Funct Integr Genomics (2003) 3:117–124
DOI 10.1007/s10142-003-0085-7

ORIGINAL PAPER

**Wei Pan · Jizhen Lin · Chap T. Le**

# A mixture model approach to detecting differentially expressed genes with microarray data

Received: 25 November 2002 / Accepted: 16 April 2003 / Published online: 1 July 2003


**Abstract** An exciting biological advancement over the past few years is the use of microarray technologies to measure simultaneously the expression levels of thousands of genes. The bottleneck now is how to extract useful information from the resulting large amounts of data. An important and common task in analyzing microarray data is to identify genes with altered expression under two experimental conditions. We propose a nonparametric statistical approach, called the mixture model method (MMM), to handle the problem when there are a small number of replicates under each experimental condition. Specifically, we propose estimating the distributions of a $t$ -type test statistic and its null statistic using finite normal mixture models. A comparison of these two distributions by means of a likelihood ratio test, or simply using the tail distribution of the null statistic, can identify genes with significantly changed expression. Several methods are proposed to effectively control the false positives. The methodology is applied to a data set containing expression levels of 1,176 genes of rats with and without pneumococcal middle ear infection.

**Keywords** Likelihood ratio · Permutation · Normal mixtures · SAM

## Introduction

Microarray technologies make it possible for the first time to simultaneously measure the expression levels of thousands of genes in a biological sample. They have been widely used over the past few years and have the potential to help advance our biological knowledge tremendously at a genomic scale (Botstein and Brown 1999; Lander 1999; Nguyen et al. 2002). However, one remaining challenge is how to analyze and interpret the resulting large amounts of data. A common task in such analyses is to detect genes with differential expression under two experimental conditions, which may refer to samples drawn from two types of tissues, tumors or cell lines, or at two time points during important biological processes. It has been known that simply using fold changes is unreliable and inefficient (Chen et al. 1997). Many statistical approaches have been proposed that aim to model some of the distributional properties of gene expression levels (e.g. Allison et al. 2002; Baggerly et al. 2001; Baldi and Long 2001; Broet et al. 2002; Chen et al. 1997; Dudoit et al. 2002; Efron et al. 2001; Guo et al. 2003; Huber et al. 2002; Ibrahim et al. 2002; Ideker et al. 2000; Kendziorski et al. 2002; Kerr et al. 2000; Kooperberg et al. 2002; Lee et al. 2002; Li and Wong 2001; Li et al. 2002; Lin et al. 2001; Newton et al. 2001, 2003; Rocke and Durbin 2001; Smyth et al. 2002; Strand et al. 2002; Thomas et al. 2001; Troyanskaya et al. 2002; Tusher et al. 2001). These approaches can be roughly classified into two categories. The first handles data from a single microarray containing only one spot for each gene, and has to depend on possibly too strong parametric assumptions. On the other hand, the second category of approaches takes advantage of the existence of multiple microarrays (or one array containing multiple spots for each gene) under each experimental condition. It has been found that due to high noise-signal ratio, a single microarray may not provide enough information to be reliably extracted in analysis (Lee et al. 2000). More importantly, multiple microarrays make it possible to assess possibly different variability of various genes. Some studies have found evidence that the variance of gene expression levels may be related with mean expression levels, hence leading to differential variability for various genes (Chen et al. 1997; Ideker et al. 2000; Newton et al. 2001; Huang and Pan 2002). In addition, an

W. Pan (✉) · C. T. Le
Division of Biostatistics, School of Public Health,
University of Minnesota, A460 Mayo, MMC 303,
420 Delaware Street SE, Minneapolis, MN 55455–0378, USA
e-mail: weip@biostat.umn.edu
Tel.: +1-612-6262705
Fax: +1-612-6260660

J. Lin
Department of Otolaryngology, School of Medicine,
University of Minnesota, Minneapolis, MN 55455-0378, USA

emerging novel idea is that with replicates of microarrays, one can estimate the distribution of random errors without strong parametric assumptions, making it possible to distinguish genuinely altered gene expression from noises with high confidence. This idea was first suggested in an empirical Bayes (EB) method of Efron et al. (2000, 2001) and a statistical method called the significance analysis of microarrays (SAM) of Tusher et al. (2001); we follow the same line in this work. In particular, we propose a mixture model method (MMM) that uses a mixture of Normal distributions as a flexible and powerful tool to estimate each of the two distributions of the test statistics and the null statistics, based on which we can use a likelihood ratio test (LRT) or the tail distribution of the null statistics to determine genes with differential expression. An advantage of this mixture model approach is that it enables us to control the number of false positives effectively. We apply the methodology to a data set containing the expression of 1,176 genes of normal rats and of those with pneumococcal middle ear infection. The results appear to be interesting and useful. We also discuss some results of using SAM.

