

12

SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays

John D. Storey Robert Tibshirani

Abstract

SAM is a computer package for correlating gene expression with an outcome parameter such as treatment, survival time, or diagnostic class. It thresholds an appropriate test statistic and reports the q -value of each test based on a set of sample permutations. SAM works as a Microsoft Excel add-in and has additional features for fold-change thresholding and block permutations. Here, we explain how the SAM methodology works in the context of a general approach to detecting differential gene expression in DNA microarrays. Some recently developed methodology for estimating false discovery rates and q -values has been included in the SAM software, which we summarize here.

12.1 Introduction

In this chapter we discuss the problem of identifying differentially expressed genes from a set of microarray experiments. Statistically speaking, this task falls under the heading of “multiple hypothesis testing.” In other words, we must perform hypothesis tests on all genes simultaneously to determine whether each one is differentially expressed or not. Recall that in statistical hypothesis testing, we test a null hypothesis versus an alternative hypothesis. In this example, the null hypothesis is that there is no change in expression levels between experimental conditions. The alternative hypothesis is that there is some change. We reject the null hypothesis if there is enough evidence in favor of the alternative. This amounts to rejecting the null hypothesis if its corresponding statistic falls into some predetermined rejection region. Hypothesis testing is also concerned with measuring the probability of rejecting the null hypothesis when it is really true (called a false positive), and the probability of rejecting the null hypothesis when the alternative hypothesis is really true (called power).

There are four important steps one must take in testing for differential gene expression. The first is that a statistic must be formed for each gene. The choice of this statistic is important in that one wants to make sure that no relevant information is

lost with respect to the test of interest, yet all measurements on the gene are condensed into one number. The second step is to calculate the null distribution(s) for the statistics. One can assume that each gene has a different null distribution or one can calculate a null distribution for each gene. The third step is choosing the rejection regions. One can take a priori symmetric or one-sided rejection regions, or one can choose them adaptively. This involves comparing the original statistics to null versions of the statistics. Asymmetric rejection regions are most appropriate because we do not know beforehand what proportion of differentially expressed genes are in the positive or negative direction. If either of these two steps is not performed well, a loss in statistical power can result. The fourth step is to assess or control the number of false positives in some fashion. If one were simply to perform uncorrected hypothesis tests at the traditional 5% level, then $6000 \times 0.05 = 300$ false positives would be expected under the full null hypothesis if we were testing 6000 genes. This is clearly unacceptable, so some procedure must be performed to control the false-positive rate in a reasonable manner.

In the remainder of this work, we discuss some simple methods for achieving these three steps in testing for differential gene expression. The methodology presented in this chapter is implemented in the SAM software.

12.2 Methods and Applications

For simplicity, we limit our discussion of the methodology to the case where there are two experimental conditions and each sample is independent. We begin by discussing multiple hypothesis testing.

12.2.1 Multiple Hypothesis Testing

We are interested in determining which genes show a statistically significant difference in gene expression between two or more conditions. Therefore, the null hypothesis for each gene is that the data we observe have some common distributional parameter among the conditions (usually the mean or average expression level). For each gene we form a *statistic* that is a function of the data (e.g., a *t*-statistic). A significance region is then defined so that if the statistic lies in the region, then the gene is said to exhibit differential gene expression; otherwise, there is no significant differential gene expression.

There are two kinds of errors that can be committed *at each particular gene* when testing for differential gene expression. The first is a *Type I error* (false positive), which is when the statistic is significant yet the gene is not truly differentially expressed. The second is a *Type II error* (false negative), and this occurs when the statistic is not significant yet the gene is truly differentially expressed. Power is defined to be 1 - Type II error rate. See Rice (1995) for an introduction to hypothesis testing.

When testing multiple hypotheses, the situation becomes much more complicated. Now each gene has possible Type I and Type II errors, and it becomes unclear how one should measure the overall error rate. Specifically, consider Table 12.1 which lists the possible outcomes when testing J hypotheses simultaneously.

For example, V is the number of Type I errors (false positives), T is the number of Type II errors (false negatives), and $R = V + S$ is the total number of significant

Table 12.1. Outcomes from m hypothesis tests

	Accept	Reject	Total
Null True	U	V	J_0
Alternative True	T	S	J_1
Total	W	R	J

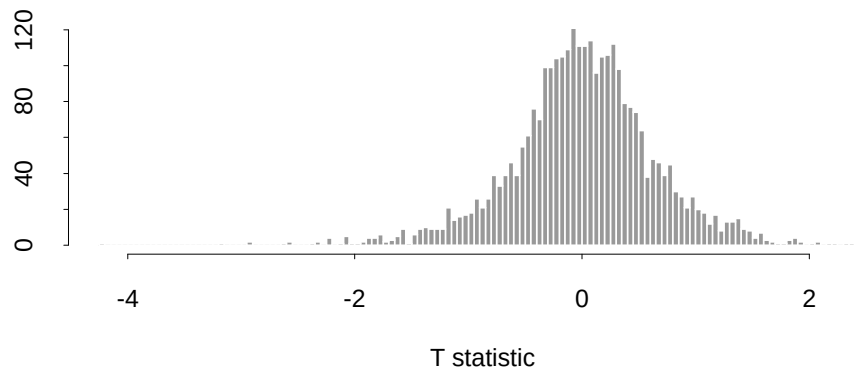
hypotheses. In order to measure the errors incurred in multiple hypothesis testing, it is convenient to define a compound error measure. Four that have been considered in the multiple hypothesis testing literature are defined as follows:

$$\begin{array}{llll}
 PCER & = & \mathbf{E}[V]/J & \text{per comparison error rate,} \\
 FWER & = & \mathbf{Pr}(V \geq 1) & \text{family wise error rate,} \\
 FDR & = & \mathbf{E}[V/R | R > 0] \cdot \mathbf{Pr}(R > 0) & \text{false discovery rate,} \\
 pFDR & = & \mathbf{E}[V/R | R > 0] & \text{positive false discovery rate.}
 \end{array}$$

