Research

How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach

Wei Pan*, Jizhen Lin† and Chap T Le*

Addresses: *Division of Biostatistics, School of Public Health, University of Minnesota, 420 Delaware Street, Minneapolis, MN 55455-0378, USA. †Department of Otolaryngology, School of Medicine, University of Minnesota, Minneapolis, MN 55455, USA.

Correspondence: Wei Pan. E-mail: weip@biostat.umn.edu

Published: 22 April 2002

*Genome **Biology*** 2002, **3(5)**:research0022.1–0022.10

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/5/research/0022

Received: 27 December 2001
Revised: 15 February 2002
Accepted: 11 March 2002

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: 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 (that is, probability) to detect it, a specified Type I error 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 changes in gene expression.

Results: The methodology is applied to a data set containing expression levels of 1,176 genes in rats with and without pneumococcal middle-ear infection. We illustrate how to calculate the power functions for 2, 4, 6 and 8 replicates.

Conclusions: The proposed method is potentially useful in designing microarray experiments to discover differentially expressed genes. The same idea can be applied to other statistical methods.

Background

Microarrays are used to measure the (relative) expression levels of thousands of genes (or expressed sequence tags). A comparison of gene expression in cells or tissues from two conditions may provide useful information on important biological processes or functions [1,2]. The challenge now is how to detect those genuine changes from noisy data. It is now known that simply using fold changes, as in the earlier days, is unreliable and inefficient [3,4]. More sophisticated statistical methods are called for. Many proposals have appeared in the literature [3-10]. 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 because of a high noise-to-signal ratio, a single array may not provide enough information that can be reliably extracted [11]. More important, multiple measurements from each gene make it possible to assess the potentially different variability of genes. The problem then 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 with 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 [8,12]. In a class of nonparametric approaches [5,9,10], 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 [3,4,6]. Therefore, it implies that the *t*-statistic being used should be based on unequal variances for the two samples.

An important and natural question often asked by biologists is how many replicates are required. For microarray experiments, unlike many other experimental contexts, this issue has rarely been discussed in the literature. To our knowledge, the only exception is the work by Black and Doerge [13], 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 [10], is used to detect differential expression. But to facilitate calculations of sample size, the formulation is slightly changed from their original one. Nonparametric methods of microarray data analysis have been pioneered by Efron and Tibshirani and co-workers [5,9]. They take advantage of the presence of replicates and thus can impose much weaker modeling assumptions. For instance, the parametric methods of Black and Doerge [13] depend on the assumption on the log-normal 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 has advantages over direct modeling of expression levels, and is a common practice in applied statistics. For example, gene-expression levels may be correlated (for example, as a result of coexpression of some genes) whereas random errors can be more reasonably assumed to be independent. This is similar to modeling longitudinal data using a linear mixed-effects model [14]: the responses from each subject (corresponding to a group of coregulated 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 canceled out from the *t*-statistic for each gene. Our proposal here also shows an attractive feature of the mixture model approach, as compared to the other two nonparametric approaches [5,9], 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 replicates required in a microarray experiment is similar to that of sample size/power calculations in clinical trials and other experiment designs; 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 (that is, probability) (β) to detect the change, and the specified Type I error rate (α). The problem of how to calculate the number of replicates for any given triplet (d, β, α) is equivalent to that of how the power β depends on the pair (d, α) 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 used in the example later), or a summary difference of the perfect match (PM) and mismatch (MM) scores from an oligonucleotide array. The gene-expression levels may have been suitably preprocessed, including dimension reduction, data normalization and data transformation [5,15-18].