## Materials and methods

### Test statistic and null statistic

We consider a generic situation that for each gene $i$, $i$ =1, 2,..., $N$, we have expression levels $X_{1i},...,X_{mi}$ from $m$ microarrays under condition 1, and $Y_{1i},...,Y_{ni}$ from $n$ arrays under condition 2. The expression level can be based on a summary measure of relative red to green channel intensities in a fluorescence-labeled cDNA array, a radioactive intensity of a radiolabeled cDNA array, or a summary difference of the perfect match (PM) and mis-match (MM) scores from an oligonucleotide array. The proposed method is not restricted to any specific microarray technology. Usually, the total number of genes $N$ is large (>1,000) whereas the number of replicates of microarrays, $m$ and $n$, are small (typically <30).

The goal here is to identify genes whose mean expression levels are different under the two conditions. Thus, it appears to be a usual two-sample comparison problem for each gene. However, some unique features of microarray data, such as the small $m$ and $n$ and the large $N$, render the traditional statistical tests, such as the $t$-test or rank-based nonparametric tests, ineffective (Thomas et al. 2001; Pan 2002). An alternative is to draw statistical inference based on the distributions of quantities related to $\{X_{1i},...,X_{mi}\}$ or $\{Y_{1i},...,Y_{ni}\}$, for $1 \leq i \leq N$, to take advantage of the large $N$.

Following Efron et al. (2000, 2001) and Tusher et al. (2001), we use a $t$-type score as the test statistic:

$$Z_i = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{v_{(1),i}/m + v_{(2),i}/n} + a_0} \quad (1)$$

for gene $i$. $\bar{X}_i = \sum_{j=1}^{m} X_{ji}/m$ and $\bar{Y}_i = \sum_{j=1}^{n} Y_{ji}/n$ are the sample means. $v_{(1),i}$ and $v_{(2),i}$ are the sample variances:

$$v_{(1),i} = \sum_{j=1}^{m} \left(X_{ji} - \bar{X}_i\right)^2/(m-1),\ v_{(2),i} = \sum_{j=1}^{n} \left(Y_{ji} - \bar{Y}_i\right)^2/(n-1)$$

The constant $a_0$ serves to stabilize the denominator of $Z_i$. Here we follow the recipe given in Efron et al. (2000) to take $a_0$ as the 90th percentile of $\left\{\sqrt{v_{(1),i}/m + v_{(2),i}/n} : i = 1, \ldots, N\right\}$. A more sophisticated method for choosing $a_0$ is used in SAM, and a Bayesian interpretation can be given (Baldi and Long 2001; Lonnstedt and Speed 2002).

The null distribution of a test statistic is defined as the distribution of the test statistic when the null hypothesis is true. Here the hypothesis is that the corresponding gene does not have differential expression. To estimate the null distribution of $Z_i$, one can use a permutation test. We first randomly permute the sample label, then apply the same $t$-type score to obtain a null score $z_i$ for gene $i$. In practice, we can permute data $B$ times, leading to $B$ sets of null scores $\{z^{(b)}{}_i : i = 1,...,N\}$ for $b$ =1,...,$B$.

Suppose that the distribution of all the $Z_i$s is $f$, and that of all the $z_i$s is $f_0$. Using the data $z_i$s and $Z_i$s, we can estimate $f_0$ and $f$ directly. A key idea is that for those genes with no differential expression, the distribution of their $Z_i$s is also $f_0$. If we assume that the distribution of $Z_i$s for genes with altered expression is $f_1$, $f$ can be expressed as a mixture of $f_0$ and $f_1$,

$$f = p_0 f_0 + p_1 f_1,$$

where $p_1$ is the proportion of the genes with altered expression and $p_0$ =1–$p_1$. Lee et al. (2000) and Newton et al. (2001) considered parametric approaches by assuming Normal or Gamma distributions for $f_0$ and $f_1$ respectively. As Efron et al. (2000) pointed out, without parametric assumptions, the parameters $p_0$, $f_0$ and $f_1$ are non-identifiable.

Here, we consider a nonparametric frequentist approach by estimating $f_0$ and $f$ directly. With a large sample size $N$, it seems unnecessary to take a parametric approach due to its possibly too strong distributional assumption. Rather, we propose estimating the density functions $f_0$ and $f$ using finite mixture models, which provide a flexible and powerful tool to model various random phenomena (e.g. Titterington et al. 1985; McLachlan and Peel 2000). For continuous data, such as gene expression data, the use of Normal components in a mixture distribution is natural. Note that a Normal mixture model is essentially a nonparametric density estimator. Compared to other popular nonparametric density estimators (such as the kernel estimator and the local likelihood estimator), a Normal mixture model also provides more stable estimates of tail probabilities; tail probabilities play a critical role in declaring statistical significance for a statistical test. It also facilitates determining the rejection region for a likelihood ratio test and a simplified test we propose to detect differential gene expression.

With a Normal mixture model, it is assumed that

$$f_0\left(z;\ \Phi_{g_0}\right) = \sum_{i=1}^{g_0} \pi_i \phi(z;\ \mu_i,\ V_i),$$

where $\phi(y;\mu_i,V_i)$ denotes the normal density function with mean $\mu_i$ and variance $V_i$, and $\pi_i$s are mixing proportions. We use $\Phi_{g0}$ to represent all unknown parameters $\{(\pi_i,\mu_i,V_i) : i = 1,...g_0\}$ in a $g_0$-component mixture model. The number of components $g_0$ can be selected based on the data. Similarly, a normal mixture model can be fitted to estimate $f$. Next we describe how to fit an Normal mixture model.

