BIOINFORMATICS

Vol. 20 no. 3 2004, pages 357–363
DOI: 10.1093/bioinformatics/btg417

Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions *in* Mycobacterium bovis

Zoltan Kutalik[1], Jacqueline Inwald[2], Steve V. Gordon[2], R. Glyn Hewinson[2], Philip Butcher[3], Jason Hinds[3], Kwang-Hyun Cho[4] and Olaf Wolkenhauer[5,]*

[1]Control Systems Centre, Department of Electrical Engineering and Electronics, P.O. Box 88, UMIST, Manchester M60 1QD, UK, [2]Veterinary Laboratories Agency, Woodham Lane, New Haw, Addlestone, Surrey KT 15 3NB, UK, [3]Bacterial Microarray Group, St George's Hospital Medical School, London SW17 0RE, UK, [4]School of Electrical Engineering, University of Ulsan, Ulsan 680-749, South Korea and [5]Department of Computer Science, University of Rostock, Albert Einstein Str. 21, 18059 Rostock, Germany.

Received on February 4, 2003; revised on May 26, 2003; accepted on July 22, 2003

ABSTRACT

Motivation: When analyzing microarray data, non-biological variation introduces uncertainty in the analysis and interpretation. In this paper we focus on the validation of significant differences in gene expression levels, or normalized channel intensity levels with respect to different experimental conditions and with replicated measurements. A myriad of methods have been proposed to study differences in gene expression levels and to assign significance values as a measure of confidence. In this paper we compare several methods, including SAM, regularized *t*-test, mixture modeling, Wilk's lambda score and variance stabilization. From this comparison we developed a weighted resampling approach and applied it to gene deletions in *Mycobacterium bovis*.

Results: We discuss the assumptions, model structure, computational complexity and applicability to microarray data. The results of our study justified the theoretical basis of the weighted resampling approach, which clearly outperforms the others.

Availability: Algorithms were implemented using the statistical programming language *R* and available on the author's web-page.

Contact: wolkenhauer@informatik.uni-rostock.de

Supplementary information: For additional material see http://www.sbi.uni-rostock.de/

*To whom correspondence should be addressed.

1 INTRODUCTION

Among the various reasons to conduct microarray experiments we may identify the following key objectives: comparison of different gene expression levels in varying conditions; identification of functionally related genes, discriminating samples through the clustering of gene expression profiles; and the unravelling of gene interaction networks from time series data. Here, we are interested in microarray data with at least three replicates under two different 'conditions'. These conditions can either be normalized channel intensities, or gene expression levels at two different time points, comparing cell lines or strains, or generally 'before' versus 'after' or 'control' versus 'treatment' comparative experiments. Since in Section 4 we apply this technique to gene deletion studies, we refer to 'normalized channel intensities' when gDNA was hybridized against gDNA, instead of using the term 'gene expression'. The aim is to reliably identify genes with significant differences in gene expression between the two conditions. This problem is non-trivial due to uncertainties arising from various sources of non-biological variation during experimentation, measurement and data pre-processing. For differences in expression levels the use of 'fold changes' is unreliable and a statistical analysis is required to distinguish true changes from random variations and to assign significance values to differences.

The data set we analysed was based on comparative genomic experiments between strains of *Mycobacterium bovis*, the agent of bovine tuberculosis, and the human pathogen *M.tuberculosis*. These organisms share >99.9% identity at the nucleotide level, but this is offset by a range of gene deletions

from *M.bovis* (Gordon *et al*., 2001). It is still unclear whether these deletions confer any phenotype on *M.bovis*, or whether the distribution of deletions differs across *M.bovis* isolates. This would have major implications for strain evolution. Furthermore, this data set should in theory have been simple, in that we were comparing the presence or absence of genes across strains. However, as we shall show, even this relatively simple data set requires a robust statistical approach to ensure the validity of the results.

An important step in pre-processing microarray data is 'normalization' (Kepler *et al*., 2002; Yang *et al*., 2002; Quackenbush, 2002), which ensures that data from different arrays are comparable. For the significance analysis investigated in this paper, we assume that data were normalized before we investigate differences in normalized channel intensities of the genes.

