BIOINFORMATICS

Vol. 20 no. 1 2004, pages 93–99
DOI: 10.1093/bioinformatics/btg382

A global test for groups of genes: testing association with a clinical outcome

Jelle J. Goeman[1,2,], Sara A. van de Geer[2], Floor de Kort[3] and Hans C. van Houwelingen[1]*

[1]Department of Medical Statistics, Leiden University Medical Center, P.O. Box 9604, 2300 RC Leiden, The Netherlands, [2]Mathematical Institute, Leiden University, The Netherlands and [3]Center for Human and Clinical Genetics, Leiden University Medical Center, The Netherlands

Received on April 1, 2003; revised on June 24, 2003; accepted on July 6, 2003

ABSTRACT

Motivation: This paper presents a global test to be used for the analysis of microarray data. Using this test it can be determined whether the global expression pattern of a group of genes is significantly related to some clinical outcome of interest. Groups of genes may be any size from a single gene to all genes on the chip (e.g. known pathways, specific areas of the genome or clusters from a cluster analysis).

Result: The test allows groups of genes of different size to be compared, because the test gives one *p*-value for the group, not a *p*-value for each gene. Researchers can use the test to investigate hypotheses based on theory or past research or to mine gene ontology databases for interesting pathways. Multiple testing problems do not occur unless many groups are tested. Special attention is given to visualizations of the test result, focussing on the associations between samples and showing the impact of individual genes on the test result.

Availability: An R-package globaltest is available from http://www.bioconductor.org

Contact: j.j.goeman@lumc.nl

1 INTRODUCTION

The popularity of microarray technology has led to a surge of new statistical methods aimed at finding differentially expressed genes. A sophisticated methodology has been developed to counter the multiple testing problem that occurs when testing thousands of genes simultaneously.

This paper looks at the problem of finding differentially expressed genes from a different point of view. It presents a global test that can be used to determine whether some pre-specified *group of genes* is differentially expressed. This allows the unit of analysis to be shifted from individual genes to groupings of genes. The question addressed is whether the gene expression pattern over the whole group of genes is related to a clinical outcome. It does not matter for the test whether the group consists of up- or downregulated genes or is a mixture of both. The clinical outcome may be a group label or a continuous measurement.

Often researchers who conduct microarray experiments have one or more specific groups of genes that they are especially interested in, e.g. certain pathways or areas on the genome. Even if this is not the case, many pathways are at least partially known from the scientific literature and it could sometimes be more worthwhile to test a limited number of pathways or gene ontology classes than an enormous number of individual genes. Other potentially interesting groups of genes to be tested include the clusters from a cluster analysis or all genes on the chip.

The first part of the paper presents the mathematical details, starting with the empirical Bayesian generalized linear model on which the test is based. Connections to other methods (esp. prediction methods) are elaborated.

In the second part two elaborate applications are presented, showing different aspects of the test. One is the well-known public data set by Golub *et al*. (1999) with Affymetrix arrays of patients with Acute Lymphoic Leukemia (ALL) and Acute Myeloid Leukemia (AML). Here the test is applied to the set of all genes to show an enormous difference in overall expression pattern. The second is a smaller in-house data set with oligonucleotide arrays of cell lines of which some were exposed to a heat shock. The test is applied to two groups of genes associated with heat shock.

In the applications, special attention is given to visualizations of the test result which make the results easier to interpret for the researcher. These include graphs to search for outlying samples and diagnostic plots to judge how much each individual gene contributes to a significant test result for the group.

2 THE DATA

Proper normalization of data is very important for a meaningful analysis of microarray data. The problem of normalization generates an enormous amount of literature and is fast becoming a statistical specialization by itself. In this paper we will

*To whom correspondence should be addressed.

simply assume that the data have been normalized beforehand in a way that fits the experimental design and that possible confounding effects of array, dye etc. have been removed as well as the experimental design allows. However, missing values are allowed (see Section 8).

We assume we have normalized gene expression measurements of n samples for p genes. Of these p genes, there is a subgroup of m ($1 \leq m \leq p$) genes, which we want to test. It is important that the clinical outcome was not used in the selection of these m genes. Define $X = (x_{ij})$ as the $n \times m$ data matrix containing only the m genes of interest. Note that we follow the statistical convention to use the rows for the samples and the columns for the genes, instead of the transposed notation which is common in microarray literature.