Results and discussion

A statistical model

We consider a generic situation that, for each gene i, $I = 1,2,...,N$, we have (relative) expression levels $X_{1i},..., X_{mi}$ from m microarrays under condition 1, and $Y_{1i},..., Y_{mi}$ 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:

$$X_{ji} = \mu_{(1),i} + \epsilon_{ji}, \qquad Y_{li} = \mu_{(2),i} + e_{li},$$

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

$$E(\epsilon_{ji}) = E(e_{li}) = 0, \qquad Var(\epsilon_{ji}) = \sigma^2_{(1),i}, \qquad Var(e_{li}) = \sigma^2_{(2),i},$$

for any $j = 1,..., m$, $l = 1,..., m$ and $i = 1,..., N$. It is assumed that random errors $\varepsilon_{ji}/\sigma_{(1),i}$ and $e_{li}/\sigma^2_{(1),i}$ 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 (that is, 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 [3,4,6] have found that the variance $\sigma^2{}_{(c),i}$ (for $c = 1,2$) of gene-expression levels may depend on the mean expression $\mu_{(c),i}$. 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

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

Note that the mean and variance of Z_i are

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

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 (that is, m) increases, the variance of the test statistic Z_i decreases. Hence, it is possible to reject H_0 (that is, detect differential expression for gene i) with any $E(Z_i) \neq 0$ if m is large enough. In other words, if the Type I error 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 housekeeping genes that are known to have non-differential expression, one can construct their Z_i scores and then estimate f_0 using the obtained Z_i scores. In practice, however, there may be only a small number of or no housekeeping genes in a given experiment. Here, following the basic idea in a class of nonparametric methods [5,9,10], 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:

$$z_i = \frac{X_{1i} - X_{2i} + \ldots + X_{m-1,i} - X_{m,i}}{m\sigma_{(1),i}} + \frac{Y_{1i} - Y_{2i} + \ldots + Y_{m-1,i} - Y_{m,i}}{m\sigma_{(2),i}}$$

$$= \frac{\epsilon_{1i} - \epsilon_{2i} + \ldots + \epsilon_{m-1,i} - \epsilon_{m,i}}{m\sigma_{(1),i}} + \frac{e_{1i} - e_{2i} + \ldots + e_{m-1,i} - e_{m,i}}{m\sigma_{(2),i}}. \quad (2)$$

Under the assumption that ε_{ji} and e_{ji} have symmetric distributions, then ε_{ji} and $-\varepsilon_{ji}$ have the same distribution, and e_{ji} and $-e_{ji}$ 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

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

Thus z_i and Z_i have the same distribution f_0 under H_0. We use all z_i values 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 values and Z_i values, on the basis of which we can estimate f_0 and f respectively. By comparing f_0 and f, we can gain insight about genes with altered expression (that is, $\mu_{(1),i} \neq \mu_{(2),i}$).

We assume that all the z_i values for $i = 1,\ldots, N$ are a random sample from f_0; thus we can use the observed z_i values to estimate f_0. Pan *et al.* [10] proposed estimating f_0 using a finite normal mixture model [19]. 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 $\phi(z; a_r, V_r)$ denotes the density function of a normal distribution $N(a_r, V_r)$ with mean a_r and variance V_r, and π_r values are mixing proportions. Ω_{g0} 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 [19-21]. The number of components can be selected adaptively using the Akaike Information Criterion (AIC) [22] or the Bayesian Information Criterion (BIC) [23]. 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 [24]. Some empirical studies seem to favor the use of BIC [24].

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 upregulated 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. As f_0 should be symmetric about its mean 0, and often we are interested in both up- and downregulated 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) Type I error rate α. As usual, $C_\alpha > 0$ is chosen such that the rejection rate under H_0 is exactly α:

$$\begin{aligned}\alpha &= \Pr(Z < -C_\alpha \text{ or } Z > C_\alpha | f_0) \\ &= \int_{-\infty}^{-C_\alpha} f_0(z)dz + \int_{C_\alpha}^{\infty} f_0(z)dz \qquad (3) \\ &= \sum_{r=1}^{g_0} \pi_r[\Phi(-C_\alpha; a_r, V_r) + 1 - \Phi(C_\alpha; a_r, V_r)].\end{aligned}$$

where $\Phi(.; a, V)$ is the corresponding cumulative distribution function for $\phi(.; a, V)$. Using a numerical algorithm, such as the bisection method [25], we can solve the above equation to obtain C_α for any given α.