Various statistical approaches are available to test whether two samples are drawn from the same population or distribution. Although non-parametric tests such as Wilcoxon's 2-sample test are applied extensively to microarray data, these are out of the scope of the present study. On the other hand, parametric tests applied to microarray data are based on the t-test at one stage or another. They can be further classified according to whether they use assumptions about gene-to-gene interactions. Regression models and mixture models require prior assumptions that some calculated quantities have special, implicit relationships (e.g. expression level and standard deviation of genes) (Baldi and Long, 2001; Huber *et al*., 2002). The other type of tools, such as the simple t-test, concentrate only on individual genes. We are going to demonstrate that these methods, concentrating on individual genes, perform poorer than methods based on regression.

In this paper we compare several methods including SAM (Tusher *et al*., 2001; Efron *et al*., 2000, http://www.stat-standford.edu/~tibs/ftp/microarrays.pdf), Bayesian regularized t-test (Baldi and Long, 2001), mixture modeling (Pan, 2002) (analyzed only theoretically), Wilk's lambda score (Hwang *et al*., 2002) and variance stabilization (Huber *et al*., 2002). We discuss implicit assumptions, the model structure, computational complexity and applicability to microarray data. From this comparison, we developed a weighted resampling approach and applied it to the study of differences in channel intensities as the result of gene deletions in *M.bovis*. The approach is derived from ideas in SAM and the regularized t-test and employs weighted resampling. We introduced a new way of replicate handling to grasp the reliability of the replicates in two steps. First, this approach detects particular outliers gene-by-gene (which reflects the local disturbance on the array), second, it weights every resampled group according to the probability of containing an outlier. The results of our comparison justified the theoretical basis of this approach, which outperforms previously published algorithms. All the methods are not restricted to specific microarray experiments, thus can be applied to any comparative experiment.

The outline of this paper is as follows. In Section 2, we provide a short summary of the methods used in the comparative study. Section 3 introduces our weighted resampling approach to identify significant differences in gene expression levels (or in normalized channel intensities). This is followed by a comparison of our approach to the others with an application to data sets obtained for gene deletions in *M.bovis*. The summary of the results and discussion are in Sections 4 and 5, respectively. In the supplementary material (http://www.sbi.uni-rostock.de/) we prove the equality of variance stabilization with a special case of the modified t-test. The web-link also includes the proof of the equivalence of Wilk's lambda score (for two-class comparisons) with the simple t-test. Note that in this paper we call two statistical tests equivalent if their test statistics are monotone function of each other, thus they give the same significance ranking for the genes. This does not necessarily mean that they pick the same group of genes at each confidence level.

2 METHODS

In this paper experimental data are assumed to be normalized and divided into two sub-groups. The process of normalization should be tailored to the particular experiment and we refer to the literature for a survey of methods (Kepler *et al*., 2002; Yang *et al*., 2002; Quackenbush, 2002). For our *M.bovis* data set, we used a simple three-step normalization procedure for two-color DNA microarrays. This procedure consisted of background subtraction, Lowess normalization and finally across replicate normalization.

Hereafter $x(g) = \{x_i(g)\}_{i=1}^{n}$ refers to the sub-group 'before treatment' and $y(g) = \{y_i(g)\}_{i=1}^{m}$ to the 'after' sub-group, where $g = 1, \ldots, G$ indexes individual genes, n and m are the number of replicates in each sub-group and G denotes the total number of genes.

For the t-test, the following notation is used for replicate averages and variances

$$\bar{x}(g) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i(g), \quad \bar{y}(g) = \frac{1}{m} \cdot \sum_{i=1}^{m} y_i(g)$$

$$s_x^2(g) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left[x_i(g) - \bar{x}(g)\right]^2,$$

$$s_y^2(g) = \frac{1}{m-1} \cdot \sum_{i=1}^{m} \left[y_i(g) - \bar{y}(g)\right]^2.$$

Now the standard t-statistic for gene g is given by

$$d(g) = \frac{\bar{x}(g) - \bar{y}(g)}{s(g)}, \tag{1}$$

where $s(g)$ is the sample estimate for the standard deviation, i.e.

$$s(g) = \sqrt{s_x^2(g)/n + s_y^2(g)/m}. \tag{2}$$

Under normality assumptions (i.e. individual gene measurements follow a normal distribution) $d(g)$ follows a students t-distribution. In the following, we briefly summarize commonly used methods for significance analysis applied to microarray data.

