How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach

Wei Pan
Division of Biostatistics, School of Public Health, University of Minnesota

Jizhen Lin
Department of Otolaryngology, School of Medicine, University of Minnesota

Chap T. Le
Division of Biostatistics, School of Public Health, University of Minnesota

Revised February 2002

1Correspondence author. Division of Biostatistics, A460 Mayo, MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455-0378, USA. Email: weip@biostat.umn.edu, phone: (612)626-2705, and fax: (612)626-0660.
How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach

Abstract
It has been recognized that replicates of arrays (or spots) may be necessary for reliably detecting differentially expressed genes in microarray experiments. However, the often asked question of how many replicates are required has barely been addressed in the literature. In general, the answer depends on several factors: a given magnitude of expression change, a desired statistical power (i.e. probability) to detect it, a specified false positive rate, and the statistical method being used to detect the change. Here, we discuss how to calculate the number of replicates in the context of applying a nonparametric statistical method, the Normal mixture model approach, to detect gene expression change. The methodology is applied to a data set containing expression levels of 1176 genes of rats with and without pneumococcal middle ear infection. We illustrate how to calculate the power functions for 2, 4, 6 and 8 replicates respectively.

Key Words: AIC; BIC; cDNA microarray; EM algorithm; Normal mixtures; Sample size/power.
**Introduction**

Microarrays are used to measure the (relative) expression levels of thousands of genes (or expressed sequence tags). A comparison of gene expression of cells or tissues from two conditions may provide useful information on important biological processes or functions (Brown and Botstein 1999; Lander 1999). The challenge now is how to detect those genuine changes from noisy data. It has now been known that simply using fold changes as in the earlier days is unreliable and inefficient (Chen et al 1997; Newton et al 2001). More sophisticated statistical methods are called for. Many proposals have appeared in the literature (e.g. Chen et al 1997; Efron et al 2000; Ideker et al 2000; Li and Hong 2001; Newton et al 2001; Thomas et al 2001; Tusher et al 2001; Pan et al 2001a). In particular, it has been noticed that it may be necessary to design an experiment that uses multiple arrays (or multiple spots on each array) containing multiple measurements for each gene under each condition. One reason is that due to a high noise-to-signal ratio, a single array may not provide enough information that can be reliably extracted (Lee et al 2000). More importantly, multiple measurements from each gene make it possible to assess possibly different variability of various genes. Then the problem seems to fall within the traditional two-sample comparison in statistics. Two of the best known two-sample statistical tests are the two-sample t-test and the Wilcoxon test (or equivalently, Mann-Whitney test). The t-test is parametric and is based on the assumption that the gene expression levels have Normal distributions. In contrast, the Wilcoxon test is nonparametric and is based on the ranks of observed gene expression levels. Although the t-test is robust to departures from Normality and the Wilcoxon test does not depend on the Normality assumption, the problem is that under non-Normal situations the t-test may be too conservative, and hence, as the Wilcoxon test, may have too low power, especially when the sample size is small, which is the case for most microarray experiments. These points have been verified in two case studies using real data (Thomas et al 2001; Pan 2001). In a class of nonparametric approaches (Efron et al 2000; Tusher et al 2001; Pan et al 2001a), a version of the two-sample t-statistic is used but its null distribution is estimated nonparametrically, rather than directly assumed to be a t-distribution. In addition, some earlier studies have suggested that the variability of gene expression may be related to the mean expression (Chen et al 1997; Ideker et al 2000; Newton et al 2001). Therefore, it implies that the t-statistic being used should be based on unequal variances for the two samples.
An important and natural question that is often asked by biologists is how many replicates are required. Unlike in many other contexts of experimental studies, this issue in microarray experiments has rarely been discussed in the literature. To our knowledge, the only exception is the work by Black and Doerge (2001), which however is for the situation where parametric statistical methods are applied to detect expression changes. In this paper, we discuss the problem when a nonparametric method, the Normal mixture model approach (Pan et al 2001a), is employed to detect differential expression, though to facilitate sample size calculations, the formulation is slightly changed from their original one. Nonparametric methods with microarray data analyses are pioneered by the works of Efron and Tibshirani and co-workers (Efron et al 2000; Tusher et al 2001). They take advantage of the presence of replicates and thus can impose much weaker modeling assumptions. For instance, the parametric methods in Black and Doerge (2001) depend on the assumption on the Lognormal or Gamma distribution of gene expression levels whereas the mixture model approach does not have such a distributional assumption and directly estimates distributions related to random errors. Note that modeling the distribution of random errors is advantageous to directly modeling that of expression levels, which is a common practice in applied statistics. For example, gene expression levels may be correlated (e.g. due to co-expression of some genes) while random errors can be more reasonably assumed to be independent. This is similar to modeling longitudinal data using a linear mixed-effects model (Diggle et al, 1994): the responses from each subject (corresponding to a group of co-regulated genes here) are in general correlated, but the measurement errors from the same subject can be considered to be independent after incorporating a random subject effect in the model. Note that the random effect will be cancelled out from the t-statistic for each gene. Our proposal here also shows an attractive point of the mixture model approach, as compared to the other two nonparametric approaches (Efron et al 2000; Tusher et al 2001), because it is still unclear how the sample size/power calculation can be done in the other two approaches.