### Model fitting

The mixture model is typically fitted by maximum likelihood using the expectation-maximization (EM) algorithm (Dempster et al. 1977). For completeness, we briefly review how to fit a Normal mixture model to estimate $f_0$ (McLachlan and Basford 1988; Titterington et al. 1985). Given $N$ observations $z_1,..., z_N$, we want to maximize the log-likelihood

$$\log L\left(\Phi_{g_0}\right) = \sum_{j=1}^{N} \log f_0\left(z_j;\ \Phi_{g_0}\right)$$

to obtain the maximum likelihood estimate $\hat{\Phi}_{g_0}$. The EM algorithm computes $\hat{\Phi}_{g_0}$ by iterating the following steps.

Suppose that at the $k$th iteration, the parameter estimates are $\pi^{(k)}{}_i$s, $\mu^{(k)}{}_i$s and $V^{(k)}{}_i$s. Then in the ($k$ +1)th iteration, the estimates are updated by

$$\pi_i^{(k+1)} = \sum_{j=1}^{N} \tau_{ij}^{(k)}/N,$$

$$\mu_i^{(k+1)} = \sum_{j=1}^{N} \tau_{ij}^{(k)} z_j / \sum_{j=1}^{N} \tau_{ij}^{(k)},$$

and

$$V_i^{(k+1)} = \sum_{j=1}^{N} \tau_{ij}^{(k)} (z_j \mu_i^{(k+1)})^2 / \sum_{j=1}^{N} \tau_{ij}^{(k)},$$

for $i$=1,..., $g_0$, where

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \phi(z_j;\ \mu_i^{(k)},\ V_i^{(k)})}{f_0\left(z_j;\ \Phi_{g_0}^{(k)}\right)}$$

is the posterior probability that $z_j$ belongs to the $i$th component of the mixture, using the current parameter estimate $\Phi_{g0}^{(k)}$ for $\Phi_{g0}$, for $i$=1,..., $g_0$ and $j$=1,..., $N$.

At convergence, we obtain $\hat{\Phi}_{g_0}=\Phi^{(\infty)}{}_{g0}$ as the maximum likelihood estimate. Since local maxima can be found by the EM algorithm, it is desirable to run the algorithm multiple times with various starting values and choose the final estimate as the one resulting in the largest log-likelihood.

One interesting but difficult problem is to determine the number of components $g_0$. This can be accomplished through using various model selection criteria, of which the most well known are the Akaike Information Criterion (AIC; Akaike 1973) and the Bayesian Information Criterion (BIC; Schwartz 1978):

$$AIC = -2\log L(\hat{\Phi}_{g_0}) + 2\nu_{g_0},$$

$$BIC = -2\log L(\hat{\Phi}_{g_0}) + \nu_{g_0}\log(N),$$

where $\nu_{g0}$ is the number of independent parameters in $\Phi_{g0}$. 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 other criteria have been studied but it does not appear that there exists a clear winner (Biernacki and Govaert 1999). Some empirical studies seem to favor the use of BIC (Fraley and Raftery 1998).

A different approach to selecting $g$ is through hypothesis testing. This could be done through the use of the likelihood ratio test (LRT) to test for the null hypothesis $H_0$: $g = g_0$ against the alternative $H_1$: $g = g_0$+1 for any given positive integer $g_0$. The LRT statistic is $2\log L(\hat{\Phi}_{g_0+1}) - 2\log L(\hat{\Phi}_{g_0})$, which, however, does not have the usual asymptotically chi-squared distribution due to the violation of required regularity conditions (e.g. the maximum likelihood estimate may lie in the boundary of its parameter space). McLachlan (1987) proposed using the bootstrap to approximate the distribution of the LRT statistic under the null hypothesis. Based on the resulting $P$-value, one can decide whether to reject $H_0$.

In the current context, we did not find selecting the number of components so critical. The reason is that our goal here is to estimate the distribution function, not $g_0$. In our experience, when the results of selecting $g_0$ based on AIC and BIC do not agree with each other, it often means that several models are reasonable and that no one can dominate the others. Based on the earlier studies (Fraley and Raftery 1998) and for simplicity, we lean to use BIC in the current context. Furthermore, we can compare a fitted mixture model with the empirical distribution (i.e. histogram) to have a visual check.

### Determining statistical significance

As discussed in Efron et al. (2000), for a given $Z$, if we want to test for the null hypothesis $H_0$ that $Z$ is from $f_0$, we can construct a likelihood ratio test (LRT) based on the following statistic:

$$LR(Z) = f_0(Z)/f(Z).$$

A large value of $LR(Z)$ gives no evidence against $H_0$, whereas a too small value of $LR(Z)$ leads to rejecting $H_0$. With the Normal mixture model, it is possible to numerically determine the rejection region. For any given false positive rate $\alpha$, we can use the bisection method (Press et al. 1992, p 353) to solve the equation

$$\alpha = \int_{LR(z)<s} f_0(z)dz.$$

to obtain a suitable $s = s(\alpha)$. Then the rejection region is $RR(\alpha) = \{Z{:}LR(Z){<}s\}$. We call the method of using the LRT in MMM as MMM-LRT.

