A simple method for assessing sample sizes in microarray experiments

Robert Tibshirani [*]

October 3, 2005

Abstract

In this short article, we discuss a simple method for assessing sample size requirements in microarray experiments. Our method starts with the output from a permutation-based analysis for a set of pilot data, e.g. from the SAM package. Then for a given hypothesized mean difference and various samples sizes, we estimate the false discovery rate and false negative rate of a list of genes; these are also interpretable as per gene power and type I error. We also illustrate our method on other kinds of response variables, for example survival outcomes.

1 Introduction

Assessment of sample sizes for microarray data is a tricky exercise. The data are complex, as are the biological questions that one might try to answer from such data. What assumptions should one make, and what quantities should be provided as output?

There have been a number of recent papers that address this problem. Lee & Whitmore (2002) utilize an ANOVA model and provides power calculations for various alternative models. Muller et al. (2004) use a decision-theoretic approach and a hierarchical Bayes model. Wei et al. (2004) examine the roles of technical and biological variability, in determining sample size. Pawitan et al. (2005) assume that the genes are independent and have equal variance, and report false discovery rates and sensitivities. The `ssize` package (Warnes & Liu 2004) also assumes that the genes are independent, but uses pilot data to estimate the variance. It focuses on power and type I error.

All of these approaches may have shortcomings, namely the assumption of equal variances or independence of genes (or both). These assumptions are often violated in real microarray data and can have a real impact on sample size calculations.

[*] Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305; `tibs@stanford.edu`

Table 1: *Possible outcomes from m hypothesis tests of a set of genes. The rows represent the true state of the population and the columns are the result a data-based decision rule.*

	Called Not Significant	Called Significant	Total
Null	U	V	m_0
Non-null	T	S	m_1
Total	$m - R$	R	m

We avoid these assumptions in our proposal. We start with the output from a permutation-based analysis for a set of pilot data. From this we estimate the standard deviation of each gene, and the overall null distribution of the genes. Then for a given hypothesized mean difference, we estimate the false discovery rate (FDR) and false negative rate (FNR) of a list of genes. Many authors now favor the FDR over the family-wise error rate (FWER) as the appropriate error measure for microarray studies. The latter is the probability of at least one false positive call, given that we expect many false positive calls among thousands of genes, the FWER does not seem to be as relevant.

Since the calculation is based on the gene scores from permutations of the data, the correlation in the genes is accounted for. Use of the permutation distribution avoids parametric assumptions about the distribution of individual genes. And by working with the scores rather than the raw data, we avoid the difficult task of simulating new data from a population having a complicated (and unknown) correlation structure.

We provide interpretation of our results both in terms of FDR and FNR, and in terms power and type I error. Our proposal is implemented in the current version of the SAM package (Chu et al. 2002)

Our main focus is on microarray experiments for determining which genes are differentially expressed across two different experimental conditions, like treatment versus control. However our approach is also applicable to other settings, for example studies that correlate survival time with gene expression.

2 The proposed method

First we need some definitions. Table 1 summarizes the outcomes of m hypothesis tests on a set of m genes.

We have FDR $= V/R$ and FNR $= T/(m-R)$, power $= S/m_1$ and type 1 error $= V/m_0$. For simplicity, for assessing sample sizes we choose our rule so that the number of genes called significant (R) is the same as the number of non-null genes in the population (m_1). This implies that $1 -$ power $=$ FDR and type I error=FNR. Hence conveniently, the FDR can be interpreted as one minus the power per gene, and similarly for the FNR.

Here are the details of the calculation for the two-class unpaired case (below we indicate changes necessary for other data types). Let x_{ij} be the expression for gene i in sample j; C_j is the set of indices for the n_j samples in group j, for $j = 1$ or 2. The two-sample unpaired t-statistic is

$$d_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i} \tag{1}$$

where

$$s_i = [(1/n_1 + 1/n_2)\{\sum_{j \in C_1}(x_{ij} - \bar{x}_{i1})^2 + \sum_{j \in C_2}(x_{ij} - \bar{x}_{i2})^2\}/(n_1 + n_2 - 2)]^{1/2}$$

