mixture model for sample size effects on TP, TN and EDR

Suppose that a researcher has fitted a mixture model to a distribution of *P*-values obtained from a pilot study or a study similar to one being planned. This model is now assumed to be fixed, meaning that the estimated model from the initial study is $f^*(\underline{p})$ from Equation (5) (estimated parameters are the true parameters in the model). Assume that there is an evidence from the model that some genes are differentially expressed. This model will be used to evaluate a chosen threshold and sample size on TP, TN and EDR, given in Equation (1).

3.1 Estimating TP, TN and EDR

Let N be the number of experimental units (e.g., N microarray chips), let $Z = \{1, 2, \ldots, k\}$ be a set of indices corresponding to the genes in the study and let T be a subset of Z representing the set of genes that have a true differential expression across two experimental groups (i.e., $T \subseteq Z$). In practice, T is unknown. One purpose of a microarray study is to identify T. Let

$$I_{\{T\}}(i) = \left\{ \begin{array}{ll} 1 & i \in T \\ 0 & i \notin T \end{array} \text{ for } i = 1, \ldots, k \right\}$$

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

A gene is declared to be differentially expressed if the *P*-value (calculated on observed data) from a statistical test falls below a predetermined threshold (τ). The resulting decision function, when equal to 1, declares a gene differentially expressed:

$$\psi_i(\underline{x}_i) = \left\{ \begin{array}{ll} 1 & p_i \leq \tau \\ 0 & p_i > \tau \end{array} \right\}$$

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

Estimates for the values in Table 1 that can be calculated in computer simulation experiments are given by

$$\begin{aligned} \hat{A} &= \sum_{i=1}^{k} (1-\psi_i)[1 - I_{\{T\}}(i)] \quad \hat{B} = \sum_{i=1}^{k} (1-\psi_i) I_{\{T\}}(i) \\ \hat{C} &= \sum_{i=1}^{k} \psi_i [1 - I_{\{T\}}(i)] \qquad\qquad \hat{D} = \sum_{i=1}^{k} \psi_i I_{\{T\}}(i) \end{aligned} \tag{6}$$

To define A, B, C and D from Table 1, the expectations of the estimates in Equation (6) are taken with respect to the mixture model (5):

$$\begin{aligned} E(\hat{D}) &= E\left[\sum_{i=1}^{k} \psi_i I_{\{T\}}(i)\right] = \sum_{i=1}^{k} E[\psi_i I_{\{T\}}(i)] \\ &= \sum_{i=1}^{k} P[\psi_i = 1,\, I_{\{T\}}(i) = 1] \\ &= \sum_{i=1}^{k} P[\psi_i = 1 | I_{\{T\}}(i) = 1]\; P(I_{\{T\}}(i) = 1) \\ &= k\lambda_1 B(\tau; r, s) \\ &= D \end{aligned}$$

Similarly,

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

It is now evident that $\mathrm{TP} = D/(C+D)$ and $\mathrm{TN} = A/(A+B)$, defined in Equation (1), have the same form as TP_i and TN_i in Equations (3) and (4), respectively, except that p_i in the latter is replaced by the threshold τ in the former. $\mathrm{EDR} = D/(B+D)$, which simplifies to $B(\tau; r, s)$. In the following simulations, estimates of TP, TN and EDR will be computed using

$$\widehat{\mathrm{TP}} = \frac{\hat{D}}{\hat{C}+\hat{D}}, \quad \widehat{\mathrm{TN}} = \frac{\hat{A}}{\hat{A}+\hat{B}}, \quad \widehat{\mathrm{EDR}} = \frac{\hat{D}}{\hat{B}+\hat{D}} \tag{7}$$

These are consistent estimators for a given mixture model meaning, essentially, that estimates of TP, TN and EDR (given the fitted model) should be 'close' to true proportions when A, B, C and D in Table 1 are large. The estimators in Equation (7) allow for the evaluation of sample size effects on TP, TN and EDR. This differs from

previous work that produced estimates of upper bounds for an FDR, for example, Benjamini and Hochberg.[13]