2.1 Regularized t-test

Because the denominator of the t-statistic in (1) uses an estimate s^2 for the real variance σ^2, the t-test is too conservative for a decreasing sample size compared to its regularized modification (Baldi and Long, 2001).

Baldi and Long (2001) followed the Bayesian approach, which resulted in the following modified t-statistic

$$t^* = \frac{\bar{x} - \bar{y}}{s^*}, \tag{3}$$

where

$$s^* = \sqrt{\frac{(n-1)\left(s_x^2/n + s_y^2/m\right) + \nu_0\sigma_0^2}{\nu_0 + n - 1}}. \tag{4}$$

Although not immediately apparent, this approach is very informative. Namely, ν_0 refers to the strength of prior belief about σ_0^2. Note that the parameter estimates suggest that one should use the weighted average of *prior knowledge* and *empirical estimate*.

By looking for connections between parameters, for example, mean expression level, $(\bar{x} + \bar{y})/2$, and empirical standard variation, s, we can establish regression models to define a functional relationship between those parameters. Baldi and Long (2001) introduced 'local' regression models. These regression estimators can be used for σ_0. Other models also exploit regression models for variance. One such approach is variance stabilization.

2.2 Variance stabilization method

The variance stabilization method is described well in Huber *et al*. (2002). The approach is based on quadratic regression, modeling the variance against the mean of untransformed but normalized microarray data. Referring to this regression function denoted by v, in Tibshirani (1988) it was shown that the best transformation function to reduce the unevenness of variance across the sample is

$$h(y) = \int_{-\infty}^{y} 1/\sqrt{v(u)}\, du. \tag{5}$$

For further details refer to Huber *et al*. (2002). Note that in the case of $v(u) = u^2$ this leads to $h(y) = \log(y)$, which is the usual transformation used for microarray data. The measure of significance is defined as the difference of transformed values. In case the dye effect is removed properly by the pre-normalization, the transformation functions v will agree between the Cy3 and Cy5 channels. Suppose now that the functions v do not differ too much and therefore Cy3 and Cy5 channels assign a common function v (and therefore a common function h). In the supplementary material we show that in case of common variance regression functions in both channels, the proposed measure of significance is equivalent to a special case of the regularized t-test. However, when it is difficult to remove the dye effect suitably the use of different h functions for each channel can deviate considerably from the t-test (1). Another equivalence to the t-test is the Wilk's lambda score as will be shown in the following section.

2.3 Wilk's lambda scoring

Wilk's lambda score is a measure of significance and the basis of the 'leave one out' cross-validation method described in Hwang *et al*. (2002). Their procedure iteratively builds test- and training sets of genes and computes the corresponding misclassification error rate. The final set of discriminatory genes will build up from the genes with the top Wilk's lambda score. The number of discriminatory genes is decided by the misclassification error rate function. In the supplementary material, we prove that the Wilk's lambda score (for two-class comparison) is equivalent to the t-statistic.

2.4 SAM

A popular method for significance analysis, called SAM (Tusher *et al*., 2001), based on a modified version of the t-statistic (1), is defined as follows:

$$d_{\mathrm{SAM}}(g) = \frac{\bar{x}(g) - \bar{y}(g)}{s(g) + s_0}, \tag{6}$$

where s_0 is defined to minimize the difference in the coefficient of variation of d_{SAM} within classes of genes with approximately equal variance (Chu *et al*., 2001, http://www.stanford.edu/~wanjen/Chu%20Lab/Papers/sam.pdf). A drawback of calculating s_0 is the computational cost. As will be shown in Section 4, using a simpler regularized t-test (Tibshirani *et al*., 2002) one obtains very similar results.

