Statistica Sinica **12**(2002), 31-46

REPLICATED MICROARRAY DATA

Ingrid Lönnstedt and Terry Speed[†]

Uppsala University, [†] *University of California, Berkeley and*
[†] *Walter and Eliza Hall Institute*

Abstract: cDNA microarrays permit us to study the expression of thousands of genes simultaneously. They are now used in many different contexts to compare mRNA levels between two or more samples of cells. Microarray experiments typically give us expression measurements on a large number of genes, say 10,000-20,000, but with few, if any, replicates for each gene. Traditional methods using means and standard deviations to detect differential expression are not completely satisfactory in this context, and so a different approach seems desirable. In this paper we present an empirical Bayes method for analysing replicated microarray data. Data from all the genes in a replicate set of experiments are combined into estimates of parameters of a prior distribution. These parameter estimates are then combined at the gene level with means and standard deviations to form a statistic B which can be used to decide whether differential expression has occurred. The statistic B avoids the problems of using averages or t-statistics. The method is illustrated using data from an experiment comparing the expression of genes in the livers of SR-BI transgenic mice with that of the corresponding wild-type mice. In addition we present the results of a simulation study estimating the ROC curve of B and three other statistics for determining differential expression: the average and two simple modifications of the usual t-statistic. B was found to be the most powerful of the four, though the margin was not great. The data were simulated to resemble the SR-BI data.

Key words and phrases: cDNA microarray, differential expression, empirical Bayes, replication, ROC curve, t-statistic.

1. Introduction

cDNA microarrays are used to compare gene expression in different samples of cells. They permit us to study the expression levels of thousands of genes simultaneously. The technique has a wide range of applications including learning how genes interact, which genes are used in different cell types, and which genes change their expression in cells due to disease or drug stimuli. Microarray experiments typically result in expression measurements from a large number of genes (usually 10,000-20,000), but with few if any replicates for each gene (usually 1-10). Throughout this paper the data from one microarray are always a

comparison of the expression levels in two cell samples. For gene i on array j, we use the value M_{ij} where

$$M_{ij} = \log_2 \frac{(\text{expression level in sample 1})_{ij}}{(\text{expression level in sample 2})_{ij}}. \tag{1}$$

The numerator and denominator are often referred to as the red and green intensities, R_{ij} and G_{ij}, because of the experimental procedure described below. If there are N genes on each array, and n replicates (arrays), the complete set of data from the experiment consists of (M_{ij}), $i = 1, \ldots, N$ and $j = 1, \ldots, n$. Our task is to determine which genes have different expression levels in the two samples, i.e., which genes have M-values genuinely different from zero. This is often expected to be true for only a small proportion (say 1%) of the genes. The variation of the rest of the genes would then be due to chance.

Natural statistics which might be used to select the differentially expressed genes are the mean and standardized mean expression levels, $(M_{i.})$ and $(t_i) = (M_{i.}/SE_i)$, where $SE_i = s_i/\sqrt{n}$ is the standard error of $M_{i.}$, s_i being the standard deviation of the values $M_{ij}, j = 1, \ldots, n$.

There are problems with both of these traditional statistics. For example, a large mean might be driven by an outlier, something which occurs quite frequently in this context. A large t might arise because of a small denominator SE, even though the mean itself is small. Because of the large number of genes on each array, there will usually be genes with very small standard errors, and some of these will have small means as well. A simple solution to this problem is to discount genes with a small absolute mean whose standard errors are in the bottom 1%, say. A more sophisticated statistic for use in this context has recently been presented (Tusher, Tibshirani and Chu (2001)), slightly tuning the t-statistic by adding a suitable constant to each SE. Here we introduce another alternative based on the empirical Bayes approach. Data from all the genes in a replicate set of experiments are combined into estimates of parameters of a prior distribution. These parameter estimates are then combined at the gene level with means and standard deviations to form a statistic B which is a Bayes log posterior odds. B can then be used to determine if differential expression has occurred. It avoids the problems of the average M and the t-statistic just mentioned. We also carry out a simulation comparison of the four different methods, relating the power to detect differentially expressed genes to the false positive rate in an ROC (receiver operating characteristic) curve.

The paper is organized as follows. Sections 1.1-1.2 describe the scientific background and procedure of the microarray experiments. In Section 2 the four different statistics are described, and they are illustrated in Section 4 using data introduced in Section 3. The comparison is based on simulated datasets presented

in Section 5 together with the analysis. Our findings are summarized in Section 6. The derivation of our statistic B is presented in the Appendix.

1.1. Background on cDNA microarrays

A cell contains a complete set of its host's genetic code (genes), stored in a DNA molecule. Depending on the function of the cell, it uses different genes, so that brain cells and liver cells express different genes. When a cell wants to use a gene, the code of that gene is copied into messenger RNA (mRNA) in a procedure called transcription. The mRNA then gets translated into a protein, which is the functional product of most genes. Transcription occurs all the time and for all the genes currently used by the cell, at different levels. Microarray experiments measure the concentration of mRNA floating around in a set of cells, and a high concentration of mRNA for a given gene reflects a high expression level of that gene. In practice, the levels of expression of a gene across two (or more) cell samples are always compared, not measured in absolute terms on a cDNA microarray.