3.2 Effects of varying sample size and threshold

For a computational look at the effect of threshold and sample size on TP, TN and EDR, again assume an experiment has been conducted with $N = 2n$ units divided into two groups of equal size, and a mixture model $f^*(\underline{p})$ (5) has been fitted to the distribution of P-values obtained from a t-test of differential expression on each gene. We use a t-test when describing the following procedure though a P-value from any valid test can be used as long as it can be back-transformed to the test statistic that produced it. The procedure is readily adaptable for a Welch corrected t-test. In cases where the validity of a P-value is questionable (due to small sample sizes and potentially skewed distributions), one can employ nonparametric randomization tests and compare this resulting distribution of P-values with that obtained from t-tests as a form of 'sensitivity check' as proposed in Gadbury *et al.*[25] Simulation study showed that t-tests performed quite well when compared with other distribution free tests when there are equal numbers of samples in each group, the situation considered here. Still, it is worth noting that the procedure, herein, depends on a test that produces valid P-values.

The following procedure is a parametric bootstrap routine[23,26] that can yield estimates of TP, TN and EDR for any given sample size and threshold. A sample $p^* = p_1^*, \ldots, p_k^*$ is randomly drawn from the mixture model $f^*(\underline{p})$, with its parameters estimated from the preliminary sample. The outcome of a Bernoulli trial first determines whether a p_i^* is generated from the uniform component with probability λ_0, that is, $I_{\{T\}}(p_i^*) = 0$ or the beta distribution component with probability $\lambda_1 = 1 - \lambda_0$, that is, $I_{\{T\}}(p_i^*) = 1$. So, values of $I_{\{T\}}(p_i^*)$ are known for each simulated P-value. From this sample of P-values, a set of adjusted P-values, $p^{**} = p_1^{**}, \ldots, p_k^{**}$, is created by transforming the p_i^* for which $I_{\{T\}}(p_i^*) = 1$ back to the corresponding t-statistic $t_i^* = t^{-1}[(1 - p^*/2), 2n - 2]$, where $t^{-1}(a,b)$ is the quantile of a t distribution with b degrees of freedom evaluated at a. An adjusted t-statistic, t_i^{**} is computed using the new sample size, that is, $t_i^{**} = t_i^* \sqrt{n^*/n}$, and a new P-value, p_i^{**}, is obtained using this new sample size, n^*. The p_i^* for which $I_{\{T\}}(p_i^*) = 0$ are left unchanged. From the new p^{**}, $\widehat{\text{TP}}$, $\widehat{\text{TN}}$ and $\widehat{\text{EDR}}$ are computed. The process is repeated M times, thus obtaining M values of $\widehat{\text{TP}}$, $\widehat{\text{TN}}$ and $\widehat{\text{EDR}}$. The value of M is chosen sufficiently large, so that Monte Carlo estimates of $E[\widehat{\text{TP}}]$, $E[\widehat{\text{TN}}]$ and $E[\widehat{\text{EDR}}]$ can be accurately estimated using the average over M values of $\widehat{\text{TP}}$, $\widehat{\text{TN}}$ and $\widehat{\text{EDR}}$, that is, in these simulations, the estimators for the three parameters of interest are $\widehat{E}_{\widehat{\text{TP}}} = \sum_{i=1}^{M} \widehat{\text{TP}}_i / M$, $\widehat{E}_{\widehat{\text{TN}}} = \sum_{i=1}^{M} \widehat{\text{TN}}_i / M$ and $\widehat{E}_{\widehat{\text{EDR}}} = \sum_{i=1}^{M} \widehat{\text{EDR}}_i / M$. Uncertainty in these estimators is attributed to what is only inherent in the simulations themselves. The original fitted mixture model has been assumed to be fixed very similarly to how effect sizes and prior variance estimates are considered fixed in traditional power calculations, although one could do otherwise.[27] Thus, estimated standard errors of $\widehat{E}_{(\cdot)}$ are readily available from the sample variance of the M values of $\widehat{\text{TP}}$, $\widehat{\text{TN}}$ and $\widehat{\text{EDR}}$, divided by M. Then, the earlier procedure described can be repeated for different values of n^* and τ.