Assuming that n and m are even numbers, the algorithm computes all $N = \binom{n}{n/2}\binom{m}{m/2}$ possible regroupings of the union of the two groups x and y. Regroupings where each of the two groups contain an equal number of elements from x and y are indexed with p. For a particular permutation p, denote the two groups with $p1$ and $p2$; the 'null score' is defined by

$$d_p(g) = \frac{\bar{x}_{p1}(g) - \bar{y}_{p2}(g)}{s(g) + s_0}. \tag{7}$$

The method uses these values to mimic the null statistics. To save space, we do not go into the details of the method, the reader is referred to Chu *et al*. (2001) for more details.

2.5 Mixture modeling

Pan (2002) used the principle idea of the SAM method (Chu *et al*., 2001; Efron *et al*., 2000), and computed only one randomly selected $d_p(g)$ value from all possibilities (for each gene). As these $d_p(g)$ null-values (for all genes) are drawn

WEIGHTED RESAMPLING

STEP 1: Regrouped t-Statistics
(a) $t_{g,p,q} = \frac{\sum_{i=1}^{k-1} x_{q_i}(g)/(k-1) - \sum_{j=1}^{l-l} y_{p_j}(g)/(l-1)}{s_{g,p,q}+s_0}$ for all $\mathbf{p},\mathbf{q} \in \{h : h \subseteq \{1,\ldots,l\}, |h| = k-1\}$
and for all genes $g = 1,\ldots,G$.
(b) $s_{g,p,q}$ is the empirical estimate for the variation of $\{x_1(g),\ldots,x_{q_{k-1}}(g), y_{p_1}(g),\ldots,y_{p_{k-1}}(g)\}$.
(c) $s_0 = \frac{\sum_{g,p,q} s_{g,p,q}}{k \cdot l \cdot G}$.

STEP 2: Weighting Replicates
(a) Fix the threshold δ
(b) $s_g = \frac{\sum_{p,q} s_{g,p,q}}{k^2}$.
(c) Define $dev_{g,p,q} = \frac{s_{g,p,q} - s_g}{s_g}$.
(d) If $dev_{g,p,q} < \delta$ then $w_{g,p,q} = [k \cdot l/2]$, otherwise $w_{g,p,q} = 1$.
(e) $w^*_{g,p,q} = \frac{w_{g,p,q}}{\sum_{p,q} w_{g,p,q}}$.

STEP 3: Final Score
$z_g = \sum_{p,q} t_{g,p,q} w^*_{g,p,q}$.

Fig. 1. Detailed description of the proposed WR method.

from the same null distribution they can be pooled together to have an estimate for the null density function, f_0. Approximating the distribution by a mixture of well-known distributions is not well justified since it imposes strong assumptions on the tail of the distribution. [Using the statistical language *R* we used several functions for the approximation (incl. simple smoother, normal, mixture-normal) but noticed that they varied widely.] Pan (2002) suggested creating a pooled empirical density function (called f), where the histogram is built up from the original $d(g)$ values. The final score of significance is therefore f_0/f. The smaller the value is, the more probable it is that a significant change is detected.

3 WR ALGORITHM

All of the previously described methods treat replicates with the same weight. Here, we propose an approach that combines the advantages from the aforementioned methods. The SAM method (Tusher *et al*., 2001; Chu *et al*., 2001; Efron *et al*., 2000) approached the replicate handling through randomized resampling while we suggest the following two-step approach. Our weighted resampling (WR) approach is based on the regularized t-test (Baldi and Long, 2001) as in SAM. We modeled the replicates' reliability by measuring local disturbance (individual genes) through outlier detection.

