COMPUTATIONAL STATISTICS & DATA ANALYSIS

Computational Statistics & Data Analysis 39 (2002) 1–20

www.elsevier.com/locate/csda

A mixture model approach for the analysis of microarray gene expression data

David B. Allison[a,*], Gary L. Gadbury[b], Moonseong Heo[c], José R. Fernández[c], Cheol-Koo Lee[d,e], Tomas A. Prolla[e], Richard Weindruch[f,g]

[a]*Department of Biostatistics and Center for Research on Clinical Nutrition, University of Alabama at Birmingham, Ryals Public Health Bldg, Suite 327, 1665 University Ave. UAB Station, Birmingham, AL 35294-0022, USA*

[b]*Department of Mathematics and Statistics, University of Missouri—Rolla, USA*

[c]*Obesity Research Center, St. Luke's/Roosevelt Hospital, Institute of Human Nutrition, Columbia University College of Physicians & Surgeons, New York, NY, USA*

[d]*Environmental Toxicology Center, University of Wisconsin, Madison, WI, USA*

[e]*Departments of Genetics and Medical Genetics, University of Wisconsin, Madison, WI, USA*

[f]*Department of Medicine and the Wisconsin Regional Primate Research Center, University of Wisconsin, Madison, WI, USA*

[g]*The Geriatric Research, Education, and Clinical Center, William S. Middleton VA Hospital, Madison, WI, USA*

Received 1 November 2000; received in revised form 1 May 2001

Abstract

Microarrays have emerged as powerful tools allowing investigators to assess the expression of thousands of genes in different tissues and organisms. Statistical treatment of the resulting data remains a substantial challenge. Investigators using microarray expression studies may wish to answer questions about the statistical significance of differences in expression of any of the genes under study, avoiding false positive and false negative results. We have developed a sequence of procedures involving finite mixture modeling and bootstrap inference to address these issues in studies involving many thousands of genes. We illustrate the use of these techniques with a dataset involving calorically restricted mice.
© 2002 Elsevier Science B.V. All rights reserved.

* Corresponding author: Tel.: +1-205-975-9169; fax: +1-205-975-2540.
E-mail address: dallison@ms.soph.uab.edu (D.B. Allison).

0167-9473/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(01)00046-9

1. Introduction

A relatively recent addition to the armamentarium of investigators studying the genetics, molecular biology, and physiology of living organisms is the use of microarrays. This technique enables simultaneous and rapid assessment of the expression of literally thousands of genes or ESTs [1] in different tissues, different groups of experimental animals, humans, or other organisms, or organisms measured under different circumstances (Kahn et al., 1999). After the gene expression is measured, one may try to identify those genes for which there is differential expression across groups, and then interpret the results.

In brief, there are two broad classes of microarrays for gene expression measurement. One class, cDNA arrays, apply spots of cDNAs to glass slides. One can than estimate mRNA by examining hybridization to the cDNA spots. These arrays, though somewhat easier to develop, may not be as specific in their measurement properties as the second class of arrays, namely oligonucleotide arrays. Oligonucleotide arrays place many thousands of gene-specific oligonucleotides in silico and allow one to examine mRNA binding to the oligonucleotides after correcting for estimates of background binding ('noise'). For more details, see Weindruch et al. (2001) and references therein.

An example of this approach can be found in Lee et al. (1999) who studied differences in gene expression in 6347 genes in three groups of mice: (a) old mice; (b) young mice; and (c) old mice that had their caloric intake restricted since weaning. Each group consisted of three mice. Using a criterion of a two-fold increase in gene expression in the group exhibiting higher expression, Lee et al. (1999) identified 113 genes that appeared to exhibit differential expression between groups *a* and *b*. The two-fold criteria is admittedly somewhat arbitrary and Lee et al. (1999) did not provide any formal statistical information regarding significance levels, confidence intervals around effect size estimates, or related statistics. This is not surprising since detailed discussions of how to approach such data from a statistical point of view are notably absent from the literature and the approach used by Lee et al. (1999) appears to be state of the art (see White et al. (1999) for a similar example taking a similar statistical approach).

Investigators, of course, may wish to answer the question "Is the difference in expression for the *such-and-such* gene statistically significant?" However, there are number of other questions that are at least equally important and interesting including: (Q1) is there statistically significant evidence that *any* of the genes under study exhibit a difference in expression across the groups?; (Q2) what is the best estimate of the number of genes for which there is a true difference in gene expression?; (Q3) what is the confidence interval around that estimate?; (Q4) if we set some threshold above which we declare results for a particular gene '*interesting'* and worthy of follow-up study, what proportion of those genes are likely to be genes for which

[1] "Expressed sequence tag (EST): A unique stretch of *DNA* within a coding region of a *gene* that is useful for identifying full-length genes and serves as a landmark for *mapping*. An EST is a *sequence tagged site* (*STS*) derived from *cDNA*." according to *Medterms Dictionary* (http://www.medterms.com/).

there is a real difference in expression and what proportion are likely to be false leads?; (Q5) what proportion of those genes not declared '*interesting*' are likely to be genes for which there is a real difference in expression (i.e., misses or false negatives)?

A key challenge to the development of statistical methods for microarray data is the fact that the sample size (e.g., the number of mice) is often small but the number of measurements per item (the number of genes) is very large. The expression levels for an individual gene or EST may not be independent. If they were, statistical models could be developed that model gene expression levels as independent measurements. The absence of such models limits one's ability to answer many important questions regarding the distribution of differential gene expression levels across two or more groups.

The purpose of this paper is to present methods for addressing questions above. We begin with only two very general assumptions and another specific assumption that we later relax. The two general assumptions are: (1) For each gene (EST), the measurements of gene expression have a finite population mean and variance; (2) For each gene under study, there is a measure of expression available for each case and this measure has sufficient reliability and validity to be useful (we use the word 'case' generically to refer to any organism or tissue on which expression is measured). The extent to which any particular measure of gene expression meets the second criterion is an important question, but beyond the scope of this paper. The more specific assumption that we later relax is that differences in gene expression levels across two groups are independent. This allows the development of a statistical model that will facilitate answers to many of the above questions (and perhaps additional questions). After presenting the methodology, we then illustrate it using a data example. A simulation study evaluates the performance of the methods, assesses the effect that "dependence" has on the interpretation of results from the model, and provides a mechanism to allow for non-independence. Finally, we provide a discussion and description of future work.