1.2. Construction of a microarray

A string of mRNA is a single-stranded copy of the DNA sequence for a gene. It is possible to construct a complementary or cDNA copy of an mRNA molecule and further experimentation with the mRNA occurs via the cDNA copy.

A microarray is a glass slide on which thousands of spots of cDNA representing different genes are printed using a robotic arrayer. The arrayer has a grid of pins, or print tips, which can pick up sets of samples from cDNA clones and print them onto the slide. Each spot will contain thousands of copies of the cDNA fragment from one gene. Normally all the spots hold cDNA from different genes, but sometimes replicate spots are printed for each gene on the same microarray.

The two cell samples to be compared for gene expression are often cells subject to some treatment and normal (non-treated) cells, tumour cells and normal tissue, or just two different kinds of tissue. For each of the two samples, the mRNA is isolated and each sequence is labeled with a fluorescent dye at the time of conversion to cDNA. Usually, the treatment (or tumour) cDNA is labeled red and the normal green.

By adding equal amounts of the two labeled cDNA samples to the microarray, the sample cDNA will *hybridize* to the cDNA spots on the glass slide, i.e., it will pair with its complementary fragments of cDNA on the slide. If a gene has a higher level of expression in the treatment sample than in the normal, there will be more red than green dye on its spot. The intensities of red and green dyes on each spot are detected using a laser scanner. The red and green intensities of the

scanned image are the measurements which are the starting point of a statistical analysis. These are combined into M-values as described by (1).

2. Statistics to be Compared

The data from one microarray experiment consist of (M_{ij}), $i = 1, \ldots, N$, $j = 1, \ldots, n$. We will assume that these are already normalized according to Yang, Dudoit, Luu and Speed (2001), that is, a smooth intensity-dependent adjustment is made to the log ratio value to remove red-green color bias. For each gene g, we ask whether its observed vector of M-values, $\boldsymbol{M}_g$, provides evidence to suggest that the true M-value of that gene is different from zero.

We compare four different methods of addressing this question. Each method assigns a score to each gene, and the putative differentially expressed genes are selected using a cutoff value for that score. For none of the methods is it obvious how to choose the cutoff in such a way as to control the type 1 error, or how to assign a p-value to a score. To do these things we would need to know the null distribution of our scores, and that is not so straightforward. Much of the time this is not a serious limitation. Often the genes that would be significant are just a few, very extreme ones, many of which are already well known to the investigators. Mostly what is wanted are a few more genes that are probably differentially expressed, and there is generally a willingness to permit several of these to be false positives in order to avoid missing too many of the true positives. A better method is the one with the lower type I and type II errors over a range of cutoffs.

The first statistic we consider is the average $(M_{g.})$ for each gene g. In practice we are interested in comparing the absolute value of the average with (positive) cutoffs. The average statistic will sometimes be referred to as M where it is obvious that we mean the statistic rather than the separate M-values for each gene and slide. M does not depend on the standard error of the genes, and hence treats a highly variable gene in the same way as a stable one.

By dividing each average by its corresponding standard error, giving $(t_g) = (M_{g.}/SE_g)$, variation across replicates can be accounted for in the usual way, though frequently the number of degrees of freedom is small. Because of the large number of genes included in microarray experiments, there will always be some genes with a very small sum of squares across replicates, so that their (absolute) t-values will be large whether or not their averages are large. Some of these will turn out to be false positives for the t-statistic. To reduce this problem in the comparison to come, we have chosen to ignore those genes whose standard errors fall in the bottom 1% if their absolute average value is smaller than 0.01.

Tusher *et al.* (2001) have proposed a refinement of t which avoids the difficulty just mentioned. By adding a constant term to the denominator of the standardized average, the denominator is prevented from getting too small. The factor, a_0, is suggested by Efron, Tibshirani, Goss and Chu (2000) to be equal to the 90th percentile of the standard errors of all the genes. This is used throughtout the current paper, although a different approach is suggested in Tusher, Tibshirani and Chu (2001). Hence we study the absolute value of $S_g = M_{g.}/(s_g + a_0)$ where s_g is the standard deviation for gene sample (M_{gj}), $j = 1, \ldots, n$.