The LRT has certain optimal properties (Lehmann 1986). However, its use is somewhat complicated. Based on our experience (see Results), a simplification can be made. We can simply choose the rejection region as the two tails of $f_0$ for a two-sided test (or obviously, one tail of $f_0$ for a one-sided test): $RR(\alpha) = \{Z{:}|Z|{>}t\}$ such that a cut-off point $t = t(\alpha)$ is determined by solving the following equation:

$$\alpha = \int_{|Z|>t} f_0(z)dz.$$

Again, the bisection method or other numerical methods can be used to solve the above equation. We call the resulting method MMM-tail.

For cases where the expression levels from an array are possibly correlated across the genes, or for simplicity, following the spirit of SAM (Tusher et al. 2001), we can estimate the numbers of false positives (FP) and total positives (TP) directly. In MMM-LRT, for any given $s$, we have:

$$\mathrm{FP}(s) = \frac{1}{B}\sum_{b=1}^{B} \#\{i : LR(z_i^{(b)}) < s\},\ TP(s) = \#\{i : LR(Z_i) < s\}.$$

Similarly, in MMM-tail we can use:

$$FP(t) = \frac{1}{B}\sum_{b=1}^{B} \#\{i : |z_i^{(b)}| > t\},\ TP(t) = \#\{i : |Z_i| > t\}.$$

In estimating FP, one can also use median, rather than mean, FP over permuted data. Based on the estimated FP and TP, one can also calculate the false discovery rate as FDR = FP/TP (Benjamini and Hochberg 1995; Storey 2001; Tusher et al. 2001).

## Results

### Data

Pneumococcal otitis media is one of the most common disease in children. Almost every child in the United States experiences at least one episode of acute otitis media by the age of 5 years. To understand the pathogenesis of otitis media, it is important to identify genes involved in response to pneumococcal middle ear infection and to study their roles in otitis media. A study was recently conducted, applying radioactively labeled DNA microarrays (Friemert et al. 1998) to the mRNA analysis of 1,176 genes in middle ear mucosa of rats with and without subacute pneumococcal middle ear infection (Pan et al. 2002a). It consisted of eight experiments: two DNA microarrays were run with controls while six were run with pneumococcal middle ear infection. A more detailed description of the experiments and on how to obtain the data has been provided in Pan et al. (2002a).

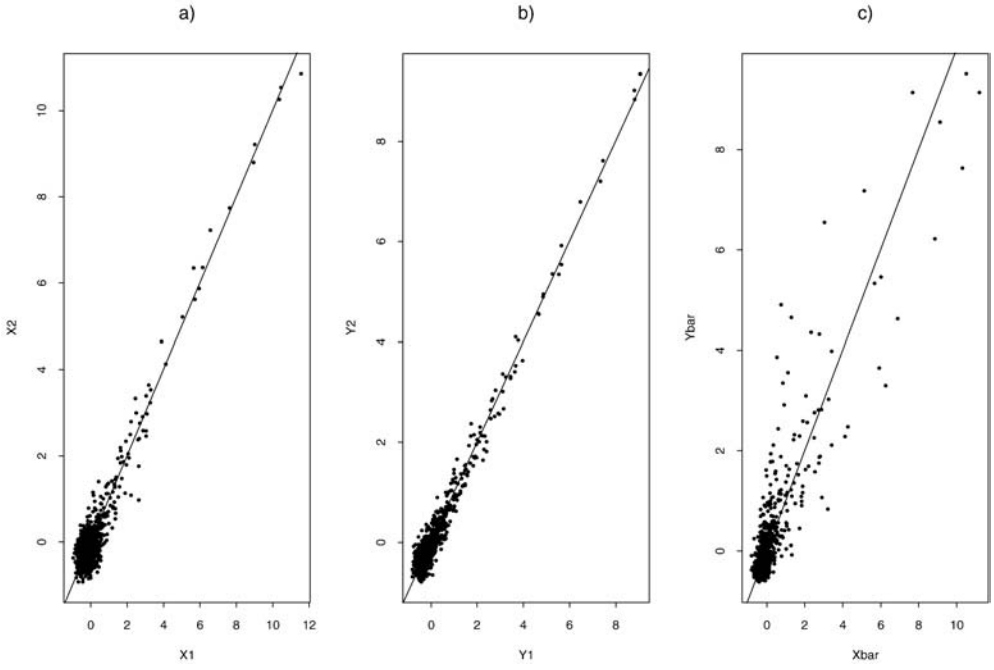


**Fig. 1a–c** Comparison of between-array agreement and between-condition discrepancy. **a** Gene expression levels from the two arrays in the pneumococcal infection group. **b** Gene expression levels from the first two two arrays in the control group. **c** The average gene expression levels in the pneumococcal infection group vs those in the control group