Using the replicates we create regularized t-statistics of smaller size. To make the rest of the procedure clear we will explain each step using an example. For instance, given a '4-against-4' experiment. We create all (16) possible '3-against-3' t-statistics as these are the largest t-statistics one can obtain without including all the replicates from the control and treatment group. Here, the prior SD σ_0 in (4) is determined as the mean of all estimated SD. The strength of the prior ν_0 in (4) is set to 0.7 based on the formula $(10 - n + 1)$ as suggested in Baldi and Long (2001). Once we have these regularized t-values (3), we should assign weights to them. The weights depend on the variance of the actual sample (in our example we have to compute all 16 3-against-3 variances). Before we continue, we test the variances for 'outliers'. We call a variance a *negative outlier* if its relative deviation from the mean variance (of all the possible sampling of one particular gene) is smaller than a fixed threshold δ. In our application δ was set to -0.5 to pick up the top third of the most reliable replicates across all genes. This means that a particular grouping exhibiting significantly low variation suggests that the replicate left out from that resampled group may be unreliable. The threshold can be altered according to the data set (e.g. to produce only a certain number of groupings with significantly low variation among genes). We found that this does not play an important role for the results in our application. Once we have chosen the groupings with greater importance, they are assigned a multiple weight (here we used the half of total number of regroupings, i.e. 8); a single weight is given to all other groupings. It does not make significant difference in the results if we use a more sophisticated weighting. The idea behind this weighting is that in the case of *particular genes* some replicates can be rather distorted or noisy as detected by the relative group variation. However, the same replicate can *in general* exhibit good correlation properties. This regular phenomenon has escaped the scope of previous studies. Thus we would like to give more weight to those groupings where the replicates are *more consistent in that particular gene's measurement* than others. Of course, eventually the weights have to be normalized to sum up to one for each gene. This weighted sum for each gene will represent its new score of significance denoted z_g. The approach described here is summarized in Figure 1. This weighting scheme seems to

be a good compromise between the t-test and more robust statistics, like

$$\frac{\text{median}(x) - \text{median}(y)}{\text{MAD}}$$

where MAD is the median SD from the median. The t-test is sensitive to outliers, which can completely impair the results. On the other hand, the mentioned robust methods are not sensitive enough to detect if some of the replicates has changed. For instance, in case of '4-against-4' data the largest (and the smallest) measurement is not taken into account at all, as long as its value is greater than the second largest replicate.

4 IMPLEMENTATION

DNA microarray technology allows the large-scale analysis of whole genomes for comparative genomics. Using this technology we can therefore rapidly screen the genomes of *M.bovis* strains for deletions, using an *M.tuberculosis* H37Rv array and exploiting the >99.9% sequence identity at the nucleotide level between the two bacilli. For further details of the experiment see the supplementary material: http://www.sbi.uni-rostock.de/. Ideally the microarray signal intensities should differ significantly in case of gene deletion. We carried out replicated measurement to boost the reliability of the results. Four replicates were used in our experiment and 3852 genes passed the quality control pre-processing. We applied the simple t-test, regularized t-test (Baldi and Long, 2001), Wilk's-lambda score (Hwang *et al*., 2002), two different SAM methods (Tusher *et al*., 2001; Efron *et al*., 2000), the mixture modeling approach used by Pan (2002) and the proposed WR method. RD deletion regions in different types of *M.bovis* are extensively explored and many of them are verified by PCR results, making these data suitable for a comparative study (Gordon *et al*., 2001). Finally, we applied all the implemented methods—including our proposed method—to the *M.bovis* data set.

Although the similarity of the two genomes is about 99%, since not the entire gene sets were used, we were looking for the top 5% of the genes with most significantly changed channel intensities since these are the most plausible candidates for deletions. Using PCR it is known that deleted RD regions consist of 74 genes (Gordon *et al*., 2001), one of which had to be excluded during normalization, as a consequence of high background noise in the data. Note that only the remaining 73 genes were examined in the further analysis. The SAM method (Efron *et al*., 2000) recognized 92%, while the regularized t-test (Baldi and Long, 2001) identified more than 90% of the proven deletions. Finally, the proposed WR method identified 92% of gene deletions. It should be mentioned that the distortion of the measurements sometimes was strong due to cross-hybridization, gene homologies and other source of non-biological noise (Dorrel *et al*., 2001). As a result, two of the 73 genes (3%) showed positive difference in channel intensities (i.e. the test DNA channel produced higher intensity than the control strain). These gene deletions could not be identified by any statistical tools. Which means that the proposed WR method missed only three genes (4%), which might have been detected somehow else. One of these three genes were among the top 10% of genes with most significant differences in channel intensities, thus just have missed to make it. The complete results for all proportion between 0–10% of the genome can be seen in Figure 2.