Ranges of values for n^* can be chosen to reflect future practical sample sizes. Ranges for τ may be chosen from very liberal (e.g., 0.1) to more conservative (e.g., 0.00001). The value of τ chosen in actual practice by the researcher will depend on the researcher's interest in balancing EDR with TP and TN. A very small τ can, in theory, make TP close to 1 (i.e., genes that are declared significant will be the ones that are truly differentially expressed), but many important genes may be excluded for further follow-up investigation (values of TN will be lower). A very small threshold can also make EDR small in some studies when *P*-values from statistical tests do not reach this level to declare a gene interesting for follow-up study. This balancing of the effect of sample sizes and chosen thresholds on values of TP, TN and EDR is described and illustrated in Section 4.

4 Results

4.1 Description of two more example data sets

In the second example, aging, the study sought genes that are differentially expressed between CD4 cells taken from five young and five old male rhesus monkeys. Statistical significance for $k = 12\,548$ genes was assessed using pooled variance *t*-tests after quantile–quantile normalization. The mixture model estimated that 20% of genes were differentially expressed. However, the log-likelihood was only 159, indicating that the signal was less than in example 1, despite the larger sample. Simulation results in Allison *et al.*[17] indicate that this value, 159, could result from a situation where a subset of genes had differential expression, no genes had differential expression but moderate to high dependence was present among genes, or possibly a combination of both. The TP posterior probability would be expected to be lower, reflecting this additional uncertainty in values meaningful to the scientist.

For the third example, obesity, the study sought differences in gene expression between adipocytes from lean and obese humans; $k = 63\,149$ gene expression levels were measured for all subjects. The mixture model estimated that 31% of genes were differentially expressed. The value of the log-likelihood is 16 860, indicating that this signal is strong, partly because of the relatively large size of the experiment (i.e., 19 subjects per group).

4.2 Illustrating sample size and threshold effects

The first row of Figure 3 shows the minimum and maximum number (from $M = 100$ simulations) of genes (out of k) determined to be differentially expressed at three chosen thresholds for different sample sizes. The second row is a plot of the average 100 $\widehat{\text{TP}}$ values for the three thresholds at each sample size. The third and fourth rows show the average of the 100 $\widehat{\text{TN}}$ and 100 $\widehat{EDR}$ values, respectively. These are $\widehat{E}_{\widehat{\text{TP}}}$, $\widehat{E}_{\widehat{\text{TN}}}$ and $\widehat{E}_{\widehat{\text{EDR}}}$ defined earlier for $M = 100$.

The relationship between sample size and number declared significant (row 1) reveals key information about TP. At smaller sample sizes and at very low thresholds, τ, very few (and sometimes 0) genes are declared significant. This quantity estimates $C + D$ in Table 1, the denominator of TP. TP is defined to be 0 when $C + D$ is 0. Estimates, $\widehat{\text{TP}}$, are not expected to be very accurate when $\widehat{C} + \widehat{D}$ is a small positive number. This effect is seen in plots for TP (row 2) at lower values of sample size. These plots also show the