2. Methods

Suppose that $N = 2n$ cases are randomly divided into two groups of size n.[2] We assume equal numbers in each group to remain consistent with the data example (Section 3) and simulation (Section 4), but it is not required. Suppose that the investigator uses some statistical test to produce a p-value for testing the null hypothesis H_0 there is no difference in gene expression between the two groups for the ith gene, $i = 1, \ldots, k$. Ordinary frequentist significance testing and estimation procedures will provide reasonable answers to all of the key questions framed in the

[2] We discuss two groups for simplicity without loss of generality. However, the rationale can be applied to any number of groups by extension to ANOVA and other tests for differences among multiple groups. Similarly, the rationale can be extended to allow for tests of genes whose expression changes over time as long as one can construct a test for changes over time (i.e., a paired t-test if there are only two time points) that yields valid p-values at the level of the individual gene.

introduction only in the unrealistic event that power is maintained at nearly perfect (e.g., 0.99) levels for the smallest effect of interest (e.g., 1% of the variance) and the experimentwise type 1 error rate is maintained at some low level (e.g., 0.05) overall. However, as stated above, the number of simultaneous tests can be large with microarray data. For example, if $k = 6347$ as in Lee et al. (1999), the per test α required is ~ 0.0000081 to keep the overall experiment-wide type I error to be 0.05. In order to have $\sim 99\%$ power to detect a group difference of 0.2 standard deviations (or 1% of the variance) in gene expression at this α level, would require approximately 2400 cases per group. Such a large number of cases is currently unrealistic for microarray data. We therefore propose a set of alternative procedures designed to circumvent this problem and provide answers to questions of interest. This set of procedures is based on the idea that when *many* statistical tests are conducted, one obtains a distribution of test statistics and corresponding p-values and that there is information available in this distribution that can be exploited (Thomas et al., 1985; Devlin and Roeder, 1999; Drigalenko and Elston, 1997; Parker and Rothenberg, 1988).

2.1. A mixture-model approach

Assuming independence of gene expression levels across genes, under the null hypothesis, the distribution of p-values is uniform on the interval [0,1] regardless of the statistical test used (as long as that test is valid) and regardless of the sample size (Donahue, 1999). In contrast, under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one (Sackrowitz and Samuel-Cahn, 1999). Thus, by referring to the entire distribution of p-values obtained in the sample, we can answer the question (Q1) by conducting an omnibus test of whether the observed distribution of p-values is significantly different from a uniform distribution. A useful way in the present context is through the use of finite mixture models (Titterington, 1990). Other approaches such as the Kolmogorov–Smirnov test (Conover, 1999, Chapter 6) are also possible but do not address the additional questions of interest as effectively as does the mixture model approach.

Parker and Rothenberg (1998) point out that any distribution on the interval [0,1] can be modeled as a mixture of V separate component distributions where the jth component ($j = 1$ to V) is a beta distribution with parameters r_j and s_j. The probability density function (PDF) for the beta distribution is

$$\beta(r,s)(x) = I_{(0,1)}(x)\frac{x^{r-1}(1-x)^{s-1}}{\mathrm{B}(r,s)},$$

where $\mathrm{B}(r,s) = \int_0^1 u^{r-1}(1-u)^{s-1}\,\mathrm{d}u$ (Evans et al., 1993). In subsequent notation, we drop the "$I_{(0,1)}(x)$" and it is taken as given that x takes the range [0,1] exclusively. The beta distribution is chosen because of its great flexibility in modeling any shaped distribution on the interval [0,1]; a uniform distribution a special form of the beta distribution when $r = s = 1$. The log of the likelihood for the collection of k p-values from a model with $v+1$ components can then be expressed as

$$L_{V+1}=\sum_{i=1}^{k}\ln\left[\lambda_0\beta(1,1)(x_i)+\sum_{j=1}^{v}\lambda_j\beta(r_j,s_j)(x_i)\right],$$

where x_i is the p-value for the ith test, λ_0 is the probability that a randomly chosen test from the collection of tests is for a gene for which there is no population difference in gene expression (i.e., a test of a true null hypothesis), and λ_j is the probability that a randomly chosen test from the collection of tests is for a gene from the jth component distribution for which there is a true population difference in gene expression (i.e., a test of a false null hypothesis). The sum from 1 to k in the above log-likelihood expression results from our independence assumption. Without this assumption, the expression would be intractable. If there is statistically significant evidence for there being more than the one uniform component, that is, if for some model with v greater than zero, the accompanying λ_j are not all zero, then the global null hypothesis can be rejected and one can conclude that there is statistically significant evidence that one or more of the genes under study exhibit a difference in expression across the groups.

For any given model with v components, maximum likelihood estimates (MLEs) of the parameters λ_j, r_j, and s_j, can be obtained by iteratively finding those values that maximize the log-likelihood expression above subject to the constraint that $1=\lambda_0+\sum_{j=1}^{v}\lambda_j$ and $0\leqslant\lambda_j\leqslant 1$ for all λ_j. The log-likelihood evaluated at the MLEs using a model with v components beyond the uniform distribution ($v=1,2,\ldots$) will be denoted by L_v. The fit of models with v components can be compared to the fit of models with $v-1$ components by means of a statistic Q, where $Q=2(L_v-L_{v-1})$. Unfortunately, in the case of mixture modeling, the statistic Q cannot be assumed to be distributed as χ^2 with 3 df under the null hypothesis as might be expected (Parker and Rothenberg, 1988; Schork, 1992; Schork et al., 1996). However, Schork (1992) and others have shown that the so-called "parametric bootstrap" can provide valid significance tests and confidence intervals in the mixture model context.

2.2. Testing for the number of components

In constructing a bootstrap approach to significance testing, it is necessary to take two things into account. First, the expectation of the statistic Q is not known under the null hypothesis (Schork et al., 1996; McLachlan, 1987). Second, the magnitude of gene expression at various genes may not be (indeed seems unlikely to be) independent. The first issue necessitates use of some variant of the so-called "parametric bootstrap" (Schork, 1992; McLachlan, 1987; Chernick, 1999) in which W bootstrap samples are generated from a distribution under a presumed null hypothesis and from these W samples, one derives a critical value of the test statistics as that value of Q corresponding to the $(1-\alpha)$ W order statistic of the collection of W values of Q.

The second issue is that of the potential non-independence of expression across genes. This implies that the p-values to which the mixture models will be fitted will not necessarily be independent. As Horowitz (in press) points out, bootstrap inference can be conducted with dependent data provided that the bootstrap samples are generated by a process that preserves the dependency in the data. The typical approach to bootstrapping in the mixture model context would take bootstrap samples directly from the data to which the mixtures are fitted, in this case, the p-values. However, once the p-values are calculated and the data on individual cases put aside, the information about the dependency in the data is lost. An alternative would be to resample from n cases in each group (with replacement) so that the correlation structure of the data would be preserved in the resampling process. It is unclear, however, that resampling n cases in each group would accurately reflect the true sampling variability of parameter estimates from the mixture model, which would have been fitted with k observations. So we proceed with the former approach, that is, resampling from the distribution of k p-values and later, in a simulation study, we assess the sensitivity of calculated results to various levels of dependence among measures of gene expression.

The following procedure is proposed for significance testing of a model with v components compared to a model with $v-1$ components.

1. Fit the models with $v-1$ components and v components to the data and calculate the statistic Q (begin with $v=1$, and note that a model with 0 components is a uniform distribution).
2. Use parameter estimates from a model with $v-1$ components to create a parametric mixture model. This is an assumed model under the null hypothesis.
3. Create W bootstrap samples by selecting k observations from the model from step 2.
4. For each of the W bootstrap samples, fit the models with v components and $v-1$ components to the sample, and calculate the statistic Q_w ($w=1$ to W).
5. Define the critical value, Q_{crit}, as the $(1-\alpha)W$ order statistic of Q_w.
6. If the observed value of Q exceeds Q_{crit}, then one can reject the null hypothesis at level α.

From this procedure, one can determine the gain in the likelihood associated with adding one or more beta components, and a p-value of the hypothesis test, H_0: there are $v-1$ beta components versus H_{a}: there are v components, can be calculated.