The PCER and FWER have been used for many years, but the FDR was more recently proposed by Benjamini and Hochberg (1995) and the pFDR was proposed by Storey (2001). In multiple hypothesis testing, the family wise error rate (FWER) has traditionally been the main quantity of interest. Also, the goal has been to *control* the error rate at a particular prechosen level. We choose a level α at which to control the error rate (e.g., we want $FDR \leq \alpha$). Then an algorithm is applied to the statistics and the rejection regions to estimate the rejection region to use. If the algorithm guarantees that *on average* we will have $FDR \leq \alpha$ regardless of J_0 , then the algorithm is said to provide *strong control*. If this only occurs when $J = J_0$ (i.e., when all null hypotheses are true), then the algorithm only provides *weak control*.

The FWER is the probability of one or more false positives occurring among all significant hypotheses. However, in microarray studies, the number of tests (genes) is large, and the nature of the analysis is exploratory; that is, we want to identify many differentially expressed genes without too many false positives resulting. Hence, we expect more than one false positive, but we do not want too many in proportion to true positives. If we estimate the FDR or pFDR to be 5% among a list of 100 significant genes, then we roughly expect that 5 of those 100 are false positives. This kind of information is very useful for choosing genes to examine more carefully, and it is much more statistically rigorous than taking the top k genes for some arbitrarily chosen k .

Therefore, the FWER is not a relevant quantity in this setting. Instead, we focus here on the FDR and pFDR. In Storey (2002), Storey (2001), Storey and Tibshirani (2001), and Storey et al. (2002) we study false discovery rates from a unified point estimation approach, making connections to Bayesian procedures, introducing the pFDR, and developing more powerful methodology than that in Benjamini and Hochberg (1995). Moreover, we show that the methodology is robust against the kind of dependence encountered in DNA microarrays.

Figure 12.1. Histogram of 3000 t -statistics from the example.

12.2.2 An Application

Here is a simple example of the methodology to detect differential gene expression. The exact methods are a bit more complicated than we present here, but this example serves as a useful reference.

We obtained expression data from Stanford collaborators on $J = 3000$ genes, comparing normal untreated samples to a set of treated samples. (Since the data are unpublished, we must omit the biological details, but many similar data sets are publicly available.) There are 15 samples in group 1 (normal) and 13 in group 2 (treated). Figure 12.1 is a histogram of the 3000 two-sample t -statistics from the genes. They range from -4.54 to 3.72 . Suppose that we decide to reject all genes whose t -statistic is greater than 2 in absolute value; there are 146 such genes. What is the FDR among these 146 genes? To assess this, we do a random permutation of the sample labels

$$(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2),$$

recompute the t -statistics and count how many exceed ± 2 . (A “random permutation” is a random scrambling of the labels so that each label configuration is equally likely.) Doing this for 100 permutations, we find that the average number is 12.3. Thus a simple estimate of the FDR is $12.3/146 = 8.42\%$.

Now it turns out that this simple estimate tends to be biased upward. The reason is clear from Table 12.1. The permutations make all J genes null, but in our data only a proportion $\pi_0 = J_0/J$ are null. Hence to improve our estimate of the FDR, we multiply it by an estimate of π_0 . To obtain the latter, we look at small values of the t -statistic (in absolute value), where null statistics are much more abundant than alternatives. Looking, for example, at t -statistics below 0.15 in absolute value, we find that 668 of

the observed t -statistics fall in that range, while on the average 750 of the t -statistics from the permutations fall in that range. (The 0.15 cutoff is arbitrary in this example, but it can be automatically chosen taking bias and variance into account as in Storey and Tibshirani (2001).) Hence our estimate of π_0 is $\hat{\pi}_0 = 668/750 \approx 0.890$. Finally, our revised estimate of the FDR is $0.890 \cdot 8.42 = 7.49\%$.

12.2.3 Forming the Test Statistics

Suppose that we have J genes measured on I arrays, under two different experimental conditions. Let $\bar{x}_{j1}$ and $\bar{x}_{j2}$ be the average gene expression for gene j under conditions 1 and 2, and let s_j be the pooled standard deviation for gene j :

$$s_j = \sqrt{\left(\frac{1}{I_1} + \frac{1}{I_2}\right) \cdot \frac{\sum_1 (x_{ji} - \bar{x}_{j1})^2 + \sum_2 (x_{ji} - \bar{x}_{j2})^2}{I - 2}}$$

Here, I_k is the number of arrays in condition k , and each summation is taken over its respective group. Then, a reasonable test statistic for assessing differential gene expression is the standard (unpaired) t -statistic:

$$t_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j}.$$

Another statistic that could be formed is the rank-sum statistic. Let r_{ji} for $j = 1, \dots, J$ be the rank of the i th expression value within gene j . Then, the rank-sum statistic for gene j is

$$r_j = \sum_1 r_{ji},$$

where the summation is taken over the genes in group 1. An extreme r_j value in either direction would indicate a difference in gene expression. The statistic t_j tests for a difference in means, whereas r_j tests for a more general difference in distributions. Usually, one is more concerned with a difference in mean gene expression, so t_j is a more powerful statistic to use for this test.

Each of these statistics is formed using only information from the gene itself. It is possible to model the data in such a way that one can “borrow strength” across the genes. For example, if we view the s_j as coming from some overall random process, then they can be jointly modeled. This can lead to reducing the overall variance of the s_j , giving the tests more power on average. Tusher et al. (2001) take a nonparametric approach to this and shrink the s_j toward an adaptively chosen s_0 . The modified t -statistic is then

$$d_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j + s_0}.$$

Specifically, s_0 is chosen as the percentile of the s_j values that makes the coefficient of variation of d_j approximately constant as a function of s_j . This has the added effect of dampening large values of d_j that arise from genes whose expression is near zero.