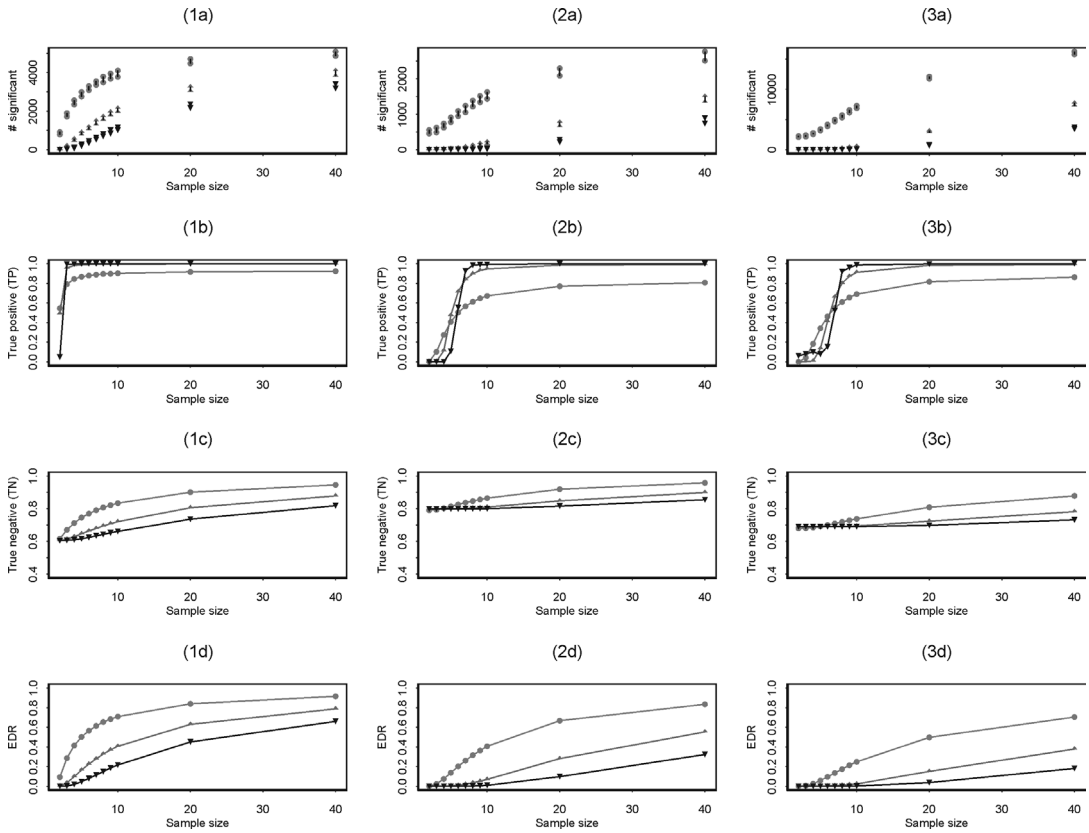

Figure 3 Effect of sample size, n^*, on estimated quantities of interest for example data sets. Row 1, number declared significant; row 2, TP; row 3, TN; row 4, EDR. Column 1, arthritis; column 2, aging; column 3, obesity. Three lines in each plot are for three selected thresholds: $\tau = 0.05$ (circles), $\tau = 0.001$ (triangles) and $\tau = 0.00001$ (inverted triangles).

crossing over of lines representing different thresholds. To illustrate, in the arthritis example, when $n^* = 2$ and $\tau = 0.00001$, $\widehat{E}_{\widehat{\text{TP}}} = 0.05$ and the estimated standard error is 0.02. At the same threshold with $n^* = 3$, $\widehat{E}_{\widehat{\text{TP}}} = 0.99$ and the estimated standard error is 0.001. In the aging example, when $\tau = 0.00001$, $\widehat{E}_{\widehat{\text{TP}}} = 0.10$ with estimated standard error of 0.03, even when $n^* = 5$. In the latter case, at such small thresholds, many simulations do not detect any genes differentially expressed until $n^* \geq 5$, thus making individual estimates of TP less stable at smaller sample sizes. In general, values of TP are higher for lower thresholds when the sample size is large enough to actually detect differentially expressed genes.

Estimates of the quantities $A + B$ and $B + D$ (i.e., the denominators of TN and EDR, respectively) are more accurate at small sample sizes and small thresholds because A and B are expected to be large for the data sets used here. However, estimates of EDR are small at these n and τ, because D is small. So, lines do not cross over in plots for EDR (row 4) because a smaller threshold makes it more difficult to detect differentially