The problem of calculating the number of required replicates in microarray experiments is similar to that of sample size/power calculations in clinical trials and other designed experiments; the to-be-determined sample size in microarray experiments refers to the number of replicates whereas the number of genes is not an issue here. As usual, we assume that the replicates are (approximately) independent with each other, whether they are drawn from the same individual
or multiple individuals. In general, the required sample size depends on several factors: the true magnitude of the change of gene expression (say, \(d\)), the desired statistical power (i.e. probability) \((\beta)\) to detect the change, and the specified false positive rate (\(\alpha\)). The problem of how to calculate the number of replicates for any given triplet \((d, \beta, \alpha)\) is equivalent to that of how the power \(\beta\) depends on the pair \((d, \alpha)\) and the number of replicates, which we consider in the paper.

The proposed method is not restricted to any specific microarray technology. From now on, the expression level can refer to a summary measure of relative red to green channel intensities in a fluorescence-labeled cDNA array, a radioactive intensity of a radiolabeled cDNA array (as to be used in the example later), or a summary difference of the perfect match (PM) and mis-match (MM) scores from an oligonucleotide array. The gene expression levels may have been suitably preprocessed, including dimension reduction, data normalization and data transformation (e.g. Dudoit et al 2000; Efron et al 2000; Li and Wong 2001; Kerr et al 2000; Yang et al 2000).

Methods

A statistical model

We consider a generic situation that, for each gene \(i, i = 1, 2, \ldots, N\), we have (relative) expression levels \(X_{1i}, \ldots, X_{mi}\) from \(m\) microarrays under condition 1, and \(Y_{1i}, \ldots, Y_{mi}\) from \(m\) arrays under condition 2. We need to assume that \(m\) is an even integer. A general statistical model is assumed for gene expression data:

\[
X_{ji} = \mu_{(1),i} + \epsilon_{ji}, \quad Y_{li} = \mu_{(2),i} + \epsilon_{li},
\]

where \(\mu_{(1),i}\) and \(\mu_{(2),i}\) are the mean expression levels for gene \(i\) under the two conditions respectively, and \(\epsilon_{ji}\) and \(\epsilon_{li}\) are independent random errors with means and variances

\[
E(\epsilon_{ji}) = E(\epsilon_{li}) = 0, \quad Var(\epsilon_{ji}) = \sigma^2_{(1),i}, \quad Var(\epsilon_{li}) = \sigma^2_{(2),i},
\]

for any \(j = 1, \ldots, m\), \(l = 1, \ldots, m\) and \(i = 1, \ldots, N\). It is assumed that random errors \(\epsilon_{ji}/\sigma_{(1),i}\) and \(\epsilon_{li}/\sigma^2_{(1),i}\) are randomly taken respectively from one of two (not necessarily equal) distributions that are symmetric about their mean 0. Note that the above assumption on the distributions of random errors, not on that of gene expression levels (i.e. \(X_{ji}\) and \(Y_{li}\)), is often reasonable and similar assumptions are common in other statistical applications. In addition, we do not assume
that the expression levels of all the genes have an equal variance, because some previous studies (e.g. Chen et al 1997; Ideker et al 2000; Newton et al 2001) have found that the variance \( \sigma^2_{(e),i} \) of gene expression levels may depend on the mean expression \( \mu_{(e),i} \). Also, we do not even need to assume that \( \sigma^2_{(1),i} = \sigma^2_{(2),i} \) unless \( \mu_{(1),i} = \mu_{(2),i} \).

A goal is to detect all genes with \( \mu_{(1),i} \neq \mu_{(2),i} \). This can be accomplished through statistical hypothesis testing.