Comparing a model with $v=1$ to a reduced (null) model with $\lambda_1=0$, allows one to answer the first question (Q1). That is, if the null hypothesis that $\lambda_1=0$ is rejected, then there is statistically significant evidence that the expression of one or more genes does differ between the groups. One can then successively compare models with greater values of v eventually selecting the “best” model. Once the best (and simplest) model is selected, the remaining questions (Q2)–(Q5) can be answered.

2.3. Interpreting the mixture model

The best estimate of the number of genes for which there is a true difference in gene expression is simply $k(1-\hat{\lambda}_0)$, where $\hat{\lambda}_0$ is the maximum likelihood estimate of λ_0. A $100(1-\alpha)\%$ confidence interval can be placed around $\hat{\lambda}_0$ by usual bootstrap methods (cf. Efron, 1982) and/or the standard error of $\hat{\lambda}_0$ can be estimated via the bootstrap or by numerically obtaining the Fisher information matrix (Lehman, 1991, Section 2.7).

Other questions can also be considered. If we set some threshold (T) below which results (e.g., p-values for differences in gene expression) for particular genes are declared '*interesting*' and worthy of follow-up study, the proportion of those genes that are likely to be genes for which there is a real difference in expression can be written as

$$P(\bar{H}_{0,i} \mid x_i \leqslant T) = 1 - P(H_{0,i} \mid x_i \leqslant T) = 1 - \frac{P(H_{0,i} \cap x_i \leqslant T)}{P(x_i \leqslant T)},$$

where $H_{0,i}$ denotes that the null hypothesis is true for the ith gene and $\bar{H}_{0,i}$ denotes that the null hypothesis is not true for the ith gene,

$$P(x_i \leqslant T) = \lambda_0 T + \sum_{j=1}^{v} \lambda_j \int_0^T \frac{x^{r_j - 1}(1-x)^{s_j - 1}}{\mathrm{B}(r,s)}\, \mathrm{d}x$$

and $P(H_{0,i} \cap x_i \leqslant T) = \lambda_0 T$. The estimated proportion of genes declared interesting that are likely to be false leads is simply

$$P(H_{0,i} \mid x_i \leqslant T) = \frac{P(H_{0,i} \cap x_i \leqslant T)}{P(x_i \leqslant T)}.$$

Similarly, the proportion of those genes not declared '*interesting*' that are likely to be genes for which there is a real difference in expression (i.e., misses or false negatives) can be written as

$$P(\bar{H}_{0,i} \mid x_i \geqslant T) = 1 - P(H_{0,i} \mid x_i \geqslant T) = 1 - \frac{P(H_{0,i} \cap x_i \geqslant T)}{P(x_i \geqslant T)},$$

where

$$P(x_i \geqslant T) = \lambda_0(1 - T) + \sum_{j=1}^{v} \lambda_j \int_T^1 \frac{x^{r_j - 1}(1-x)^{s_j - 1}}{\mathrm{B}(r,s)}\, \mathrm{d}x$$

and $P(H_{0,i} \cap x_i \geqslant T) = \lambda_0(1 - T)$.

Once again, confidence intervals around these estimates can be calculated using a bootstrap technique. The value of T can be adjusted up or down to achieve a balance of minimizing the false positive rate (i.e., maximizing specificity) and minimizing the false negative rate (i.e., maximizing sensitivity) in the manner of receiver operating characteristic (ROC) methodology (Hanley, 1989).

2.4. Remarks on small samples

Very small sample sizes present two potential complications. First, parametric statistical tests of the differences between the mean levels of gene expression for each of the genes will be more sensitive to assumed distributional forms of the expression data, and resulting p-values may not be accurate. This could be resolved by use of an appropriate non-parametric test at the first stage when the differences between the mean levels of gene expression for each of the genes are tested. Here we would again recommend use of a bootstrap test rather than a permutation test or traditional non-parametric test such as the Mann–Whitney U or Kruskal–Wallace tests because, unlike these latter tests, the bootstrap need not assume homogeneity of variance (Good, 1999) and is therefore less restrictive. If one chooses the bootstrap as an alternative method to "non-parametrically" produce the distribution of p-values, a second complication arises when resampling from very few cases, that is, the maximum number of different bootstrap samples is only $W_{\max} = [(2n-1)!/n!(n-1)!]^2$ (Horowitz, in press). If n is very small (e.g., $n < 5$), p-values will be affected by the discreteness of the bootstrapped distribution and there will be a limited number of "distinct" p-values among the k reported p-values. The resulting mixture model that would be fitted to such data might be unreliable since the mixture model approach attempts to fit a continuous model to the data. An alternative could employ a smoothed bootstrap (Chernick, 1999) to produce the k p-values. This is a subject for further research.

3. An illustrative example

To illustrate the methods that we are proposing, we analyze data described by Lee et al. (2000). Two groups of mice were considered: each group contains three mice. [3] A distribution of 6347 t-distribution based p-values was obtained. Each p-value was obtained from the test, for a specified gene, H_0: there is no difference in mean gene expression versus H_a: there is a difference in mean gene expression.

A uniform distribution ($v = 0$), a mixture of a uniform and one beta distribution ($v = 1$) and a mixture of a uniform and two beta distributions were fitted to the distribution of p-values using numerical routines in S-Plus. The graphical display is shown in Fig. 1. Fig. 1 suggests that p-values are clustering toward smaller values than would be expected if H_0 was true. The two mixture models using the beta distributions (a uniform plus 1 beta vs. a uniform plus 2 betas) do not appear very different. The next step is to determine some best model to represent the distribution of p-values, keeping in mind that the primary purpose is to model the peak near zero. To proceed with a statistical test, 6347 p-values were generated from a uniform distribution on the interval 0 to 1 and a mixture model with a uniform component and one beta distribution was fit to the simulated data and L_1 was recorded. This was done 500 times. In 389 of the 500 simulations, the algorithm correctly identified the uniform distribution, that is, it estimated λ_0 to be one. In the remaining 111 simulations, a beta distribution component was identified that was very close to being a uniform distribution (i.e., r_1 and s_1 were close to one). The maximum value of L_1 from the 500 simulations was 8.8. Thus, noting that L_0 is equal to zero, the

[3] We use t-tests for illustration, but any test producing a valid p-value can be used. The selection of the test may depend on distributional assumptions and/or sample size.

