Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

Is there an alternative to increasing the sample size in microarray studies?

Lev Klebanov[1, 2] **and Andrei Yakovlev**[2*]
[1]Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630,
Rochester, New York 14642; [2]Department of Probability and Statistics, Charles University, and Institute of Informatics and
Control of the National Academy of Sciences, Sokolovska 83, Praha-8, CZ-18675, Czech Republic;
Andrei Yakovlev* - Email: Andrei_Yakovlev@URMC.Rochester.edu; * Corresponding author

received October 05, 2006; accepted January 29, 2007; published online April 10, 2007

Abstract:

Our answer to the question posed in the title is negative. This intentionally provocative note discusses the issue of sample size in microarray studies from several angles. We suggest that the current view of microarrays as no more than a screening tool be changed and small sample studies no longer be considered appropriate.

Keywords: sample size; microarray; alternative

Background:

As obvious from recent literature, in the decade since the advent of microarray technology, the enthusiasm about this technology has substantially subsided. The titles of papers like "An Array of Problems **[1]** or "Getting the Noise out of Gene Arrays" **[2]** published in high profile journals speak for themselves. A growing number of such publications reflects a frustration among biologists who spend too much effort and money pursuing false leads while missing many important findings. Is it the microarray technology or the way it has been used that is to blame for the current attitude towards the as yet emerging methodologies for the generation and analysis of high throughput data in genomics and proteomics?

Notwithstanding the fact that contemporary microarray technology still calls for substantial improvements in both the quality of measurements and accuracy of probe set definitions, this powerful technology provides a rich source of multidimensional information on the functioning of the whole genome machinery at the level of transcription. Nonetheless, it is typically employed as a simplistic screening tool with a focus on individual gene profiling. Unfortunately, even this limited goal cannot be achieved with currently practiced sample sizes for the following two reasons. First, all multiple testing procedures are very unstable in the presence of correlations between gene expression levels, which is the main factor causing instability of gene lists. By and large the more liberal the procedure, the more unstable the adjusted *p*-values. **[3, 4]** This effect is exacerbated in small samples. Therefore, the actual number of false discoveries is not well controlled even if strong control is guaranteed in terms of expected values. Second, follow-up confirmatory studies can handle only Type 1 errors while a lack of power is a much more serious problem. When the sample size is small and the number of tests is large, the power of multiple testing procedures is extremely poor, so we tend to report only unusually strong effects while missing an uncontrollable number of biologically significant findings.

It is still quite common in biological publications to report microarray data on a small number of subjects. All papers claiming "consistency" or, conversely, "inconsistency" of the results produced by different microarray platforms draw their conclusions from just 3-6 arrays (subjects) per group. We would like to emphasize that it is not the technical noise that represents the main hurdle, it is the biological variability that calls for larger samples. A recently published report of the MicroArray Quality Control (MAQC) Consortium **[5]** provides direct measurements of the technical noise specific for the Affymetrix platform in the absence of biological variability. We estimated standard deviations of log-transformed expression levels associated with all genes (probe sets) from technical replicates produced by the MAQC study. They appear to be symmetrically distributed across genes. The resultant average (over genes) equaling 0.11 is at least three times as small as a typical minimal (let alone the mean) standard deviation in the presence of biological variability. A log-additive random noise as small as this exerts almost no effect on the results of statistical analysis in general and estimates of correlation coefficients in gene pairs in particular. This is definitely good albeit belated news to the scientific community. However, the inter-subject variability is beyond the control of the manufacturer and can be accounted for only through sampling from a general population of subjects.

The use of small samples significantly diminishes the utility of microarray analysis. The following simplistic experiment with real biological data illustrates this point. We used microarray data reporting expression levels of 12558 probe sets in two groups of patients with different types of childhood leukemia (Hyperdip and TALL) identified through the St. Jude Children's Research Hospital Database. **[6]** There were 88 and 45 subjects in these groups, respectively. To mimic a small-sample setting, five hundred (B=500) pseudo-independent subsamples of n=5 arrays were drawn repeatedly without replacement from each group. Differentially expressed genes were selected from each pair of subsamples by a *t*-test with Bonferroni adjustment at a nominal level of the family-wise error rate of 0.05. This experiment resulted in the mean number of rejections equaling 3.63, while the standard deviation was equal to 7.69, indicating a very high variability of the results of testing.

To remedy the situation, investigators resort to bioinformatics tools that utilize prior biological knowledge, such as partially known pathways, for prioritization of candidate genes. However, the current biological knowledge is still limited and inaccurate. This way of validation serves as a reasonable underpinning for statistical inference (limited to Type 1 errors) but not a rigorous method. When the biologist has some preliminary idea of what specific set of genes to look at, the significance analysis becomes confirmatory, thereby dramatically reducing the magnitude of multiple testing problems. This is the basic idea behind the Gene Set Enrichment Analysis (see **[7]** and references therein). However, this approach is limited to pre-defined gene sets and does not offer an alternative to the much needed exploratory tools.