Sometimes one wants to test for a difference between three or more conditions or for a trend in time-course data. Statistics that can be used for these and other situations are described in Tusher et al. (2001) and Section 12.3.

12.2.4 Calculating the Null Distribution

Concentrating on the two-sample case, the null distribution can most easily be calculated by permuting the group labels as we did in the example, or one can use the bootstrap. For the latter method, one can use the standard methodology described in Efron and Tibshirani (1993). The permutation method has a strength in that if the null hypothesis is true, then we are able to calculate the null distribution exactly. Westfall and Young (1993) discuss the differences between these two approaches more thoroughly, and Romano (1989) shows that they are asymptotically equivalent (in I , the number of arrays) in the two-sample case as well as many other cases.

As in forming the statistic, one can decide to treat each gene individually or borrow strength across the genes. If we calculate the null distribution individually for each gene, then this has two drawbacks. The first is that one suffers from a granularity problem: the null distribution has a resolution on the order of the number of permutations. For example, if we perform 1000 permutations, then the genes' resulting p -values will be on a resolution of $1/1000$. If we test 30,000 genes with 1000 permutations, then we will reject about 30 at a time. If we assume that each gene has the same null distribution, and combine them for each permutation, then the resolution of the null distribution is much finer. With J genes and B permutations, the resolution is on the order of $1/B$ for individual null distributions, but $1/(JB)$ for a pooled null distribution.

The second drawback of calculating a separate null distribution for each gene is that one is not then able to estimate better rejection regions. With individual null distributions, one essentially treats each gene as a different experiment. Therefore, for each "experiment," we have B observations from the null distribution and one observation from the original data. One cannot compare the null distribution to the observed statistic to derive more powerful, asymmetric rejection regions. This can lead to a loss of power and a waste of information in the data. On the other hand, if we pool the statistics and consider them to follow a mixture distribution as in Efron et al. (2001), then we now have many observations from a mixture of the null and affected distributions, as well as from the pure null distribution. We can use these observations jointly to derive more powerful rejection regions as in SAM and Efron et al. (2001).

The drawback of pooling the null statistics is that one can end up using different distributions to estimate an overall error rate. This is actually not a problem for using false discovery rates if the conditions discussed in Storey et al. (2002) hold. The null statistics just have to converge to some overall "null" distribution in J , the number of genes. This is likely met because J is typically very large, and one can easily see that upon repeats of the experiments, the null distribution is similar. Therefore, in the methodology that we describe here, the null statistics are pooled across the genes, making for more accurate and powerful procedures as well as cutting down on the computation time.

12.2.5 The SAM Thresholding Procedure

The SAM (significance analysis of microarrays) method (Tusher et al. 2001) and software forms cut points based on the data. Rather than using a standard rule of the form $|d_j| > t$ to call genes significant (i.e., having symmetric cut points $\pm t$), SAM derives cut points t_1 and t_2 and uses the rejection rule $d_j < t_1$ or $d_j > t_2$. This can lead to a more powerful test in situations where more genes are over expressed than under expressed, or vice-versa. Given its cut points, the SAM method estimates the false discovery rate

quantities described below in SubSection 12.2.6. The derivation of the cut points is illustrated in Figure 12.2 and detailed below:

The SAM Procedure

1. Compute the ordered statistics $d_{(1)} \leq d_{(2)} \cdots \leq d_{(J)}$
2. Take B sets of permutations of the group labels. For each permutation b compute statistics d_j^{*b} and corresponding order statistics $d_{(1)}^{*b} \leq d_{(2)}^{*b} \cdots \leq d_{(J)}^{*b}$. From the set of B permutations, estimate the expected order statistics by $\bar{d}_{(j)} = (1/B) \sum_{b=1}^B d_{(j)}^{*b}$ for $j = 1, 2, \dots, J$.
3. Plot the $d_{(j)}$ values versus the $\bar{d}_{(j)}$. For a fixed threshold Δ , starting at the origin, and moving up to the right, find the first $j = j_2$ such that $d_{(j)} - \bar{d}_{(j)} \geq \Delta$. All genes past j_2 are called “significant positive.” Similarly, start at the origin, move down to the left and find the first $j = j_1$ such that $\bar{d}_{(j)} - d_{(j)} \leq \Delta$. All genes past j_1 are called “significant negative.” For each Δ , define the upper cut point $t_2(\Delta)$ as the smallest d_j among the significant positive genes, and similarly define the lower cut point $t_1(\Delta)$.
4. If $t_1(\Delta) > t_2(\Delta)$, then we set $t_2(\Delta) = t_1(\Delta) = 0$.

In Figure 12.2, we used $\Delta = 0.531$ to yield the same number of genes (252) that are obtained using the cutoffs ± 1.5 . As the figure shows, the resulting cut points from SAM are asymmetric; they are -1.45 and 1.71. As a result, the number of genes called significant at the negative and positive ends are quite unequal – they are 167 and 85 (and they are 157 and 95 for the cut points ± 1.5). As mentioned earlier, the actual SAM procedure adds (the same) constant s_0 to the denominator of each statistic d_j . Here we set s_0 equal to the median value of the s_j , which is 17.7.

Note that under this procedure, if $\Delta' \geq \Delta$, then $t_2(\Delta') \geq t_2(\Delta)$ and $t_1(\Delta') \leq t_1(\Delta)$, giving a nested set of rejection regions as is most useful in hypothesis testing.