expressed genes regardless of sample size. Because the denominator for EDR is expected to be large, as estimates for D increase with increasing sample size, the corresponding increase in estimates for EDR will be less dramatic than for TP. In the obesity example, for instance, even when $n^* = 20$ at $\tau = 0.00001$, $\widehat{\widehat{E}}_{\widehat{\text{EDR}}}$ is only 0.04 with estimated standard error equal to 0.0001. Estimates of standard error are always small (with respect to the mean) for this quantity, as the denominator for $\widehat{\widehat{E}}_{\widehat{\text{EDR}}}$ stays large in these simulations. Standard errors for $\widehat{\widehat{E}}_{\widehat{\text{TN}}}$ are also very small because of a combination of the large numbers of genes (i.e., translating into large numbers of expected values for A and B) and the chosen value for M. Because the original fitted mixture model was considered fixed, the standard errors in these simulations reflect simulation uncertainty rather than uncertainty in the fitted model and, thus, are expected to be relatively small with respect to the estimated mean. Moreover, in theory, these standard errors can be driven to any arbitrarily small value (effectively 0) by increasing M, the number of simulations.

The strongest signal in the three data sets is seen for Example 1 (arthritis) even though it had the smallest sample size. Initial results from the mixture model for Example 3 (obesity, recall the high value, 16 860, from the likelihood function for these data) might have indicated that these data were the strongest, but Figure 3 shows that the patterns seen in Example 3 (all plots) are similar to those for Example 2 (aging and all plots), which had the weakest signal from the mixture model. The key difference is in the samples sizes: $n = 5$ for Example 2 and $n = 19$ for Example 3. Thus, Figure 3 shows what might have happened in an experiment using a larger sample and also what might have happened if a smaller sample had been used.

From the three examples, it might be tempting to conclude that $n = 3$ per group is adequate with cell lines, as appears the case in Example 1. Although studies with cell lines may require fewer replicates given their presumed greater homogeneity, the information here is relevant only for one study, and generalizing based only on these data would be inappropriate. Similarly, concluding that studies of cell lines require fewer replicates than do studies of laboratory-housed monkeys which, in turn, require fewer replicates than do studies of free-living humans generalizes far beyond what the data here can establish. Finally, the data sets used here are not meant to represent all data sets that we have seen. Occasionally, the mixture model has detected a mode in the distribution of P-values away from 0, thus producing a posterior probability curve that is not monotonic. This effect may be due to strong dependence among some genes or unusually high variance in genes that are differentially expressed. This is a subject of continuing research.

5 Discussion

Advances in genomic technology offer new challenges and possibilities as data increase massively. Statistical science continues to evolve, partly in response to advances in other scientific disciplines. The 1970s witnessed the beginning of a revolution in statistical inference provoked by the wide availability of computing power to scientists. Efron described this revolution as 'thinking the unthinkable' by building inferential confidence

on computer-generated distributions rather than a priori selected theoretical distributions and analytic derivations.[28] 'Omic' technologies (genomic, proteomic and so on) that offer the possibility of testing thousands of hypotheses in single experiments create challenges and opportunities that may require another radical alteration in thinking.[29] One of the greatest challenges is the difficulty in estimating and obtaining sample sizes needed to ensure that interesting effects can be detected with some desired probability and without many false positives. The traditional approach to power and sample size estimation involves testing a single or very few hypotheses and selecting a sample size and significance threshold to ensure that power is high and type I error is low. This approach seems well suited to, for example, the context of clinical trials, where basic discovery has already been done, public application of findings may be immediate and cost of inferential error is high. Dramatically different is the context of basic research in the age of HDB, where promising discoveries are followed with further laboratory research. When there are potentially thousands of false null hypotheses, ensuring that the probability of *any* inferential error remains low seems less important than ensuring the generation of a sufficient list of findings meriting further investigation by having a low expected proportion of items mistakenly included. The approach offered, herein, allows achievement of that goal.

Experimenters may often have additional information at their disposal, either from other experiments, prior beliefs or knowledge about the differential measurement reliability for each of the dependent variables under consideration. A number of approaches, often Bayesian in nature, are potentially available for utilizing such information.[30–32] Incorporating such information into a procedure such as that described herein is a topic of future investigation.

Acknowledgements

This research was supported in part by NIH grants T32AR007450 R01DK56366, P30DK56336, P01AG11915, R01AG018922, P20CA093753, R01AG011653, U24DK058776 and R01ES09912; NSF grants 0090286 and 0217651; a grant from University of Alabama Health Services Foundation; and CNGI grant U54CA100949.