The problem of differential expression is neither unique nor the most important one in the analysis of microarrays. The magnitude of differential expression does not necessarily indicate biological significance, so that the price for non-discoveries is difficult to assess. By limiting the use of microarrays to screening purposes, we do not unveil the true potential of this resourceful technology. It is imperative for statisticians to be able to extract more information from microarray data in order to prioritize genes in a more meaningful way. In particular, additional information can be provided by gene pairs rather than individual genes. It is noteworthy that recent years have seen a growing interest in correlations between gene expression levels in statistical methodologies for microarray analysis (**[8-11]** and many others). We suggest that the focus of future efforts be switched to the formation of a vector of different attributes that can be assigned to each gene in order to provide more information for gene prioritization beyond changes in the marginal distribution across phenotypes. The components of this vector might be adjusted *p*-values resulting from various statistical tests, prevalence of a specific type of correlation with other genes, relevance to the known pathways, etc. This will allow the investigator to initially increase the target set of genes by including more biologically meaningful features and then to narrow it down by putting such pieces of information together and generating a final output in an automated fashion. Such an endeavor is feasible only if larger samples become more readily available.

Statisticians have never insisted on increasing the sample size vigorously enough. Instead, many attempts have been made to overcome the sample size and cost limitations by means of mathematics. Such methodological endeavors invariably resort to the idea of pooling the information on gene expression across genes. While some of them are quite elegant, it has become clear that the actual correlation structure of microarray data is a barrier to their real world applications and this barrier seems to be insurmountable at this point in time. We have contributed to the discussion of this issue with several publications [12-14], providing evidence that the variability of the results of testing based on such methods may be extremely high. This variability manifests itself in the number of rejected hypotheses and estimated values of the false discovery rate. As a consequence, one may declare 1500 genes differentially expressed while there are none. **[13]** It is correlations between gene expression signals that cause this kind of instability because they are not only strong but also long-ranged, involving thousands and sometimes tens of thousands of genes that form pairs with each particular gene. **[10, 15]** The long-range strong correlation prevails in a huge proportion of randomly selected genes. Pooling strategies such as the Empirical Bayes method may work for cluster dependent data **[16]**, but not in the presence of long-range dependencies.

Unfortunately, there is no theoretical way to justify the required minimal sample size. We share the opinion of Yang and Speed **[17]** that power calculations are of little utility in microarray studies. The main point is that microarray analysis is exploratory (not confirmatory!) by nature and the most essential components of the standard power calculations (such as preliminary information on the expected effect sizes, variability, and the number of affected genes) are absent. **[17]**

It is our strong conviction that small sample sizes in microarray studies are a serious handicap to the progress of modern genomics. However trivial the above statement may sound, its importance remains unrealized by practitioners. We have expressed this concern before in connection with the MAQC study. **[18]** At the same time, there is a growing understanding of the importance of replication in microarray experiments and many large databases are being created in different areas of biomedical research. Now we are facing a new era in this field of data analysis, an era of large data sets. The future of microarray technology hinges on our ability to respond to this challenge.

References:

[01] S. Frantz, *Nat. Rev. Drug Discov.,* 4:362 (2005) [PMID: 15861563]
[02] E. Marshall, *Science,* 306:630 (2004) [PMID: 15499004]
[03] X. Qiu, *et al., BMC Bioinformatics,* 7:50 (2006)
[04] A. Gordon, *et al., Ann. Appl. Statist.*
[05] L. Shi, *et al., Nat. Biotechnol.,* 9:1151 (2006) [PMID: 16964229]
[06] E. J. Yeoh, *et al. Cancer Cell,* 1:133 (2002) [PMID: 12086872]
[07] L. Tian, *et al., Proc. Natl. Acad. Sci.,* 102:13544 (2005)]
[08] M. Dettling, *et al, Genome Biol,* 6:R88 (2005) [doi:10.1186/gb-2005-6-10-r88]
[09] Y. Lai, *et al., Bioinformatics,* 20:3146 (2004) [PMID: 15231528]
[10] A. Almudevar, *et al., NeuroRx.,* 3:384 (2006) [PMID: 16815221]
[11] L. Klebanov, *et al., Statist. Appl. Genet. Mol. Biol.,* 5:7 (2006) [PMID: 16646871]
[12] L. Klebanov & A. Yakovlev, *Statist. Appl. Genet. Mol. Biol.,* 5:9 (2006) [PMID: 16646873]
[13] X. Qiu, *et al., Statist. Appl. Genet. Mol. Biol.,* 4:34 (2005) [PMID: 16646853]
[14] X. Qiu, & A. Yakovlev, *J. Bioinformatics Comp. Biol.,* 4:1057 (2006) [PMID: 17099941]
[15] X. Qiu, *et al., BMC Bioinformatics,* 6:120 (2005)
[16] S. Datta & S. Datta, *Bioinformatics,* 21:1987 (2005)

[17] Y. H. Yang & T. Speed, *Nat. Rev. Genet.*, 3:579 (2002) [PMID: 12154381]

[18] L. Klebanov, *et al*., *Nat. Biotechnol*., 25:25 (2007) [PMID: 17211383]

Edited by Susmita Datta
Citation: Klebanov & Yakovlev, Bioinformation 1(10): 429-431 (2007)