Note that this is the gene score used in the SAM method; see the Remark below regarding the exchangability constant. If σ_i is the true within-group standard deviation for gene i (assumed to be the same for each group), then ${s_i}^2$ estimates

$$\text{var}(\bar{x}_{i2} - \bar{x}_{i1}) = \sigma_i^2(1/n_1 + 1/n_2)$$

Hence a shift of δ units in one gene for each sample in group 2 causes an average increase in the score d_i of $\delta/(\sigma_i\sqrt{1/n_1 + 1/n_2})$ (we assume that the proportion of samples in groups 1 and 2 remains the same as we vary the sample size). This suggests the following procedure for assessing sample sizes:

1. Estimate the null distribution of the scores, and the per gene standard deviation σ_i, by randomly permuting the class labels and recomputing the gene scores for the permuted data.

2. For k (the number of truly changed genes) running from (say) 10 to $m/2$, do the following:

   - Sample a set of m scores from the permutation distribution of the scores
   - Add $\delta/(\hat{\sigma}_i\sqrt{1/n_1 + 1/n_2})$ in class 2 to a randomly chosen set of k of these scores.
   - Find the cutpoint c equal to the kth largest score in absolute value
   - Estimate the FDR and FNR of the rule $|d_i| > c$. This is straightforward since we know which genes are truly non-null (they are the ones that were incremented above).

3. Repeat Step 2 B times and report the median result for each k. We also report the 10th and 90th percentiles of the FDR across the B permutations.

In our examples we use a relatively small number of repetitions ($B = 20$); this makes the procedure fast and gives sufficiently accurate estimates.

The results of this process provide information on how the FDR and FNR will improve if the sample size were to be increased. To get an idea of what values of the mean difference δ are appropriate or reasonable, one can look at the values $\bar{x}_{i2} - \bar{x}_{i1}$ among the significant genes in the pilot data.

This approach can be easily applied to other designs and other types of response parameters. For paired data, we take $n_1 = n_2 = n/2$ (remember n is the total sample size). and all of the above recipe is the same. For one class data var $= \sigma_i^2/n$. For survival data with r_i equal to the numerator of the Cox score statistic, we assume that $\text{var}(r_i) = \sigma_i^2/n$ and we interpret δ relative to r_i. That is for example, if in our pilot data the genes that we call significant have $|r_i| > 100$ (roughly), we might set $\delta = 100$ in our sample size assessment.

Remark: In the SAM approach, the denominator s_i in the score (1) is replaced by $s_i + s_0$, where s_0 is an exchangeability constant. It shrinks the scores of genes with expression near 0 (having $s_0 \approx 0$).

3 An example

We generated some pilot data in two classes: there were a total of 1000 genes and 20 samples, with 10 samples in each of class. Each measurement was standard Gaussian (i.e. there was no difference between the groups in the pilot data). We ran a SAM permutation analysis, assuming the data are in a log base 2 scale and specifying a mean difference of $\log_2 2 = 1.0$. This corresponds to a mean difference of 2 fold for class 1 versus class 2. The results are shown in Figure 1.

Remember that the quantity on the horizontal axis— `number of genes`— refers to both the hypothesized number of truly non-null genes, and the number of genes called significant.

We see that, depending on the number of genes truly changed at 2-fold, the sample size should be increased to 60 or 100, in order to get the FDR down to 10% or 5%. The false negative rate is consistently low throughout, when $n = 60$ or 100.

Does our approach provide accurate estimates of FDR and FNR? For the setup of the previous example, we estimated FDR and FNR directly from repeated simulations of data from the underlying model. The results are shown in Figure 2. Note the similarity between Figures 1 and 2. Of course with real data, the second method— generating data from the underlying model— would not be available, since the underlying model is unknown.

Figure 3 shows a second example. Here there are 20 samples, and 10 blocks of 100 genes, with genes having pairwise correlation 0.5 in in each block. The mean structure is the same as in the previous example. We see that the FDR and FNR curves are similar to those in Figure 1, but the 10% and 90% curves are much wider. With less certainty in the estimate, it would be advisable to take a large sample size to ensure a reasonably low FDR. This illustrates the importance of preserving the correlation structure of the genes, i.e. it is not safe to make the (unrealistic) assumption of independence between the genes.