Our empirical Bayes log posterior odds statistic is called B and is similar to S above, although the argument justifying it is more complex. We regard the (M_{ij}) as random variables from a normal distribution with mean (μ_i) and variance (σ_i^2), something which is found empirically to be roughly the case. More fully, we suppose that, independently for all i and k,

$$M_{ik}|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2). \tag{2}$$

Most genes have $\mu_i = 0$, but a small proportion p of genes have some $\mu_i \neq 0$, indicated by $I_i = 1$ as opposed to $I_i = 0$. The parameters (μ_i, σ_i^2) are treated as i.i.d. realizations of random parameters with some prior distributions. We calculate the log posterior odds for gene g to be differentially expressed, $B_g = \log \frac{Pr(I_g=1|(M_{ij}))}{Pr(I_g=0|(M_{ij}))}$. By assuming conjugate prior distributions for the variances of the genes, as well as for their means where not zero, an explicit formula for B is found to be

$$B_g = \log \frac{p}{1-p} \frac{1}{\sqrt{1+nc}} \left[\frac{a + s_g^2 + M_{g.}^2}{a + s_g^2 + \frac{M_{g.}^2}{1+nc}} \right]^{\nu + \frac{n}{2}}. \tag{3}$$

Here a and ν are hyperparameters in the inverse gamma prior for the variances, and c is a hyperparameter in the normal prior of the nonzero means. For details see the Appendix (where we also include an application of B to replicated spots within as well as between microarray slides). In particular, s_g^2 is in this case the sum of squares over n, rather than over $n - 1$ as in t and S.

The only gene-specific part of B is the last ratio, which is always ≥ 1 since $1/(1 + nc) < 1$. We deduce that increasing relative gene expression (and hence increasing $M_{g.}^2$) increases B_g, and more so if the variance is small. If $M_{g.}^2$ is small too, a ensures that the ratio cannot be expanded by a very small variance. B will be illustrated further below.

3. Data

For the illustrations and simulations in this paper we restrict ourselves to one dataset, the experiment comparing gene expression in liver cells from scavenger receptor transgenic (SR-BI) mice to those of the corresponding wild-type control

mice, see Callow, Dudoit, Gong, Speed and Rubin (2000). Eight SR-BI mice are all compared to a reference sample of pooled cDNA from the control mice, resulting in 8 microarray data sets. After image processing and normalization as described in Yang et al. (2001) and Buckley (2000), our M-values consisted of the 8 sets of log ratios as in (1). There were 6,356 genes on each slide. The resulting data are displayed in an average M vs average A plot in Figure 1a, where for each gene and array the A-value is the log of the geometric mean of the two intensity channels ($A = 0.5(\log_2 R + \log_2 G)$). The M vs A plot is often used for scientific reasons when displaying and analysing microarray data. Figure 1b displays the average M-values vs the log sample variances, to be compared to simulated data further below.

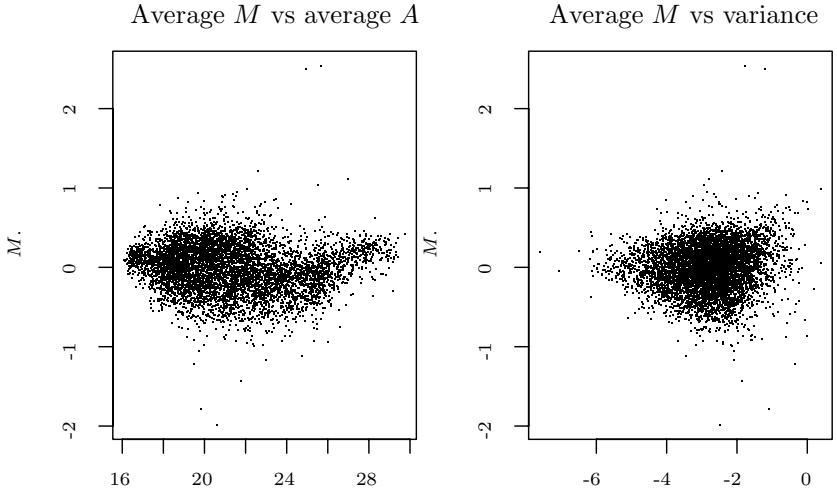

Figure 1. The first plot is the $M_{\cdot}$ vs $A_{\cdot}$ plot of the SR-BI data, showing the average M-value vs the average mean log dye intensity for each gene. The second plot displays $M_{\cdot}$ vs the *log variance*.

4. Illustration of the Statistics

Below we illustrate the function of B and how it differs from M and t. For illustrations of S we refer to Tusher, Tibshirani and Chu (2001) and Efron, Tibshirani, Goss and Chu (2000).

The shape of a plot of (B) vs ($M_{i\cdot}$) is usually parabolic, as we see for SR-BI in Figure 2. Most genes have an average value around zero, and these have the minimum values of B. The larger the average expression level of a gene, the larger is the chance of a high B. The actual B-level of genes with high averages depends on the variance, so that the outer ends of the parabola will be rather fuzzy. Two genes with exactly the same average might be far apart in B.

Ideally we would like to be able to say that genes with $B > 0$ have a significantly changed expression, but since we cannot be sure about p a priori, or estimate it, this cutoff cannot be relied upon. The threshold has to be judged on a case by case basis. Figure 2 shows that the rationale for our method seems valid. We now take a closer look at the details and assess the number of false negatives and false positives in $M_{i\cdot}$ and $t_i = M_{i\cdot}/SE_i$.