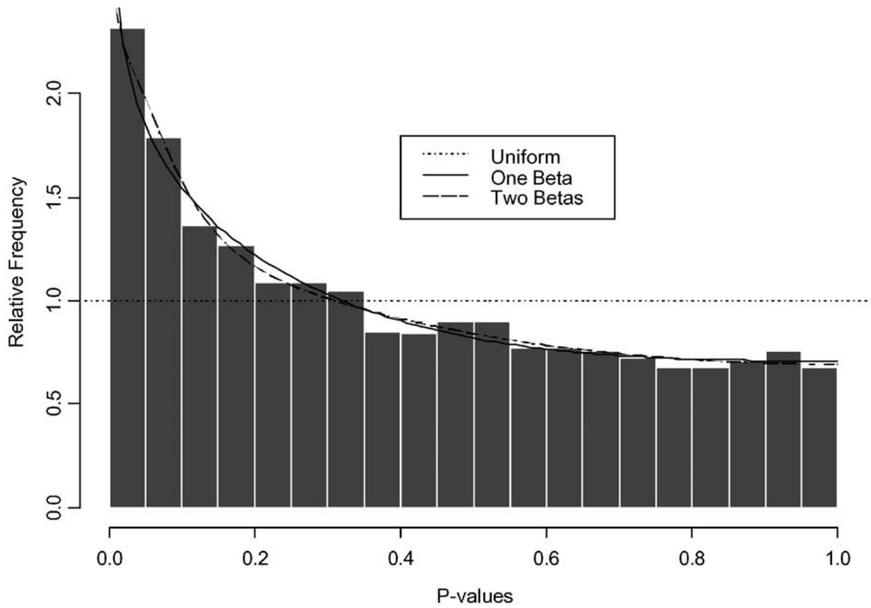

Fig. 1.

critical statistic Q_{crit} would be estimated less than $2(8.8) = 17.6$. The value of the maximum log-likelihood, L_1, from fitting the mixture model to the actual data was 454.2 so the observed value of Q would be twice 454.2 or 908.4 since, again, $L_0 = 0$. An estimated p-value for the test of H_0: the distribution is uniform versus H_a: the distribution is not uniform, would be < 0.005. This p-value is "estimated" since all possible bootstrap samples were not taken.

Of course, it must be recognized that the critical value used in the example was based on simulations under the null hypothesis when all genes were independent. In a small number of preliminary simulations under the null hypothesis with data from genes with different degrees of dependency (data not shown) the critical values necessary for significance were, not surprisingly, higher. Thus, to derive more appropriate critical values with applied data, it might be preferable to generate the initial simulated critical values from a dataset with the same covariance structure as the observed data. This is something currently under investigation, but the computational burden of estimating the 6347×6347 covariance matrix among the genes and conducting simulations therewith is challenging. However, in simulations, to be described in the next section, with relatively strong dependencies in the block diagonal matrices (i.e., pairwise correlations equal to 0.8), we did not observe a Q-statistic larger than 666 (with $n = 3$), with most values being well below 150. In contrast, the cortex data yield a Q statistic of 908.4 suggesting that the significance of the omnibus test is not in doubt. Moreover, we computed the $(6347(6347 - 1))/2$ pairwise correlations in the actual data after centering the group means around zero. The average correlation was 0.001, the average squared correlation was 0.266, and the average absolute correlation was 0.440. We then computed the same statistics on

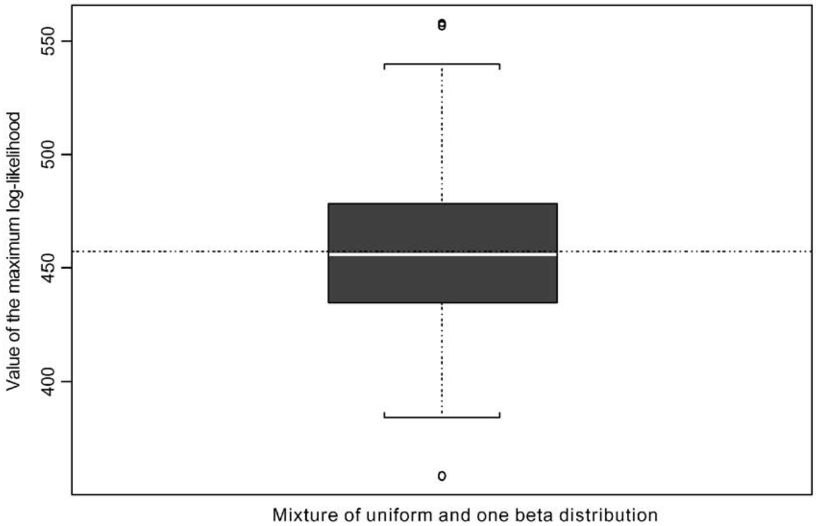

Fig. 2.

1000 samples of simulated data from 6 mice with 6347 genes for which expression values were simulated from a multivariate normal distribution with the covariance matrix being a 6347×6347 identity matrix; that is with no dependence. Across all simulated data sets, the average correlation was 0.000, the average squared correlation was 0.227, and the average absolute correlation was 0.383. This indicates that our observed data appears rather like data for which there is little dependency once again suggesting that the significance of the omnibus test is not in doubt.

As another illustration, we obtained a simulated sampling distribution of L_1 by drawing 500 bootstrapped samples from the original distribution of p-value. For each sample we fit the mixture model with one uniform and one beta distribution and computed L_1. This distribution is shown in Fig. 2. The horizontal dotted line at 458 is the value of L_2, the maximum log-likelihood when fitting a uniform and “two” beta distributions to the actual data. Fig. 2 suggests that fitting a second beta distribution does little to increase the value of the likelihood since the value of 458 is well within the sampling distribution of L_1. A p-value of the test of H_0: one beta distribution suffices versus H_a: two beta distributions are needed can be approximated by calculating how many values of the simulated sampling distribution fall equal to or above 458, the value of the maximum log-likelihood when using two beta distributions. This p-value is 0.476. We conclude that a mixture of a uniform and one beta distribution is the best model for the distribution of the 6347 p-values. We cannot say that more than one beta distribution in addition to the uniform will never be required to adequately fit a dataset, but in our experience with multiple datasets, we have yet to need more that one beta beyond the uniform.

Fitting this model results in MLEs for λ_0, r_1, and s_1 of 0.712, 0.775, and 3.862, respectively. The bootstrap was then implemented by resampling from the distribution of p-values. Simulated sampling distributions of statistics of interest are shown in Fig. 3. These distributions represent 500 statistics obtained from the bootstrap

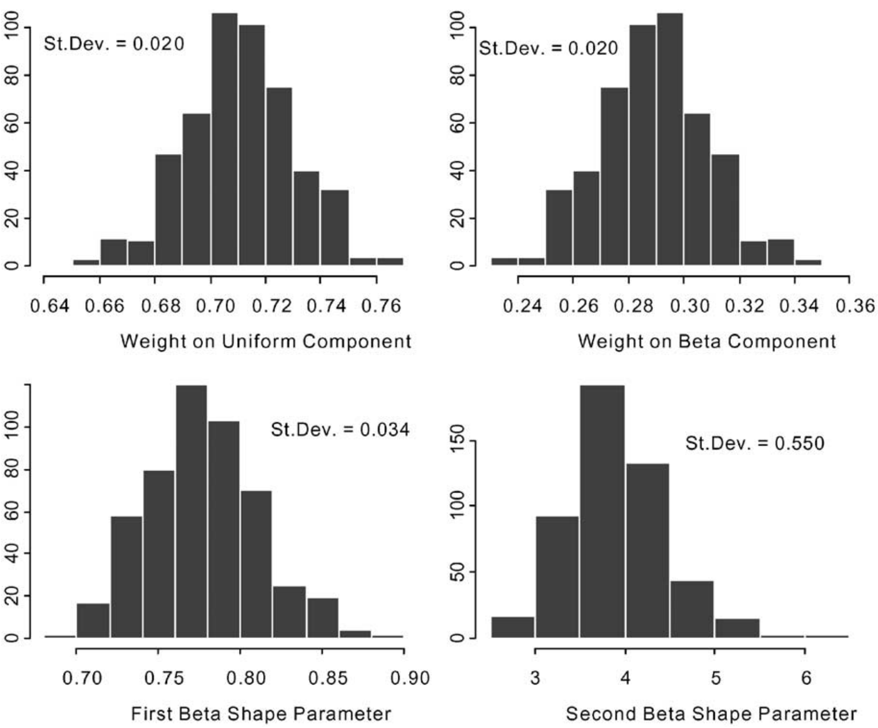

Fig. 3.