12.2.6 Estimating False Discovery Rates

For a fixed rejection region (fixed Δ), the FDR and pFDR are

$$\begin{aligned} FDR(\Delta) &= \mathbf{E} \left[\frac{V(\Delta)}{R(\Delta)} \middle| R(\Delta) > 0 \right] \mathbf{Pr}(R(\Delta) > 0), \\ pFDR(\Delta) &= \mathbf{E} \left[\frac{V(\Delta)}{R(\Delta)} \middle| R(\Delta) > 0 \right], \end{aligned}$$

where

$$V(\Delta) = \#\{d_j : \text{gene } j \text{ unchanged and } d_j \leq t_1(\Delta) \text{ or } d_j \geq t_2(\Delta)\},$$

$$R(\Delta) = \#\{d_j : d_j \leq t_1(\Delta) \text{ or } d_j \geq t_2(\Delta)\}.$$

Therefore, $V(\Delta)$ is the number of false positive genes and $R(\Delta)$ is the number of significant genes for the SAM threshold indexed by Δ .

In Storey (2002) we develop the following estimates of the FDR and pFDR for a given Δ :

$$\begin{aligned} \widehat{FDR}_{\Delta'}(\Delta) &= \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{R(\Delta) \vee 1}, \\ \widehat{pFDR}_{\Delta'}(\Delta) &= \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{\mathbf{Pr}(R^0(\Delta) > 0) \cdot [R(\Delta) \vee 1]}. \end{aligned}$$

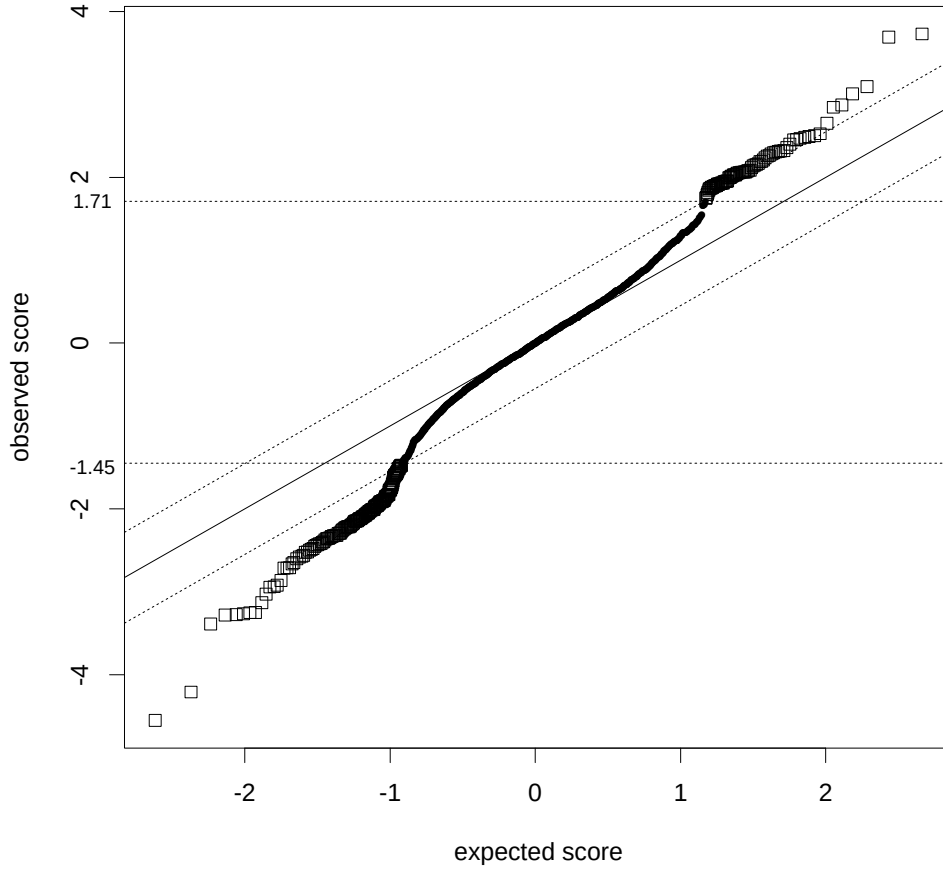


Figure 12.2. SAM plot for microarray example. We draw a band of two parallel (broken) lines a distance $\Delta = 0.531$ from the 45° (solid) line. Moving up and to the right, we find the first time that the points go outside the band. All genes to the right of that point are called significant (even if they fall inside the band). We do the same thing in the bottom left corner. The upper and lower excursion values imply upper and lower cut points— here 1.71 and -1.45.

$R(\Delta)$ is defined as before, but

$$R^0(\Delta) = \frac{\sum_{b=1}^B \#\{d_j^{b*} : d_j^{b*} \leq t_1(\Delta) \text{ or } d_j^{b*} \geq t_2(\Delta)\}}{B},$$

$$\Pr(R^0(\Delta) > 0) = \frac{\#\{b : \#\{d_j^{b*} : d_j^{b*} \leq t_1(\Delta) \text{ or } d_j^{b*} \geq t_2(\Delta)\} > 0\}}{B}.$$

Lastly, $\hat{\pi}_0(\Delta')$ is an estimate of the overall proportion of true null hypotheses (unchanged genes). This estimate depends on us choosing another Δ' . The SAM methodology takes Δ' such that $R^0(\Delta') = J/2$ (i.e., half the null statistics fall in the rejection region defined by Δ'), but Δ' can be chosen more adaptively to minimize the bias and the variance of the estimates (Storey and Tibshirani 2001). Nevertheless the estimate is

$$\hat{\pi}_0(\Delta') = \frac{J - R(\Delta')}{J - R^0(\Delta')}.$$

The denominator represents the number of genes not significant at Δ' when all null hypotheses are true. The numerator is the number of observed genes that are not significant at Δ' . For a reasonably chosen Δ' , this is going to consist mostly of unchanged genes, so $\hat{\pi}_0(\Delta') \approx J_0/J$. It is easy to show that $\mathbf{E}[\hat{\pi}_0(\Delta')] \geq J_0/J$.