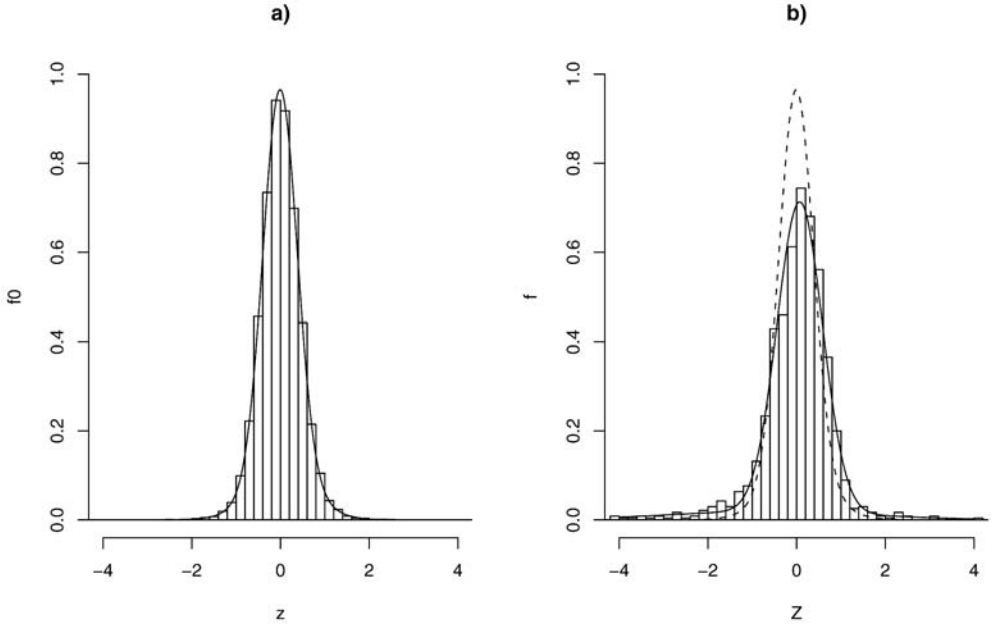


**Fig. 2a, b** Histograms and estimated distributions. **a** Histogram of $z_i$s and estimated $f_0$. **b** Histogram of $Z_i$s, estimated $f$ (*solid line*) and estimated $f_0$ (*dotted line*)

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. Then, for each microarray, we standardize the transformed gene expression levels by subtracting their mean and dividing by their standard deviation. Some scatter plots showing between-experiment comparisons are presented in Fig. 1. It can be seen that, in general, there is a good agreement as well as some variation between two arrays under the same condition, either for the control group or infected group. Also it appears that expression levels of some genes are altered with pneumococcal infection. The goal here is to identify those genes with a statistically significant expression change.

## Identifying significant genes

We generated all $B$ =28 sets of the null scores from all the possible permutations. Five mixture models were fitted to estimate $f_0$ with $g_0$ ranging from 1 to 5. Based on BIC, $g_0$=3 was selected. The parameter estimates are $\hat{\pi}$=(.836, .159, .003), $\hat{\mu}$=(–0.0090, 0.0403, –0.4773), and $\hat{V}$=(0.1462, 0.4891, 6.4069). Figure 2a presents the

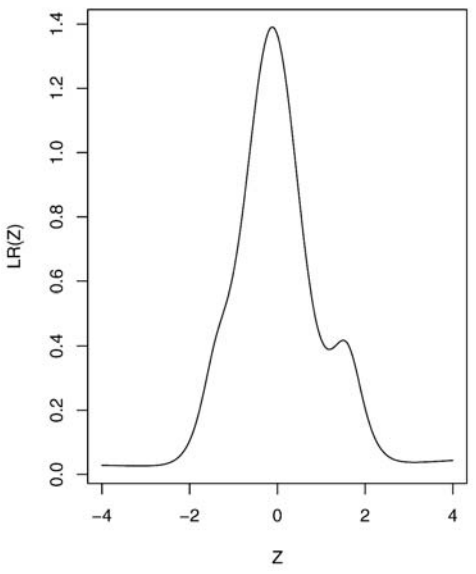


**Fig. 3** Likelihood ratio statistic as a function of $Z$ value

histogram of $z_i$s and the fitted $f_0$, which do not indicate strong discrepancy.

Next, we fitted five mixture models for $f$ with $g$ =1,..., 5. A two-component mixture model was selected using BIC. The resulting mixture model has maximum likelihood estimates $\hat{\pi}$=(0.893, 0.107), $\hat{\mu}$=(0.0712, –0.6476) and $\hat{V}$=(0.2627, 5.0873). The histogram of $Z_i$s and the fitted mixture model agree well (Fig. 2b).

The constructed LR statistics are plotted in Fig. 3. It is not surprising to see that as $Z$ moves away from 0, $LR(Z)$ decreases. It also indicates that the rejection region of the LRT will be in the two tails of $f_0$, which motivated our use of MMM-tail. Table 1 presents the number of significant genes (TP) and estimated FP using various cut-off points $s$ in the LRT. As a comparison, we also used the MMM-tail with various $\alpha = FP/1176$ where $FP$ are the same as the mean FP in MMM-LRT. It can be seen that the results from the two implementations are close, though there may be a slight loss of power (i.e. reduced TP) in using MMM-tail as compared to the LRT. The reason is that the LRT takes advantage of the available information on $f$.

