A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments

Wei Pan

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

August 3, 2001; Revised October 8, 2001

1Division 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.
A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments

ABSTRACT

Motivation: A common task in analyzing microarray data is to determine which genes are differentially expressed across two kinds of tissue samples or samples obtained under two experimental conditions. Recently several statistical methods have been proposed to accomplish this goal when there are replicated samples under each condition. However, it may not be clear how these methods compare with each other. Our main goal here is to compare three methods, the t-test, a regression modeling approach (Thomas et al, 2001) and a mixture model approach (Pan et al, 2001) with particular attention to their different modeling assumptions.

Results: It is pointed out that all the three methods are based on using the two-sample t-statistic or its minor variation, but they differ in how to associate a statistical significance level to the corresponding statistic, leading to possibly large difference in the resulting significance levels and the numbers of genes detected. In particular, we give an explicit formula for the test statistic used in the regression approach. Using the leukemia data of Golub et al (1999), we illustrate these points. We also briefly compare the results with those of several other methods, including the empirical Bayesian method of Efron et al (2000) and the Significance Analysis of Microarray (SAM) method of Tusher et al (2001).

Contact: weip@biostat.umn.edu

Key Words: Gene expression; Microarray; Mixture modeling; Regression modeling; t-test.
INTRODUCTION

An exciting development in genomics is the use of microarray technology to simultaneously monitor the expression levels of thousands of genes (or expressed sequence tags) (Brown and Botstein 1999; Lander 1999). A common task is to compare the expression levels of genes in samples drawn from two different tissues or at two different time points or conditions. Specifically, it is of interest to detect genes with differential expression under the two conditions. In early days, the simple method of fold changes was used and now it is known to be unreliable (Chen et al 1997) because statistical variability was not taken account. Since then, many more sophisticated statistical methods have been proposed (e.g. Chen et al 1997; Efron et al 2000; Ideker et al 2000; Newton et al 2001; Tusher et al 2001; Lin et al 2001; Pan et al 2001a). It has also been noticed that data based on a single array may not be reliable and may contain high noises (Lee et al 2000). As the technology advances, microarray experiments are becoming less expensive, which makes the use of multiple arrays (or multiple spots on each array) feasible. In this paper, we consider the detection of differentially expressed genes with replicated measurements of expression levels of each gene under each condition.

A straightforward method is to use the traditional two-sample t-test (e.g. Devore and Peck, 1997). Thomas et al. (2001) proposed a regression modeling approach. Pan et al. (2001a) suggested a mixture model approach, which follows the basic idea of Efron et al (2000) and Tusher et al (2001). However, it is not clear how these methods compare with each other. For practitioners to choose a method, it is important to elucidate various modeling assumptions underlying each method. In this paper, we comparatively review the three methods, the t-test, the regression approach of Thomas et al. (2001), and the mixture model approach of Pan et al. (2001a). In particular, we give an explicit form of the test statistic in the regression method, facilitating the discussion of the connections and differences among the three methods. We apply the three methods to the leukemia data of Golub et al (1999). We also briefly discuss the results of applying the empirical Bayesian (EB) method of Efron et al (2000) and the Significance Analysis of Microarray (SAM) method of Tusher et al (2001) to the same leukemia data.

All the methods are 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, or a sum-
mary 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). To focus on the main issue, we assume that all the methods use the data preprocessed in the same way.

**METHODS**

**Data**

Suppose that \( Y_{jk} \) is the expression level of gene \( j \) in array \( k \) \((j = 1, \ldots, n; \ k = 1, \ldots, K_1, K_1 + 1, \ldots, K_1 + K_2)\). Suppose that the first \( K_1 \) and last \( K_2 \) arrays are obtained under the two conditions respectively.

A general statistical model is

\[
Y_{jk} = a_j + b_j x_k + \epsilon_{jk},
\]

where \( x_k = 1 \) for \( 1 \leq k \leq K_1 \) and \( x_k = 0 \) for \( K_1 + 1 \leq k \leq K_1 + K_2 \), and \( \epsilon_{jk} \) are random errors with mean 0. Hence \( a_j + b_j \) and \( a_j \) are the mean expression levels of gene \( j \) under the two conditions respectively. Determining whether a gene has differential expression is equivalent to testing for the null hypothesis

\[
H_0: b_j = 0 \text{ against } H_1: b_j \neq 0.
\]

A statistical test consists of two parts. The first is to construct a summary test statistic. The second is to determine the significance level or p-value associated with the test statistic. The p-value is usually calculated based on the null distribution of the test statistic (i.e. the distribution of the test statistic under \( H_0 \)), which may be specified or estimated via modeling assumptions.