A natural question to ask is how to pick the rejection region (i.e., the Δ value). We can use these point estimates in one of the following three options:

1. Fix Δ and calculate $\widehat{FDR}_{\Delta'}(\Delta)$ or $p\widehat{FDR}_{\Delta'}(\Delta)$. Then, for large J and weak dependence, or if the unchanged genes are independent, we have $\mathbf{E}[\widehat{FDR}_{\Delta'}(\Delta)] \geq FDR(\Delta)$ and $\mathbf{E}[p\widehat{FDR}_{\Delta'}(\Delta)] \geq pFDR(\Delta)$.
2. Choose a level α beforehand at which we wish to control the FDR. Then, take the smallest Δ such that $\widehat{FDR}_{\Delta'}(\Delta) \leq \alpha$. For large J and weak dependence, this guarantees that $FDR(\Delta) \leq \alpha$. If the unchanged genes are independent and we limit the Δ that we consider to $\Delta \geq \Delta'$, then regardless of the size of J , we have $FDR(\Delta) \leq \alpha$.
3. Calculate $\widehat{FDR}_{\Delta'}(\Delta)$ and $p\widehat{FDR}_{\Delta'}(\Delta)$ over all $\Delta > 0$ simultaneously. Then, for large J and weak dependence, we have $\min_{\Delta} [\widehat{FDR}_{\Delta'}(\Delta) - FDR(\Delta)] \geq 0$ and $\min_{\Delta} [p\widehat{FDR}_{\Delta'}(\Delta) - pFDR(\Delta)] \geq 0$; that is, the estimates are asymptotically simultaneously conservative.
4. For each gene j , $j = 1, \dots, J$, estimate its q -value as

$$\hat{q}\text{-value}(\text{gene } j) = \min_{\Delta: \text{gene } j \text{ significant}} p\widehat{FDR}_{\Delta'}(\Delta).$$

The true q -value(gene j) is calculated analogously with the true $pFDR$. Then, for large J and weak dependence, we have $\min_{\text{all gene } j} [\hat{q}\text{-value}(\text{gene } j) - q\text{-value}(\text{gene } j)] \geq 0$; that is, the estimated q -values are asymptotically simultaneously conservative.

In an attempt to keep the mathematical details within the scope of this book, we have not stated these options in precise detail. See Storey et al. (2002) for a rigorous mathematical justification of the above claims as well as the precise conditions of “weak dependence.” (For example, we must bound $\Delta \geq \delta$ for some pre-chosen arbitrarily small δ .) These results are valid for fixed rejection regions; that is, those chosen independently of the data. Since SAM chooses the rejection regions and does false discovery rate calculations from the same data, some bias may be incurred. This is yet to be explored. See Storey et al. (2002) and Storey and Tibshirani (2001) for arguments as to why weak dependence holds for the problem of detecting differential gene expression in DNA microarrays, including simulation results that show that $J \geq 3000$ usually suffices. Notice that option 2 does not include the $pFDR$ estimate. This is because the $pFDR$ is 1 when all genes are unchanged, and therefore it cannot be controlled in the sense described in that option.

The SAM software allows one to interactively change Δ , so option 3 is potentially the most useful. It is argued in Storey (2001) that the pFDR has a more sound interpretation, so it is perhaps the better quantity to use. In single hypothesis testing, one can assign a p -value to a statistic, which is usually defined to be the probability of a statistic more extreme than the one observed given that the null hypothesis is true. The more technical definition of the p -value for an observed statistic is the minimum Type I error rate that can be attained over all rejection regions containing that observed statistic. In Storey (2001), a natural false discovery rate analog of the p -value is introduced, which is called the q -value. The q -value of a statistic is the minimum pFDR that can be attained over all rejection regions containing that observed statistic. Its name is derived from its special relationship to the p -value when considering the Bayesian posterior probability form of the pFDR. Whereas the p -value is the probability of observing a statistic as extreme or more extreme given that the null hypothesis is true, the q -value is the probability that the null hypothesis is true given a statistic as extreme or more extreme than the one observed. It is easy to see how the q -value can be interpreted among many tests, whereas the p -value cannot.

The q -value of a particular gene can be estimated by taking the minimum $\widehat{pFDR}_{\Delta'}(\Delta)$ over all Δ for which the gene is found to be significant. The q -value estimate is conservatively consistent under the conditions that we assumed above (Storey et al. 2002). In testing for differential gene expression, we suggest estimating the q -value for each gene. It gives a measure of the strength of evidence for differential gene expression in terms of the pFDR. This is an individual measure for each gene that simultaneously takes into account the multiple comparisons. Note that by using the q -value, it is not necessary to pick the rejection region or the desired error rate beforehand. The q -value estimate is included in the SAM software.

12.3 Software

SAM is a package for implementing the significance analysis of microarrays technique of Tusher et al. (2001). The false discovery rate methodology of Storey (2002) was recently included. The software was written by Balasubramanian Narasimhan.

The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. The response variable may be a grouping such as untreated/treated (either unpaired or paired), a multiclass grouping (e.g., breast cancer, lymphoma, colon cancer, etc.), a quantitative variable (e.g., blood pressure) or a possibly censored survival time. SAM computes a statistic d_j for each gene j , measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response. The cutoff for significance is determined by a tuning parameter Δ , chosen by the user based on the false positive rate. One can also choose a fold change parameter, to ensure that called genes change by at least a pre-specified amount.