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 Type I error rate at the usual 5% level, then the Bonferroni-adjusted gene-specific (that is, test-specific) Type I error rate is $\alpha = 0.05/N$, where N is the total number of genes to be tested.

Once C_α 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

$$\begin{aligned}\beta(d,\alpha) &= \Pr(Z - d < -C_\alpha \text{ or } Z - \mathrm{d} > \mathrm{C}_\alpha | f_0) \\ &= \int_{-\infty}^{d-C_\alpha} f_0(z)dz + \int_{d+C_\alpha}^{\infty} f_0(z)dz \qquad (4) \\ &= \sum_{r=1}^{g_0} \pi_r[\Phi(d-C_\alpha; a_r, V_r) + 1 - \Phi(d + C_\alpha; a_r, V_r)].\end{aligned}$$

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 $f_{0,m}(z;\Omega_{g0})$ of $z_{m,i}$ values as a normal mixture

$$f_{0,m}(z; \Omega_{g_0}) = \sum_{r}^{g_0} \pi_r \phi(z; a_r, V_r). \qquad (5)$$

From now on, we treat $f_{0,m}$ as known in Equation (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 $z_{mk,i}$ values from $z_{m,i}$ values. Note that we can draw random realizations of $z_{m,i}$ from the estimated $f_{0,m}$ (see Pan *et al.* [10] or the example below). Suppose $z_{m,i}^{(j)}$ values (for $j = 1,2,\ldots, k$) are k independent realizations of $z_{m,i}$, then it is easy to show that

$$z_{mk,i} = \sum_{j=1}^{k} z_{m,i}^{(j)}/k \qquad (6)$$

have the distribution $f_{0,mk}$. Thus, the density function for $z_{mk,i}$ values is

$$f_{0,mk}(z; \Omega_{g_0}) = \sum_{r1,\ldots,r_k=1}^{g_0} \pi_{r_1}\ldots\pi_{r_k} \phi(z; \sum_{j=1}^{k} a_{r_j}/\mathrm{k}, \sum_{j=1}^{k} V_{r_j}/k^2). \qquad (7)$$

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

$$f_{0,3m}(z;\, \Omega_{g_0}) = \sum_{r1,r2,r3=1}^{g_0} \pi_{r_1}\pi_{r_2}\pi_{r_3}\, \phi\,(z;\, (a_{r_1}+a_{r_2}+a_{r_3})/3,\ (V_{r_1}+V_{r_2}+V_{r_3})/9).$$

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 have a negligible role, hence the form of $f_{0,mk}$ may be simplified. In the extreme situation, as $mk \to \infty$, 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 $f_{0,mk}$.

On the basis of the mixture model $f_{0,m}$ in Equation (5), we can generate a random sample of $z_{m,i}^{(j)}$ values [10], from which we can calculate $z_{mk,i}$ values using Equation (6). Using $z_{mk,i}$ values 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^k_0, as dictated in Equation (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 Type I error rate α, determine the cutoff point C_α 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 formula (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 Type I error rate α, determine the cutoff point C_α 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 formulae (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 Type I error 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 1,176 genes in middle-ear mucosa of rats with and without subacute pneumococcal middle-ear infection [26]. The data are available for the control group and for the pneumococcal middle-ear infection group. A more detailed description of how the data were collected and their public availability was provided in Pan *et al.* [26]. 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 (that is, 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.

Because of 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.* [5]. Here we follow the idea of Lin *et al.* [27] and use a loess smoother [28] 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 $z_{2,i}$ values 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 as 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 purposes of general illustration, 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 2a presents the histogram of z_i values 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.