To focus on the main issue, we use \( \alpha = 0.01 \) as the genome-wide significance level, and we use the Bonferroni adjustment to deal with multiple comparisons. Hence the test- or gene-specific significance level (for a two-sided test) is \( \alpha^* = \alpha / (2n) \). We do not consider other possibly better adjustment methods for multiple comparisons (e.g. Dudoit et al 2000; Thomas et al 2001). Most, if not all, statistical tests can be modified accordingly for a multiple comparison adjustment.

In the following, we review the three methods along the line.
The t-test

There are several versions of the two-sample t-test, depending on whether the sample size (i.e. \( K_1 \) and \( K_2 \)) is large and whether it is reasonable to assume that gene expression levels have an equal variance under the two conditions (Devore and Peck, 1997, Sections 10.1-10.2). Because usually both \( K_1 \) and \( K_2 \) are small, and there is evidence to support unequal variances (e.g. Thomas et al, 2001), we will only discuss the t-test with two independent small Normal samples with unequal variances.

Let the sample means and variances of \( Y_{jk} \)'s for gene \( j \) under the two conditions be

\[
\bar{Y}_{j(1)} = \frac{\sum_{k=1}^{K_1} Y_{jk}}{K_1}, \quad \bar{Y}_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2} Y_{jk}}{K_2}
\]

and

\[
s_{j(1)}^2 = \frac{\sum_{k=1}^{K_1} (Y_{jk} - \bar{Y}_{j(1)})^2}{K_1 - 1}, \quad s_{j(2)}^2 = \frac{\sum_{k=K_1+1}^{K_1+K_2} (Y_{jk} - \bar{Y}_{j(2)})^2}{K_2 - 1}.
\]

The t-statistic is

\[
Z_j = \frac{\bar{Y}_{j(1)} - \bar{Y}_{j(2)}}{\sqrt{s_{j(1)}^2/K_1 + s_{j(2)}^2/K_2}}.
\]

Under the Normality assumption for \( Y_{jk} \), \( Z_j \) approximately has a \( t \)-distribution with degrees of freedom

\[
d_j = \frac{(s_{j(1)}^2/K_1 + s_{j(2)}^2/K_2)^2}{(s_{j(1)}^2/K_1)^2/(K_1 - 1) + (s_{j(2)}^2/K_2)^2/(K_2 - 1)}.
\]

This \( t \)-test was proposed by Welch (1937). Its method of calculating the degrees of freedom is similar to the idea of the Satterthwaite approximation. The good performance of Welch’s \( t \)-test, compared with many other alternatives (e.g. approximating \( d_j \) by \( K_1 + K_2 - 2 \)), has been well documented (e.g. Scheffe 1970; Best and Rayner 1987).

We will use the Welch \( t \)-test throughout. Note that the Welch \( t \)-test is specifically designed to handle the possibility of having unequal variances. If we ignore the part of multiple comparison adjustment, Dudoit et al (2000) adopt the same \( t \)-statistic, but calculate the p-value by permutation.

A regression modeling approach

Thomas et al (2001) proposed a regression modeling approach. In their original formulation, data preprocessing and testing for differential expression are coupled together. To focus on the
main issue, we only consider the testing part of their approach throughout the paper. Then their model is the same as (1) (after ignoring the preprocessing step). Treating (1) as a regression model, they proposed to estimate \((a_j, b_j)\) using the weighted least square method, and then to estimate the variance of \(\hat{b}_j\) using the robust or sandwich variance estimator, say \(\text{Var}(\hat{b}_j)\). Their test statistic is

\[
Z_j' = \frac{\hat{b}_j}{\sqrt{\text{Var}(\hat{b}_j)}},
\]

with reference to the (asymptotically) Normal distribution.