Table 12.2. Response formats

Response type	Coding
Quantitative	Real number (e.g., 27.4 or -45.34)
Two class (unpaired)	Integer 1, 2
Multiclass	Integer 1, 2, 3, ...
Paired	Integer -1, 1, -2, 2, etc. (e.g., - means Before treatment, + means after treatment) -1 is paired with 1, -2 is paired with 2, etc.
Survival data	(Time, status) pair such as (50,1) or (120,0) First number is survival time, second is status (1=died, 0=censored)
One class	Integer, every entry equal to 1

12.3.1 Obtaining the Software

SAM can be downloaded at <http://www-stat.stanford.edu/~tibs/SAM>. QVALUE, a program for calculating q-values from a list of p-values, can be downloaded at <http://www.stat.berkeley.edu/~storey/qvalue/>.

12.3.2 Data Formats

The data should be put in a Microsoft Excel spreadsheet. The first row of the spreadsheet has information about the response measurement; all remaining rows have gene expression data, one row per gene. The columns represent the different experimental samples. The first line of the file contains the response measurements, one per column, starting at column 3. This is further described below in SubSection 12.3.3. The remaining lines contain gene expression measurements, one line per gene. We describe the format below.

Column 1 This should contain the gene name. It is for the user's reference.

Column 2 This should contain the gene ID for the user's reference. Note that the gene ID column is the column that is linked to the SOURCE web site by SAM. Hence, a unique identifier (e.g. clone ID, accession number or gene name/symbol) should be used in this column if SOURCE web site gene lookup is desired.

Remaining columns These should contain the expression measurements as numbers. Missing expression measurements should be coded as **NA**. This is done easily in good editor or in Excel. In Excel, to change blank fields to **NA**, choose all columns, pull down the Edit menu, choose Replace and then **nothing (Blank)** with **NA**.

12.3.3 Response Format

Table 12.2 shows the formats of the response for various data types. A look at the example files is also informative. A *quantitative* response is real-valued, such as blood pressure. *Two class (unpaired)* groups are two sets of measurements, in which the experiment units are all different in the two groups (for example, control and treatment groups, with samples from different patients). With a *multiclass* response there are more than two groups, each containing different experimental units. This is a generalization

of the *unpaired* setup to more than 2 groups. *Paired* groups are two sets of measurements in which the same experimental unit is measured in each group (for example, samples from the same patient, measured before and after a treatment). *Survival data* consists of a time until an event (such as death or relapse), possibly censored. In the *one class* problem we are testing whether the mean gene expression differs from zero (for example, each measurement might be the log(red/green) ratio from two labeled samples hybridized to a cDNA chip, with green denoting before treatment and red after treatment). Here the response measurement is redundant and is set equal to all 1s.

Sometimes, it is difficult to enter blocking information (see SubSection 12.3.5) without confusing Excel. Excel thinks such entries are formulae. Therefore, SAM allows any response to be enclosed within quotes (not apostrophes) and strips the quotes off before doing any computation.

12.3.4 Example Input Data File for an Unpaired Problem

The response variable is 1 = *untreated*, 2 = *treated*. The columns are gene name, gene ID and then by the expression values. The first row contains the response values. Note that there are two blank cells at the beginning of line 1. The gene expression measurements can have an arbitrary number of decimal places.

Table 12.3. Example data set for an unpaired problem.

		1	1	2	2	1	1	2	2
GENE1	101	7.64	-0.50	-1.95	10.12	-10.77	-4.47	-7.65	7.58
GENE2	102	38.10	4.86	7.87	-13.59	-9.79	-13.46	-8.91	-5.07
GENE3	103	21.15	5.96	3.20	-4.74	-3.70	-12.35	-10.17	0.63
GENE4	104	187.21	-23.81	16.76	14.10	-99.76	-89.11	-10.92	5.52

12.3.5 Block Permutations

Response labels can be specified to be in blocks by adding the suffix *blockN*, where N is an integer, to the response labels. Suppose for example that in the two-class data of Section 12.3, samples 1,3,5, and 7 came from one batch of microarrays, and samples 2,4,6, and 8 came from another batch. We call these batches “blocks.” Then, we might not want to mix up the batches in our permutations of the data, in order to control for the array differences; that is, we would like to allow permutations of the samples within the set 1,3,5, and 7 and within the set 2,4,6, and 8, but not across the two sets. We indicate the blocks (batches) as follows:

Table 12.4. Example data set for a blocked unpaired problem.

		1block1	1block2	2block1	2block2	1block1	1block2	2block1	2block2
GENE1	101	7.64	-0.50	-1.95	10.12	-10.77	-4.47	-7.65	7.58
GENE2	102	38.10	4.86	7.87	-13.59	-9.79	-13.46	-8.91	-5.07
GENE3	103	21.15	5.96	3.20	-4.74	-3.70	-12.35	-10.17	0.63
GENE4	104	187.21	-23.81	16.76	14.10	-99.76	-89.11	-10.92	5.52

For example, “1block1” means treatment 1, block (or batch) 1. “1block2” means treatment 1, block (or batch) 2. In this example, there are $4! = 24$ permutations within

block 1 and $4! = 24$ permutations within block 2. Hence the total number of possible permutations is $24 \cdot 24 = 196$. If the block information is not indicated in line 1, all permutations of the eight samples would be allowed. There are $8! = 40,320$ such permutations. Please note that block permutations cannot be specified with paired response as there is an implicit blocking already in force.

12.3.6 Normalization of Experiments

Different experimental platforms require different normalizations. Therefore, *the user is required to normalize the data from the different experiments (columns) before running SAM*. For cDNA data, centering the columns of the expression matrix (that is, making the column means equal to zero) is often sufficient. For oligonucleotide data, a stronger calibration may be necessary: for example, a linear normalization of the data for each experiment versus the row-wise average for all experiments.

12.3.7 Handling Missing Data

There are currently two options for imputing missing values in SAM.

Row average Each value is imputed with the average of non-missing values for that gene.