**A test statistic**

To test the null hypothesis \( H_0: \mu_{(1),i} = \mu_{(2),i} \), we use a t-type test statistic or score

\[
Z_i = \frac{\sum_{j=1}^{m} X_{ji}/m}{\sigma_{(1),i}} - \frac{\sum_{j=1}^{m} Y_{ji}/m}{\sigma_{(2),i}} = \frac{\mu_{(1),i}}{\sigma_{(1),i}} - \frac{\mu_{(2),i}}{\sigma_{(2),i}} + \frac{\sum_{j=1}^{m} \epsilon_{ji}}{m\sigma_{(1),i}} - \frac{\sum_{l=1}^{n} \epsilon_{li}}{m\sigma_{(2),i}}.
\]

(1)

Note that the mean and variance of \( Z_i \) are

\[
E(Z_i) = \frac{\mu_{(1),i}}{\sigma_{(1),i}} - \frac{\mu_{(2),i}}{\sigma_{(2),i}}, \quad \text{Var}(Z_i) = \frac{2}{m},
\]

whereas the mean \( E(Z_i) = 0 \) under \( H_0 \). Hence, it can be seen that a large absolute value of \( Z_i \), \( |Z_i| \), gives evidence against \( H_0 \). As the number of arrays (i.e. \( m \)) increases, the variance of the test statistic \( Z_i \) decreases. Hence, it is possible to reject \( H_0 \) (i.e. detect differential expression for gene \( i \)) with any \( E(Z_i) \neq 0 \) if \( m \) is large enough. In other words, if the false positive rate and other parameters are fixed, then the statistical power of the test will increase as \( m \) increases. This is the key point that motivates the discussion on sample size calculations.

To determine the cut-off point for \( |Z_i| \) to reject \( H_0 \), we need to know or estimate the distribution of \( Z_i \) under \( H_0 \), the null distribution \( f_0 \). In a parametric approach, based on some full distributional assumptions for \( X_{ji} \) and \( Y_{ji} \), one may derive the null distribution \( f_0 \), such as in a two-sample t-test. However, the validity of such a parametric method critically depends on the correctness of assumed distributions, which of course is not guaranteed. Here, we consider a nonparametric approach: a finite Normal mixture model is used to estimate \( f_0 \) nonparametrically.

**Estimating the null distribution**

There may be various ways to estimate the null distribution \( f_0 \). For instance, using expression levels of some house-keeping genes that are known to have non-differential expression, one can construct their \( Z_i \) scores and then estimate \( f_0 \) using thus obtained \( Z_i \) scores. However, in practice,
there may not be house-keeping genes in a given experiment, or there are only a small number of house-keeping genes if any. Here, following the basic idea in a class of nonparametric methods (Efron et al 2000; Tusher et al 2001; Pan et al 2001a), we construct a null score \( z_i \) for each gene and then use these null scores to estimate \( f_0 \) nonparametrically. The null score is constructed from the same observed gene expression data as used in \( Z_i \):

\[
\begin{align*}
    z_i &= \frac{X_{1i} - X_{2i} + \ldots + X_{m-1,i} - X_{m,i}}{m\sigma_{(1),i}} + \frac{Y_{1i} - Y_{2i} + \ldots + Y_{m-1,i} - Y_{m,i}}{m\sigma_{(2),i}} \\
    &= \frac{\epsilon_{1i} - \epsilon_{2i} + \ldots + \epsilon_{m-1,i} - \epsilon_{m,i}}{m\sigma_{(1),i}} + \frac{\epsilon_{1i} - \epsilon_{2i} + \ldots + \epsilon_{m-1,i} - \epsilon_{m,i}}{m\sigma_{(2),i}}.
\end{align*}
\]

(2)

Under the assumption that \( \epsilon_{ji} \) and \( \epsilon_{ji} \) have symmetric distributions, then \( \epsilon_{ji} \) and \( -\epsilon_{ji} \) have the same distribution, and \( \epsilon_{ji} \) and \( -\epsilon_{ji} \) have the same distribution. Thus, by comparing the form of \( z_i \) with that of \( Z_i \), we know that the distribution of \( z_i \) is exactly \( f_0 \), the null distribution for \( Z_i \) (under \( H_0 \)). Note that under \( H_0 \), \( \mu_{(1),i} = \mu_{(2),i} \), and hence \( \sigma_{(1),i} = \sigma_{(2),i} \) (since we assume that \( \sigma_{(c),i} \) only depends on \( \mu_{(c),i} \)), then

\[
Z_i = \frac{\epsilon_{1i} + \epsilon_{2i} + \ldots + \epsilon_{m-1,i} + \epsilon_{m,i}}{m\sigma_{(1),i}} + \frac{\epsilon_{1i} + \epsilon_{2i} + \ldots + \epsilon_{m-1,i} + \epsilon_{m,i}}{m\sigma_{(2),i}}.
\]

thus \( z_i \) and \( Z_i \) have the same distribution \( f_0 \) under \( H_0 \). We use all \( z_i \)'s across all genes to estimate \( f_0 \).