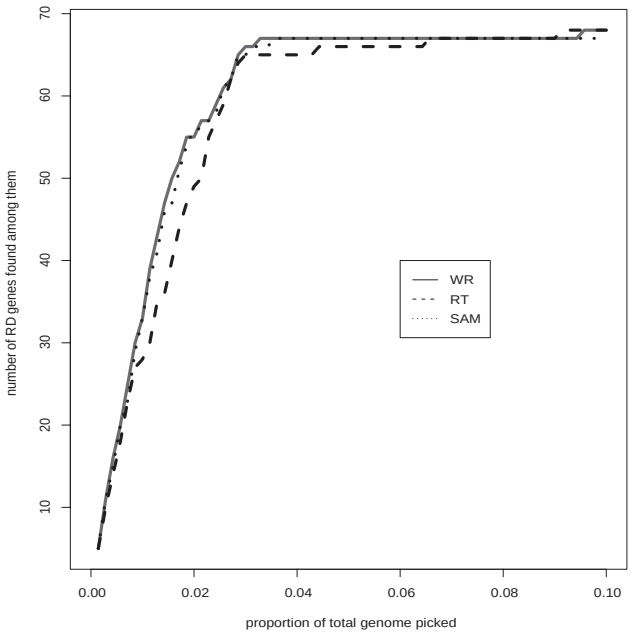

Fig. 2. The proportion of the selected genome is plotted against the ratio of picked up RD genes among them.

Other important approach is when not using the information about the *M.bovis* genome (namely that only 5% of the genes can be deleted) and controlling the false discovery rate (FDR). For SAM method we used their approach to control the FDR as described in Chu *et al*. (2001). There is a wide range of alternative techniques (Storey and Tibshirani, 2001; Yekutieli and Benjamini, 1999; Reiner *et al*., 2003; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Storey, 2001, http://faculty.washington.edu/~jstorey/papers/dep.pdf), which can be used to control the FDR. For the RT method and our WR approach we picked and implemented a less conservative approach (Storey and Tibshirani, 2001). The results are summarized in Table 1. In this context SAM method picked 55 RD genes, our WR method found 57 RD genes at 5% confidence level, while the regularized t-test identified 50 genes in the RD regions. The overall comparison of the methods can be seen in the receiver operating characteristic (ROC) curve (Fig. 3), where the FDR was plotted against the true positive rate method by method. One disadvantage of SAM is that it changes unevenly and cannot reflect minor changes. However

Table 1. Detected gene deletions comparing the different approaches

	WR (%)	RT (%)	SAM (%)
% of RD genes with FDR <5%	78	68	75
% of RD Genes in the top 5%	92	90	92

The first row represents the proportion of identified RD genes when the FDR is controlled to be <5% according to different methods. The second row shows the proportion of RD genes within the top 5% according to their significance scores. RT refers to the regularized t-test.

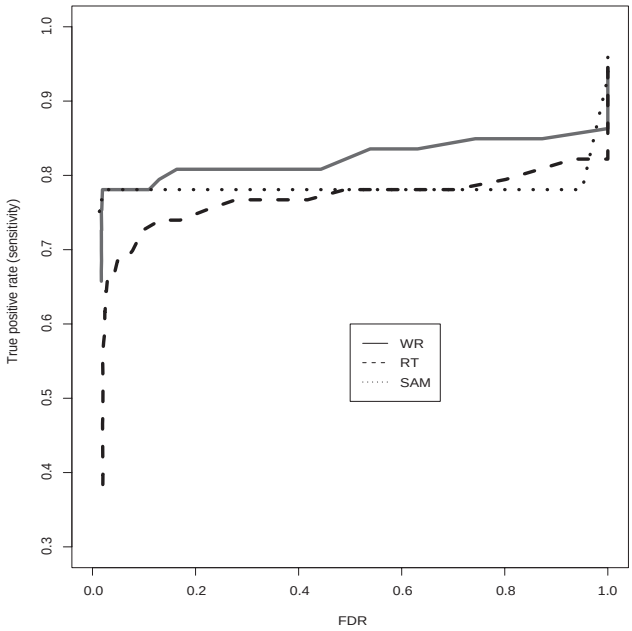

Fig. 3. Receiver Operating Characteristic curve (ROC curve) of the different methods applied to *M.bovis*.

for the lowest FDRs it performs inarguably well. The advantage of the WR method is that it is a globally reliable method either combined with FDR control, or using simply the top z% of the significance values as its ROC curve clearly dominates that of the other methods.