Define Y as the clinical outcome (an $n \times 1$ vector). Usually Y will be a 0/1 group label (e.g. AML versus ALL), but it may also be a continuous measurement.

3 THE MODEL

There is a close connection between finding differentially expressed genes and predicting the clinical outcome. If a group of genes can be used to predict the clinical outcome, the gene expression patterns must differ for different clinical outcomes. This duality will be used to derive the test.

Modelling the way in which Y depends on X, we adopt the framework of the generalized linear model (McCullagh and Nelder, 1989), which includes linear regression and logistic regression as special cases. In this model there is an intercept α, a length m vector of regression coefficients β and a link function h (e.g. the logit function), such that

$$E(Y_i|\beta) = h^{-1}\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\right). \qquad (1)$$

Here β_j is the regression coefficient for gene j ($j = 1, \ldots, m$).

Testing whether there is a predictive effect of the gene expressions on the clinical outcome is equivalent to testing the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0,$$

that all regression coefficients are zero. It is not possible to test this hypothesis in a classical way (with β non-stochastic) because m might be large relative to n. In this case there are too few degrees of freedom.

However, it is possible to test H_0 if it is assumed that $\beta_1, \ldots, \beta_m$ are a sample from some common distribution with expectation zero and variance τ^2. Then a single unknown parameter τ^2 determines how much the regression coefficients are allowed to deviate from zero. The null hypothesis becomes simply

$$H_0 : \tau^2 = 0.$$

Note that the choice of $\tau^2 I_m$ (with I_m the $m \times m$ identity matrix) as the covariance matrix of the stochastic vector β is not imperative. It is the most convenient choice which will yield a test that treats all genes on an equal footing. Any other $m \times m$ covariance matrix may be used to replace I_m, if desired, resulting in a different test with power against different alternatives. For example a different diagonal matrix can be taken to reflect prior beliefs in the greater reliability of certain genes. Assuming positive correlations between the elements of β results in more power against alternatives where all β coefficients have the same sign.

The model (1) with β random may be looked at in various ways. Firstly the distribution of β can be seen as a prior, with unknown shape and with a variance depending on an unknown parameter. Viewed in this way the model (1) is an empirical Bayesian model.

A second interpretation is to view the model as a penalized regression model, in which the estimated coefficients are shrunk towards a common mean. The log likelihood of Y can be written

$$\text{loglik}(Y, \beta) = \text{loglik}(Y|\beta) + \text{loglik}(\beta),$$

where the first term on the right is the likelihood of the ordinary generalized linear model and the second term is known as the *penalty*. Well-known examples of penalized regression include ridge regression (Hoerl and Kennard, 1970), which arises when β is normally distributed and the Lasso (Tibshirani, 1996), which is a variant where β has a double exponential distribution. Ridge regression with a logistic link function has been described by Le Cessie and Van Houwelingen (1992) and applied on microarray data by Eilers *et al.* (2001) with promising results.

There is a third interpretation which will be the basis for the test in the next section. For this we write $r_i = \sum_j x_{ij}\beta_j$, $i = 1, \ldots, n$. Then r_i is the linear predictor, the total effect of all covariates for person i. Let $\boldsymbol{r} = (r_1, \ldots, r_n)$, then $\boldsymbol{r}$ is a random vector with $E(\boldsymbol{r}) = 0$ and $\text{Cov}(\mathbf{r}) = \tau^2 XX'$. The model (1) simplifies to

$$E(Y_i|r_i) = h^{-1}(\alpha + r_i). \qquad (2)$$

This is a simple random effects model, in which each sample has a random effect that influences its outcome. The covariance matrix between the random effects is known and is determined by the gene expression levels. If $\tau^2 > 0$, two samples i and j with similar gene expression patterns have correlated random effects r_i and r_j and therefore have a greater probability of having similar outcomes Y_i and Y_j than samples with less similar expression patterns.

4 THE SCORE TEST

A test for testing H_0 in the model (2) is discussed in Le Cessie and Van Houwelingen (1995) and Houwing-Duistermaat *et al.* (1995). The marginal likelihood of Y in this model depends on

only two or three parameters. These are α and τ^2 and sometimes, depending on the specific model, an extra dispersion parameter (e.g. the residual variance σ^2 of the outcome Y in an ordinary linear regression model).