Thomas et al observed that the result based on using \(Z_j'\) is close to that of using the \(t\)-statistic \(Z_j\). But because they did not give an explicit formula for \(Z_j'\), a theoretical explanation was unavailable. Next we give an explicit formula for \(Z_j'\), which will shed light on the nature of \(Z_j'\). Note that (1) can be formulated as a linear regression problem. If we denote the corresponding design matrix as \(X\), it is easy to verify that \(X'X\) is block-diagonal. Hence, the least-squares estimates of \((a_j, b_j)\) are independent with each other for different genes. In fact, it is easy to verify that the least-squares estimate of \(b_j\) is

\[
\hat{b}_j = \bar{Y}_j(1) - \bar{Y}_j(2),
\]

Then, as pointed out by Drum and McCullagh (1993), the robust variance estimator of \(\hat{b}_j\) is

\[
\text{Var}(\hat{b}_j) = \frac{s_j^2(1)}{K_1} \left\{ \frac{K_1 - 1}{K_1} \right\} + \frac{s_j^2(2)}{K_2} \left\{ \frac{K_2 - 1}{K_2} \right\},
\]

which can be also verified directly.

Hence, it can be seen that the statistic \(Z_j'\) of Thomas et al has a similar form to the usual \(t\)-statistic \(Z_j\) with the minor difference in how to estimate the variances: rather than the unbiased sample variances as used in \(Z_j\), the maximum likelihood estimator of the variance (under the Normality assumption for \(Y_{jk}\)'s) is used in \(Z_j'\). It is obvious that \(Z_j'\) and \(Z_j\) are equivalent as both \(K_1\) and \(K_2\) tend to the infinity. However, for small \(K_1\) and \(K_2\), \(Z_j\) is preferred due to the unbiasedness of its variance estimator involved.

Furthermore, using a standard Normal distribution to calculate the \(p\)-value for \(Z_j'\) is based on the assumption that both \(K_1\) and \(K_2\) are large, which however does not hold in many microarray experiments. Therefore, as acknowledged by Thomas et al, the Normality assumption for \(Z_j'\) may be too strong and may not work well in practice.
Note that the use of the sandwich variance estimator proposed by Thomas et al (2001) is novel. It works asymptotically even if random errors $\epsilon_{jk}$'s in (1) have different variances for different $j$, or even for the same gene $j$ under the two conditions. More often, for a linear regression model like (1), it is assumed that $\epsilon_{jk}$'s are a random sample (iid) from the same distribution, and hence have the same variance. If we ignore the preprocessing that corrects array and other effects, the latter assumption on iid $\epsilon_{jk}$'s is taken by Kerr et al (2000).

A mixture modeling approach

A common problem with the above t-test and the regression approach is their strong assumptions on the null distributions of the test statistics. In contrast, following the idea of Efron et al (2000) and Tusher et al (2001), Pan et al (2001a) proposed to estimate the null distribution directly. The method takes full advantage of the existence of replicated data, but it does require that both $K_1$ and $K_2$ are even numbers.

A key step is to construct the following null statistics:

$$z_j = \frac{Y_{j[1]}p_j/K_1 - Y_{j[2]}q_j/K_2}{\sqrt{s_j^2/[K_1 + s_j^2/K_2]^1}}, \quad (3)$$

where $Y_{j[1]} = (Y_{j,1}, ..., Y_{j,K_1})$, $Y_{j[2]} = (Y_{j,K_1+1}, ..., Y_{j,K_1+K_2})$, $p_j$ is a random permutation of a column vector containing $K_1/2$ 1's and $-1$'s respectively, and $q_j$ is a random permutation of a column vector containing $K_2/2$ 1's and $-1$'s respectively. Suppose that the distribution density functions of $z_j$ and $Z_j$ are respectively $f_0$ and $f$. Under the weak assumption that the random errors $\epsilon_{jk}$ have a distribution symmetric about its mean 0, then under $H_0$, $z_j$ and $Z_j$ have the same distribution $f_0 = f$.

Using $z_j$'s and $Z_j$'s we can estimate the distributions $f_0$ and $f$ respectively. We will discuss how to estimate $f_0$ and $f$ later. For the moment, suppose $f_0$ and $f$ are known (or more precisely, taken at their estimates). For any given $Z$, we can use the likelihood ratio test statistic

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

to test for $H_0$. A small value of $LR(Z)$, say $LR(Z) < c$, provides evidence to reject $H_0$. The cut-off point $c$ is determined such that the type I error rate is

$$\frac{\alpha}{n} = \int_{LR(z)<c} f_0(z)dz, \quad (4)$$
where $\alpha$ is the genome-wide significance level.