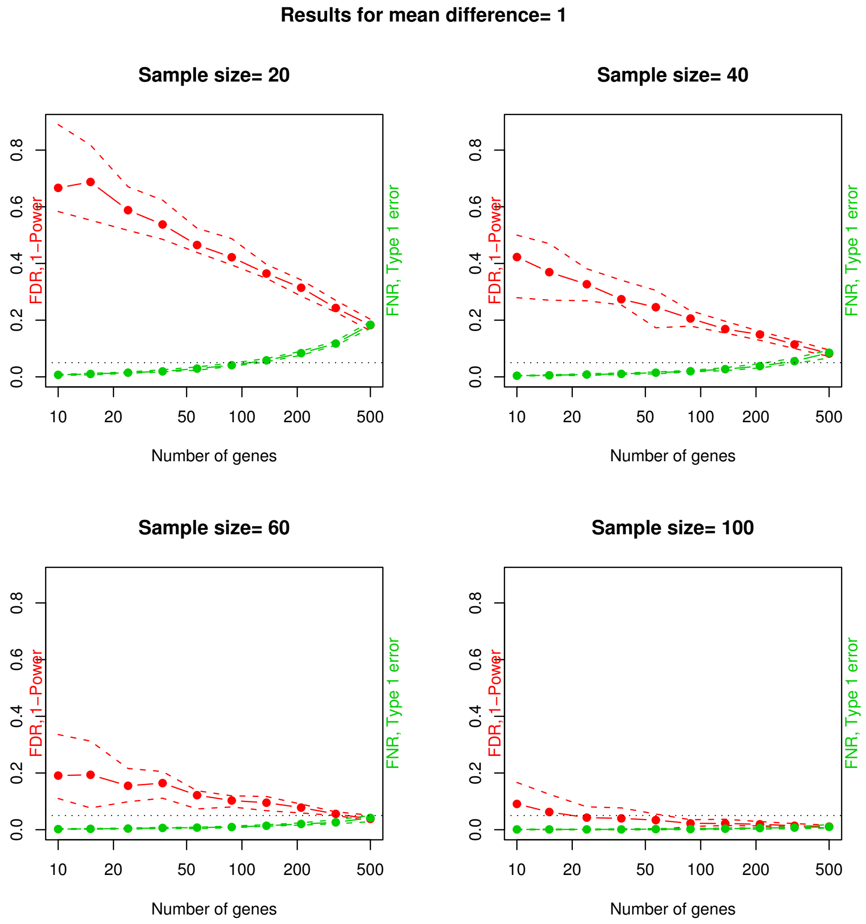

Figure 1: *Results for simulated data. The genes are generated independently. Each panel shows the estimated FDR and FNR (solid red and green curves) as well as the 10 and 90th percentiles, using the proposed method (remember that in our setup FDR=1-power and FNR=type I error). A horizontal line is drawn at 0.05. The quantity on the horizontal axis—* `number of genes` *— refers to both the hypothesized number of truly non-null genes, and the number of genes called significant. We see that the FDR is probably too high for the pilot data sample size of 20, but improves considerably when the sample size is doubled to 40.*