In this section, we first suppose that α and the dispersion parameter are known (the case where they are unknown is dealt with in Section 6). In this case a score test for $\tau^2 = 0$ can be calculated by taking the derivative of the loglikelihood with respect to τ^2 at $\tau^2 = 0$, divided by the standard deviation of this derivative under H_0. This yields the test statistic

$$T = \frac{(Y-\mu)'R(Y-\mu) - \mu_2 \text{trace}(R)}{\left[2\mu_2^2 \text{trace}(R^2) + (\mu_4 - 3\mu_2^2)\sum_i R_{ii}^2\right]^{1/2}},$$

where $R = (1/m)XX'$ is an $n \times n$ matrix proportional to the covariance matrix of the random effects $\boldsymbol{r}$, $\mu = h^{-1}(\alpha)$ is the expectation of Y under H_0 and μ_2 and μ_4 the second and fourth central moments of Y under H_0.

It can be shown that if H_0 is true, T is asymptotically normally distributed. However, it is often more convenient to use the equivalent, much simpler test statistic

$$Q = \frac{(Y-\mu)'R(Y-\mu)}{\mu_2}$$

which has expectation

$$E(Q) = \text{trace}(R) \tag{3}$$

and variance

$$\text{Var}(Q) = 2\text{trace}(R^2) + \left(\frac{\mu_4}{\mu_2^2} - 3\right)\sum_i R_{ii}^2. \tag{4}$$

The statistic Q is also asymptotically normally distributed, but it is a quadratic form which is non-negative, because R is non-negative definite. Therefore for small sample sizes a better approximation to the distribution of Q is a scaled χ^2 distribution $c\chi_\nu^2$, where c is a scaling factor and ν is the number of degrees of freedom. This has been shown using simulations in Le Cessie and Van Houwelingen (1995). Equating the mean and variance of $c\chi_\nu^2$ and Q yields $c = \text{var}(Q)/[2E(Q)]$ and $\nu = 2[E(Q)]^2/\text{var}(Q)$.

Note that the statistic Q and its distribution are easy to calculate for high-dimensional data because they only involve the small $n \times n$ covariance matrix $R = (1/m)XX'$ between the samples and never the potentially large $m \times m$ covariance matrix $(1/n)X'X$ between the genes. Testing a large number of genes therefore never gives computational problems.

5 PROPERTIES OF THE TEST

There are two ways of rewriting the test statistic Q to gain a better intuitive understanding of the test. The first can be used to show the influences of the genes, the second the influence of the samples. These two decompositions of Q will be the basis of various illustrative graphs in Sections 9 and 10. Furthermore, the fact that the test is a score test also gives the test a nice optimality property.

For the first interpretation rewrite

$$Q = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\mu_2}[X_i'(Y-\mu)]^2$$

where $X_i (i = 1, \ldots, m)$ is the $n \times 1$ vector of the gene expressions of gene i. Note however that the expression $Q_i = (1/\mu_2)[X_i'(Y-\mu)]^2$ is exactly the test statistic that would have been calculated for a group of genes consisting only of the i-th single gene in the group of interest. Therefore the test statistic Q for a group of m genes is just the average of the statistics $Q_1, \ldots, Q_m$, calculated for the m single genes that the group consists of.

Each Q_i can again be written as (a multiple of) the squared covariance between the expression pattern of the gene and the clinical outcome. Because the averaging is done at this squared covariance level, genes with large variance have much more influence on the outcome of the test statistic Q than genes with small variance. This is a nice property in the context of microarray analysis, because low-variance genes are generally seen as uninteresting, as it usually implies that there is little biological variation in these genes.

For a different look at the test the statistic Q can be written in the following way

$$Q = \frac{1}{\mu_2}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}(Y_i-\mu)(Y_j-\mu) \tag{5}$$

as the sum over all terms of the Hadamard (term-by-term) product of the matrices R and $(Y-\mu)(Y-\mu)'$. The matrix $R = (1/m)XX'$ is the covariance of the gene-expression patterns between the samples, and the matrix $(Y-\mu)(Y-\mu)'$ is the covariance matrix of the clinical outcomes of the samples. The statistic Q therefore has a high value whenever the terms of these two matrices are correlated, that is when the covariance structure of the gene-expressions between samples resembles the covariance structure between their outcomes. The score test can therefore be seen as a test to see whether samples with similar gene-expressions also have similar outcomes.