Now we discuss how to estimate $f_0$ and $f$. Pan et al proposed using a Normal mixture to model each distribution:

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

where $\phi(z; \mu_i, V_i)$ denotes the normal density function with mean $\mu_i$ and variance $V_i$, and $\pi_i$’s are mixing proportions. $\Omega_{g_0}$ represents all unknown parameters $\{(\pi_i, \mu_i, V_i) : i = 1, \ldots, g_0 \}$ in a $g_0$-component mixture model. The number of components can be selected adaptively. Similarly, a Normal mixture model can be used for $f$.

The Normal mixture model is flexible and powerful, and is widely used in many applications. It is usually fitted by maximum likelihood using the expectation-maximization (EM) algorithm (Dempster et al. 1977). To determine the number of components $g_0$, we can use various model selection criteria, of which the Bayesian Information Criterion (BIC) (Schwartz 1978) is favored in some empirical studies (Fraley and Raftery 1998):

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

where $L(\hat{\Omega}_{g_0})$ is the maximized likelihood function and $\nu_{g_0} = 3g_0 - 1$ is the number of independent parameters in $\Omega_{g_0}$. In using the 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 BIC (Fraley and Raftery 1998).

We used the EMMIX, a stand-alone Fortran program for fitting a Normal mixture model using maximum likelihood method. It was implemented by McLachlan et al. (1999) and is freely available from the web at


It has many interesting features, including multiple starts of the EM algorithm and calculation of model selection criteria.

**RESULTS**

**Data**

We apply the methods to the leukemia data of Golub et al (1999), which consists of 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples. The goal
is to find genes with differential expression between ALL and AML. Thomas et al also analyzed this data set. Since the mixture model approach requires even numbers of samples, we randomly take the first $K_1 = 26$ ALL samples and last $K_2 = 10$ AML samples for use. There are $n = 7129$ genes in each sample. As mentioned earlier, we take the genome-wide significance level at the usual $\alpha = 0.01$ level, and use the Bonferroni method to adjust for multiple comparisons. Hence the gene-specific significance level being used is $\alpha^* = 0.01/(7129 \times 2) = 7.014 \times 10^{-7}$.

Data preprocessing is accomplished for each sample by subtracting its median and dividing by its quartile range (the difference between the first and the third quartiles). Note that rather than the commonly used mean and standard deviation, we used the median and quartile range because the latter two are more robust estimators for the center and the dispersion of a distribution respectively. All the methods are applied to thus preprocessed data.

**Fitted mixture models**

We fitted three mixture models for either $f_0$ or $f$ with 1 to 3 components (Table 1). Based on BIC, we can see that a two-component mixture is the best for each and the fitted models are:

$$f_0(z) = 0.479 \phi(z; -0.746, 0.697) + 0.521 \phi(z; 0.739, 0.641),$$

and

$$f(z) = 0.518 \phi(z; -0.318, 1.803) + 0.482 \phi(z; 0.781, 4.501).$$

Figure 1 presents the histograms and the fitted models. For comparison, the density function of a $t$-distribution with degrees of freedom 34 is imposed in Figure 1(a). It can be seen that the $t$ distribution has much heavier tails than the estimated $f_0$. For Figure 1(b), we present both the estimated $f_0$ and $f$. It can be seen that, unsurprisingly, $f$ has heavier tails than $f_0$.

The LR function is depicted in Figure 2. Using the bisection method (Press et al 1992, p.353) to solve equation (4), we obtain the cut-off point $c = 0.0003437$, yielding a rejection region of $\{Z : Z < -4.8877 \text{ or } Z > 4.4019\}$ for $H_0$.
The test statistics $Z_j$ and $Z'_j$ are easy to calculate. But the three methods have different rejection regions for $H_0$.

For t-test, because the degrees of freedom $d_j$ vary with $j$, the rejection region is gene-specific: \( \{ Z : |Z| > t(\alpha^*, d_j) \} \), where $t(\alpha^*, d_j)$ is the upper $\alpha^*$-percentile for a $t$-distribution with degrees of freedom $d_j$. The distribution of $n$ calculated $d_j$’s ranges from 9 to 34 with mean and median 17.8 and 19.6 respectively. Note that the smaller the $d_j$, the large the $t(\alpha^*, d_j)$. To give a rough idea of the rejection region, we consider the conservative situation $d_j = 34$: the rejection region is \( \{ Z : |Z| > 5.8369 \} \).