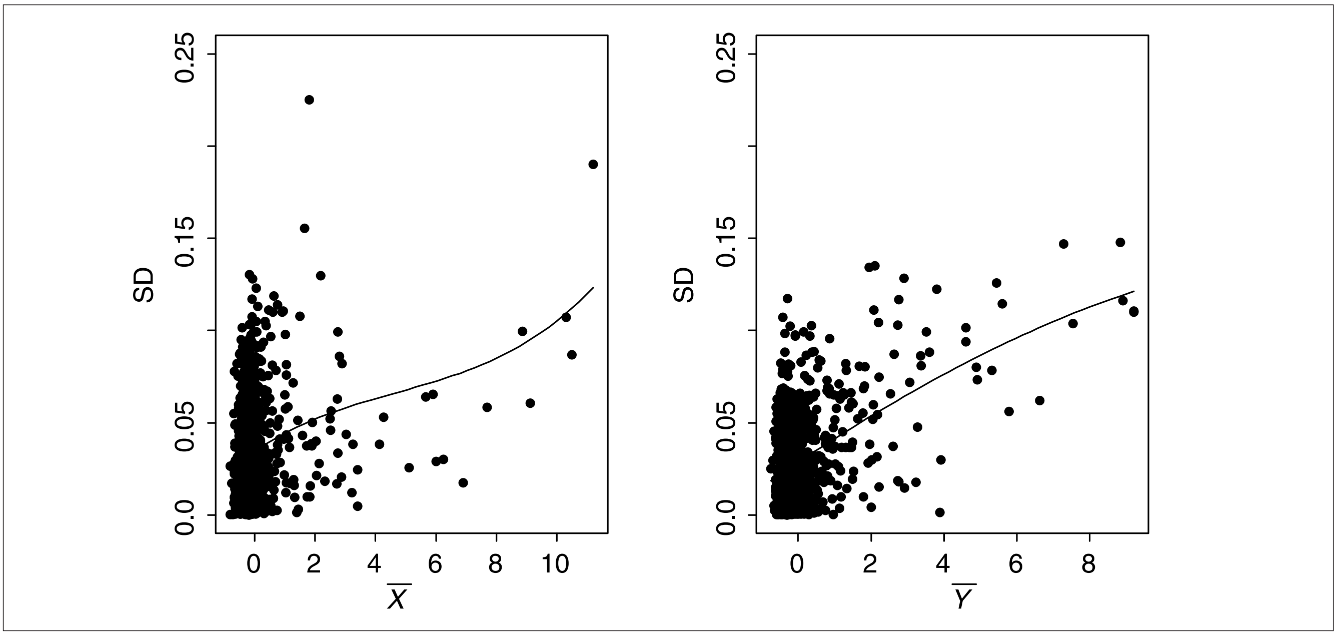

Figure 1
Sample standard deviations of expression levels and their loess smoothers as a function of the average expression levels for the two conditions respectively.

A realization of $z_{2,i}$ can be simulated in the following two steps. First, we draw a random number p_i from {1, 2} with probability 0.76 and 0.24 respectively. Second, 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 $\phi(z; 0.0700, 2.6970)$. From the generated $z_{2,i}$ values, following expression (6) we generated three simulated data sets: $z_{2k,i}$ values, I = 1,..., 1,176 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 Equation (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 Equation (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 1,176 non-differentially expressed genes, the gene-specific (or test-specific) Type I error rate is α = 1/1176 = 0.09%. Using formula (3) and fitted-mixture model $f_{0,2k}$, the cut-off points C_α 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 α =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

Table 1

AIC and BIC for fitted mixture models with various number of components g_0

	Two replicates		Four replicates		Six replicates		Eight replicates	
g_0	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
1	3928.10	3938.24	3111.75	3121.89	2612.98	2623.12	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