An interesting property of a score test in general is that it maximizes the average power against all alternatives where the true value of the parameter is small. Equivalently, in this case it has optimal power against the range of alternatives $R_t = \{\|\beta\|^2 \leq t^2\}$ as $t^2 \to 0$. As R_t is an m-ball it contains relatively many alternatives with all β's non-zero but small, therefore the test is focussed mostly on detecting alternatives where many genes play a part. This is a desirable property because the test is designed to say something about the group of genes as a whole.

6 SOME TECHNICAL ADJUSTMENTS

In the previous section it was assumed that α (and therefore μ) was known and that the dispersion parameter (if any) was also known. In practice this is never true. In this section some adjustments of the test are presented which is correct for using estimated parameters.

First suppose that μ is unknown, but μ_2 and μ_4 are known. It is easily verified that

$$Y - \hat{\mu} = (I - H)(Y - \mu),$$

where $H = (1/n)\mathbf{1}\mathbf{1}'$ is the hat matrix for estimation of the mean μ of Y and $\mathbf{1}$ is a length n column vector of ones. Therefore calculating Q using $\hat{\mu}$ instead of μ results in calculating

$$\begin{aligned} Q &= \frac{1}{\mu_2}(Y - \hat{\mu})' R (Y - \hat{\mu}) \\ &= \frac{1}{\mu_2}(Y - \mu)'(I - H) R (I - H)(Y - \mu). \end{aligned}$$

The mean and variance of Q are therefore simply given by (3) and (4) with R replaced by $\tilde{R} = (I - H)R(I - H)$. This is equivalent to centering the genes so that the average value of each gene over the samples is set to zero.

Correction for estimation of μ_2 is not so easy. Simply replacing μ_2 by its estimate $\hat{\mu}_2$ would generally lead to a test that is too conservative, because the numerator $(Y - \hat{\mu})' R (Y - \hat{\mu})$ and the denominator $\hat{\mu}_2 = (1/n)(Y - \hat{\mu})'(Y - \hat{\mu})$ of Q are positively correlated, so that the variance of Q is overestimated if this dependency is not taken into account.

Corrections for the variance of Q are available from Houwing-Duistermaat *et al.* (1995) for the linear regression model (continuous clinical outcome) and for the logistic regression model (two groups). For a linear regression $Q = (Y - \hat{\mu})' R (Y - \hat{\mu})/\hat{\sigma}^2$, which has $E(Q) = \text{trace}(\tilde{R})$ and variance

$$\text{Var}(Q) = \frac{2}{n+1}\left[(n-1)\text{trace}(\tilde{R}^2) - \text{trace}^2(\tilde{R})\right].$$

For the logistic regression model $Q = (Y - \hat{\mu})' R (Y - \hat{\mu})/[\hat{\mu}(1 - \hat{\mu})]$. This also has $E(Q) = \text{trace}(\tilde{R})$ and its variance can be approximated by

$$\begin{aligned} \text{Var}(Q) \approx \frac{1 - 6\mu + 6\mu^2}{\mu(1-\mu)} \left[\sum_{i=1}^{n} \tilde{R}_{ii}^2 - \frac{1}{n}\text{trace}^2(\tilde{R})\right] \\ + 2\text{trace}(\tilde{R}^2) - \frac{2}{n-1}\text{trace}^2(\tilde{R}). \end{aligned} \quad (6)$$

7 HANDLING SMALL SAMPLE SIZE

If the sample size is small the asymptotic formula's used to calculate the p-value may not be correct. In this case a different approach could be to find the p-value using a permutation method. The empirical distribution of Q can be found by calculating Q for all permutations of the outcome Y or a random sample from these. The permutation method also works for other distributions of Y than normal or Bernoulli.

A drawback of the permutation method is that it is hard to demonstrate low p-values. Showing that a p-value is lower than 10^{-7} for example, needs at least 10^7 permutations. Often if the sample size is small, the total number of permutations is not large enough to attain very low significance levels. The minimum sample size needed to attain $\alpha = 0.05$ can be calculated as 2×4 samples if Y takes two values and five samples if Y is continuous. The permutation method is illustrated in Section 9.

