The effect of replication on gene rankings: a practical comparison of methods for detecting differential expression in microarray experiments

Laura Elo [1,2], Tero Aittokallio [1], Sanna Filén [2], Riitta Lahesmaa [2]

[1] Department of Mathematics, University of Turku, FIN-20014, Turku, Finland
[2] Turku Centre for Biotechnology, FIN-20521, Turku, Finland

Abstract

One of the principal goals of microarray analyses is to rank genes in order of evidence for differential expression under experimental conditions. As microarray experiments are costly and involve mRNA samples that can be difficult to obtain, techniques to estimate in advance the performance of different ranking approaches with respect to replication number are of major importance when designing experiments. We study several recent ranking methods in a practical research setting using real microarray experiments. We consider the stability of the results as the level of replication changes and investigate the effect of different methods on the ranking at each level.

1 Introduction

Gene expression profiling with DNA microarrays has become a widely used method in functional genomics. Typical goal of a microarray experiment is to identify for further studies a set of candidate genes that characterize the differences between various experimental conditions. The identification of candidate genes is traditionally divided into two parts: (i) selection of a statistic that ranks genes in order of evidence for differential expression, and (ii) determination of a critical cut-off value for the ranking statistic, which defines the set of genes that are considered as changed. This work was motivated by practical questions that researchers face when analyzing their microarray experiments: how many biological replicates are needed to achieve reliable results; which method should be used among all the approaches available for ranking the genes; and how much the results can change if some other method is used instead? The main focus here is on ranks of genes instead of their statistics' values, as ultimately the ranking alone determines whether or not a gene is selected for further studies. The problem of inferring statistical significance is of secondary importance if the researcher is interested in following a selected number of top ranking genes regardless of variation in data.

A wide range of approaches has been introduced for detecting differential gene expression in microarray experiments. Early studies typically used the simple fold change approach, which defines a gene as changed if its mean difference between the conditions is larger than a constant cut-off value (typically 2). A frequently used parametric method is the ordinary t-test, which takes into account the variability of expression levels among replicates. As replication numbers are often rather small (typically 2 to 5), however, gene-specific variability can easily be underestimated leading to false positive findings. Alternatively, nonparametric approaches are sometimes preferred, such as Wilcoxon rank sum test and permutation t-test [1]. With a limited number of replicates, however, statements concerning false positive rates based on the discrete distribution obtained from relabeled measurements may become less effective. Besides gene-specific statistics, more advanced approaches have been developed particularly for situations where a large number of genes is tested based on small sample sizes. To compensate for the lack of replicates, several authors have modified the ordinary t-statistic to provide more reliable estimates of gene-specific variability by combining information across genes or arrays [2-7]. Another empirical Bayesian approach identifies differentially expressed genes by calculating posterior odds of change within a mixture model [8].

A standard approach for estimating the required replication number is power calculation. In the context of microarray experiments, however, rigid statistical determination of sufficient sample size

is difficult, as the vital elements of standard power calculations are absent [9]. Black and Doerge [10], for instance, used t-test for power calculation, but they did not consider subject level variation. Yang *et al.* [11] considered the sufficient number of replications necessary to minimize false discovery rate (FDR), but the number of candidate genes must be specified before their power calculation. To our knowledge, only two studies have investigated the trade-off between cost and reliability in a practical comparison setting. Pavlidis *et al.* [12] examined how replication number affects the ability to find genes that meet particular statistical criteria with the ordinary t-test. Kim and Park [7] compared recently their modified t-statistic to the ordinary and penalized t-statistics using replication numbers from one to five. In the present work, we combine the good properties of the previous approaches to give a more comprehensive view of the influence of replication number and ranking method on the detection of differentially expressed genes. We investigate these effects in an in-house microarray experiment monitoring gene expression levels in adult asthma and in two public data sets.