In practice, \( \sigma_{(c),i} \) (for \( c = 1, 2 \)) are unknown, and can be estimated using the sample standard deviations (SDs) \( s_{(c),i} \). Although the sample SD \( s_{(c),i} \) is asymptotically unbiased, if \( m \) and \( n \) are small, \( s_{(c),i} \) may not be stable, and some modifications may be necessary. In any case, substituting \( \sigma_{(c),i} \) by any suitable estimates, we can calculate the scores \( z_i \)'s and \( Z_i \)'s, based on which we can estimate \( f_0 \) and \( f \) respectively. By comparing \( f_0 \) and \( f \), we can gain insight about genes with altered expression (i.e. \( \mu_{(1),i} \neq \mu_{(2),i} \)).

We assume that all the \( z_i \)'s for \( i = 1, \ldots, N \) are a random sample from \( f_0 \), thus we can use the observed \( z_i \)'s to estimate \( f_0 \). Pan et al (2001a) proposed estimating \( f_0 \) using a finite Normal mixture model (e.g., Titterington et al. 1985). Specifically, it is assumed that

\[
f_0(z; \Omega_{g_0}) = \sum_{r=1}^{g_0} \pi_r \phi(z; a_r, V_r),
\]

where \( \phi(\cdot; a_r, V_r) \) denotes the density function of a Normal distribution \( N(a_r, V_r) \) with mean \( a_r \) and variance \( V_r \), and \( \pi_r \)'s are mixing proportions. \( \Omega_{g_0} \) represents all unknown parameters \( \{ (\pi_r, a_r, V_r) \) :
\( r = 1, \ldots, g_0 \) in a \( g_0 \)-component mixture model. Among others, a Normal mixture is essentially nonparametric and flexible, and easy to use with stable tail probabilities.

A mixture model can be fitted by maximum likelihood using the expectation-maximization (EM) algorithm (Dempster et al. 1977; see also McLachlan and Basford 1988; Titterington et al. 1985). The number of components can be selected adaptively using the Akaike Information Criterion (AIC) (Akaike 1973) or the Bayesian Information Criterion (BIC) (Schwartz 1978). In using the AIC or BIC, one first fits a series of models with various values of \( g_0 \), then picks up the \( g_0 \) corresponding to the first local minimum of AIC or BIC (Fraley and Raftery 1998). Some empirical studies seem to favor the use of BIC (Fraley and Raftery 1998).

**Determining the cut-off point**

Once we obtain an estimate of the null distribution \( f_0 \), we can determine the cut-off point of the rejection region for testing \( H_0 \). In general, as for a two-sample test, the rejection region can be selected in the tails of \( f_0 \) because, under the null hypothesis \( Z_i \) should be close to the center of \( f_0 \), whereas if there is differential expression for gene \( i \), \( Z_i \) is likely to be in one of the two tails of \( f_0 \). The specific choice may depend on the goal of the analysis. For example, if we are only interested in detecting up-regulated genes, we can choose the rejection region at the right-tail of \( f_0 \). Our proposed method works for any specified way of determining the rejection region. Since \( f_0 \) should be symmetric about its mean 0, and often we are interested in both up- and down-regulated genes, we propose to take the rejection region at the two tails of \( f_0 \), \( \{ z : f_0(z) < C_\alpha \} \), where the constant \( C_\alpha > 0 \) is the cut-off point and depends on the specified (gene-specific) false positive rate \( \alpha \). As usual, \( C_\alpha > 0 \) is chosen such that the rejection rate under \( H_0 \) is exactly \( \alpha \):

\[
\alpha = \Pr(Z < -C_\alpha \text{ or } Z > C_\alpha | f_0) \\
= \int_{-\infty}^{-C_\alpha} f_0(z)dz + \int_{C_\alpha}^{\infty} f_0(z)dz \\
= \sum_{r=1}^{g_0} \pi_r \Phi(-C_\alpha; a_r, V_r) + 1 - \Phi(C_\alpha; a_r, V_r),
\]

where \( \Phi(\cdot; a, V) \) is the corresponding cumulative distribution function for \( \phi(\cdot; a, V) \). Using a numerical algorithm, such as the bisection method (Press et al 1992, p.353), we can solve the above equation to obtain \( C_\alpha \) for any given \( \alpha \).

For microarray data, because we are testing \( H_0 \) for each gene, the multiple test problem arises
and some control on it is necessary. Usually, we can use Bonferroni’s method. For instance, if we want to maintain the genome-wide false positive rate at the usual 5% level, then the Bonferroni-adjusted gene-specific (i.e., test-specific) false positive rate is \( \alpha = 0.05/N \), where \( N \) is the total number of genes to be tested.