It is important to note that using permutations one calculates the distribution of Q under H_0 conditional on the set of observed outcomes in Y. For Y a group label this means that the sizes of the groups are taken as fixed; for a continuous outcome each value in the observed vector Y is assumed to occur exactly once. Therefore the permutation version is strictly speaking a different test (although asymptotically equivalent). The expectation and variance of Q under the null hypothesis and the p-value can therefore be systematically different, although in practice the difference is usually small except for very small samples.

8 HANDLING MISSING VALUES

Missing values for some genes in the data set are not a problem. If some genes with missing values are too important to be left out of the analysis, the missing values can be handled by simply imputing the mean expression value of the same gene from the other samples (or the K-nearest samples). This allows the matrix $\tilde{R}$ of covariance between the gene expression patterns of the samples to be calculated using all available information. A nice property of this imputation is that genes or samples with many missing values get a small variance and are therefore automatically given less weight in the analysis.

9 APPLICATION: AML/ALL

The first application is the well-known data set by Golub *et al.* (1999). These data were collected for the purpose of distinguishing between AML and ALL on the basis of gene expression.

There were microarray data of 7129 genes from 27 ALL and 11 AML patients. A pre-selection of genes was made in the same manner as in earlier publications on this data set (Golub *et al.*, 1999; Eilers *et al.*, 2001), truncating very high and very low expression levels and removing genes whose truncated expression showed no variation. This left 3571 genes. There were no missing values.

This data set will be used here to illustrate the use of the score test on all genes. The null hypothesis to be tested here is whether AML and ALL patients are different with respect to their overall gene expression pattern.

9.1 Test result

The ALL patients were coded 0 and the AML patients 1. Now $\hat{\mu} = 11/38$, which was used to calculate

$$Q \approx 13.2.$$

Under the null hypothesis H_0 the distribution has $E(Q) \approx 2.88$ and s. e. $(Q) \approx 0.78$, calculated using (6). This results in a rejection of H_0 with a p-value $\approx 1.6 \times 10^{-14}$, calculated on the $c\chi^2_\nu$-distribution with $c \approx 0.11$ and $\nu \approx 27.0$.

This shows that AML and ALL patients do indeed differ enormously with respect to their overall gene expression signature. The extremely low p-value here can be seen as an explanation why many people using many different methods have been able to find good discriminating rules between AML and ALL on the basis of these data.

9.2 The permutation method

Because the p-value is so extreme, it is prudent to check the distribution of Q empirically. We do this by randomly taking 100 000 permutations of the vector Y of outcomes, calculating Q and making a histogram. The result is given in Figure 1, with the observed value of Q in the real data set indicated by an arrow. The empirical mean and standard deviation are $\bar{Q} \approx 2.96$ and s.e. $(Q) \approx 0.80$, which are not very far from the theoretical values.

The empirical p-value is the number of times the Q for the permuted Y is as least as large as the 'true' Q, divided by the number of permutations. In principle, because there are about 3.3×10^{29} possible permutations of Y, this can be calculated to almost any desired accuracy. But taking only 10^5 permutations (about 10 s on a normal computer) we can only say that the p-value is most probably below 10^{-5}, although Figure 1 suggests that it is much lower than that.

9.3 The regression and checkerboard plots

It has already been explained using (5) that the test statistic Q evaluates the resemblance between the covariance between the gene expressions of all pairs of samples and the covariance between their clinical outcomes. This comparison might also be done by inspection. Figure 2 is an image of the symmetric matrix $\tilde{R}$, with white denoting that an entry is larger than the median off-diagonal element and black that it is smaller.

From this image it is easy to recognize that the true outcomes Y had been sorted, starting with the 27 ALL patients and continuing with the 11 AML patients. The block-like structure of the matrix $\tilde{R}$ strongly resembles the block structure of the covariance matrix between the outcomes Y. This can be used as an illustration of the low p-value that was found.

This method of visualization works best when the outcome is a group indicator. For continuous outcomes, two images of $\tilde{R}$ and $S = (Y - \hat{\mu})(Y - \hat{\mu})'$ might be placed side by side for comparison, perhaps with the samples sorted by their outcomes to simplify the structure of the two matrices. In that case a multi-color plot might be preferred over a black and white one.