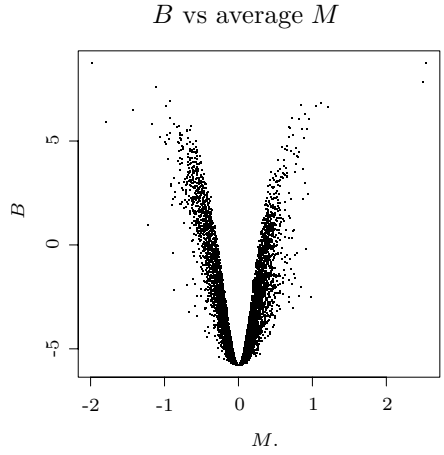

Figure 2. B vs $M_{\cdot}$ plot for SR-BI. B is our proposed statistic, the log posterior odds for each gene to be differentially expressed. It depends on the average as well as on the standard error for each gene.

The statistics M, t and B are illustrated for the SR-BI data in Figure 1, where we have labelled sets of genes as in Table 1.

Genes are selected as extreme for M if $|M_{i\cdot}| > 0.5$, for t if $|t_i| > 4.5$ and for B if $B_i > -0.5$. These cutoffs are chosen so that there will be several genes

Table 1. Sets of genes. A 1 indicates that the genes in the set are extreme for that statistic.

Set	M	t	B
S1	0	0	1
S2	0	1	0
S3	0	1	1
S4	1	0	0
S5	1	0	1
S6	1	1	0
S7	1	1	1

in each set (except for S6) in order to illustrate the ideas. If our purpose is to identify differentially expressed genes with reasonable certainty, the cutoffs would probably need to be a bit larger.

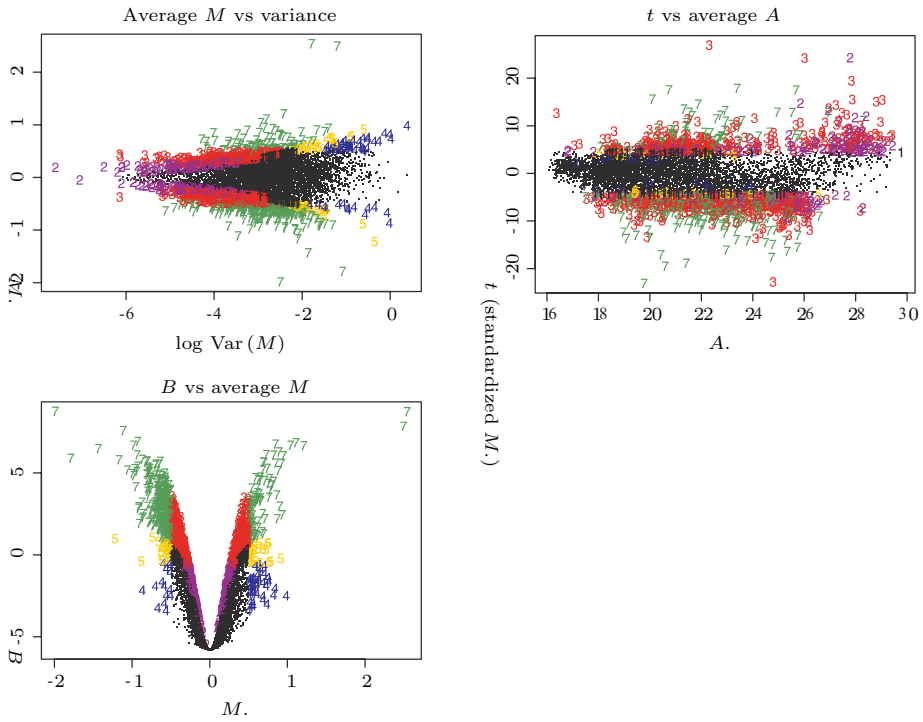

Figure 3. SR-BI for different statistics. The three plots are of average M vs the log variance, standardized average t vs the mean intensity, and the log posterior odds B vs the average expression level. The different sets of genes display the dependencies between the three statistics as well as the variance. 1 means extreme in B ($B > -0.5$) only; 2 means extreme in t ($|t| > 4.5$) only; 3 means extreme in B and t only; 4 means extreme in M only ($|M.| > 0.5$); 5 means extreme in M and B only; 6 means extreme in M and t only; 7 means extreme in all thee statistics M, t and B.

Compare the locations of the different sets of genes in the plots in Figure 3. The plots display the average M vs the variance (in fact the logarithm of the sum of squares), t vs A (the average intensity), and B vs the average M again. S7 is the set of genes which are high for all the three statistics, and this shows clearly in the graphs. The set S3 is also clearly high for B due to the small variances (large t's), although the means are only moderately enhanced. These genes would not have been selected by M, being false negatives there, but are

readily detected with B. S5 are genes that are extreme in M and B, but not in t. This makes sense when we note that their t-values are actually rather large, probably just below our cutoff. Also, these are genes just above the borderline in B, but there are not that many of them. Similarily, S1 are just above the cutoff for B but neither for M nor for t. Both S1 and S5 are what we call false negatives in t : they would not have been detected with t, but they are assumed to have a true large (but not extreme) variance and a large (but not too large) average. There are not always any genes in these sets, but there sometimes are, and they should be selected. S2 are false positives for t: they have a small average but are driven by tiny variances. It is reassuring to see that these are not extreme in B. Finally, S4 is a large set of genes with a high average but with standard errors that are too large for the result to be trusted. They are false positives for M, and are properly downweighted in B. Note that the absence of genes in S6 confirms that genes high in both M and t are also high in B. It follows that by using B we have avoided the main problems with false positives and false negatives in M and t.