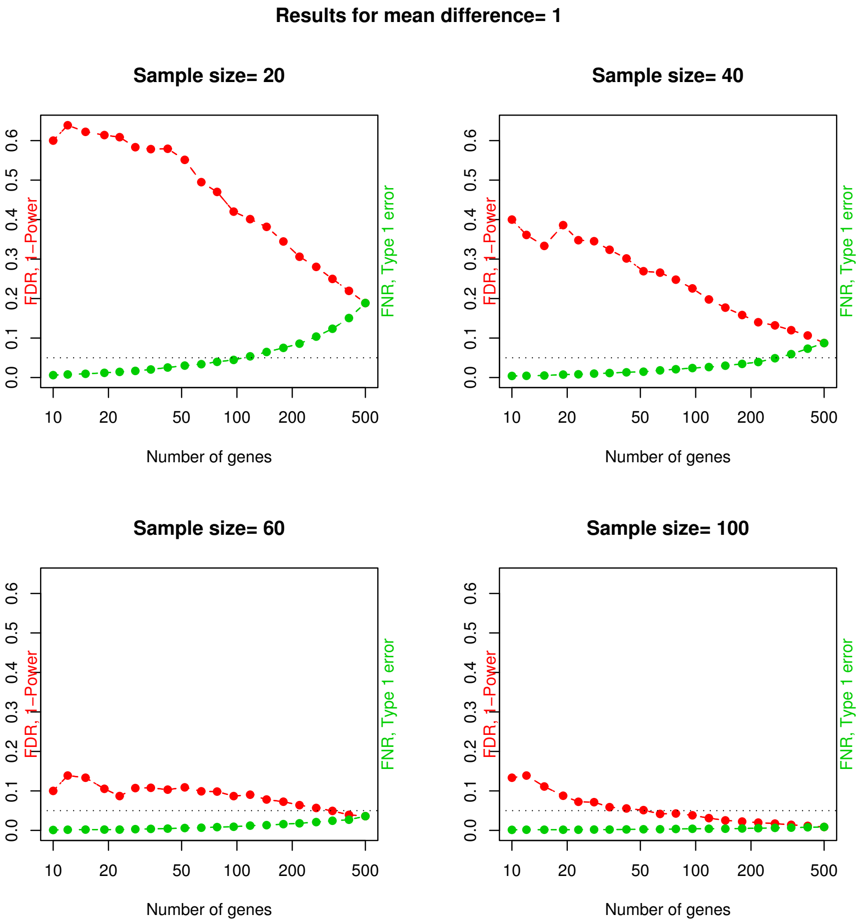

Figure 2: *Results for first simulation study. Here the FDR and FNR are estimated by direct simulation from underlying model.*

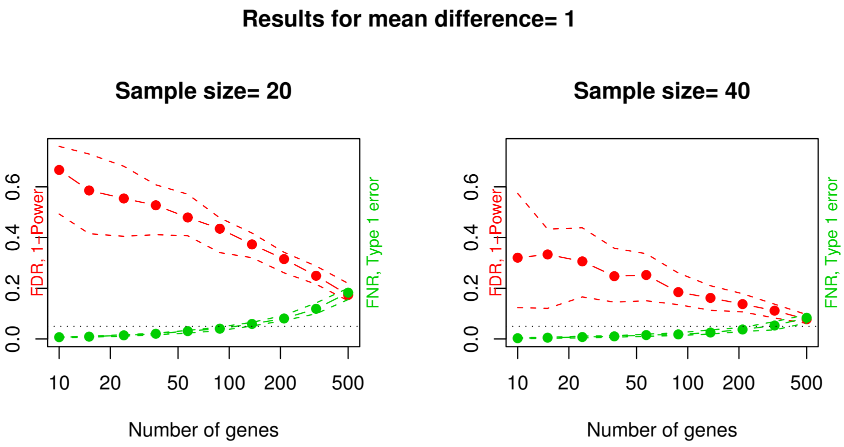

Figure 3: *Results for second simulated example (correlated genes).*

4 Discussion

We have presented a simple method for assessing sample sizes, that starts with a permutation-based analysis for some pilot data. The method gives reasonably accurate estimates of false discovery rates and false negative rates, as a function of the total number of samples. Our proposal is implemented in the SAM package- the Excel add-in and the R package `samr` (Chu et al. 2002).

Acknowledgments: The author was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

References

Chu, G., Narasimhan, B., Tibshirani, R. & Tusher, V. (2002), Significance analysis of microarrays (sam) software. Available: http://www-stat.stanford.edu/˜tibs/SAM/ via the Internet. Accessed 2003 July 16.

Lee, M.-L. & Whitmore, G. (2002), 'Power and sample size for microarray studies', *Statistics in Medicine* (21), 3543–3570.

Muller, P., Parmigiani, G., Robert, C. & Rousseau, J. (2004), Optimal sample size for multiple testing: the case of gene expression microarrays. Working paper.

Pawitan, Y., Michiels, S., Koscielny, S. Gusnanto, A. & Ploner, A. (2005), 'False discovery rate, sensitivity and sample size for microarray studies', *Bioinformatics* (21), 3017–24.

Warnes, G. & Liu, P. (2004), Sample size estimation for microarray experiments. submitted to Bioinformatics; ssize package in R.

Wei, C., Li, J. & Bumgartner, R. (2004), 'Sample size for detecting differentially expressed genes in microarray experiments', *BMC Genomics* (5), 1–10.