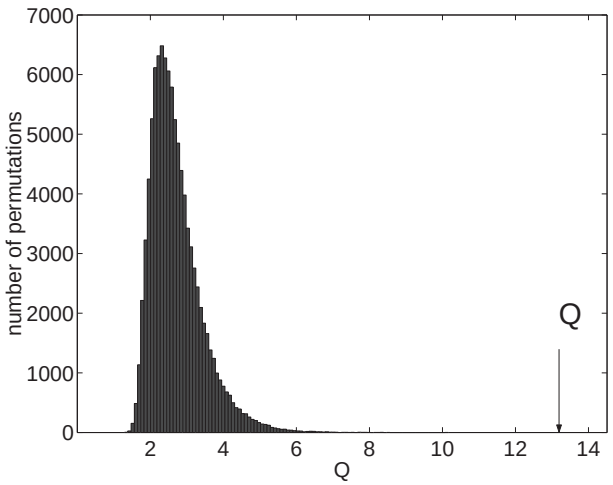

Fig. 1. Histogram of values of the test statistic Q for 100 000 permutations of Y, compared with the observed value.

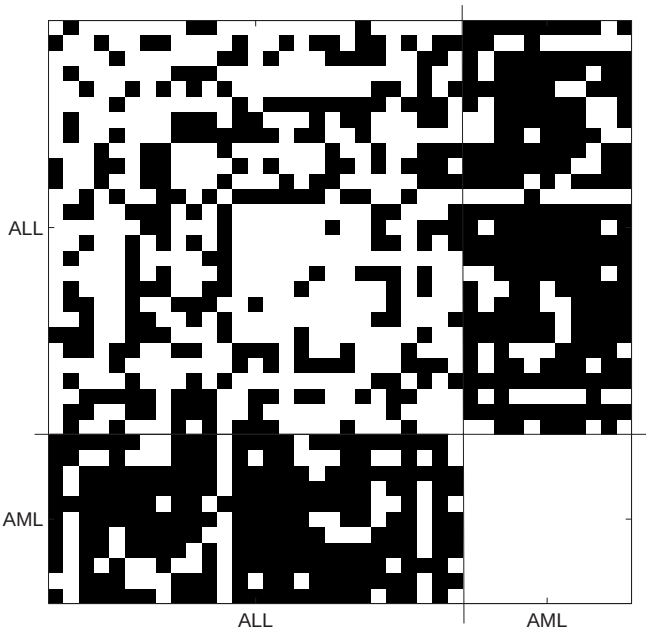

Fig. 2. Checkerboard plot for the AML/ALL data set, showing the matrix $\tilde{R}$ of covariance between the gene expressions of all pairs of samples. White = above median; black = below median.

Some interesting things can be learned from the plot in Figure 2. In the first place it can be seen from the image that the AML group is much more homogeneous than the ALL group. Another thing that can be noticed is that some arrays do not seem to fit very well into the block-like structure. The ALL arrays #2 and #12 for example (2nd and 12th row/column) seem at least as similar to the AML group as to the ALL group.

These arrays could have been wrongly classified or be of poor quality.

A second way of visualizing the test is by plotting the off-diagonal entries of R against those of $S = (Y - \hat{\mu})(Y - \hat{\mu})'$. This is a way of representing Q, because Q is proportional to the covariance between the plotted entries and can therefore be represented by the slope of the regression line of the off-diagonal entries of R on those of S. This type of plot is also very useful when the outcome Y is continuous.

For the AML/ALL data set, the plot shown in Figure 3. Because Y takes only the values 0 and 1, the matrix S takes only three values. From left to right on the x-axis, these are ALL versus AML, ALL versus ALL and AML versus AML. The AML/AML comparisons have a higher covariance between outcomes than the ALL/ALL comparisons because there are fewer AML (so that $Y_i - \hat{\mu} = 27/38$ for the AML and $Y_i - \hat{\mu} = -11/38$ for the ALL). The large value of Q is seen from the steep slope of the regression line.

Using this type of plot the possibly outlying arrays can be investigated further. In Figure 4 all points corresponding to pairs of arrays that involve array #12 have been replaced by crosses. An extra dotted regression line is drawn for reference, which is the least squares fit only through the crosses. This way it can be seen that ALL array #12 actually resembles the AML arrays better than it resembles the other ALL arrays. This is not suggestive of bad data quality (in which case #12 would resemble none of the arrays very well) so it either indicates a misclassification of #12, or perhaps it might be that ALL is quite diverse and some forms are genetically closer to AML.