2 Testing procedure

We compared the performance of seven different ranking approaches with respect to replication number: fold change, ordinary t-statistic, three modified versions of t-statistic [3, 4, 6], Wilcoxon rank sum statistic, and a mixture modeling approach [8]. To assess the performance of the ranking methods, we sampled 200 subsets of replicates from the pre-processed expression data at each possible replication level starting from two. In all sampled subsets, genes were ranked in order of evidence for differential expression according to each procedure. The obtained rankings were compared with a proper approximation of the true ranking (explained below) by calculating correlation between the orderings. To punish for inconsistencies in fold change directions, up-regulated genes were separated from down-regulated genes by considering also the sign of change when ordering the genes. No cut-off levels were used for the statistics but the relative importance of the most up- and down-regulated genes was considered by introducing a weight function into Spearman's rank correlation coefficient. The weighted version of the coefficient is defined by

$$\rho_u = \frac{\sum_{g=1}^{G} u_g (r_g - \frac{G(G+1)}{2})(q_g - \frac{G(G+1)}{2})}{\sqrt{\sum_{g=1}^{G} u_g (r_i - \frac{G(G+1)}{2})^2 \sum_{g=1}^{G} u_g (q_i - \frac{G(G+1)}{2})^2}},$$

where (r_g, q_g) denotes the pair of ranks assigned to gene g in the two rankings being compared, $g = 1, \ldots, G$. With weights $u_g = 1$ for all g, this gives the ordinary Spearman's rank correlation coefficient. It can be shown that the correlation equals 1 if the rankings are in perfect agreement and -1 if they are in perfect disagreement. We applied exponential weight function of the form $u_g = \max\{e^{-a(r'_g - 1)}, e^{-a(q'_g - 1)}\}$, where a is a small positive constant that determines how steep the decrease in weight is and (r'_g, q'_g) denotes the pair of ranks assigned to gene g when the signs of expression changes are not taken into account. In the current study, we used the value $a = 0.01$, which essentially weights approximately 200 of the most significant genes. We empirically tested several values of constant a and investigated how it affects the results. As a increases, the overall level of correlations increases. However, it does not dramatically affect the relative performance of the methods.

The first approach to investigate the performance of a ranking method was to assess the stability of its results. The stability of a method was determined by calculating at each replication level the average correlation with the "gold standard" ranking obtained by applying the given method to the whole data set. Although the true ranking of genes was unknown, as always in real microarray experiments, with sufficiently large material, this reference ranking was considered as a close approximation to the true ranking of genes. The second approach to investigate the performance of the different ranking methods was cross-comparison between the methods at each replication level. In the cross-comparison analysis, the effect of different methods on the ranking at a particular replication level was examined in terms of correlation with consensus ranking among the methods. The consensus ranking was constructed as a combination of the individual "gold standard" rankings with a simple summing procedure called Borda Count, which orders the genes on the basis of their average rank across the different rankings [13]. Such a combined "gold standard" was presumed to be more fair and less biased reference than relying on a single method only.

3 Results and discussion

The expression data primarily used for the comparisons were from our cDNA microarray experiment monitoring asthma-related changes in gene expression. Peripheral blood lymphocytes (PBL) were isolated from 22 patients with asthma and atopy and 22 healthy controls. Cells were polarized towards T helper 1 (Th1) and T helper 2 (Th2) subsets using IL-12 and IL-4, respectively, and samples were collected after 48 hours of polarization. Each sample was hybridized once against a pooled reference sample prepared from PBL from 41 donors (Finnish Red Cross). We compared patients with controls separately in Th1 and Th2 conditions (two-sample data), and Th1 and Th2 conditions separately for patients and controls (paired data). To avoid relying on a single experiment and platform only, we repeated the analyses on two public data sets, both of which contained data from Affymetrix oligonucleotide arrays and involved comparing different tumor types. The first data set contained samples from ALL and MLL patients [14], and the second data set consisted of samples from DLBCL and FL patients [15]. The maximum number of replicate samples included was determined by the size of the smaller group and was 20 in the former case and 19 in the latter.

According to the results from the asthma study and the two cancer studies, near maximal effectiveness with respect to both within-method stability and cross-comparisons with consensus ranking is typically obtained with approximately 10 replications, when Wilcoxon statistic or the different t-statistics are applied. In some cases, sufficient performance was reached already at 5 replications with Wilcoxon statistic or the modified t-statistics. The greatest enhancements in effectiveness were usually achieved by increasing the replication number from 2 to 5. On the other hand, increasing the sample size beyond 15 yielded relatively small improvements in any of the cases. In general, ordinary t-statistic and mixture modeling statistic required more replications to produce reliable rankings than Wilcoxon statistic and the modified t-statistics. The behavior of the fold change approach was highly dependent on the data, while the problem of Wilcoxon statistic with small sample sizes was the huge amount of tied rank values. Figure 1 shows the results of stability analysis and cross-comparisons in two cases, one between Th1 and Th2 conditions in asthma patients and the other between asthma patients and controls in Th2 condition. The other comparisons yielded similar results.