Once \( C_\alpha \) is determined, we can calculate the power as a function of \( d \), the magnitude of the expression change targeted to be detected. Note that

\[
d = \frac{\mu_{[1],i}}{\sigma_{[1],i}} - \frac{\mu_{[2],i}}{\sigma_{[2],i}}
\]

is the difference of the coefficients of variation under the two conditions. If \( \sigma_{[1],i} = \sigma_{[2],i} \), \( d \) can be interpreted as the change of the mean expression levels from condition 1 to condition 2. Otherwise, it can be regarded as the difference of (variation) standardized mean expression levels. Specifically, we have the power function

\[
\beta(d, \alpha) = \Pr(Z - d < -C_\alpha \text{ or } Z - d > C_\alpha | f_0) \\
= \int_{d-C_\alpha}^{-\infty} f_0(z)dz + \int_{d+C_\alpha}^{\infty} f_0(z)dz \\
= \sum_{r=1}^{g_0} \pi_r [\Phi(d - C_\alpha; a_r, V_r) + 1 - \Phi(d + C_\alpha; a_r, V_r)].
\]

Unsurprisingly, we can see that \( \beta(d, \alpha) \) will increase as \( |d| \) increases. The effects of having more replicates will reduce the variability of \( f_0 \), leading to larger \( \beta(d, \alpha) \) for any given \( d \).

**Calculation of replicate numbers**

Now we describe how to calculate replicate numbers based on some pilot data taken from earlier studies. We use \( z_{m,i} \) to explicitly denote the \( z_i \) scores in (2) with \( m \) replicates. Based on the data we can estimate the density function \( f_{0,m}(z; \Omega_{g_0}) \) of \( z_{m,i} \)'s as a Normal mixture

\[
f_{0,m}(z; \Omega_{g_0}) = \sum_{r=1}^{g_0} \pi_r \phi(z; a_r, V_r). \tag{5}
\]

From now on, we treat \( f_{0,m} \) as known in (5).

With estimated \( f_{0,m} \), we want to estimate the density function \( f_{0,mk} \) for \( z_{mk,i} \), the \( z_i \) scores based on \( mk \) replicates (with \( k > 1 \)). If we can have an estimate of \( f_{0,mk} \), then we can obtain the corresponding power function \( \beta(d, \alpha) \) for \( mk \) replicates in the same way as described earlier for \( m \) replicates. Of course, we assume that our pilot data are drawn from only \( m \) arrays under
each of the two experimental conditions, and thus we do not observe any \( z_{mk,i} \) based on \( mk \) arrays. However, we show next that it is possible to generate \( z_{mk,i} \)'s from \( z_{m,i} \)'s. Note that we can draw random realizations of \( z_{m,i} \) from the estimated \( f_{0,m} \) (see Pan et al (2001a) or the example below). Suppose \( z^{(j)}_{m,i} \)'s (for \( j = 1, 2, ..., k \)) are \( k \) independent realizations of \( z_{m,i} \), then it is easy to show that

\[
 z_{mk,i} = \sum_{j=1}^{k} \frac{z^{(j)}_{m,i}}{k}
\]

(6)

have the distribution \( f_{0,mk} \). Thus, the density function for \( z_{mk,i} \)'s is

\[
f_{0,mk}(z; \Omega_{g_0}) = \sum_{r_1, ..., r_k=1}^{g_0} \prod_{r_1} \frac{1}{\sqrt{2\pi V_r}} \exp \left( -\frac{1}{2} \frac{a_{r_1}^2}{V_{r_1}} \right), \quad \sum_{j=1}^{k} \frac{a_{r_j}}{V_{r_j}} / k^2.
\]

(7)

For example, if we triple the number of replicates, the resulting density function is

\[
f_{0,3m}(z; \Omega_{g_0}) = \sum_{r_1, r_2, r_3=1}^{g_0} \prod_{r_1} \frac{1}{\sqrt{2\pi V_r}} \exp \left( -\frac{1}{2} \frac{a_{r_1}^2}{V_{r_1}} \right), \quad \frac{1}{9} \left( \frac{a_{r_1} + a_{r_2} + a_{r_3}}{V_{r_1} + V_{r_2} + V_{r_3}} \right).
\]

The number of components of \( f_{0,mk} \) may be too large. For example, if the number of components is \( g_0 = 3 \) for \( m = n = 2 \), the corresponding numbers of components for \( m = n = 4 \), \( m = n = 6 \) and \( m = n = 8 \) are respectively \( g_0^2 = 9 \), \( g_0^3 = 27 \) and \( g_0^4 = 81 \). In fact, some of these components may be very similar or play a negligible role, hence the form of \( f_{0,mk} \) may be simplified. In the extreme situation, as \( mk \rightarrow \infty \), by the Central Limit Theorem, the mixture model will reduce to a single component Normal distribution. Hence, we propose a simulation-based method to select a more parsimonious model for \( f_{0,mk} \).