10 APPLICATION: HEAT SHOCK

The second data set contains six replicates each of a cell line treated with a heat shock (hs+) and untreated (hs−). These samples were labelled with two different fluorescent dyes and cohybridized in hs+/hs− pairs on six spotted oligonucleotide microarrays containing 20 160 genes. Normalization on the 12 samples was carried out using the variance stabilizing method VSN (Huber *et al.*, 2002).

In this data set two groups of genes were of specific interest. One was a group of 27 genes which were classified for biological process as heat shock response genes by the Gene Ontology Consortium (http://www.geneontology.org). Another group of 17 genes belonged to different biological processes but their gene names referred to heat shock.

The test on the total group of all 20 160 genes gave a non-significant result ($p = 0.94$). Looking at all genes, it could not be proved that any gene was affected: the overall expression pattern was not notably different between the hs+ and hs− groups.

However, using the global test on the selected genes gave a different picture. The global test on the 27 genes known to function in heat shock response had an empirical p-value of 0.017. The expression pattern of this group of genes was

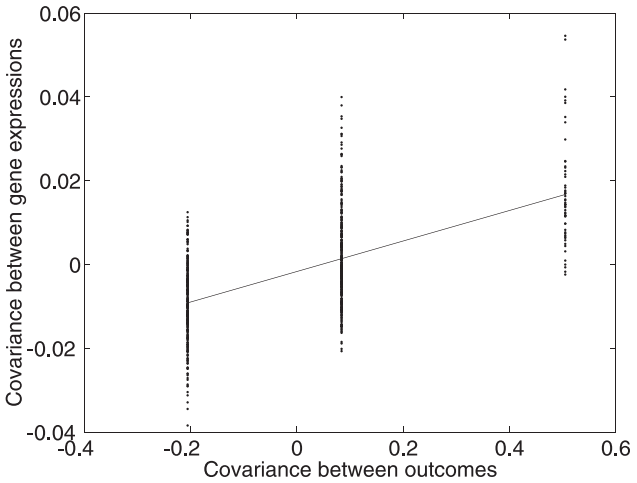

Fig. 3. Regression plot I: visualization of Q as a regression between off-diagonal entries of S and $\tilde{R}$.

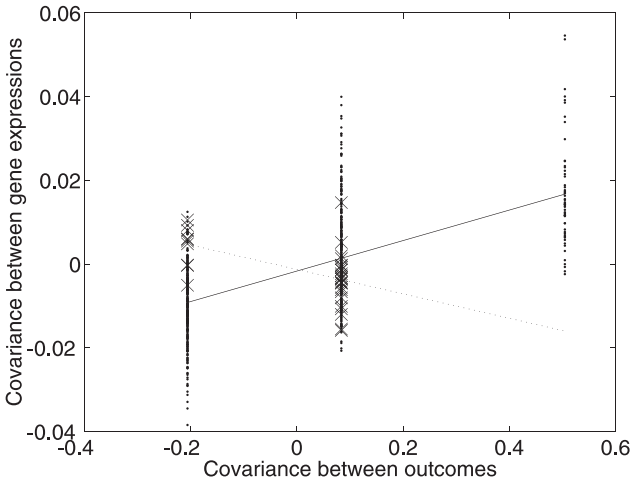

Fig. 4. Regression plot II: visualization of Q as a regression between off-diagonal entries of S and $\tilde{R}$. Crosses involve array #12.

therefore different between the two experimental conditions. The other group of 17 genes with heat shock' in the name only had a non-significant p-value of 0.25.

As an informal comparison, we did an analysis using SAM (Tusher *et al.*, 2001). On the optimal false discovery rate which was 11%, we could find only a small set of nine differentially expressed genes. This set contained a single gene from the group of 27 heat shock genes (no. 31 in Fig. 5).