5. Methods

One hundred different datasets were simulated and analysed for differentially expressed genes using each of the four different methods presented in Section 2. The aim is to compare the type I and type II errors of the methods without having to involve p-values, for the reasons mentioned above.

5.1. Simulation of datasets

The SR-BI data was used as a model for the simulated dataset, so that they all contained $n = 8$ replicates for each of $N = 6,356$ genes. The genes were modeled as independent, see the Discussion. Given the parameters, the replicates for each gene i were produced as independent observations from a distribution $N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, N$, as in (2). For the variances, an inverse gamma prior distribution was used, common to all genes. For most genes, the expectation μ was fixed at zero, whereas for a proportion p of the genes, a normal prior distribution was used instead, to produce truly differentially expressed genes. See (6) (Appendix) for details on the prior distributions. The hyperparameters were estimated from the SR-BI dataset according to the procedure in A.2 (Appendix), giving $\nu = 2.8$, $a = 0.040$ and $c = 1.2$. The proportion p of differentially expressed genes was fixed at $p = 0.01$ throughout the simulations, as well as in the following analysis. An example of a simulated dataset is shown in Figure 4, where the genes with a true non-zero expectation are highlighted.

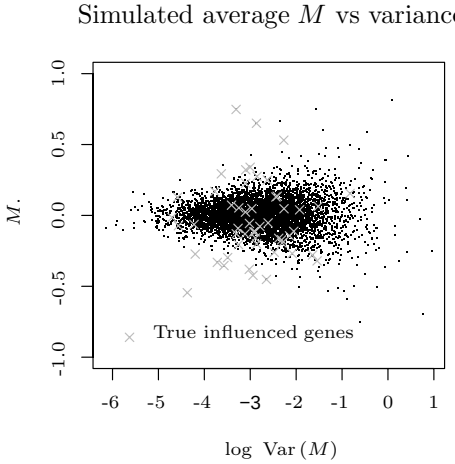

Figure 4. This is the $M_\cdot$ vs *log variance* plot corresponding to Figure 1b for one of the 100 simulated datasets. Highlighted genes are truly differentially expressed, i.e., they were simulated from a normal distribution with a non-zero mean.

5.2. Analysis

For each gene in each of the 100 simulated datasets, the four statistics M, t, S and B (Section 2) were calculated. For B the estimation method in the Appendix was used separately for each of the datasets, ignoring the known parameters used in the simulations. For a range of cutoff values for each statistic, the numbers of false positive and false negative genes could be deduced for each of the 100 datasets. The observed numbers of false positive and false negative genes were then averaged over the datasets for each cutoff value of each statistic. If $c_M = (c_1, \ldots, c_{n_M})$ denotes the vector of cutoffs for the statistic M, the results for M is summarized by the vectors $r_M^+ = (r_1^+, \ldots, r_{n_M}^+)$ and $r_M^- = (r_1^-, \ldots, r_{n_M}^-)$, containing the average numbers of false positive and negative genes respectively. Similarly for t, S and B.

5.3. Results

The cutoff vectors were chosen to give reasonable ranges in the results. For each statistic, the number of false positives ranges from 0 to approximately 200 (out of the 6,356 genes) and the number of false negatives ranges from 30 to 60. In the model used to simulate the datasets, the expectations of the activated genes are normally distributed around zero. This implies that many of these genes will have a negligible average, as do the false genes. We will not be able to detect them using any method, unless we allow the number of false positive genes to be unrealisticly large. Thus the range of the numbers of false negatives starts around 30 rather than 0.

The average numbers of false positive and false negative genes for all four statistics are plotted against one another in Figure 5. This Reciever Operating Characteristic (ROC) curve allows us to compare the type I and type II errors for the statistics without involving p-values. The better of two methods is the one with lower scores on both axes.

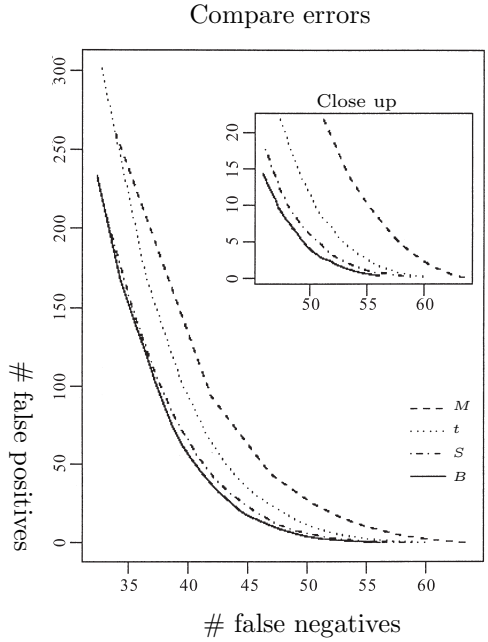

Figure 5. Comparison of the four different statistics M, t, S and B for the 100 simulated datasets. For a certain cutoff value, each method defines the numbers of false negative and false positive genes in each of the simulated datasets. The lines reflect the averages of these numbers over a range of cutoffs.