For the regression modeling method, since the null distribution is assumed to be a standard Normal, the rejection region is \( \{ Z' : |Z'| > 4.8246 \} \), where 4.8246 is the upper $\alpha^*$-percentile for a standard Normal distribution.

As described earlier, the rejection region for the mixture model method is \( \{ Z : Z < -4.8877 \text{ or } Z > 4.4019 \} \).

Comparing the rejection regions of the three methods and the estimated null distribution and a $t$ distribution in Figure 1(a), we can see that the t-test is probably too conservative. Unsurprisingly, the t-test detects total 20 genes with significant expression changes, whereas the regression method and the mixture model method find 157 and 187 genes respectively.

Thomas et al provided some biological justifications for many identified genes.

**Top genes with differential expression**

Tables 2 and 3 list top 25 genes which are more highly expressed in AML and in ALL respectively. In general, the regression and the mixture model methods give very similar rankings; This can be explained by the closeness of $Z_j$ to $Z'_j$. In particular, it is reassuring that the two spots of the same gene, TCF3 (E2A), in Table 3 are ranked closely as number 14 and number 16 by both methods. The results of the three methods are in good agreement in Table 3. However, only the top 6 genes in Table 2 are identified as such by all the three methods. This may be due to the fact that the absolute values of the test statistics in Table 3 are much larger than those in Table 2. In other words, there is stronger evidence to suggest differential expression for the top genes more highly expressed in ALL than that for those more highly expressed in AML. In fact, according to the mixture model method, only the top 16 genes in Table 2 are identified as having significant
expression change.

| Table 2 about here |
| Table 3 about here |

Note that the ranking of the t-test is based on the corresponding p-values. Since the degrees of freedom of the null distribution in the t-test are gene-specific, the resulting ranking is different from that based on the test statistics $Z_j$. Thomas et al (2001) reported a good agreement between the t-test and the regression approach. The reason is that they used a fixed number (36) for the degrees of freedom in the t-test (LP Zhao, personal communications). If we do it that way, then the same conclusion can be drawn. Also, since the null distributions for the regression method and the mixture model method are fixed (i.e. non-gene-specific), the ranking based on $Z_j$ or $Z_j'$ should be the same as that based on the corresponding p-values.

For the purpose of comparison, we also give the ranking results taken from Thomas et al (2001). Most of the genes listed are also reported by them. However, the specific ranking may be very different. This may be largely due to the different methods used in preprocessing the data. This demonstrates the importance of data preprocessing.

**DISCUSSION**

A *comparative summary of the three methods*

We have given an explicit expression of the test statistic $Z_j'$ for the regression approach of Thomas et al, from which we can see that it has a similar form to the t-statistic, which is also used in the mixture model approach. Hence, the three methods usually give similar results in terms of the test statistics. However, they differ in how to determine the statistical significance level (or rejection region). For small sample sizes, both the t-test and the regression approach depend on the strong parametric assumptions, the t-distribution of $Z_j$ (or equivalently, the Normality assumption on the random errors) and the Normal distribution of $Z_j'$ respectively. It is possible that these parametric assumptions are violated in practice when small sample sizes are more common, though the two methods are asymptotically valid (with large sample sizes). In contrast, the mixture model approach estimates the null distribution directly. It takes advantage of the existence of multiple samples to construct the null scores $z_j$’s, and the large number of genes makes it feasible to estimate...
the null distribution $f_0$ (and $f$) nonparametrically. Note that the null distribution $f_0$ is for random errors, not for the gene expression levels. Of course, the mixture model method (as the EB method of Efron et al and SAM) also has its own modeling assumptions: it is assumed that the random errors have symmetric distributions, and after a suitable standardization (here we divide them by the sample variances), the random errors from all the genes have a common distribution. We believe that these assumptions are weak and reasonable. In particular, they are weaker than the Normality assumption used in the t-test.