10.1 A gene diagnostics plot

When testing a small group of genes for differential expression of the group, it is often interesting to look at the single genes, even if the group is the main focus of interest. A group of genes can yield a significant test result because a few genes are very much differentially expressed or because most genes

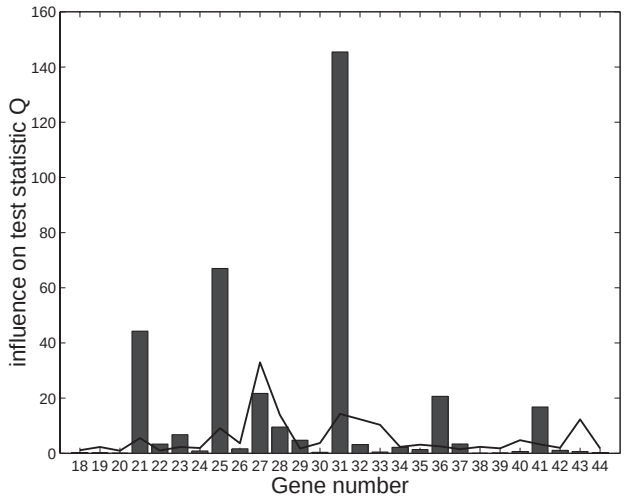

Fig. 5. Gene influence plot for the heat shock data. High bars indicate influential genes. Reference line is the expected influence under the null hypothesis.

are a little differentially expressed. This can be an interesting biological difference. In other cases single genes within the group may be of interest in themselves.

The influence of single genes on the test result can be evaluated in a gene influence plot, as shown for the group of 27 genes in Figure 5. The bars in the figure indicate the values of Q_i for each gene, the value of the test statistic if the group only consisted of this gene. A line is drawn for reference to indicate the expected length of the bar under the null hypothesis.

The interpretation of the figure is that it can be seen which genes contribute positively to a high value of the test statistic and which do not contribute. The difference in expected contribution arises because genes which have greater variance among all arrays are naturally expected to also have a greater discriminating power. In this data set we can see that really only a minority of five or six genes out of 27 is clearly above the reference line and that the majority of the genes do not show any effect. The biological interpretation of this observation, however, is beyond the scope of this paper.

11 DISCUSSION

The test presented in this paper is a useful new tool for the analysis of microarray data. It allows researchers to use prior information on groupings of genes and to specifically investigate groups of genes that interest them from a biological point of view.

In cases where there is a single candidate group of interest, the global test opens the door to real inference: testing hypotheses about biological mechanisms based on theory or past research. In other cases, when researchers have many candidate pathways, available e.g. from gene ontology databases (http://www.geneontology.org) or programs like GenMAPP (http://www.genmapp.org), the global test can be used to find promising pathways. Alternatively the clusters from a cluster analysis can be assigned a p-value to mark how much the genes are coregulated with the clinical outcome.

Test results for groups of different sizes are fully comparable. However, when many groups of genes are to be tested, multiple testing procedures come back into play (Benjamini and Hochberg, 1995). Nested groups may be tested without adjustments to the α-level. Always keep in mind that groups of genes may never be chosen with reference to the clinical outcome.

Furthermore using the test on all genes could be a useful preliminary data quality check. If the test is not significant, samples with a similar clinical outcomes do not have very similar gene expression patterns. In this case it is unlikely that there are many genes highly differentially expressed and it is unlikely that a good classification rule can be found on the basis of all genes. Because of the close connection of the global test to penalized regression methods, the p-value that results from the test can be used as a quality label for the classification rule found with these methods.

REFERENCES

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

Le Cessie,S. and Van Houwelingen,H.C. (1992) Ridge estimators in logistic regression. *Appl. Stat.*, **41**, 191–201.

Le Cessie,S. and Van Houwelingen,H.C. (1995) Testing the fit of regression models via score tests in random effects models. *Biometrics*, **51**, 600–614.

Houwing-Duistermaat,J.J., Derkx,B.H.F., Rosendaal,F.R. and Van Houwelingen,H.C. (1995) Testing familial aggregation. *Biometrics*, **51**, 1292–1301.

Eilers,P.H.C., Boer,J.M., Van Ommen,G.J.B. and Van Houwelingen,H.C. (2001) Classification of microarray data with penalized logistic regression. *Proceedings of SPIE Volume 4266: Progress in Biomedical Optics and Imaging*, **2**, 187–198.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfeld,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.

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

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*, Chapman and Hall, Boca Raton, USA.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B*, **58**, 267–288.

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