Figure 4 shows the deletions (with dark dots) on the genome suggested by the analyzed methods. The outer circle refers to WR algorithm suggested here, the next inner ring is the SAM (Efron *et al.*, 2000), and third is regularized t-test (Baldi and Long, 2001).

5 DISCUSSION

We compared methods for significance analysis theoretically, and through the application to experimental data. While a theoretical comparison might suggest differences or the formal equivalence, depending on the experiment conducted there may be practical arguments for and against these methods. In this section, we complete our discussion by looking at the structure and statistical validity of models using the experimental data set introduced in the previous section.

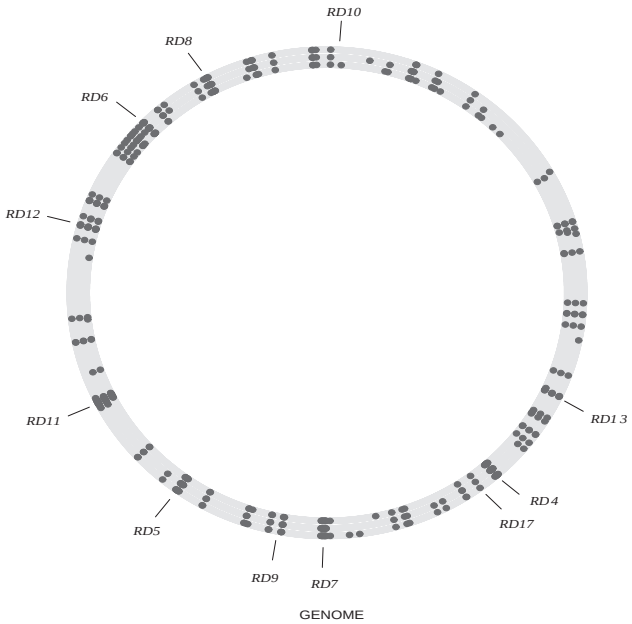

Fig. 4. Genome of *M.bovis* Type 12. Dark dots represent RD deletions.

First, it is obvious that the regularized t-test incorporates the basic t-test as a special case for which we do not intend to use any prior assumptions (i.e. the weight $\nu_0 = 0$). The randomized SAM method is clearly better than the t-test; it takes into consideration not only the actual t-statistics, but can control the *type I* error more accurately by the re-sampling technique. To gain an insight into the comparison of different significance scores we can scatter-plot the significance scores against each other, as shown in Figure 5. Here one can compare how the different methods agree on scoring. If the plot shows a monotone increasing relationship, we can be confident that for that particular data set those two methods of scoring are similar. The more monotone-like the cloud of points is the more the two methods agree. Figure 5 shows how the different methods agree in this sense. Considering expression profiles for thousands of genes, it is important to remember that each gene is a random variable and intensity measurements are therefore not necessarily drawn from a common distribution. To create a histogram for d-values as done in Pan (2002), is therefore questionable, especially if from the purpose or context of the experiments we expect significant differences in channel intensities. In such experiments we should have doubts about Pan's f function where the f_0/f score for significance does not measure exactly what we need. It is irrelevant for us whether the t-statistic [i.e. the d-value (1)] of a significantly changed gene is rare among the t-statistics of the entire gene set. To put this concern into another form, the Neyman–Pearson lemma cannot be used in this case. Function f is not a density function, although in the Neyman-Pearson lemma the alternative hypothesis is that a particular test statistic belongs to an alternative (f_1)