Note that the Normality assumption in the t-test is required only for small $K_1$ and $K_2$. If both $K_1$ and $K_2$ tend to the infinity, $d_j$ also goes to the infinity, implying that the null distribution reduces to a standard Normal. Thus, if both $K_1$ and $K_2$ are large, such as $> 30$ as suggested by many introductory statistics textbooks (e.g. Devore and Peck, 1997, p.352), one can use a standard Normal as the null distribution for the t-test. The asymptotic Normality assumption for the null distribution of $Z_j^2$ in the regression approach also requires large $K_1$ and $K_2$. Hence, with large $K_1$ and $K_2$, both methods are essentially nonparametric.

With practical numbers of samples (i.e small $K_1$ and $K_2$), however, the power of the t-test is limited (due to the too small degrees of freedom), whereas the Normality assumption for the regression approach is more likely to be seriously violated. These are the situations where the mixture model approach and other similar approaches (Efron et al 2000, Tusher et al 2001) are more attractive. An advantage of the regression approach is its flexibility: it can be extended to model more complex biological processes (Zhao et al., 2001). An attractive point of the mixture model approach is its use for sample size/power calculations (Pan et al 2001b).

A brief comparison with other approaches

The Wilcoxon rank sum test (equivalent to Mann-Whitney test) has also been used as an alternative to the t-test in two-sample comparisons with microarray data. Because it is nonparametric, it avoids the possibly questionable parametric assumption used in the t-test. However, as demonstrated by Thomas et al., the price we pay for the robustness of the Wilcoxon test is the loss of power: when applied to the leukemia data, it does not find any gene with significant expression change. This is also related to another often neglected issue: the Wilcoxon test requires that the two samples have distribution functions with the same shape (with the only difference in their lo-
cation parameters). This implies, strictly speaking, that it is not applicable if the expression levels of a gene may have unequal variances under the two conditions, which is exactly the same reason why we prefer the t-test with unequal variances to that with an equal variance. These same issues remain with the use of other permutation-based nonparametric tests.

The mixture model approach follows the novel idea of the EB approach of Efron et al (2000) and of the SAM of Tusher et al (2001): estimating the null distribution using $z_j$’s. They belong to the same family with the same basic modeling assumptions. Here, we give a brief comparison of these methods by applying them to the same leukemia data. The general conclusion below is similar to that of Pan et al (2001a). To save space, we do not go to the details of the EB and SAM methods; the reader is referred to the above references for more details.

For the EB approach, Efron et al (2000) proposed a logistic regression method to estimate the likelihood ratio statistic $LR(Z)$ (Figure 3(a)), which is close to that obtained by the mixture method (Figure 2). Using $LR(Z)$, Efron et al derived a lower bound of the posterior probability that a gene with the $t$-statistic $Z_j$ has differential expression, $Pr(\text{Event} | Z_j)$. The posterior probability is drawn as a function of $Z$ in Figure 3 (b). The qualitative conclusion is the same as other methods: as $|Z_j|$ increases, there is stronger evidence to reject $H_0$. Corresponding to the rejection region by the mixture model method, the estimated posterior probabilities are $Pr(\text{Event} | -4.9) = 0.978$ and $Pr(\text{Event} | 4.4) = 0.982$. The posterior probability is closely related to the so-called false discovery rate (FDR) (Efron et al, 2001). The FDR is used as an alternative to controlling the false positive rate (i.e., Type I error rate) in handling multiple comparisons (Benjamini and Hochberg 1995). A potential problem is that, since only the lower bound of the posterior probability is actually estimated and given, the interpretation of the result in terms of significance level may be conservative. Nonetheless, there are many interesting features in the EB approach.

As pointed out by Efron et al, SAM is best suited to detecting a small number of genes with differential expression, which is not true for the leukemia data. Pan et al (2001a) also pointed out a problem with SAM under these situations. Although the SAM was originally designed to control the FDR, if desired, it can be also directly applied to control the Type I error. Using $B = 20$ versions of random permutations of $z_j$ scores, one can calculate the expected order statistics of $z_j$’s,
If we use $s = 2$, the estimated false positives and true positives are 0.1 and 297 respectively. If we use $s = 2.1$, the estimated false positives and true positives are 0 and 267 respectively. Figure 4 presents the results of SAM using $s = 2$, where the identified 297 genes with differential expression are those satisfying $|Z_{(j)} - \bar{Z}_{(j)}| > s$ with $Z_{(j)}$'s being the order statistics of $Z_j$'s. Since simulation is used in SAM, it is in general difficult to obtain results for a given Type I error $\alpha^*$. On the other hand, there are many attractive points of SAM. For instance, it does not have strong parametric assumptions and does not involve any complex estimation procedures (i.e. only order statistics are involved). In particular, it compares $Z_j$'s of all the genes collectively with their $z_j$'s. In contrast, all the other methods test gene by gene, which may be less efficient.