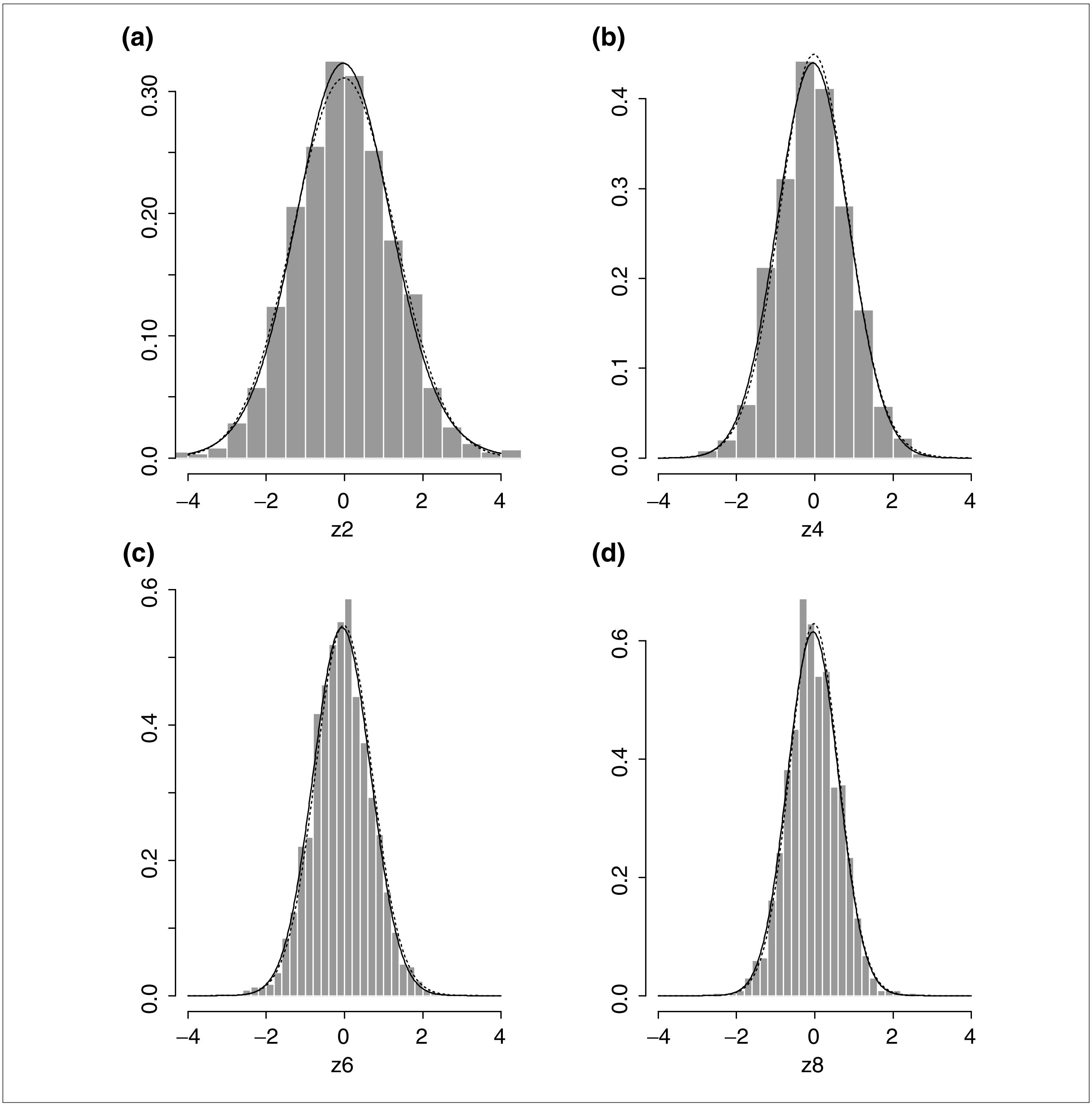

Figure 2
Histograms and estimated distribution density functions. **(a-d)** Two, four, six and eight replicates (z2 - z8), respectively. In (a), the solid and dotted lines are the fitted one- and two-component mixtures. In (b-d), the solid and dotted lines are the fitted and the theoretically derived mixtures.

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 4-6 give the results for testing N = 1,000, 5,000 and 10,000 genes, respectively, while controlling the genome-wide Type I error 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 = 10,000 (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.