procedure. The standard deviations of the distribution could be used as standard errors of the estimators from the mixture distribution. Since the sampling distributions appear fairly symmetric, approximate 95% confidence intervals could be obtained from the original estimate plus/minus two standard deviations. Taking this approach, an approximate 95% confidence interval for λ_0 is (0.673, 0.751), for r_1 it is (0.708, 0.842), and for s_1 it is (2.784, 4.940). Any function of parameters from the mixture distribution can be estimated using the corresponding statistics that are calculated for each bootstrap sample. The validity of point estimates and confidence intervals will depend on the effect that correlation among gene expression levels might have on results. This is investigated in the next section.

Given these parameter estimates, our best estimate for the number of genes that have a real difference in mean expression across the two groups is, therefore, $6347(1-0.712)=1828$. Suppose we believed p-values from the distribution of p-values that are less than 0.10 are interesting and worthy of follow-up (the p-values are from a 2-tailed test). The estimated proportion of these genes that are likely to be false leads is $(0.712^{*}0.10)/[0.712^{*}0.10+0.288\,\beta(0.10,0.775,3.862)]$, where $\beta(a,r,s)$ is the cumulative beta distribution with parameters r and s, evaluated at a. This proportion is 0.356. That is, there is about a 36% chance that any randomly selected gene with an ordinary p value less than 0.10 will be a gene for which there is no real difference. The proportion not declared interesting (using an ordinary p-value of 0.10 as the cutoff for interesting) that are likely to be genes for which there is a true

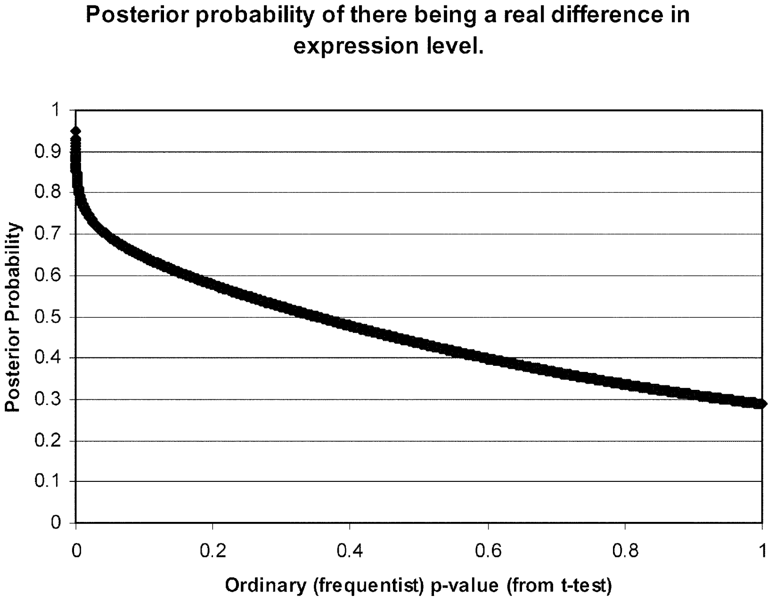

Fig. 4.

significant difference in expression is 0.199. Furthermore, the mean of a beta distribution is $r/(r+s)$, which in the current example is $0.775/(0.775+3.862)=0.167$. This implies that even among genes for which there is a real difference in expression, we only expect the p-values to be about 0.17 on average. This indicates that, not surprisingly, given the sample size, power is very low with an n of 3 per group and conventional significance testing with an alpha level of 0.05 or smaller would lead to many false negatives (i.e., misses).

Fig. 4 displays the posterior probabilities of the 6347 genes along with their corresponding p-values. As can be seen, as long as the p-value is smaller than about 0.35, there is more than a 50% chance that the gene is a gene for which there is a real difference in expression. Thus, if one used the unconventionally large alpha level of 0.35 as indicative of genes for which one guessed there was a real effect, one would be correct more often than not in this case.

Finally, we highlight just a few specific genes and show how the mixture model can guide our inferences. The gene identified by accession number L06451 was more highly expressed among restricted animals and had an ordinary (frequentist) p-value of 0.06 and posterior probability of 0.68. Thus, although this gene would not be "significant" at the conventional 0.05 alpha level, there is still a 68% chance that it is a gene with a real difference in expression induced by caloric restriction (CR). This gene encodes a protein that is homologous to agouti signaling protein (Miller et al., 1993). This gene product is believed to be involved in body weight and appetite regulation and it is therefore quite plausible that it is affected by CR. As another example, consider the gene identified by accession number M74180. This gene had, by Affymetrix's definition, a "fold-change" of 2.7 (increased expression among CR animals) which, while large, was hardly the largest of those reported by Lee et al (2000). Nevertheless, this gene had the highest posterior probability (0.95) meaning that there is a 95% chance (in the Bayesian sense; Savage, 1951) that there is a true difference in gene expression for this gene. This gene encodes a protein that is homologous to mouse hepatocyte growth factor-like or macrophage stimulating protein (MSP) (Degen et al., 1991). Finally, consider the gene identified by accession number W75705 which codes a protein that is over 80% homologous to mouse cyclophilin (Hasel and Sutcliffe, 1990). This gene had a "fold-change" of 3.3 (by Affymetrix's definition; increased expression with CR) which many investigators would consider clearly significant (Glynne et al., 2000). Nevertheless, we estimate the posterior probability for that gene to be only 0.44 indicating that it is *at least* equally reasonable to guess that, based on these data, this gene is not differentially expressed as a function of CR. The reason that our posterior probabilities do not 'agree' (i.e., have a 1 : 1 correspondence) with the fold-change metric is that the latter does not take the within group variability in gene expression into account. That is, fold-change is a measure of magnitude effect (and not necessarily an optimal one) and not a direct measure of strength of evidence for an effect. These examples illustrate how the mixture model can guide the interpretation of the overall suite of group gene expression differences as well as differences for individual gene expression levels.

4. Simulation study

To remain consistent with the previous example we use the term "mouse" to refer to a case or an experimental unit. We generated gene expression levels for $2n$ mice (n mice per experimental group, where $n = 5$, 10, 20, and 40), and $k = 3000$ genes. The data for the $2n$ mice are multivariate normal and generated independently from a 3000 dimensional normal distribution. That is, measurements for a mouse are generated from

$$X \sim N_{3000}(\underline{\mu}, \Sigma),$$

where μ was a constant vector of length 3000 and equal to 10 and $\Sigma = \sigma^2 \mathbf{B} \otimes \mathbf{I}_6$, $\mathbf{B} = \mathbf{1}_{500}\mathbf{1}'_{500}\rho + (1-\rho)\mathbf{I}_{500}, \mathbf{1}_{500} = (1, 1, \ldots, 1)'$ with length 500, and $\mathbf{I}_m$ is the $m \times m$ identity matrix. For the simulations, the common variance was $\sigma^2 = 4$. We varied ρ over the three values of 0 (independence), 0.4 (moderate dependence), 0.8 (strong dependence). The covariance structure seems plausible since groups of genes are likely to be co-expressed but it is unlikely that a particular gene expression is correlated with ALL other genes. In fact, empirical studies of resulting sample correlation matrices from simulated data suggested that even $\rho = 0.4$ tended to produce higher correlations among gene expressions than were present in the actual example data set. We only considered positive values of ρ though a negative correlation could also be plausible. Finally, for 20% of the genes (600 randomly selected genes), a true mean difference in expression between the two groups of n mice was incorporated by adding d to the gene expression measurements of mice $(n+1)$ through $2n$.