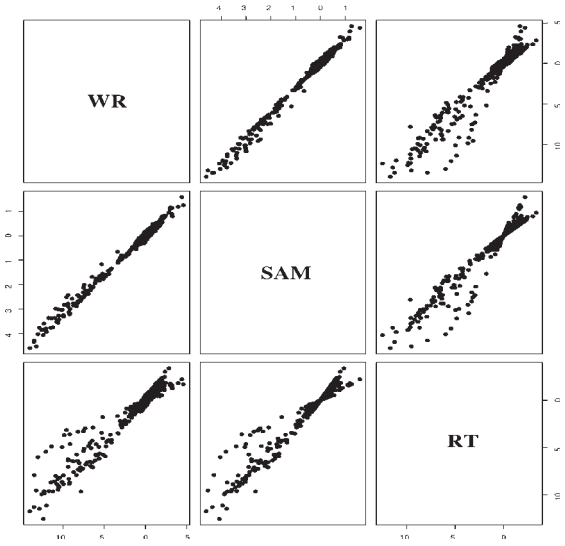

Fig. 5. Scatter plot compares methods for significance analysis applied to a TB data set.

distribution. Here it is not the question of whether a gene belongs to one distribution or another, but whether they are identically or differentially expressed. Identical expression can be translated into belonging to a particular distribution, but differential expression cannot.

A further advantage of our WR method is that due to its reduced set of assumptions, it can easily deal with missing data, odd number of replicates, and uneven number of replicates in the 'treatment' and 'control' groups, unlike SAM. Naturally a major drawback of our proposed method is computational cost. To compare the computational complexity with other methods let us consider G genes with n control and m treatment replicates and C random resampling for the FDR computation, the complexity of the SAM method, the regularized t-test is $\mathcal{O}(C \cdot G)$, while WR needs $\mathcal{O}(C \cdot G \cdot n \cdot m)$ steps. Our WR has by far the largest computational needs. The complexity of all other methods are independent of the sample size. These extended running times are still below 20 min (with a 2.8 GHz Pentium 4 CPU). Furthermore, where high data quality is apparent, scores for all methods coincide. However, in our experience it is still a major challenge to obtain very high-quality microarray data.

ACKNOWLEDGEMENTS

The authors would like to thank Fatima Sanchez-Cabo for her advice on microarray data normalization. This work was funded as a part of the DEFRA project 'Application of post-genomics to *M.bovis*'.

REFERENCES

Baldi,P. and Long,A. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

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

Benjamini,Y. andYekutieli,D. (2001) The control of the false discovery rate under dependency. *Ann. Stat.*, **29**, 1165–1188.

Chu,G., Narasimhan,B., Tibshirani,R., and Tusher,V. (2001) SAM 'Significance Analysis of Microarrays'. *User Guide and Technical Document*.

Dorrel,N., Mangan,J., Laing,K., Hinds,J., Linton,D., Al-Ghusein,H., Barrell,B., Parkhill,J., Stoker,N., Karlyshev,A., Butcher,P. and Wren,B. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using low-cost microarray reveals extensive genetic diversity. *Genome Res.*, **11**, 1706–1715.

Efron,B., Tibshirani,R., Goss,V., and Chu,G. (2000) Microarrays and their use in comparative experiment. *Technical Report*.

Gordon,S., Eiglmeier,K., Garnier,T., Brosch,R., Parkhill,J., Barrell,B., Cole,S. and Hewinson,R. (2001) Genomics of *Mycobacterium bovis*. *Tuberculosis*, **81**, 157–163.

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

Hwang,D., Schmitt,W. and Stephanopoulos,G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**, 1184–1193.

Kepler,T., Crosby,L. and Morgan,K. (2002) Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.*, **3**, research 0037.1–0037.12.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**(suppl.), 496–501.

Reiner,A., Yekutieli,D. and Benjamini,Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Storey,J. (2001) The positive false discovery rate: a bayesian interpretation and the q-value. *Technical Report 2001-12*. Department of Statistics, Stanford University.

Storey,J. and Tibshirani,R. (2001) Estimating false discovery rates under dependence, withapplications to DNA microarrays. *Technical Report 2001-28*. Department of Statistics, Stanford University.

Tibshirani,R. (1988) Estimating transformations for regression via additivity and variance stabilization. *J. Am. Stat. Assoc.*, **83**, 394–405.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

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

Yang,Y., Dudoit,S., Lin,D., Peng,V., Ngai,J. and Speed,T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Planning Inference*, **82**, 171–196.