It turns out that B is the best of the four statistics in the sense that it minimizes the two error rates in our analysis. However, it is not very different from S. If we allow the number of false positives to exceed 200, the statistic S is preferred to B, but following such a large number of genes up is hardly intresting in practice. On the other hand, M and t are clearly found to have a larger error rate than B and S.

6. Discussion and Conclusions

The problem of identifying differentially expressed genes using data from replicated microarray experiments has been addressed. In contrast to the approaches of Roberts et al. [10] and Ideker et al. [7], both of whom have explicit

error-models for their data, we impose a lighter modelling structure on our observations. Our method uses information from all the genes in a series of experiments to estimate a posterior odds score for each gene, indicating whether a gene has changed its expression level. The resulting statistic B deals properly with difficulties met by well-known statistics, such as the average or the standardized average t. It is easy to use and computationally cheap.

We now need to make some remarks about our modelling assumptions and philosophy. Although the discussion has been framed in either Bayesian or classical testing terms, we present our analysis primarily as a way of ranking genes. The normality and independence assumptions we make are not meant to be taken literally, rather, they are to be regarded as devices leading to a formula which we believe improves upon the average and the t statistic in this context. While the normality assumption for log ratios is probably a good first approximation, it would not be wise to base formal inference on it. We do not see our discussion as doing so because we make no probabilistic claims for our procedure. Assuming independence of genes is clearly unrealistic. In reality, the unknown dependence structure will include near complete dependence between essentially duplicate genes, through varying degrees of dependence among genes which are biologically related, to total independence. None of this implies that a statistic such as B derived using independence assumptions is without value. We hope that the way in which B is seen to overcome real problems with the average and t support our claims.

One drawback in using B is that we need a value for the prior proportion of differentially expressed genes (see the Appendix). Although the rankings of genes by their B values change only marginally for genes with large B when the parameter estimates are altered, the actual scale changes. Thus we cannot rely on any standard cutoff value, such as $B = 0$, for the selection of differentially expressed genes. However, the same difficulty arises when we use more traditional statistics such as average M or t.

We have found B is an improvement over the average and t methods. This conclusion is based on data simulated from the same model that we used to derive our statistic B. Since this model describes the data produced in our microarray experiments reasonably well, giving only a small proportion p of genes with non-zero expectation, we are not too worried it works in favour for B. However, there are two points that deserve comment. First, since the simulation model does not explicitly produce outliers in the data, the simulated data will contain fewer such genes than real data. It follows that the gap between the average M and the other methods can be expected to be even greater in practice than what we see in Figure 5. Second, it is possible that the fact that p is the same in the simulations and the analysis might improve B slightly relative to the other statistics. To investigate this we carried out the analysis with $p = 0.005$ as well

as $p = 0.05$, although the data sets were simulated with $p = 0.01$. The range of cutoffs we had to choose for B in these new cases differed from those used initially, but in the plot corresponding to Figure 5, the three lines (analysed using $p = 0.005$, $p = 0.01$ and $p = 0.05$ respectively) were hardly separable by eye. Each of them was best compared to the other three statistics. Note that this observation also suggests that the statistic B is not very dependent on the choice of the parameter p, as long as we can choose the cutoff to suit the p-specific results. In a way, we can regard B as providing a ranking of the genes with respect to the posterior probability of each gene being differentially expressed. A suitable cutoff for the detection of these genes can then be chosen by the ranking in combination with experimental preferences, e.g., so that it selects the top 50, 100 or 150 genes. The number of genes selected can depend on the size, aim, background and follow-up plans of the experiment.

The log posterior odds method has a large potential for extensions beyond the application to spots within and between microarray slides (see (8) in the Appendix). It can be modified to apply to linear models across experiments, including several cell samples. ANOVA could also be considered, for example, in comparing the effects of different treatments.

The software producing these analyses is available at http://cran.r-project.org/src/contrib/PACKAGES.html#sma.

Acknowledgements

Our thanks go to the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), to David Freedman for many valuable discussions, and to Tom Britton.

A. The Statistic B

A.1. Posterior odds B of differential expression

Let N be the number of genes in a microarray experiment, n the number of replicates for each gene (the number of microarray slides), and $M_{ij} = \log R_{ij} - \log G_{ij}$, $i = 1, \ldots, N$ and $j = 1, \ldots, n$, the log ratios of our green and red intensities for each gene. We regard the M_{ij} as random variables from a normal distribution with mean μ_i and variance σ_i^2, so that, independently and identically,

$$M_{ij}|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2) \qquad \text{for all } i. \tag{4}$$

Let I indicate whether the gene g is differentially expressed ($\mu_g \neq 0$). For each gene g we are interested in $Pr(I_g = 1|(M_{ij}))$, equivalently, in the log odds B for this, $B_g = \log \frac{Pr(I_g=1|(M_{ij}))}{Pr(I_g=0|(M_{ij}))}$. Thus $P\left(I_g = 1|(M_{ij})\right) > P\left(I_g = 0|(M_{ij})\right)$ if

and only if $B_g > 0$. (The parameters (μ_i, σ_i^2) are treated as i.i.d. realizations of random parameters with some prior distributions.)