References

1 Wolfsberg TG, Wetterstrand KA, Guyer MS, Collins FS, Baxevanis AD. A user's guide to the human genome. *Nature Genetics* 2002; **32** (suppl.): 1–79.

2 Knudsen, S. *A biologists guide to analysis of DNA microarray data*. New York: John Wiley & Sons, Inc., 2002.

3 Speed T. ed. *Statistical analysis of gene expression microarray data*. London: Chapman and Hall/CRC, 2002.

4 Lee C, Klopp RG, Weindruch R, Prolla TA. Gene expression profile of aging and its retardation by caloric restriction. *Science* 1999; **285**: 1390–93.

5 Allison DB. Statistical methods for microarray research for drug target identification. In *2002 proceedings of the american statistical association [CD ROM]*. Alexandria, VA: American Statistical Association, 2002.

6 Ideker T, Thorsson V, Siehel AF, Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* 2000; **7**: 805–17.

7 Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving

statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 2001; **8**: 37–52.
8 Long AD, Mangalam HJ, Chan BYP, Tolleri L, Hatfield GW, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *Journal of Biological Chemistry* 2001; **276**: 19937–44.
9 Kerr KM, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 2000; **6**: 819–37.
10 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science, USA* 2001; **98**: 5116–21.
11 Allison DB, Coffey CS. Two-stage testing in microarray analysis: what is gained? *Journal of Gerontology, Biological Sciences* 2002; **57**: B189–92.
12 Hochberg Y, Tamhane A. *Multiple comparison procedures*. New York: Wiley, 1987.
13 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.
14 Keselman HJ, Cribbie R, Holland B. Controlling the rate of type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology* 2002; **55**: 27–39.
15 Howard GS, Maxwell SE, Fleming KJ. The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods* 2000; **5**: 315–32.
16 Lee M-LT, Whitmore GA. Power and sample size for DNA microarray studies. *Statistics in Medicine* 2002; **21**: 3543–70.
17 Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee C-K, Prolla TA, Weindruch R. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* 2002; **39**: 1–20.
18 Everitt BS, Bullmore ET. Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping* 1999; **7**: 1–14.
19 Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* 2002; **3**: 0022.1–0022.10.
20 Titterington DM, Smith AFM, Makov UE. *Statistical analysis of finite mixture distributions*. Chichester: John Wiley & Sons, 1985.
21 Everitt BS. An introduction to finite mixture distributions. *Statistical Methods in Medical Research* 1996; **5**: 107–127.
22 Parker RA, Rothenberg RB. Identifying important results from multiple statistical tests. *Statistics in Medicine* 1988; **7**: 1031–43.
23 Schork NJ. Bootstrapping likelihood ratios in quantitative genetics. In LePage R, Billard L, eds *Exploring the limits of bootstrap*. New York: Wiley, 1992: 389–93.
24 McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* 1987; **36**: 318–24.
25 Gadbury GL, Page GP, Heo M, Mountz JD, Allison DB. Randomization tests for small samples: an application for genetic expression data. *Applied Statistics* 2003; **52**: 365–76.
26 Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman and Hall, 1993.
27 Taylor DJ, Muller KE. Computing confidence bounds for power and sample size of the general linear univariate model. *The American Statistician* 1995; **49**: 43–47.
28 Efron B. Computers and the theory of statistics: thinking the unthinkable. *SIAM Revue* 1979; **21**: 460–80.
29 Donoho D, Candes E, Huo X, Stoschek A, Levi O. *Mathematical challenges of the 21st century*. Online presentation last accessed 12 May 2004 at: http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/MathChallengeSlides2*2.pdf
30 Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 2003; Suppl. 1: I84–90.
31 Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proceedings of the national academy of science, USA* 2003; **100**: 8348–53.
32 Savoie CJ, Aburatani S, Watanabe S, Eguchi Y, Muta S, Imoto S, Miyano S, Kuhara S, Tashiro K. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. *DNA Research* 2003; **10**: 19–25.