Based on the mixture model \( f_{0,m} \) in (5), we can generate a random sample of \( z^{(j)}_{m,i} \)'s (Pan et al 2001a), from which we can calculate \( z_{mk,i} \)'s using (6). Using \( z_{mk,i} \)'s we can fit a Normal mixture model for \( f_{0,mk} \). As we shall show later, we find such a fitted mixture model often contains a smaller number of components than \( g_0^k \), as dictated in (7), leading to a simplified form of \( f_{0,mk} \).

Summary of the proposed method

In summary, our proposed method of calculating the required replicate number works in the following steps.

Step 1. Suppose that we have pilot gene expression data \( X_{ji} \) and \( Y_{ji} \) from \( m \) arrays under each condition. Use formula (2) to calculate the scores \( z_{i,m} \).
Step 2. Use $z_{i,m}$ and the Normal mixture model (5) to estimate $f_{0,m}$.

Step 3. For a specified false positive rate $\alpha$, determine the cut-off point $C_\alpha$ for the rejection region using formula (3), in which $f_0$ is replaced with the estimated $f_{0,m}$.

Step 4. For any specified $d$, calculate the power function $\beta(d, \alpha)$ using formulas (4), in which $f_0$ is replaced with the estimated $f_{0,m}$.

Step 5. For any given $k > 1$, use formula (7) or (6) to estimate $f_{0,mk}$.

Step 6. For a specified false positive rate $\alpha$, determine the cut-off point $C_\alpha$ for the rejection region using formula (3), in which $f_0$ is replaced with the estimated $f_{0,mk}$.

Step 7. For any specified $d$, calculate the power function $\beta(d, \alpha)$ using formulas (4), in which $f_0$ is replaced with the estimated $f_{0,mk}$.

Step 8. Repeat Steps 5 to 7 until all $k > 1$ of interest have been tried.

After the power functions for many possible $mk$ replicates have been obtained, we can determine an appropriate number of replicates by considering all the factors involved, the desired power and false positive rate, the targeted expression changes and other experimental constraints.

An Example

To understand the pathogenesis of otitis media, a study was conducted to identify genes involved in response to pneumococcal middle ear infection and to study their roles in otitis media. Radioactively labeled DNA microarrays were applied to the mRNA analysis of 1176 genes in middle ear mucosa of rats with and without subacute pneumococcal middle ear infection (Pan et al 2001b). The data are available for the control group and for the pneumococcal middle ear infection group. A more detailed description on how the data were collected and on the public availability of the data was provided in Pan et al (2001b). For the purpose of sample size calculations and to mimic many practical situations with only a small number of replicates, we only use $m = n = 2$ arrays from each group. We first take a natural logarithm transformation for all the observed gene expression levels (i.e. radioactive intensities) so that the resulting distributions are less skewed (which will reduce the number of components of a fitted mixture model). Then, for each microarray, we standardize the transformed gene expression levels by subtracting their median.
Due to the small $m = 2$, the sample SDs may not be stable. One way is to add a small constant as suggested by Efron et al (2000). Here we follow the idea of Lin et al (2001) and use a loess smoother (Cleveland and Devlin 1988) to nonparametrically model the sample SDs in terms of the mean expression levels (Figure 1). Then we plug-in the smoothed SD to calculate $z_{2,i}$. Note that an alternative use of SD or its modification in calculating $z_{2,i}$'s will not change the basic idea and the following steps in sample size calculations.

We fitted three mixture models for $f_{0,2}$ with $g_0$ ranging from 1 to 3. Table 1 summarizes the model fitting results. $g_0 = 1$ was selected since both AIC and BIC achieve their minima there. So the fitted $f_0$ is a Normal distribution, $N(-0.0013, 0.1278)$. However, for the purpose of illustration to be general, we choose $g_0 = 2$ as the fitted model:

$$f_{0,2}(z) = 0.76\phi(z; -0.0415, 1.3117) + 0.24\phi(z; 0.0700, 2.6970).$$

Figure 2(a) presents the histogram of $z_i$'s and the fitted $f_0$ with $g_0 = 1$ and 2. There is not much difference between the two fitted $f_{0,2}$, both of which fit the data well. In particular, $f_{0,2}$ does not look like a t-distribution with small degrees of freedom, as predicted from the t-test.