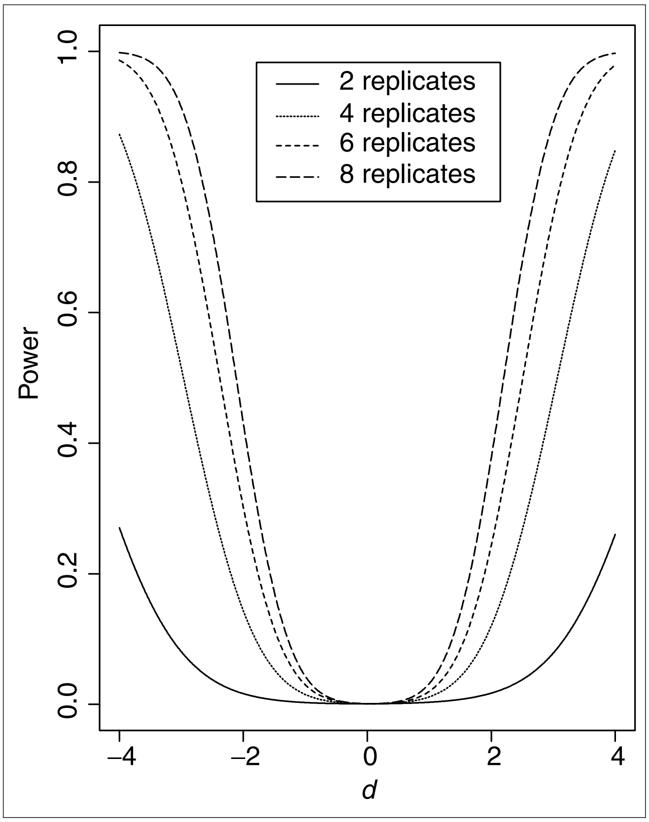

Figure 3
Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes *d* and the number of replicates, with the gene-specific Type I error rate α = 0.09% for the middle-ear data.

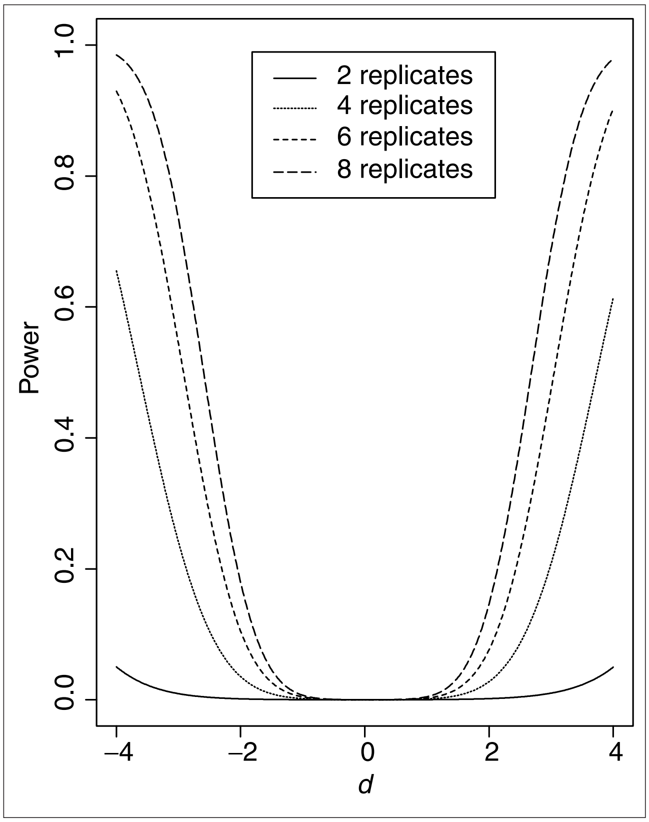

Figure 4
Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes *d* and the number of replicates, with the gene-specific Type I error rate α = 0.05/1,000 for the middle-ear data.

Conclusions

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, because of the close relation between the mixture approach and the other two recently proposed nonparametric approaches - the empirical Bayes method [5] and the statistical analysis of microarray (SAM) method [9] - 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 of estimating f_0, such as using an alternative nonparametric method (for example, 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 housekeeping 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 of some parameters needed for subsequent calculations. An alternative is to obtain reasonable estimates from other similar studies in the literature. However, because of the rapid development of microarray technology, the latter is not likely and we expect a researcher will have to do his or her own pilot study. This was the situation we considered in the example. A particular challenge 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 and data are available at [29].