K nearest neighbor In the other (default) option, missing values are imputed using a k -nearest neighbor average in gene space (default $k = 10$):

1. For each gene j having at least one missing value:
 - (a) Let S_j be the samples for which gene j has no missing values.
 - (b) Identify all other genes G_j having no missing values for samples S_j .
 - (c) Find the k nearest neighbors to gene j among genes G_j , using only samples S_j to compute the Euclidean distance.
 - (d) Impute the missing values in gene j , using the averages of the non-missing from the k nearest neighbors for that sample.
2. If a gene still has missing values after the steps above, impute them with the average available expression values for that gene.

12.3.8 Running SAM

To begin, you highlight an area of the spreadsheet that represents the data by first clicking on the top left-hand corner and then shift-clicking on the bottom right-hand corner of the rectangle. Then, click on the **SAM** button in the toolbar. A dialog form shown in Figure 12.3 (see color insert) now pops up. You have to select the type of response variable, and if desired, change any of values of the default parameters. For two-class and paired data, one has to specify whether the data are in the logged (base 2) scale or not. Click the **OK** button to do the analysis.

If you had any missing data in your spreadsheet, a new worksheet named “SAM imputed data set” containing the imputed data set is added to the workbook. This data can be used in subsequent analyses to save time. If there is no missing data, this worksheet is not added. The software adds three more worksheets to the workbook. There is one which is hidden called “SAM plot data” and should be left alone. The

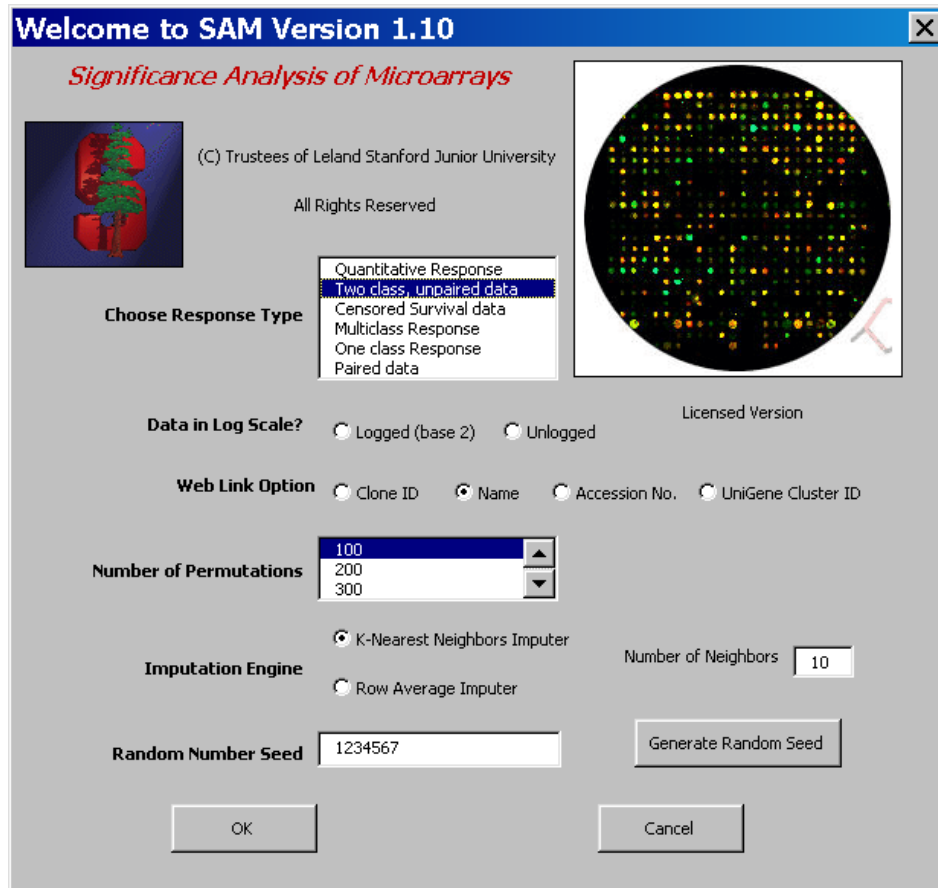


Figure 12.3. The SAM dialog box.

sheet named “SAM plot” contains the plot that the user can interact with. The sheet named “SAM output” is used for writing any output.

Initially, a slider pops up along with the plot shown in Figure 12.4 (see color insert) that allows one to change the Δ parameter and examine the effect on the false discovery rate. If you want a more stringent criterion, try setting a nonzero fold-change parameter. Positive significant genes are labelled in red on the SAM plot, negative significant genes are green. When you have settled on a value for Δ , click on the “List significant genes” button for a list of significant genes. The “List delta table” button lists the number of significant genes and the false positive rate for a number of values of Δ . Please note that all output tables are sent to the worksheet named “SAM output” erasing whatever