A realization of $z_{2,i}$ can be simulated in the following two steps: (i) We draw a random number $p_i$ from $\{1, 2\}$ with probability 0.76 and 0.24 respectively. (ii) If the drawn $p_i = 1$, $z_i$ is randomly drawn from a Normal distribution $\phi(z; -0.0415, 1.3117)$; otherwise, it is drawn from $\phi(z; 0.0700, 2.6970)$. Based on the generated $z_{2,i}$'s, following expression (6) we generated three simulated data sets $z_{2k,i}$'s, $i = 1, ..., 1176$ for $k = 2, 3$ and 4. Then a Normal mixture model was fitted to each data set. From Table 1, it can be seen that a single-component Normal distribution was selected in each case. In Figure 2, each of the fitted Normal distributions, $N(-0.0494, 0.8226)$, $N(-0.0644, 0.5383)$ and $N(-0.0438, 0.4206)$, is compared with its theoretically derived mixture model in (7); they are all very close. Here we see that using simulated data to fit a mixture model results in a much simplified model. For example, for $k = 4$, it is a fitted single component model versus a $2^4 = 16$ component model in (7). Note that as predicted, all the means of the fitted models are all essentially 0, and their variances decrease as $k$ increases.
If we want to have only one expected false positive result from testing each of 1176 genes, the gene-specific (or test-specific) false positive rate is $\alpha = 1/1176 = 0.09\%$. Note this is a Bonferroni correction for multiple comparisons for a two-sided test. Using formula (3) and fitted mixture model $f_{0,2k}$, the cut-off points $C_\alpha$ are determined. Then the power functions $\beta(d, \alpha)$ are drawn in Figure 3, which may help make a decision on the required number of replicates. For instance, if we want to detect an expression change $d = 3$ with probability at least 80% and with $\alpha = 0.09\%$, then six replicates are needed. Also, with just two replicates, the power to detect a change as high as 4 is very low, smaller than 30%. Note that the choice of $d$ may depend on some prior knowledge. For instance, based on the pilot data, we can estimate the $d$ values for some selected genes (with the sample means and sample SDs substituting the true means and SDs in the formula for $d$), from which one can determine a range of $d$ values of interest.

Figures 4 to 6 give the results for testing $N = 1000$, 5000 and 10000 genes respectively while controlling the genome-wide false positive rate at the usual 5% level. It can be seen that as $N$ increases, we also need a larger number of arrays to maintain the power of the statistical test when other parameters are fixed. For instance, for $N = 10000$ (Figure 6), even eight replicates cannot detect a change as large as $d = 3$ with 80% power, but six replicates can detect a change $d = 4$ with 80% power.

Discussion
We have described a method for calculating the number of replicates in microarray experiments. This method is designed for the situation where the mixture approach is going to be taken to analyze the data. Note that any method for sample size/power calculations has to depend on a specific statistical test to be used in data analysis; this explains why there is a huge literature on the topic for clinical trials. However, due to the close relation between the mixture approach and the other two recently proposed nonparametric approaches, the empirical Bayes method (Efron et al, 2000) and the Statistical Analysis of Microarray (SAM) method (Tusher et al, 2001), our proposed method can be also applied to provide some useful guideline for designing microarray experiments even when one of the latter two approaches (or other approaches) is planned to be used for data analysis in a later stage. For instance, even though the null distribution $f_0$ is estimated using
the null scores $z_i$ in our proposal, there may be alternative ways to estimate $f_0$, such as using an alternative nonparametric method (e.g. kernel or local likelihood), rather than the finite Normal mixture model, to estimate $f_0$, or using the test statistics $Z_i$ of a large number of house-keeping genes to estimate $f_0$. Some modifications to the test statistic $Z_i$ and the null statistic $z_i$ are also possible, especially when we consider differential gene expression across more than two conditions. These are all interesting topics we are investigating now.

In most sample size/power calculations, some pilot data are needed to provide reasonable estimates on some parameters needed for subsequent calculations. An alternative is to obtain reasonable estimates from other similar studies appeared in the literature. However, due to the rapid development of microarray technology, the latter is not likely and we expect that a researcher has to do his/her own pilot study. This was the situation we considered in the example. A particularly challenging issue is how to obtain good estimates of the variances of gene expression levels from a small number of replicates. In our example, we considered a nonparametric method to smooth sample variances. Some alternative smoothing methods have also appeared in the literature. But it is not clear which one is the most desirable. This is a topic for future study.

The proposed method is straightforward to statisticians and can be implemented in many existing statistical packages. Our sample S-Plus program being used will be available once the paper is accepted.

ACKNOWLEDGMENTS

We thank the two referees for helpful and constructive comments. This research was partially supported by NIH.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory (eds. B. N. Petrov and F. Csaki), 267-281, Akademiai Kiado, Budapest.

