Determining the number of replicates needed to detect differentially expressed genes in DNA array experiments

Mario Medvedovic

Mario Medvedovic is a Research Assistant Professor,
Department of Environmental Health,
University of Cincinnati Medical Center,
3223 Eden Av.,
Cincinnati, OH, 45267-0056

Email: medvedm@email.uc.edu
Phone: 513-5589-8564
Fax: 513-558-4838

Abstract

The practice of identifying differentially expressed genes by comparing the observed differential expressions to variability expected due to random fluctuations in observed differential expression is becoming predominant in the analyses of DNA array data. In order to estimate the total variability in the system, more than one experimental replicate is needed. Generally, the probability of detecting differentially expressed genes is increasing as the number of replicates increases. Other factors influencing the probability of detecting differentially expressed genes are the magnitude of differential expression, the magnitude of the random fluctuations in the observed differential expressions and the homogeneity of the level of random fluctuations across different genes. We have constructed curves representing the magnitude of the differential expression levels that will result in 80% chance of detecting the corresponding differential expression as a function of the number of experimental replicates and the standard deviation. We described how to use these curves to determine the needed number of experimental replicates based on a "pilot" experiment. From characteristics of these curves we concluded that the homogeneity of the level of random variation in observed differential expressions is of utmost importance for being able to detect differential expressions of moderate magnitudes by using a reasonable number of experimental replicates. In the case when the variability is homogeneous, there is a substantial benefit in using 3 vs 2 microarrays but not much benefit can be seen in further increasing the number of replicates. In the situation when the variability is not homogenous, the substantial improvement can be seen with each additional experimental replicate.

Introduction

One of the most commonly used approaches for simultaneous measuring expressions of a large number of genes consists of competitive hybridization of differentially labeled cDNA probes on cDNA-printed glass slides (Schena et al. 1996; Cheung et al. 1999) called cDNA microarrays. The process of obtaining actual readings of expression levels from a cDNA microarray is subject to several sources of variability (biologic variability, printing of slides, preparation of fluorescent probes, measurement of the fluorescent light intensity) that will all be represented in the random fluctuations of observed expression levels. To discern the true differentially expressed genes from differences generated by random fluctuations in the data, observed differences need to be assessed in the context of expected distribution of differences for genes that are not differentially expressed. Statistical hypothesis testing allows us to quantify the confidence in conclusions made about observed differences in expression levels.

In the context of statistical hypothesis testing, the conclusion that a gene is differentially expressed is based on rejecting the null hypothesis

H_0: *Gene is not differentially expressed in two cell cultures*

in favor of the alternative hypothesis

H_1: *Gene is differentially expressed in two cell culture*

at a specified significance level α. Significance level α is the upper limit for the probability of rejecting H_0 in favor of H_1 when H_0 is actually true (Type I error). The commonly used significance level of α=0.05 translates into 95% confidence of making a correct conclusion that a gene is differentially expressed when H_0 is rejected. Identifying all differentially expressed genes in a microarray corresponds to testing this type of hypothesis for each gene. While the probability of committing a Type I error for each gene separately is equal to α, the probability of falsely concluding that at least one of the genes on the array is differentially expressed when actually none of them are differentially expressed (experiment-wise Type I error) will be much higher. Statistical multiple hypothesis (or multiple comparison) procedures are designed to control the experiment-wise Type I error rate in a situation when multiple hypotheses are tested.

Chen, et al. 1997 developed statistical procedure aimed at identifying differentially expressed genes using a single microarray data and by using differential expressions of "house-keeping" genes to construct the reference distribution of non-differentially expressed genes. More recently, Newton et al present a bayesian approach to identifying differentially expressed genes using data from a single microarray. Kerr and Churchill, 2000A presented a whole family of experimental designs in the framework of linear models for microarray experiments involving experimental replicates. Same authors (Kerr and Churchill, 2000B) also described how to analyze data obtained in such experiments.

Claverie, 1999, has first noted the importance of multiple comparison procedures when identifying differentially expressed genes. He proposed t-test with the Sidak's multiplicity adjustment as the appropriate approach to identify differentially expressed genes with a specified confidence. Dudoit et al., 2000 provided a more complete treatment of the problem by proposing the use of the permutation-based step-down multiple comparison procedure (Westfall and Young, 1993) as the method of choice for identifying differentially expressed genes.