It is interesting to note that in Table 1, $N\alpha$ for various $\alpha$ are always very close to the observed numbers of false positives; that is, $FP \approx N\alpha$. Hence, it lends support to the (approximate) independence among the gene-specific scores. This is related to the controversy about the use of the Bonferroni method to adjust multiple tests for microarray data. A popular view on the conservativeness of the Bonferroni adjustment may not be applicable here because, if the gene-specific scores are independent, the Bonferroni adjustment is almost exact. Specifically, if the $N$ gene-specific scores are independent, using the gene-specific significance level $\alpha/N$ in the Bonferroni method leads to a genome-wide (or family-wise) significance level $1–(1–\alpha/N)^N$, whose first-order Taylor approximation is $\alpha$, the specified genome-wide significance level. Thus the Bonferroni method should work well for a large $N$, which is typically the case with microarray data.

## Discussion on some significant genes

Many of the top 90 significant genes are of interest. First, the genes selected include those engaged in inflammatory reactions such as tumor necrosis factor receptor 1, acute phase reaction such as acute phase response factor, molecular switches or transcription factors for cellular growth and growth-arrest such as Id3 and Gax, mitogenic signaling such as rac-beta serine/threonine kinase, hyper-secretion activity such as Von Ebner's gland protein, cellular protection reaction such as heat shock proteins, water transport such as water channel aquaporin 3, mitogenic response such as platelet-derived growth factor alpha receptor, and anti-proliferation activity such as transforming growth factor beta 3. These genes are important in pneumococcal middle ear infection because they contribute to the development of otitis media with effusion. Detailed information regarding these genes' functions refers to the GenBank references in the third column of Table 2. The rest of the genes are not listed because most are not specifically related to pneumococcal middle ear infection. These genes encode basic cellular proteins or house-keeping molecules whose functions are for the survival of cells, production of energy, synthesis of proteins, and maintenance of cellular structure. They are responsive not only to pneumococcal middle ear infection but also to other non-inflammatory stimuli such as immune reactions.

## Comparison with SAM

Tusher et al. (2001) proposed a novel method called Significance Analysis of Microarray (SAM). We applied

**Table 1** Estimated numbers of total positives (*TP*) and false positives (*FP*) in two implementations of the mixture model method (*MMM*): MMM-LRT and MMM-tail

| MMM-LRT | | | | MMM-tail | | | |
|---|---|---|---|---|---|---|---|
| $s$ | Median FP | Mean FP | TP | $\alpha$ | Median FP | Mean FP | TP |
| 0.15 | 1.00 | 2.75 | 44 | 2.75/1,176 | 1.00 | 3.18 | 43 |
| 0.30 | 2.00 | 5.04 | 58 | 5.04/1,176 | 2.00 | 5.36 | 58 |
| 0.35 | 4.00 | 6.04 | 61 | 6.04/1,176 | 3.00 | 6.36 | 63 |
| 0.40 | 14.50 | 20.75 | 90 | 20.75/1,176 | 18.50 | 21.04 | 100 |
| 0.45 | 24.50 | 32.79 | 119 | 32.79/1,176 | 30.00 | 32.43 | 120 |
| 0.60 | 70.50 | 75.18 | 204 | 75.18/1,176 | 68.50 | 76.14 | 195 |
| 0.70 | 101.00 | 111.90 | 269 | 111.90/1,176 | 103.00 | 112.4 | 253 |

**Table 2** Some of the genes with altered expression following pneumococcal ear infection

| GenBank accession no. | Gene/Protein name | Function |
|---|---|---|
| M63122 | Tumor necrosis factor receptor 1 | Inflammatory reaction |
| X91810 | Stat3, signal transducer and activator of transcription 3 | Acute phase response factor |
| Z17223 | Gax, growth-arrest-specific protein | Transcription factor, growth arrest |
| D10864 | Id3, DNA-binding protein inhibitor | Cell cycle progression, growth |
| X74806 | Von Ebner's gland protein | Middle ear gland protein |
| D30041 | rac-beta serine/threonine kinase (rac-PK-beta); AKT2 | Mitogenic signaling |
| D30040 | rac-alpha serine/threonine kinase (rac-PK-alpha); protein kinase B | Mitogenic signaling |
| M86389 | Heat shock 27-kDa protein (HSP27) | Cellular protection |
| Z27118 | Heat shock 70-kDa protein (HSP70) | Cellular protection |
| D17695 | Water channel aquaporin 3 (AQP3) | Water transportation |
| M63837 | Platelet-derived growth factor alpha receptor (PDGFRa) | Proliferation |
| U03491 | Transforming growth factor beta 3 (TGF-beta3) | Anti-proliferation |