2. Biernacki, C. and Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. Journal of Statistical Computation and Simulation, 64, 49-71.
3. Black, M.A. and Doerge, R.W. (2001). Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Manuscript. Department of Statistics, Purdue University.

4. Brown, P and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics (Suppl.)*, **21**, 33-37.

5. Chen Y, Dougherty ER and Bittner ML (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomedical Optics*, 2, 364-367.

6. Cleveland W, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 1988; **83**:596-610.

7. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1-38.

8. Diggle, P.J., Liang, K.-Y. and Zeger, S.L. *Analysis of Longitudinal Data*. Oxford: Oxford (1994).

9. Dudoit S, Yang YH, Callow MJ and Speed TP (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Tech Rept, Stat Dept, UC-Berkeley, 2000.

10. Efron, B., Tibshirani, R., Goss, V. and Chu, G. (2000). Microarrays and Their Use in a Comparative Experiment. Manuscript. Available at [http://www-stat.stanford.edu/~tibs/research.html](http://www-stat.stanford.edu/~tibs/research.html).

11. Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering methods? - Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578-588.

12. Ideker, T., Thorsson, V., Siehel, A.F. and Hood, L.E. (2000). Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805-817.

13. Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819-837.

14. Lander, E.S. (1999). Array of hope. *Nature Genetics (Suppl.)*, **21**, 3-4.
15. Lee, M-L T., Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Nat. Acad. Sci, 97, 9834-9839.

16. Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. 2ed ed. Wiley, New York.

17. Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS*, 98, 31-36.

18. Li, H. and Hong, F. (2001), Cluster-Rasch models for microarray gene expression data. *Genome Biology*, 2(8), research0031.1-0031.13.

19. Lin, Y., Nadler, S. T., Attie, A. D., and Yandell, B. S. (2001). Mining for Low-abundance Transcripts in Microarray Data. Manuscript. Available at http://www.stat.wisc.edu/~yilin/.

20. McLachlan, G.L. and Basford, K.E. (1988). *Mixture models: Inference and Applications to Clustering*. Marcel Dekker, New York.

21. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8, 37-52.

22. Pan, W. (2001). A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. To appear in *Bioinformatics*. Available at http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

23. Pan, W., Lin, J. and Le, C. (2001a). A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data. Manuscript. Available at http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

24. Pan, W., Lin, J. and Le, C. (2001b). Model-Based Cluster Analysis of Microarray Gene Expression Data. *Genome Biology*, 2002/3/2. Available at http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

25. Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C, The Art of Scientific Computing*. 2nd ed. Cambridge: New York.

26. Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6, 461-464.
27. Statistical Sciences (1993). *S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.2.* StatSci, a division of MathSoft, Inc., Seattle.

28. Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research, 11*, 1227-1236.

29. Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

30. Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci, 98*, 5116-5121.

31. Yang YH, Buckley MJ, Dudoit S and Speed TP (2000). Comparison of methods for image analysis on cDNA microarray data. Tech Rept, Stat Dept, UC-Berkeley, 2000.
Table 1: AIC and BIC for fitted mixture models with various number of components $g_0$.

| $g_0$ | 2 replicates | 4 replicates | 6 replicates | 8 replicates |
|-------|--------------|--------------|--------------|--------------|
|       | AIC          | BIC          | AIC          | BIC          | AIC          | BIC          | AIC          | BIC          |
| 1     | 3928.10      | 3938.24      | 3111.75      | 3121.89      | 2612.98      | 2632.12      | 2322.85      | 2332.99      |
| 2     | 3928.54      | 3953.89      | 3116.40      | 3141.75      | 2617.65      | 2643.00      | 2327.03      | 2352.38      |
| 3     | 3932.67      | 3973.23      | 3122.20      | 3162.76      | 2622.61      | 2663.17      | 2331.92      | 2372.48      |
Figure 1: Sample standard deviations of expression levels and their loess smoothers as a function of the average expression levels for the two conditions respectively.
Figure 2: Histograms and estimated distribution density functions. (a)–(d) are for two, four, six and eight replicates (z2–z8) respectively. In (a), the solid and dotted lines are the fitted 1- and 2-component mixtures. In (b)–(d), the solid and dotted lines are the fitted and the theoretically derived mixtures.
Figure 3: Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes $d$ and the number of replicates, with the gene-specific false positive rate $\alpha = 0.09\%$ for the middle ear data.
Figure 4: Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes $d$ and the number of replicates, with the gene-specific false positive rate $\alpha = 0.05/1000$ for the middle ear data.
Figure 5: Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes $d$ and the number of replicates, with the gene-specific false positive rate $\alpha = 0.05/5000$ for the middle ear data.
Figure 6: Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes $d$ and the number of replicates, with the gene-specific false positive rate $\alpha = 0.05/10000$ for the middle ear data.