ACKNOWLEDGMENTS

The author thanks Drs Brad Efron, Rob Tibshirani and Lue Ping Zhao for helpful communications. In particular, Dr Zhao read an earlier version. The author is also grateful to the referees and editor for constructive comments.

REFERENCES


Available at http://www.stat.wisc.edu/~yilin/.


29. Welch, B.L. (1947). The generalization of ‘Students' problem when several different population variances are involved. *Biometrika*, 34, 28-35.

Table 1: BIC for fitted mixture models with various numbers of components $g$.

<table>
<thead>
<tr>
<th>$g$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_0$</td>
<td>21656.7</td>
<td>21585.4</td>
<td>21598.2</td>
</tr>
<tr>
<td>$f$</td>
<td>28986.1</td>
<td>28833.3</td>
<td>28857.1</td>
</tr>
</tbody>
</table>
Table 2: Top 25 genes more highly expressed in AML than in ALL.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Probe</th>
<th>$Z_j$</th>
<th>$Z'_j$</th>
<th>$t$</th>
<th>Regression rank</th>
<th>Thomas et al rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAH</td>
<td>M55150</td>
<td>-7.875</td>
<td>-8.248</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Neuromedin B</td>
<td>M21551</td>
<td>-5.767</td>
<td>-5.963</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CDC25A</td>
<td>M81933</td>
<td>-5.540</td>
<td>-5.778</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>NAB50</td>
<td>U63289</td>
<td>-5.508</td>
<td>-5.724</td>
<td>3</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>FTL</td>
<td>M11147</td>
<td>-5.426</td>
<td>-5.638</td>
<td>4</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Metargidin</td>
<td>U41767</td>
<td>-5.394</td>
<td>-5.617</td>
<td>5</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>LYN</td>
<td>M16038</td>
<td>-5.284</td>
<td>-5.563</td>
<td>-</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>LTC4S</td>
<td>U50136</td>
<td>-5.264</td>
<td>-5.541</td>
<td>-</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Metallothionein I-B</td>
<td>M13485</td>
<td>-5.096</td>
<td>-5.343</td>
<td>17</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>IGIF</td>
<td>D49950</td>
<td>-5.042</td>
<td>-5.296</td>
<td>23</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>PPlase</td>
<td>M80254</td>
<td>-5.023</td>
<td>-5.281</td>
<td>-</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>Inositol 1,3,4-trisphosphate</td>
<td>U51336</td>
<td>-5.001</td>
<td>-5.216</td>
<td>10</td>
<td>13</td>
<td>-</td>
</tr>
<tr>
<td>5/6-kinase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zyxin</td>
<td>X95735</td>
<td>-4.976</td>
<td>-5.242</td>
<td>-</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>ATP6C</td>
<td>M62762</td>
<td>-4.933</td>
<td>-5.174</td>
<td>22</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>CMKBR7</td>
<td>L08177</td>
<td>-4.894</td>
<td>-5.139</td>
<td>-</td>
<td>15</td>
<td>-</td>
</tr>
<tr>
<td>Chloride channel</td>
<td>Z30644</td>
<td>-4.884</td>
<td>-5.089</td>
<td>11</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>(putative) 2163bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thrombospondin 1</td>
<td>U12471</td>
<td>-4.847</td>
<td>-5.051</td>
<td>14</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>MDU1 Antigen</td>
<td>M21904</td>
<td>-4.795</td>
<td>-4.996</td>
<td>16</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>Proto-oncogene BCL3</td>
<td>U05681</td>
<td>-4.783</td>
<td>-5.002</td>
<td>18</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>GST-II</td>
<td>U77604</td>
<td>-4.765</td>
<td>-4.918</td>
<td>7</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>Calnexin</td>
<td>D50310</td>
<td>-4.723</td>
<td>-4.914</td>
<td>13</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>Polyadenylate binding protein II</td>
<td>Z48501</td>
<td>-4.712</td>
<td>-4.895</td>
<td>9</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>Sodium channel protein</td>
<td>M81758</td>
<td>-4.628</td>
<td>-4.789</td>
<td>8</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>HoxA9</td>
<td>U82759</td>
<td>-4.592</td>
<td>-4.824</td>
<td>-</td>
<td>23</td>
<td>13</td>
</tr>
<tr>
<td>PLCB2</td>
<td>M95678</td>
<td>-4.558</td>