**Table 3** Estimated numbers of total positives (*TP*) and false positives (*FP*; i.e. median FP and 90th percentile FP) in SAM (significance analysis of microarrays; version 1.21)

| Δ | Median FP | 90% FP | TP |
|---|---|---|---|
| 1.500 | 0.69 | 3.46 | 41 |
| 1.350 | 1.38 | 4.15 | 45 |
| 1.143 | 2.77 | 9.69 | 60 |
| 0.933 | 5.54 | 17.30 | 86 |
| 0.862 | 7.61 | 23.53 | 94 |
| 0.760 | 13.15 | 35.99 | 121 |
| 0.592 | 35.30 | 78.91 | 207 |
| 0.473 | 65.76 | 116.29 | 274 |

the SAM software version 1.21 (Chu et al. 2003) to the data, and the results are listed in Table 3. SAM outputs the median FP and the 90th percentile of FP, but not the mean FP; in the following, as usual, we will use the median FP as the estimate of the true FP. The tuning parameter Δ serves to control the FP in SAM. It can be seen that when the TP is small, the results of SAM are similar to that of MMM. However, when the detected TP is large, the estimated FP from SAM tends to be much smaller than that from MMM. On the other hand, the 90th percentile of FP given in SAM is in general in better agreement with the mean/median FP in MMM. This is probably related to a new formula used in SAM to estimate an FP: an estimated FP based on $z_i$s is multiplied by an estimated proportion of genes with no expression change (Chu et al. 2003, p 20).

## Discussion

We have proposed estimating the distributions of a test statistic and a null statistic using Normal mixture models. With fitted mixture models, we can declare statistical significance using the LRT or with the tail distribution of the null statistic detect genes with differential expression. The methodology can effectively control the false positive rate or false discovery rate. Its application to real data yields interesting and useful results.

Here we only considered use of permuting class labels to generate null scores as used in SAM (Tusher et al. 2001) and EB (Efron et al. 2000, 2001). This permutation method is both simple and general. However, it can lead to too conservative inference (Pan 2003). Thus, other methods of permuting the data have been proposed (Pan 2002, 2003; Zhao and Pan 2003); they can be equally applied to MMM, as well as to SAM and EB. Among these alternative methods, the one proposed in Pan (2002) can lead to too liberal inference (Zhao and Pan 2003), and the others have their own limitations, such as requiring that the number of replicates under each condition be no smaller than 4, which does not hold for our middle ear infection data.

The Normal mixture model can be considered as a nonparametric estimator of a distribution function. In addition to its flexibility and closed form solution, it is particularly desirable here for the stability of its tail probabilities, which play an important role in assessing statistical significance. Zhao and Pan (2003) did some simulation studies to show its good performance.

In spirit, our proposed MMM is similar to SAM. However, there are some potential advantages with MMM. First, with an estimated null distribution $f_0$, we can identify a significance/rejection region for any given Type I error rate $\alpha$. In SAM, because FP is estimated by a finite number of simulated null scores, it may not be able to estimate some small FP well. Second, it is possible to do sample size/power calculations for microarray experiments in the framework of MMM (Pan et al. 2002b), whereas it is still unclear how to do so in SAM.

We did not elaborate on data preprocessing, such as summary statistics of gene expression levels and data normalization. These are important topics many authors have addressed (e.g. Bolstad et al. 2003; Dudoit et al. 2002; Efron et al. 2000; Irizarry et al. 2003; Li and Wong 2001; Kerr et al. 2000; Naef et al. 2003; Quackenbush 2002; Yang et al. 2002a, 2002b; Zhou and Abagyan

2002). We also did not touch on other important issues (Chuaqui et al. 2002), such as experimental design (Kerr and Churchill 2001; Churchill 2002), gene network inference (Halfon and Michelson 2002) and pattern recognition (Valafar 2002) with microarray data. In general, we believe that there are many open and interesting questions to be tackled with microarray data, which provide exciting and tremendous opportunities for statistics to play an important role.

**Acknowledgements** W.P. was supported by an NIH grant (R01-HL65462) and a Minnesota Medical Foundation grant. The authors are grateful to two referees for many helpful comments and suggestions.

## References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) 2nd international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281

Allison DB, Gadbury GL, Heo M, Fernandez J, Lee K-C, Prolla TA, Weindruch R (2002) A mixture model approach for the analysis of microarray gene expression data. Comput Stat Data Anal 39:1–20

Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV, Zhang W (2001) Identifying differentially expressed genes in cDNA microarray experiments. J Comput Biol 8:639–659

Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17:509–519

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300

Biernacki C, Govaert G (1999) Choosing models in model-based clustering and discriminant analysis. J Stat Comput Simul 64:49–71

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193

Botstein D, Brown P (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet Suppl 21:33–37

Broet P, Richardson S, Radvanyi F (2002) Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. J Comput Biol 9:671–683

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

Chu G, Narasimhan B, Tibshirani R, Tusher V (2003) SAM users guide and technical document (SAM 1.21). http://www-stat.stanford.edu/~tibs/SAM/index.html

Chuaqui RF, Bonner RF, Best CJM, et al (2002) Post-analysis follow-up and validation of microarray experiments. Nat Genet Suppl 32:509–514