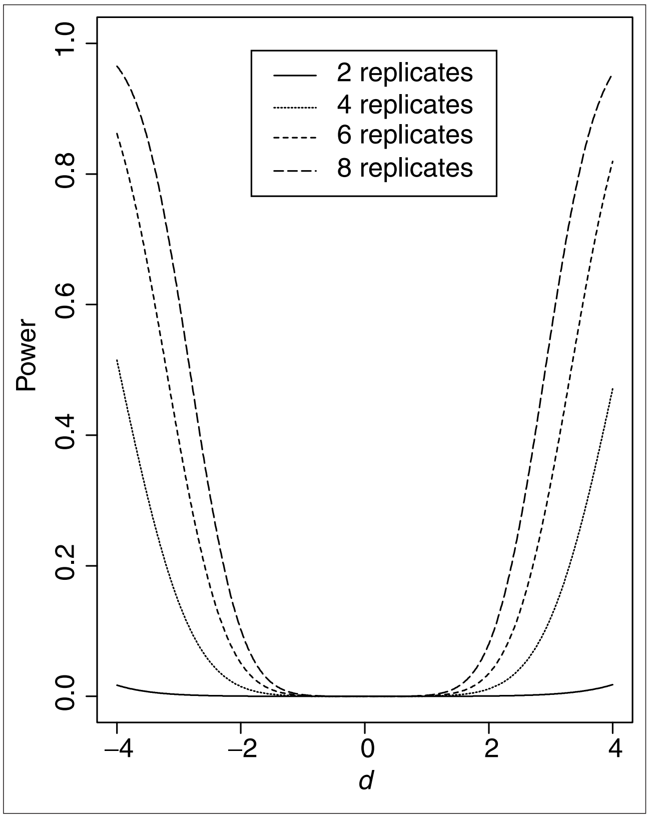

Figure 5
Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes d and the number of replicates, with the gene-specific Type I error rate α = 0.05/5,000 for the middle-ear data.

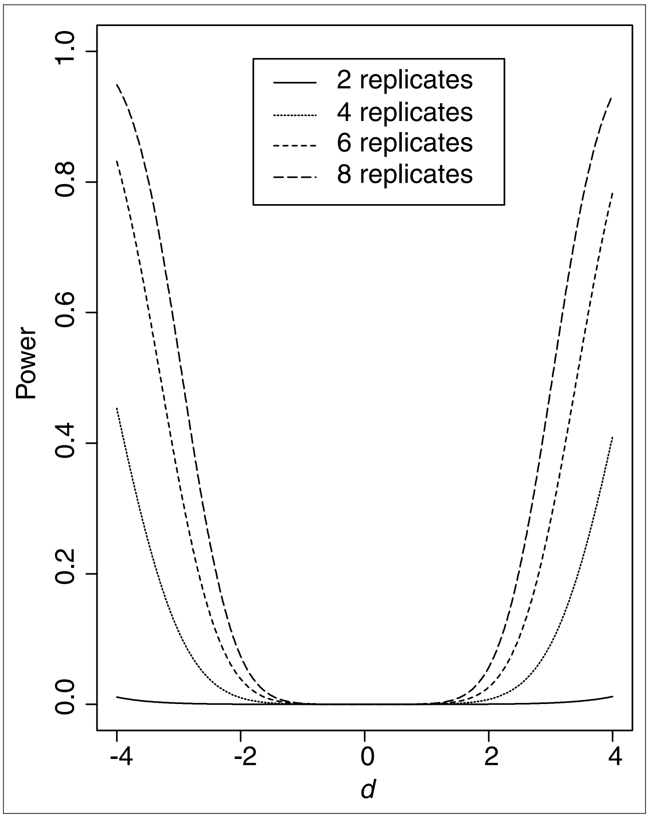

Figure 6
Power $\beta(d, \alpha)$ as a function of the magnitude of expression changes d and the number of replicates, with the gene-specific Type I error rate α = 0.05/10,000 for the middle ear-data.