<td>-4.779</td>
<td>-</td>
<td>25</td>
<td>16</td>
</tr>
</tbody>
</table>
Table 3: Top 25 genes more highly expressed in ALL than in AML.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Probe</th>
<th>$Z_j$</th>
<th>$Z_j'$</th>
<th>$t$ rank</th>
<th>Regression rank</th>
<th>Thomas et al rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>P48</td>
<td>X74262</td>
<td>8.083</td>
<td>8.286</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ACDM</td>
<td>M91432</td>
<td>7.113</td>
<td>7.278</td>
<td>3</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Macmarks</td>
<td>HG1612</td>
<td>6.945</td>
<td>7.149</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>HT1612</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYL1</td>
<td>M31211</td>
<td>6.929</td>
<td>7.151</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Proteasome</td>
<td>X59417</td>
<td>6.722</td>
<td>6.896</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>iota chain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenosine triphosphatase,</td>
<td>Z69881</td>
<td>6.683</td>
<td>6.830</td>
<td>8</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>calcium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-myb</td>
<td>U22376</td>
<td>6.587</td>
<td>6.742</td>
<td>7</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>IEF SSP 9302</td>
<td>L07758</td>
<td>6.535</td>
<td>6.742</td>
<td>6</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>Inducible protein</td>
<td>L47738</td>
<td>6.467</td>
<td>6.624</td>
<td>9</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>hdlc</td>
<td>U32944</td>
<td>6.394</td>
<td>6.532</td>
<td>13</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>DHPS</td>
<td>U26266</td>
<td>6.364</td>
<td>6.573</td>
<td>10</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>Cyclin D3</td>
<td>M92287</td>
<td>6.303</td>
<td>6.457</td>
<td>11</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>MB-1</td>
<td>U05259</td>
<td>6.250</td>
<td>6.378</td>
<td>18</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>TCF3 (E2A)</td>
<td>M65214</td>
<td>6.169</td>
<td>6.355</td>
<td>12</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>CRYZ</td>
<td>L13278</td>
<td>6.047</td>
<td>6.222</td>
<td>14</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>TCF3 (E2A)</td>
<td>M31523</td>
<td>6.007</td>
<td>6.141</td>
<td>21</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Nucleolysin TIA-1</td>
<td>M77142</td>
<td>5.986</td>
<td>6.109</td>
<td>25</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>Thymopoietin beta</td>
<td>U09087</td>
<td>5.968</td>
<td>6.124</td>
<td>15</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>MCM3</td>
<td>D38073</td>
<td>5.948</td>
<td>6.130</td>
<td>16</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>SRB</td>
<td>U38846</td>
<td>5.928</td>
<td>6.103</td>
<td>17</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>SPTAN1</td>
<td>J05243</td>
<td>5.868</td>
<td>6.026</td>
<td>20</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>Transcriptional activator hSNF2b</td>
<td>D26156</td>
<td>5.850</td>
<td>6.027</td>
<td>19</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>ALDR1</td>
<td>X15414</td>
<td>5.822</td>
<td>6.071</td>
<td>-</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>HKR-T1</td>
<td>S50223</td>
<td>5.807</td>
<td>6.013</td>
<td>-</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>T-complex protein</td>
<td>X74801</td>
<td>5.743</td>
<td>5.878</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

1. gamma subunit
Figure 1: Histograms and fitted mixture models (solid lines) for $z_j$ in (a) and for $Z_j$ in (b). In (a) the dotted line is a $t$-distribution with 34 degrees of freedom. In (b) the dotted line is the fitted $f_0$. 
Figure 2: The likelihood ratio function.
Figure 3: The results of the empirical Bayesian method: (a) the likelihood ratio function using the logistic regression method; (b) the estimated posterior probability function.
Figure 4: The results of SAM. Any gene corresponding to a point out of the bounds of the dotted lines is interpreted as having a significant expression change.