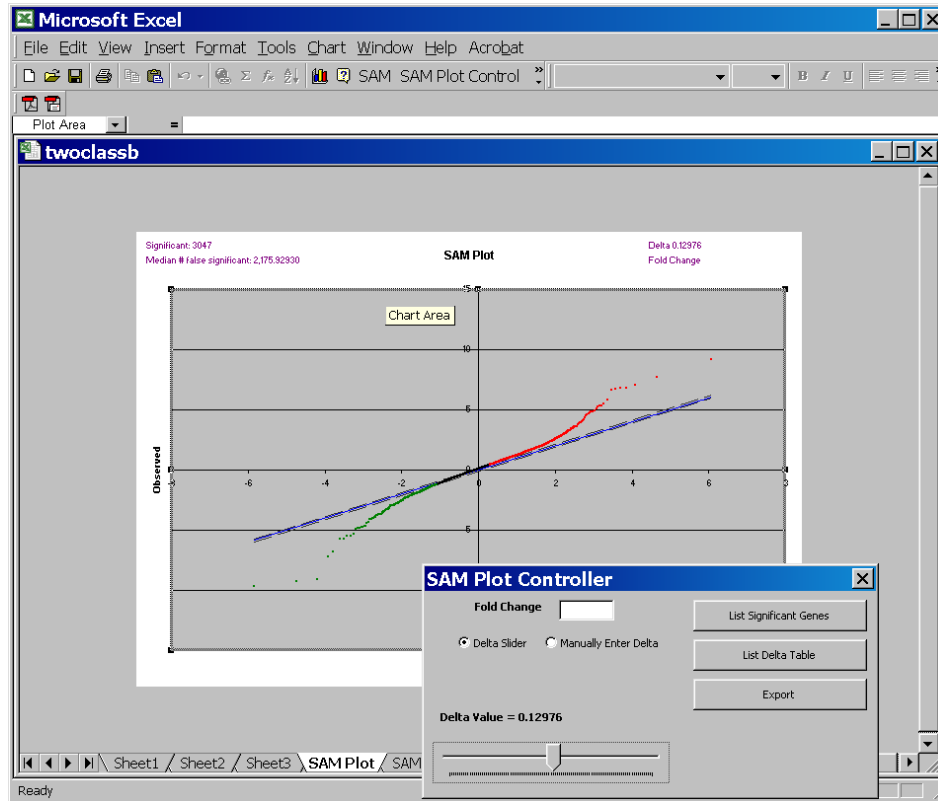


Figure 12.4. The SAM plot.

was previously present in the worksheet. While the slider is present, all interaction with the workbook is only possible via the slider. It can be killed anytime and recreated by clicking on the **SAM plot control** button.

12.3.9 Format of the Significant Gene List

For reference, SAM numbers the original genes, in their original order, as 1, 2, 3, and so on. In the output, this is the row number. The output for the list of significant genes has the following format:

Row Number The row in the selected data rectangle.

Gene Name The gene name specified in the first column selected data rectangle. This is for the user's reference.

Gene ID The gene ID specified in the second column selected data rectangle. This is for the user's reference, but is also linked to the SOURCE web-site for gene information.

SAM score(d) The t -statistic value.

Numerator The numerator of the t -statistic.

Denominator($s + s_0$) The denominator of the t -statistic.

Bayesian probability This is an univariate measure of the evidence that a gene is significant, computed using a Bayesian mixture model and ignoring the multiple comparisons. See Efron et al. (2001) for details.

q -value This is the lowest pFDR at which the gene is called significant. It is analogous to the familiar p -value, but interpreted as the probability a gene is a false positive given its statistic is as or more extreme than the observed statistic. (The p -value has the opposite interpretation.) The q -value measures how significant the gene is: as $d_j > 0$ increases, the corresponding q -value decreases. See Storey (2002), Storey (2001), and SubSection 12.2.6 for details.

The numerator, denominator and q -value were explained in Section 12.2. The list is divided into positive and negative genes, having positive or negative score d_j . A positive score means positive correlation with the response variable: e.g., for group responses 1 then 2, a positive score means expression is higher for group 2 than group 1. For a survival time response, a positive score means people with higher expression have longer survival times. All statements are reversed for negative scores.

12.4 Discussion

There have been several other methods suggested for detecting differential gene expression, including those discussed in Dudoit and Yang (Chapter 3, this volume) and Newton and Kendziorski (Chapter 11, this volume). A Bayesian method was developed in Newton et al. (2001), although the multiple hypothesis testing was not taken into account. In fact, only posterior probabilities are reported according to this methodology. They do suggest taking into account both fold change and the variance of the fold change (analogous to a t -statistic). An empirical Bayes approach is suggested in Efron et al. (2001). Posterior probabilities are calculated as well, and no distributional assumptions are made. In both of these methods, it is important to realize that a posterior probability of differential gene expression given the value of the t -statistic for only the gene at hand is not sufficient for looking at several genes at once. In Dudoit et al. (2002), symmetric rejection regions and the FWER as the multiple hypothesis testing measure are used. They use the adjusted p -value methodology of Westfall and Young (1993).

A strength of the approach we have presented is that the rejection regions are chosen to accommodate the proportion and degree to which the effected genes are over-expressed or under-expressed. Also, the multiple comparisons are taken into account via a multiple hypothesis testing error rate that is appropriate for finding many significant genes, while limiting the proportion of false positives among those called significant. Of course other methods may exist for choosing rejection regions or for assessing the false positives that provide a more powerful, but just as efficient detection of differential gene expression. This is an active area of research, and some additional, interesting statistical methodology is bound to emerge.

References

- Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* 85: 289–300.
- Dudoit S., Yang Y., Callow M. and Speed T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* 12: 111–139.
- Efron B. and Tibshirani R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.
- Efron B., Tibshirani R., Storey J. D. and Tusher V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* 96: 1151–1160.
- Newton M., Kendzierski C., Richmond C., Blatter F. and Tsui K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology* 8: 37–52.
- Rice J. A. (1995). *Mathematical Statistics and Data Analysis*, 2nd edn, Duxbury Press, Belmont, CA.
- Romano J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses, *Annals of Statistics* 17: 141–159.
- Storey J. D. (2001). The positive false discovery rate: A Bayesian interpretation and the q -value. Submitted. Available at <http://www.stat.berkeley.edu/~storey/>.
- Storey J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B* 64: 479–498.
- Storey J. D., Taylor J. E. and Siegmund D. (2002). A unified estimation approach to false discovery rates. Submitted. Available at <http://www.stat.berkeley.edu/~storey/>.
- Storey J. D. and Tibshirani R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Submitted. Available at <http://www.stat.berkeley.edu/~storey/>.
- Tusher V., Tibshirani R. and Chu C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proceedings of the National Academy of Sciences* 98: 5116–5121.
- Westfall P. H. and Young S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p -value Adjustment*, Wiley.