Now, by Bayes' Theorem and independence across genes,

$$\begin{aligned} B_g &= \log \frac{p}{1-p} \frac{Pr((M_{ij})|I_g = 1)}{Pr((M_{ij})|I_g = 0)} \\ &= \log \frac{p}{1-p} \frac{Pr(\boldsymbol{M_g}|I_g = 1)}{Pr(\boldsymbol{M_g}|I_g = 0)} \frac{\prod_{i \neq g} Pr(\boldsymbol{M_i}|I_g = 1)}{\prod_{i \neq g} Pr(\boldsymbol{M_i}|I_g = 0)} \\ &= \log \frac{p}{1-p} \frac{Pr(\boldsymbol{M_g}|I_g = 1)}{Pr(\boldsymbol{M_g}|I_g = 0)}, \end{aligned} \tag{5}$$

where $\boldsymbol{M_g}$ is the vector of the n measurements for gene g and p is the proportion of differentially expressed genes in the experiment, $p = Pr(I_i = 1)$, for any i in $1, \ldots, N$. We need to calculate $f_{I_i=1}(\boldsymbol{M_i})$ and $f_{I_i=0}(\boldsymbol{M_i})$.

We usually have very few replicates for each gene (sometimes $n = 2$), but we are investigating many genes simultaneously (e.g., $N = 10,000$). To use all our knowledge about the means and variances we collect the information gained from the complete set of genes in estimated joint prior distributions for them. We let the prior distribution of $1/\sigma_i^2$ be gamma, and that of μ_i given σ_i^2 be normal. This is a conjugate prior distribution, allowing us to calculate B_i explicitly. See e.g., Hartigan (1983). For integer degrees of freedom ν and scale parameters $a > 0$, $c > 0$, we set $\tau_i = na/2\sigma_i^2$ and suppose that

$$\tau_i \sim \Gamma(\nu, 1), \tag{6a}$$

$$\mu_i|\tau_i = \begin{cases} 0 & \text{if } I_i = 0 \\ N(0, cna/2\tau_i) & \text{if } I_i = 1, \end{cases} \tag{6b}$$

for all $i = 1, \ldots, N$. The parameter c expresses dependence between the priors for μ_i and τ_i and is necessary for the calculations. Our densities are then

$$\begin{aligned} f(\tau_i) &= \frac{1}{\Gamma(\nu)} \tau_i^{\nu-1} e^{-\tau_i}, \\ f_{I_i=1}(\mu_i|\tau_i) &= (2\pi)^{-1/2} c^{-1/2} \left(\frac{na}{2\tau_i}\right)^{-1/2} e^{-\frac{1}{2}\frac{2\tau_i}{cna}\mu_i^2}, \\ f_{I_i=0}(\mu_i) &= \delta(0), \\ f(\boldsymbol{M_i}|\mu_i, \tau_i) &= (2\pi)^{-n/2} \left(\frac{na}{2\tau_i}\right)^{-n/2} e^{-\frac{1}{2}\frac{2\tau_i}{na}\Sigma_j (M_{ij}-\mu_i)^2} \\ &= (2\pi)^{-n/2} \left(\frac{na}{2\tau_i}\right)^{-n/2} e^{-\frac{1}{2}\frac{2\tau_i}{na}\left(\Sigma_j (M_{ij}-M_{i.})^2 + n(M_{i.}-\mu)^2\right)}, \\ f_{I_i=1}(\boldsymbol{M_i}) &= \int\int f_{I_i=1}(\boldsymbol{M_i}, \mu_i, \tau_i)\, d\mu_i d\tau_i \end{aligned}$$

$$= \int\int f\left(\boldsymbol{M_i}|\mu_i,\tau_i\right) f_{I_i=1}\left(\mu_i|\tau_i\right) f\left(\tau_i\right) d\mu_i d\tau_i,$$

$$f_{I_i=0}\left(\boldsymbol{M_i}\right) = \int\int f\left(\boldsymbol{M_i}|\mu_i,\tau_i\right) f_{I_i=0}\left(\mu_i|\tau_i\right) f\left(\tau_i\right) d\mu_i d\tau_i$$

$$= \int f\left(\boldsymbol{M_i}|\mu_i=0,\tau_i\right) f\left(\tau_i\right) d\tau_i.$$