Table 1
Simulation results for 9 cases as described in Section 4. Group size = 5. 500 Simulation were conducted for each case. $\bar{\hat{\lambda}}_1$ is the mean of 500 estimates of λ_1, $S(\hat{\lambda}_1)$ is the standard deviation, and $\overline{S^*}(\hat{\lambda}_1)$ is the mean of 500 estimates of standard deviation obtained from 100 bootstrapped samples. The last column are percentiles from the 500 maximum log-likelihood values

$n=5$ per group	λ_1	$\bar{\hat{\lambda}}_1$	$S(\hat{\lambda}_1)$	$\overline{S^*}(\hat{\lambda}_1)$	$L_1(5,50,95)$th percentiles
Case 1: $\rho=0.0$, $d=0$	0.0	0.147	0.335	0.285	(0, 0, 2.8)
Case 2: $\rho=0.0$, $d=2$	0.2	0.147	0.092	0.087	(69.2, 87.8, 107.4)
Case 3: $\rho=0.0$, $d=4$	0.2	0.178	0.008	0.024	(592, 644, 703)
Case 4: $\rho=0.4$, $d=0$	0.0	0.210	0.353	0.127	(0, 0, 150.1)
Case 5: $\rho=0.4$, $d=2$	0.2	0.143	0.118	0.072	(0, 75, 328)
Case 6: $\rho=0.4$, $d=4$	0.2	0.206	0.089	0.038	(393, 641, 1020)
Case 7: $\rho=0.8$, $d=0$	0.0	0.206	0.312	0.039	(0, 0, 452)
Case 8: $\rho=0.8$, $d=2$	0.2	0.212	0.218	0.042	(0, 84, 766)
Case 9: $\rho=0.8$, $d=4$	0.2	0.313	0.206	0.035	(244, 662, 1626)

So when $d > 0$, the true mean difference in gene expression levels for 20% of the genes is equal to d, and it is zero for the other 80% of the genes. If $d=0$, then there is no difference in mean gene expression levels across the two groups for *any* of the 3000 genes. The ability of the mixture model method to detect a distribution of p-values different from a uniform will depend, of course, on the magnitude of d, on n, and on ρ. Our focus with the simulation was to assess the performance of the mixture model for modeling the distribution of p-value and to assess the meaningful interpretation of $\hat{\lambda}_1=(1-\hat{\lambda}_0)$, the estimated proportion of genes for which there is a true difference in expression. The following cases were considered: Case 1, $\rho=0.0$ and $d=0$; Case 2: $\rho=0.0$ and $d=2$; Case 3: $\rho=0.0$ and $d=4$; Case 4: $\rho=0.4$ and $d=0$; Case 5: $\rho=0.4$ and $d=2$; Case 6: $\rho=0.4$ and $d=4$; Case 7: $\rho=0.8$ and $d=0$; Case 8: $\rho=0.8$ and $d=2$; Case 9: $\rho=0.8$ and $d=4$. For each case, a group size of $n=5$, 10, 20, and 40 were used (Tables 1–4, respectively). In cases 1, 4, and 7, the true λ_1 is equal to 0, but in all other cases it is equal to 0.2. We investigated the mixture model procedure by generating 500 sets of gene expression data, fitting a mixture model to the distribution of p-values using a uniform and one beta component, and estimating model parameters. We then recorded the mean of the 500 estimated λ_1 ($\bar{\hat{\lambda}}_1$ in the tables) for each of the nine cases described above and for each group size. We also recorded the standard deviation of the 500 values ($S(\hat{\lambda}_1)$ in the tables), and the 5th, 50th, and 95th percentiles of the value of the log-likelihood, L_1. Then, as a second check of the bootstrap method

Table 2
As for Table 1 except that the group size is 10

$n=10$ per group	λ_1	$\bar{\hat{\lambda}}_1$	$S(\hat{\lambda}_1)$	$\overline{S^*}(\hat{\lambda}_1)$	$L_1(5,50,95)$th percentiles
Case 1: $\rho=0.0,\ d=0$	0.0	0.153	0.338	0.277	(0, 0, 2.7)
Case 2: $\rho=0.0,\ d=2$	0.2	0.164	0.033	0.045	(303, 348, 394)
Case 3: $\rho=0.0,\ d=4$	0.2	0.195	0.002	0.013	(2182, 2271, 2362)
Case 4: $\rho=0.4,\ d=0$	0.0	0.219	0.364	0.127	(0, 0, 112.4)
Case 5: $\rho=0.4,\ d=2$	0.2	0.195	0.048	0.105	(143, 345, 798)
Case 6: $\rho=0.4,\ d=4$	0.2	0.233	0.088	0.025	(1798, 2261, 2842)
Case 7: $\rho=0.8,\ d=0$	0.0	0.253	0.346	0.055	(0, 0, 451)
Case 8: $\rho=0.8,\ d=2$	0.2	0.271	0.213	0.039	(0, 329, 1193)
Case 9: $\rho=0.8,\ d=4$	0.2	0.300	0.132	0.018	(1541, 2290, 3194)

of evaluating standard errors, we bootstrapped the resulting 3000 p-values that were obtained from each of the 500 sets of gene expression data. For each set of 3000 p-values, we took 100 bootstrapped samples and, for each bootstrap sample, fitted the mixture model. We then computed the bootstrap standard deviation, $S^*(\hat{\lambda}_1)$. We recorded the mean of the 500 $S^*(\hat{\lambda}_1)$ for each of the nine cases ($\overline{S^*}(\hat{\lambda}_1)$ in the Tables).

The simulation results (Tables 1–4) provide some insight into the performance of the mixture model approach. In Case 1, the distribution of L_1 is very near zero suggesting that the distribution of p-values is nearly uniform. This is true for all group sizes. However, $\hat{\lambda}_1$ is not near zero since the model often fit a uniform component mixed with a beta distribution that was nearly uniform. The high variability in $\hat{\lambda}_1$ can be seen in $S(\hat{\lambda}_1)$. However, Cases 2 and 3 show that for fixed n and as d becomes larger, values of $\hat{\lambda}_1$ approach the true 0.2 and $S(\hat{\lambda}_1)$ becomes smaller. Also, for fixed $d>0$, values of $\hat{\lambda}_1$ approach the true 0.2 and $S(\hat{\lambda}_1)$ becomes smaller as n becomes larger. Cases 1, 2, and 3 are differentiated by the location of the distribution of L_1.