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32:490–495

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38

Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin 12:111–139

Efron B, Tibshirani R, Goss V, Chu G (2000) Microarrays and their use in a comparative experiment. http://www-stat.stanford.edu/~tibs/research.html

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Fraley C, Raftery AE (1998) How many clusters? Which clustering methods?—Answers via model-based cluster analysis. Comput J 41:578–588

Friemert C, Erfle V, Strauss G (1998) Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. Methods Mol Cell Biol 1:143–153

Guo X, Qi H, Verfaillie CM, Pan W (2003) Statistical significance analysis of longitudinal gene expression data. Bioinformatics (in press). Available at http://www.biostat.umn.edu/cgi-bin/rrs?print+2003

Halfon MS, Michelson AM (2002) Exploring genetic regulatory networks in metazoan development: methods and models. Physiol Genomics 10:131–143

Huang X, Pan W (2002) Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. Funct Integr Genom 2:126–133

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18:S96–S104

Ibrahim JG, Chen M-H, Gray RJ (2002) Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc 97:88–99

Ideker T, Thorsson V, Siehel AF, Hood LE (2000) Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. J Comput Biol 7:805–817

Irizarry RA, Hobbs B, Colin F, Beazer-Barclay YD, Antonellis K, Scherf U, Speed TP (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. Biostatistics (in press)

Kendziorski CM, Newton MA, Lan H, Gould MN (2002) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Stat Med (in press) Available at http://www.biostat.wisc.edu/ ~ kendzior/

Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2:183–202

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Computal Biol 7:819–837

Kooperberg C, Sipione S, LeBlanc ML, Strand AD, Cattaneo E, Olson JM (2002) Evaluating test-statistics to select interesting genes in microarray experiments. Hum Mol Genet 11:2223–2232

Lander ES (1999) Array of hope. Nat Genet Suppl 21:3–4

Lee M-LT, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci 97:9834–9839

Lee M-LT, Lu W, Whitmore GA, Beier D (2002) Models for microarray gene expression data. J Biopharmaceut Stat 12:1–19

Lehmann EL (1986) Theory of point estimation. Wiley, New York

Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci 98:31–36

Li H, Luan Y, Hong F, Li Y (2002) Statistical methods for analysis of time course gene expression data. Frontiers Biosci 7:a90–a98

Lin Y, Nadler ST, Attie AD, Yandell BS (2001) Mining for low-abundance transcripts in microarray data. http://www.stat.wisc.edu/ ~ yilin/

Lonnstedt I, Speed T (2002) Replicated microarray data. Stat Sin 12:31–46

McLachlan GL (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl Stat 36:318–324

McLachlan GL, Basford KE (1988) Mixture models: inference and applications to clustering. Dekker, New York

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

Naef F, Socci ND, Magnasco M (2003) A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations Bioinformatics 19:178–184

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 8:37–52

Newton MA, Noueiry A, Sarkar D, Ahlquist P (2003) Detecting differential gene expression with a semiparametric hierarchical mixture method. Technical report 1074, Department of Statistics, UW Madison. http://www.stat.wisc.edu/ ~ newton/papers/publications/

Nguyen DV, Arpat AB, Wang N, Carroll RJ (2002) DNA microarray experiments: biological and technical aspects. Biometrics 58:701–717

Pan W (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 12:546–554

Pan W (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. Bioinformatics (in press) http://www.biostat.umn.edu/cgi-bin/rrs?print+2002

Pan W, Lin J, Le C (2002a) Model-based cluster analysis of microarray gene expression data. Genome Biol 3(2):research009.1–research009.8

Pan W, Lin J, Le C (2002b) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol 3(5):research0022.1–research0022.10

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C, the art of scientific computing, 2nd edn. Cambridge University Press, New York

Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32:496–501

Rocke DM, Durbin B (2001) A model for measurement error for gene expression arrays. J Comput Biol 8:557–570

Schwartz G (1978) Estimating the dimensions of a model. Ann Stat 6:461–464

Smyth GK, Yang YH, Speed T (2002) Statistical issues in cDNA microarray data analysis. http://www.stat.Berkeley.EDU/users/terry/zarray/Html/papersindex.html

Storey JD (2001) The positive false discovery rate: a Bayesian interpretation and the q-value. Technical Report, Department of Statistics, Stanford University, Stanford, Calif.

Strand AD, Olson JM, Kooperberg C (2002) Estimating the statistical significance of gene expression changes observed with oligonucleotide arrays. Hum Mol Genet 11:2207–2221

Thomas JG, Olson JM, Tapscott SJ, Zhao LP (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res 11:1227–1236

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York

Troyanskaya OG, Garber ME, Brown PO, et al (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18:1454–1461

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98:5116–5121

Valafar F (2002) Pattern recognition techniques in microarray data analysis—a survey. Ann NY Acad Sci 980:41–64

Yang YH, Buckley MJ, Dudoit S, Speed TP (2002a) Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 11:108–136

Yang YH, et al (2002b) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30:e15

Zhao Y, Pan W (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. Bioinformatics (in press) http://www.biostat.umn.edu/cgi-bin/rrs?print+2002

Zhou Y, Abagyan R (2002) Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. BMC Bioinformatics 3:3