Our results also showed that the relative performance of a method may change if the rankings are compared to "gold standard" ranking defined from the whole data set with the particular method or to consensus ranking constructed as a combination of individual "gold standard" rankings from a variety of methods. This can be noted, for instance, by looking at the performance of Wilcoxon statistic and t_E-statistic in the comparison between asthma patients and controls in Th2 condition (Figures 1B and 1D). Because confidence in ranking is strengthened when two or more approaches provide consistent results, such a combined "gold standard" should be more fair reference than relying on a single method. The superiority of a voting method in optimizing the results was proposed also in [7].

Pavlidis *et al.* [12] suggested that with the ordinary t-test 10 to 15 replicates usually produce quite stable results, while stable results are typically not obtained with sample sizes smaller than 5. However, they concentrated only on genes with significant expression changes at FDR of 0.05. Our results showed that stability of the rank order of the most differentially expressed genes can be obtained between 10 to 15 replicates even if no cut-off levels for the statistics are used. We also studied the extent to which the performance of a ranking method at different replication levels depends on the proportion of genes identified as differentially expressed. In the comparison between Th1 and Th2 conditions in asthma patients and in both cancer comparisons, 7 to 12 % of the genes showed differential expression according to the ordinary t-test calculated from the whole data set at FDR of 0.05. In the rest three asthma comparisons, the changes in expression levels were more subtle and none of the genes was identified as differentially expressed at FDR of 0.05. In the former cases, correlation was high for most of the methods already with 5 replications. In the latter cases, instead, the correlations did not stabilize until the sample size was increased to over 10. This is in agreement with the intuitive expectation that large and consistent differences can be identified even with small sample sizes, while identification of smaller differences is difficult in spite of a large number of replicates.

In practice, only 2 to 3 replicates are collected for most microarray experiments. None of the ranking methods performed well with such small sample sizes. A possible solution could be to integrate data across multiple studies, which is likely to be one of the most challenging efforts during the next years. It is also the subject of our current work.