The integrations of the joint densities are performed by identifying the posterior normal distribution of $\mu_i|\tau_i$, $N(\frac{nc}{1+nc}M_{i.}, \frac{a}{2\tau_i}\frac{nc}{1+nc})$ for the case $I_i = 1$, and the posterior gamma distribution of τ_i, $\Gamma(\nu+\frac{n}{2}, 1+\frac{1}{a}(s_i^2+\frac{M_{i.}^2}{1+nc}))$, $s_i^2 = \frac{1}{n}\Sigma_j(M_{ij}-M_{i.})^2$, so that both of these are integrated out. The results are scaled t-statistics,

$$f_{I_i=1}(\boldsymbol{M_i}) = \frac{\Gamma(\nu+\frac{n}{2})}{\Gamma(\nu)}(2\pi)^{-\frac{n}{2}}\Big(\frac{na}{2}\Big)^{-\frac{n}{2}}(1+nc)^{-1/2}\Big[1+\frac{1}{a}(s_i^2+\frac{M_{i.}^2}{1+nc})\Big]^{-(\nu+\frac{n}{2})},$$

$$f_{I_i=0}(\boldsymbol{M_i}) = \frac{\Gamma(\nu+\frac{n}{2})}{\Gamma(\nu)}(2\pi)^{-\frac{n}{2}}\left(\frac{na}{2}\right)^{-\frac{n}{2}}\left[1+\frac{1}{a}(s_i^2+M_{i.}^2)\right]^{-(\nu+\frac{n}{2})}.$$

Hence for a gene g (from (5))

$$B_g = \log\frac{p}{1-p}\frac{1}{\sqrt{1+nc}}\Big[\frac{a+s_g^2+M_{g.}^2}{a+s_g^2+\frac{M_{g.}^2}{1+nc}}\Big]^{\nu+\frac{n}{2}}. \tag{7}$$

The only gene specific part of B is the last ratio, which is always ≥ 1 since $1/(1+nc) < 1$. We can deduce that an increasing differential expression (and hence an increasing $M_{g.}^2$) increases B_g, all the more if the variance is small. If $M_{g.}^2$ is small too, a ensures that the ratio cannot be expanded by a very small variance.

Expression (7) is referred to as B in this paper. It is possible to generalize (7) so that it is valid for microarray experiments where each gene has m replicated spots within each of n slides, i.e., where we have a variance component within slides and one between slides. The result is then

$$B_g' = \log\frac{p}{1-p}\frac{1}{\sqrt{1+mnc}}\Big[\frac{a+\frac{1}{mn}\left(SSB_g+kSSW_g\right)+M_{g..}^2}{a+\frac{1}{mn}\left(SSB_g+kSSW_g\right)+\frac{M_{g..}^2}{1+mnc}}\Big]^{\nu+\frac{mn}{2}}, \tag{8}$$

where $M_{g..}$ are the overall averages for each gene, SSB_g and SSW_g are the sums of squares between and within slides respectively, and k is an unavoidable extra global parameter reflecting the ratio of SSB to SSW.

A.2. Estimation of hyperparameters of B

There are four global parameters in the model for B: p, ν, a and c. Unfortunately, it is difficult to estimate (p,ν,a,c). Therefore, we fix p and estimate

$\nu, a|p$ and $c|p, \nu, a$. The parameters ν and a are such that $\tau_i = \frac{na}{2\sigma_i^2} \sim \Gamma(\nu, 1)$ for all i. We use the observed variance estimates $(\frac{1}{n-1}\Sigma_j(M_{ij} - M_{i\cdot})^2)$ to estimate ν and a by the method of moments.

The difficulty in estimating c is that it only occurs in the distribution of genes which are differentially expressed ($I_i = 1$), and we do not know which ones these are. We have chosen to compare the observed normal density of the averages $(M_{i\cdot})_{i\in T}$, where T is the given top proportion p of genes with respect to B, with the observed normal density of all averages $M_{i\cdot}$. This leads naturally to an estimate of c.

In the absence of a satisfactory estimate, p is fixed at some sensible value such as 0.01 or 0.001. The choice of p does not change the shape of the B vs $M_{\cdot}$-plot very much, but it does move it up and down the y-axis. A consequence of this is that we cannot fix a cutoff (e.g., $B = 0$, as suggested for (5)), to select all the genes with a higher score as differentially expressed.

References

Buckley, M. J. (2000). The Spot user's guide. CSIRO Mathematical and Information Sciences. http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.* **10**, 2022-2029.

Efron, B., Tibshirani, R., Goss, V. and Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical report, Stanfond University.

Hartigan, J. A. (1983). *Bayes Theory.* Springer-Verlag, New York.

Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. E. (2000). Testing for differentially expressed genes by maximal likelihood analysis of microarray data. *J. Computat. Biology* **7**, 805-817.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Computat. Graph. Statist.* **5**, 299-314.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C. and Friend, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873-880.

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

Wang, Y. W., Dudoit, S., Luu, P. and Speed, T. P. (2001). Normalization for cDNA microarray data. SPIE BiOS, San Jose, California.

Department of Mathematics, Uppsala University, Box 480 S-751 06 Uppsala, Sweden.

E-mail: ingrid@math.uu.se

Department of Statistics, University of California, Berkeley, U.S.A.

E-mail: terry@stat.berkeley.edu

Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Melbourne

E-mail: terry@wehi.edu.au

(Received July 2001; accepted October 2001)