Cases 4–6, and Cases 7–9 show a similar pattern except that the effect of $\rho>0$ is apparent. However, this effect is less pronounced for larger n. For small n,d must be sufficiently large for the mixture model to detect a set of genes differentially expressed between two groups. For example, when $n=5$ and $\rho=0.4$, $\hat{\lambda}_1$ is close to the true 0.2 with low standard deviation when $d=4$, but not when $d=2$. However, when $n=20$ and $\rho=0.4$, $\hat{\lambda}_1$ is close to the true 0.2 with low standard deviation when d is only 2. When $n=10$, 20 or 40, Cases 4–6 are differentiated by the location of the distribution of L_1. Cases 7–9 are differentiated when $n=20$ or 40. When $n=5$, there is some overlap in the distribution of L_1 for Cases 4–6 and Cases 7–9. When

Table 3
As for Table 1 except group size is 20

$n=20$ per group	λ_1	$\bar{\hat{\lambda}}_1$	$S(\hat{\lambda}_1)$	$\overline{S^*}(\hat{\lambda}_1)$	$L_1(5,50,95)$th percentiles
Case 1: $\rho=0.0$, $d=0$	0.0	0.149	0.336	0.276	(0, 0, 2.8)
Case 2: $\rho=0.0$, $d=2$	0.2	0.178	0.014	0.026	(1116, 1197, 1283)
Case 3: $\rho=0.0$, $d=4$	0.2	0.208	0.008	0.009	(5873, 6004, 6156)
Case 4: $\rho=0.4$, $d=0$	0.0	0.194	0.339	0.132	(0, 0, 110.6)
Case 5: $\rho=0.4$, $d=2$	0.2	0.220	0.096	0.033	(777,1186,1800)
Case 6: $\rho=0.4$, $d=4$	0.2	0.234	0.051	0.015	(5087, 5997, 6878)
Case 7: $\rho=0.8$, $d=0$	0.0	0.238	0.332	0.046	(0, 0, 604)
Case 8: $\rho=0.8$, $d=2$	0.2	0.309	0.166	0.027	(608, 1225, 2255)
Case 9: $\rho=0.8$, $d=4$	0.2	0.263	0.085	0.012	(4724, 5988, 7839)

Table 4
As for Table 1 except group size is 40

$n=40$ per group	λ_1	$\bar{\hat{\lambda}}_1$	$S(\hat{\lambda}_1)$	$\overline{S^*}(\hat{\lambda}_1)$	$L_1(5,50,95)$th percentiles
Case 1: $\rho=0.0$, $d=0$	0.0	0.137	0.327	0.273	(0, 0, 2.6)
Case 2: $\rho=0.0$, $d=2$	0.2	0.199	0.002	0.013	(3335, 3467, 3604)
Case 3: $\rho=0.0$, $d=4$	0.2	0.204	0.001	0.007	(13476, 13638, 13810)
Case 4: $\rho=0.4$, $d=0$	0.0	0.224	0.365	0.136	(0,0,102)
Case 5: $\rho=0.4$, $d=2$	0.2	0.231	0.068	0.020	(2679, 3429, 4317)
Case 6: $\rho=0.4$, $d=4$	0.2	0.214	0.025	0.009	(12615,13584,14645)
Case 7: $\rho=0.8$, $d=0$	0.0	0.233	0.335	0.042	(0, 0, 486)
Case 8: $\rho=0.8$, $d=2$	0.2	0.273	0.106	0.014	(2284, 3449, 4856)
Case 9: $\rho=0.8$, $d=4$	0.2	0.222	0.042	0.009	(12295, 13749, 15075)

$n=10$, there is overlap for Cases 7–9. These results suggest that as n increases, the mixture model will detect a difference in mean gene expression between two groups even when there is strong dependence among gene expression measurements.

The bootstrap estimates of standard deviation ($\overline{S^*}(\hat{\lambda}_1)$) appear to estimate the standard deviation of $\hat{\lambda}_1$ when $\rho=0$, but underestimate it when $\rho>0$. This is sensible since the bootstrap operates on a fixed distribution of p-values, and variation due to correlation among gene expression is not accounted for. This issue, also discussed in the next section, may be addressed by modifying the bootstrap so that correlation among gene expression is maintained in the resampling procedure.

The mixture model of interest to this procedure is one that models a peak of the p-value distribution that is close to zero. This would imply that the mean of the beta component is less than $\frac{1}{2}$. This condition was difficult to assess and control in the automated simulations that we conducted. In an actual data analysis, one has the added advantage of visually interpreting the distribution of p-values, as we had in the earlier data example (Fig. 1).

5. Issues, future work, and general discussion

In the present paper, we have developed a set of procedures for analyzing microarray gene expression data that are intended to not only take into account but indeed to capitalize on the fact that many thousands of genes may be studied. This set of procedures should lead to the ability to answer important questions about group or condition differences in gene expression, can be adapted to allow for non-normality and heteroscedasticity, and may be used with small sample sizes. Nevertheless, it must be emphasized that though these procedures are usable even with small sample sizes, this does not justify their use when sample sizes are very small. Indeed, when very small samples are used, it is likely that for any given threshold for "interestingness" selected either the proportion of misses or the proportion of false positives will be undesirably high. In such cases, it may be desirable to use two thresholds where genes with scores below one threshold are "ruled out" as uninteresting, genes above the higher threshold are declared interesting, and genes between the two thresholds are declared indeterminate. This may help to prioritize efforts for future research.

In some simulations that we conducted and in analyses of other microarray gene expression data sets, we noticed that when there were differences in gene expression across two groups, a single beta distribution beyond the uniform captured this difference adequately and that a second beta distribution was not needed. A second beta distribution was only significant in the mixture model under a certain type of gene expression data. This occurred when we simulated gene expression data with a high correlation parameter, i.e., $\rho = 0.9$ (not reported in this paper), and when $d > 0$. This high dependency between gene expression levels created a bimodal distribution of p-values. The first beta distribution modeled the peak near zero, and the second beta distribution modeled the second peak that occurred, typically, between 0.5 and 1.0. For the 80% of genes that were generated with no difference in mean expression between the two groups, the dependency among these genes created a cluster that resulted in the second peak. Furthermore, this second beta distribution took much of the weight off of the uniform component, and estimates of λ_0 are not interpretable. This indicates that caution must be exercised if using this procedure when the distribution of p-values is clearly multi-modal, or if there is only one mode but that mode is *not* on the left side (nearer to zero) of the distribution. In such cases, the beta distribution is not modeling the genes for which there is a significant difference in expression. This could be controlled somewhat by restricting the fitting algorithm to only use beta component(s) with a mean less than 0.5

Much work remains to be done in this area. Future research should evaluate these procedures under a variety of circumstances including different types of non-normality and other types of correlation among gene expression patterns such as negative correlation among groups of genes. In such endeavors, robust modeling with t-distributions in place of the beta distributions may be useful to consider (Peel and McLachlan, 2000). The simulation code that we used was written using the statistical package S-Plus, and it can be easily adapted to simulate more complex dependency relationships between genes and different values of d.

The simulation study focused on the validity of the mixture modeling approach. Further simulations would need to be performed to evaluate other procedures based on the model, such as estimating the number of false positives in gene expression data.