Acknowledgements
This research was partially supported by NIH.

References

1. Brown P, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21(Suppl):**33-37.
2. Lander ES: **Array of hope.** *Nat Genet* 1999, **21(Suppl):**3-4.
3. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *J Biomed Optics* 1997, **2:**364-367.
4. 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.** *J Comput Biol* 2001, **8:**37-52.
5. Efron B, Tibshirani R, Goss V, Chu G: **Microarrays and their use in a comparative experiment.** Technical Report, Department of Statistics, Stanford University, 2000. [http://www-stat.stanford.edu/~tibs/research.html].
6. Ideker T, Thorsson V, Siehel AF, Hood LE: **Testing for differentially-expressed genes by maximum likelihood analysis of microarray data.** *J Comput Biol* 2000, **7:**805-817.
7. Li H, Hong F: **Cluster-Rasch models for microarray gene expression data.** *Genome Biol* 2001, **2(8)**:research0031.1-0031.13.
8. Thomas JG, Olson JM, Tapscott SJ, Zhao LP: **An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles.** *Genome Res* 2001, **11:**1227-1236.
9. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
10. Pan W, Lin J, Le C: **A mixture model approach to detecting differentially expressed genes with microarray data.** Technical Report 2001-011, Division of Biostatistics, University of Minnesota, 2001. [http://www.biostat.umn.edu/cgi-bin/rrs?print+2001].
11. Lee MLT, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* 2000, **97:**9834-9839.
12. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.** *Bioinformatics*, in press
. [http://www.biostat.umn.edu/cgi-bin/rrs?print+2001]
13. Black MA, Doerge RW: **Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments.** Technical Report, Department of Statistics, Purdue University, 2001.
14. Diggle PJ, Liang KY, Zeger SL: *Analysis of Longitudinal Data.* Oxford: Oxford University Press, 1994.
15. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** Technical Report, Statistics Department, University of California at Berkeley, 2000. [http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html/.
16. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98:**31-36.
17. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7:**819-837.
18. Yang YH, Buckley MJ, Dudoit S and Speed TP: **Comparison of methods for image analysis on cDNA microarray data.**

Technical Report, Statistics Department, University of California at Berkeley, 2000.
[http://www.stat.berkeley.edu/users/terry/zarray/Html/image.html].

19. Titteringto DM, Smith AFM, Makov UE: *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.
20. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *J Roy Stat Soc Ser B* 1977, **39:**1-38.
21. McLachlan GL, Basford KE: *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.
22. Akaike H: **Information theory and an extension of the maximum likelihood principle.** 2nd International Symposium on Information Theory. Edited by Petrov BN, Csaki F. Budapest: Akademiai Kiado, 1973; 267-281.
23. Schwartz G: **Estimating the dimensions of a model.** *Annls Statistics* 1978, **6:**461-464.
24. Fraley C, Raftery AE: **How many clusters? Which clustering methods? - Answers via model-based cluster analysis.** *Computer J* 1998, **41:**578-588.
25. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C, The Art of Scientific Computing*. 2nd edn. New York: Cambridge University Press, 1992.
26. Pan W, Lin J, Le C: **Model-based cluster analysis of microarray gene expression data.** *Genome Biol* 2002, **3(2)**:research009.1-research009.8.
27. Lin Y, Nadler ST, Attie AD, Yandell BS: **Mining for low-abundance transcripts in microarray data.** Technical Report, Department of Statistics, University of Wisconsin-Madison, 2001. [http://www.stat.wisc.edu/~yilin/].
28. Cleveland W, Devlin SJ: **Locally weighted regression: an approach to regression analysis by local fitting.** *J Am Stat Assoc* 1988; **83:**596-610.
29. **Statistical analysis of microarray data**
[http://www.biostat.umn.edu/~weip/ge.html]