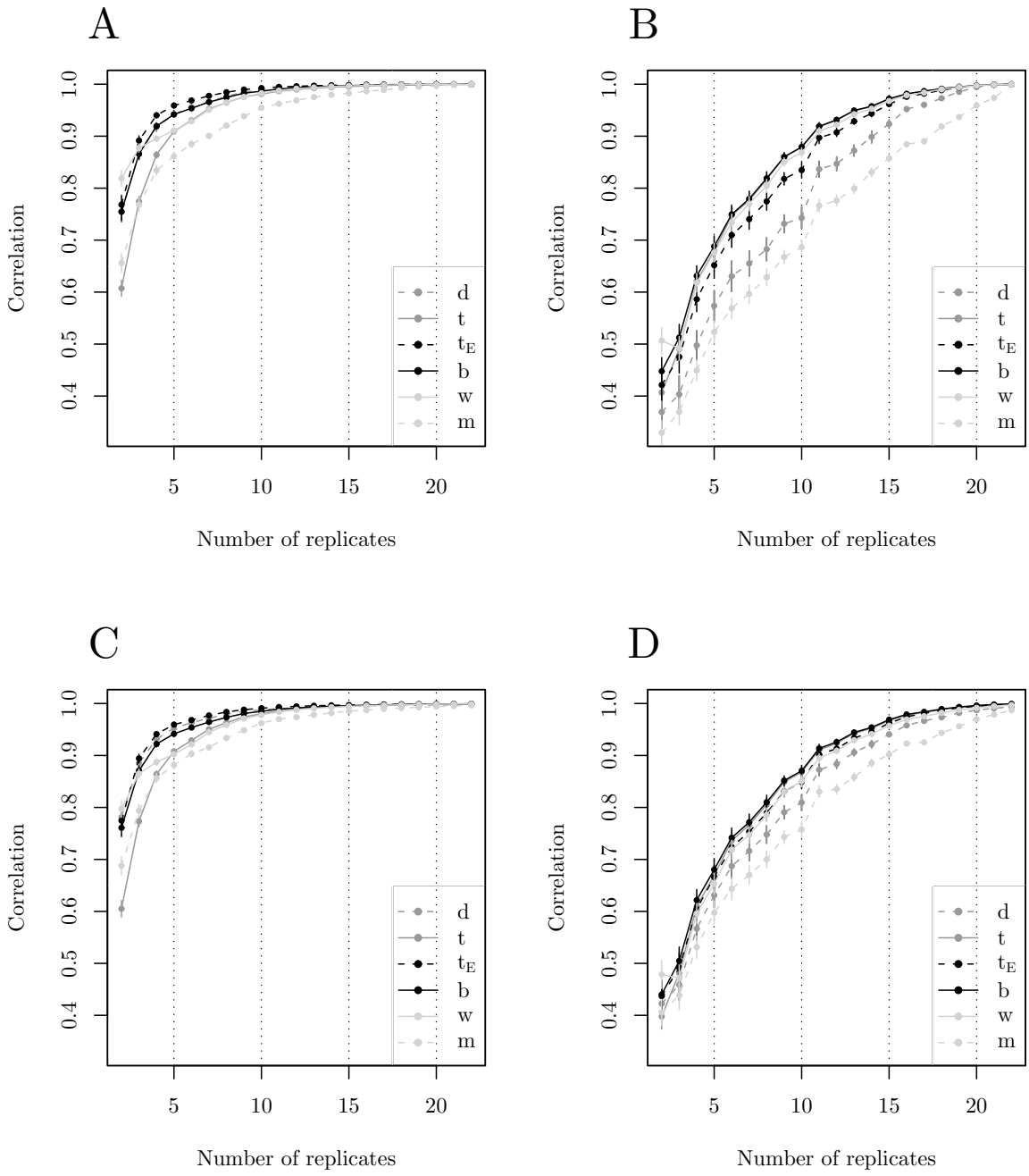

Figure 1: (A) Stability in the comparison between Th1 and Th2 conditions in asthma patients. (B) Stability in the comparison between asthma patients and controls in Th2 condition. (C) Cross-comparisons in the comparison between Th1 and Th2 conditions in asthma patients. (D) Cross-comparisons in the comparison between asthma patients and controls in Th2 condition. The average values over 200 sampled subsets are calculated, except for replication level of 21 in the paired comparisons, where there are only 22 possibilities. Error bars indicate the standard error of the mean. The results are shown for fold change (d), ordinary t-statistic (t), modified t-statistic by Efron *et al.* [3] (t_E) and Smyth [6] (b), Wilcoxon statistic (w), and mixture modeling statistic (m). The modified t-statistic by Tusher *et al.* [4] behaved similarly to b-statistic and is omitted for clarity.

References

[1] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D. and Altman, R. B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-1461.

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

[3] Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.*, **96**, 1151-1160.

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

[5] Lönnstedt, I. and Speed, T. (2002) Replicated microarray data. *Statist. Sinica*, **12**, 31-46.

[6] Smyth, G. K. (2003). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, No. 1, Article 3.

[7] Kim, R. D. and Park, P. J. (2004). Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol.*, **5**, R70.

[8] Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.*, **22**, 3899-3914.

[9] Yang, Y. H. and Speed T. (2002). Design issues for cDNA microarray experiments. *Nature Rev. Genet.*, **3**, 579-588.

[10] Black, M. A. and Doerge, R. W. (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, **18**, 1609-1616.

[11] Yang, M. C. K., Yang, J. J., McIndoe, R. A. and She, J. X. (2003). Microarray experimental design: power and sample size considerations. *Physiol. Genomics*, **16**, 24-28.

[12] Pavlidis, P., Li, Q. and Noble, W. S. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620-1627.

[13] Saari, D. G. (1995). *Basic geometry of voting.* Springer, Berlin.

[14] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, M. E., Lander, E. S., Golub, T. R. and Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41-47.

[15] Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C. and Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68-74.