It is likely that the type 1 error rate and the power of a test of a difference in mean gene expression levels for a single gene are related to the ability of the mixture model to detect a departure from the null hypothesis that the distribution of p-values is not uniform. This should be investigated (analytically and/or through simulation) under varying distributional assumptions. This could lead to a method to determine a recommended sample size (number of cases per group) for which the mixture model method would be reliable in estimating the number of genes for which there is a true difference in mean expression levels. The simulation results suggest that the mixture model method improves as sample size increases, even in the presence of moderate correlation among gene expression.

We simulated a sampling distribution of mixture model parameter estimates by bootstrapping from the empirical distribution of p-values. In the simulation, we estimated standard errors of mixture model parameter estimates using this bootstrap procedure. This method of resampling is valid when the p-values are treated as independent observations. An alternative procedure would be to resample cases (i.e., mice) and recompute the p-values for each bootstrapped sample, and then fit the mixture model. This would preserve the dependency among gene expression patterns. It is unclear if this method would induce the appropriate and valid sampling variability in parameter estimates in the mixture model. This is a topic of ongoing research. In addition, this alternative resampling approach will require modifications when the sample size (i.e., number of mice) is very small since there would be a limited number of unique resamples.

A popular approach to the analysis of microarray gene expression data is to cluster the genes into subsets of co-expressed genes. Using such approaches might yield clusters of genes that, across clusters, are largely independent. It might then be possible to obtain p-values for the effect of the independent variable (e.g., age) on gene expression at the level of the cluster. Such p-values might be more likely to be independent than p-values obtained at the level of the individual gene, Future research might evaluate whether this can alleviate some of the challenges posed by non-independent data.

Finally, in the absence of near-perfect power, ordinary estimates of effect size calculated only for the statistically significant results will be biased toward showing a greater difference than really exists even when those estimates are maximum likelihood estimates and asymptotically unbiased when considered across all results regardless of significance. This bias occurs because one is estimating effect sizes only for those results that are statistically significant and has been described elsewhere (Beavis, 1998; Thomas et al., 1985). This bias may be reduced and estimates with smaller loss functions produced by use of Empirical Bayes techniques (Morris, 1983; Samaniego and Vestrup, 1999) could be considered. Incorporating this into the mixture model approach is a topic of current research.

Acknowledgements

This research was supported in part by NIH grants R29DK47256, R01DK51716, P30DK26687, P01AG11915, R01ES09912 and 5 U24DK58776. We are grateful to Drs. Nicholas Schork, Joel Horowitz, and Michael C. Neale for their helpful comments.

References

Beavis, W.D., 1998. QTL analysis: power, precision, and accuracy. In: Paterson, A.H. (Ed.), Molecular Dissection of Complex Traits. CRC Press, FL, pp. 145–173.

Chernick, M.R., 1999. Bootstrap Methods. Wiley Series in Probability and Statistics. Wiley, New York.

Conover, W.J., 1999. Practical Nonparametric Statistics, Wiley Series in Probability and Statistics, 3rd Edition. Wiley, New York.

Degen, S.J., Stuart, L.A., Han, S., Jamison, C.S., 1991. Characterization of the mouse cDNA and gene coding for a hepatocyte growth factor-like protein: expression during development. Biochemistry 30 (40), 9781–9791.

Devlin, B., Roeder, K., 1999. Genomic control for association studies. Biometrics 55, 997–1004.

Donahue, R.M.J., 1999. A note on information seldom reported via the p value. Amer. Statist. 53, 303–306.

Drigalenko, E.I., Elston, R.C., 1997. False discoveries in genome scanning. Genet. Epidemiol. 14 (6), 779–784.

Efron, B., 1982. The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia.

Evans, M., Hastings, N., Peacock, B., 1993. Statistical Distributions, A Wiley-Interscience Publication. Wiley, New York, USA.

Glynne, R.J., Ghandour, G., Goodnow, C.C., 2000. Genomic-scale gene expression analysis of lymphocyte growth, tolerance and malignancy. Curr. Opinion Immunol. 12, 210–214.

Good, P.I., 1999. Resampling Methods. A Practical Guide to Data Analysis. Birkhauser, Boston.

Hanley, J.A., 1989. Receiver operating characteristic (ROC) methodology: the state of the art. Crit. Rev. Diagnostic Imaging 29, 307–335.

Hasel, K.W., Sutcliffe, J.G., 1990. Nucleotide sequence of a cDNA coding for mouse cyclophilin. Nucleic Acids Res. 18 (13), 4019.

Horowitz, J.L. The bootstrap. In: Handbook of Econometrics, Vol. 5. Elsevier, New York, in press.

Kahn, J., Saal, L.H., Bittner, M.L., Chen, Y., Trent, J.M., Meltzer, P.S., 1999. Expression profiling in cancer using cDNA microarrays. Electrophoresis 20, 223–229.

Lee, C., Klopp, R.G., Weindruch, R., Prolla, T.A., 1999. Gene expression profile of aging and its retardation by caloric restriction. Science 285, 1390–1393.

Lee, C., Weindruch, R., Prolla, T., 2000. Gene-expression profile of the ageing brain in mice. Nature Genetics 25, 294–297.

Lehman, E.L., 1991. The Theory of Point Estimation. Wadsworth & Brooks/Cole.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl. Statist. 36, 318–324.

Miller, M.W., Duhl, D.M., Vrieling, H., Cordes, S.P., Ollmann, M.M., Winkes, B.M., Barsh, G.S., 1993. Cloning of the mouse agouti gene predicts a secreted protein ubiquitously expressed in mice carrying the lethal yellow mutation. Genes Dev. 7 (3), 454–467.

Morris, C.N., 1983. Parametric empirical Bayes inference: theory and applications. J. Amer. Statist. Assoc. 78, 47–55.

Parker, R.A., Rothenberg, R.B., 1988. Identifying important results from multiple statistical tests. Statist. Med. 7, 1031–1043.

Peel, D., McLachlan, G.J., 2000. Robust mixture modelling using the t distribution. Statist. Comput. 10, 339–348.

Sackrowitz, H., Samuel-Cahn, E.P., 1999. Values as random variables—expected p values. Amer. Statist. 53 (4), 326–331.

Samaniego, F.J., Vestrup, E., 1999. On improving standard estimators via linear empirical Bayes methods. Statist. Probab. Lett. 44, 309–318.

Savage, L.J., 1951. The theory of statistical decision. J. Amer. Statist. Assoc. 46, 57–67.

Schork, N.J., 1992. Bootstrapping likelihood ratios in quantitative genetics. In: LePage, R., & Billard, L. (Eds.), Exploring the Limits of Bootstrap. Wiley, New York, pp. 389–393.

Schork, N.J., Allison, D.B., Theil, B., 1996. Mixture distributions in human genetics research. Statist. Methods Med. Res. 5, 155–178.

Thomas, D.C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., Armstrong, B.G., 1985. The problem of multiple inference in studies designed to generate hypotheses. Amer. J. Epidemiol. 122, 1080–1095.

Titterington, D.M., 1990. Some recent research in the analysis of mixture distributions. Statistics 21 (4), 619–641.

Weindruch, R., Kayo, T., Lee, C., Prolla, T.A., 2001. Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. J Nutr. 131, 918S–923S.

White, K.P., Rifkin, S.A., Hurban, P., Hogness, D.S., 1999. Microarray analysis of Drosophila development during metamorphosis. Science 286, 2179–2184.