Assuming that an appropriate statistical model is identified, the probability of successfully identifying a differentially expressed gene (statistical power) for a fixed confidence level depends on three factors: The magnitude of the true differential expression, the magnitude of random fluctuations (random noise) in the experimental system and the number of times the experiment is replicated. The statistical power generally increases with the increase of the magnitude of the true differential expression and with the increase in the number of experimental replicates. It decreases with an increase in the magnitude of random fluctuations in the experimental system. In this article we provided means for determining the needed number of experimental

replicates needed to detect differential expressions under a realistic statistical model and for a range of random fluctuation levels likely to be encountered in experimental systems that are currently used.

Statistical Model and Analysis

In general, the data to be analyzed consists of expression levels of N genes observed under two different experimental conditions by replicating the experiment n times (n microarrays used). z_{ijv} represents the fluorescence intensity observed for the i^{th} gene from the j^{th} cell culture on the v^{th} microarray (i=1,…,N; j=1,2; v=1,…,n), and $x_{ijv}=\ln(z_{ijv})$ is the corresponding log-transformed expression level. Traditionally, the "differential expression" refers to the gene specific ratio of the observed fluorescence intensities for two cell cultures under investigation

$$R_{iv} = \frac{z_{i1v}}{z_{i2v}}; i = 1,..., N; v = 1,..., n.$$

Differential expression measures the relative representation of a gene in the two samples and it is relative insensitive to various sources of variability associated with the variability in features of the DNA "spots", hybridization conditions, etc (Eisen and Brown, 1999). Making conclusions about the differential expressions is equivalent to making conclusions about differences between the log-transformed fluorescence intensities

$$L_{iv} = x_{i1v} - x_{i2v}; \; i = 1,\ldots,N; \; v=1,\ldots,n.$$

We are assuming that that each of the observations in the data set can be written as

$$x_{ijv}=\mu_{ijv}+\eta_{ijv},$$

where μ_{ijv} corresponds to the mean log transformed expression level for the i^{th} gene and the j^{th} on the v^{th} microarray, cell line while η_{ijv} is the random error associated with the whole process of obtaining the expression measurement. Log-transformed observed differential expression can then be written as

$$L_{iv}=(\mu_{i1}- \mu_{i2}) + \varepsilon_{iv}, \quad i=1,\ldots,N,$$

where $\mu_{i1}- \mu_{i2}$ represents the "true" underlying log-transformed differential expression and ε_{iv} represents the random error resulting from the biologic variations in gene expressions and the random noise introduced in the experimental procedure. In our model we assume that ε_{iv}'s are distributed as independent normal random variables with the mean zero and the variance σ_i^2. By using L_{iv}'s in our analysis, instead of the log-transformed fluorescence measurements we are actually extracting the effects of differences in morphology and DNA content of spots representing same genes on different arrays as well as other experimental factors that might affect magnitude of fluorescence measurements but do not affect their ratios. This approach corresponds to treating

spots on individual arrays as "Blocks" (Kerr et al. 2000A) and determines the paired t-test as the appropriate statistical analysis for identifying statistically significant differential expressions.

Testing Statistical Hypothesis

The i^{th} gene is said to be differentially expressed in the two cell cultures if $\mu_{i1} - \mu_{i2} \neq 0$. Whether this is the case is established be testing the null statistical hypothesis:

$H_0: \mu_{i1} - \mu_{i2} = 0$

vs. the alternative hypothesis

$H_1: |\mu_{i1} - \mu_{i2}| > 0.$

This test of hypothesis is performed by comparing the paired t-test statistic

$$t_i^* = \left| \frac{\bar{L}_{i\bullet}}{s_i / \sqrt{n}} \right|$$

where

$$\bar{L}_{i\bullet} = \frac{\sum_{v=1}^{n} L_{iv}}{n},$$

and s_i is the estimate of the standard deviation of L_{iv}, to its theoretical distribution under the assumption that the null hypothesis is true (null distribution). In this case, t-distribution is the appropriate null distribution. The null hypothesis is rejected at the significance α if $t^* > t_{df,\gamma/2}$. $t_{df,\gamma/2}$ is $(\gamma/2)*100^{th}$ percentile of the t-distribution with the df degrees of freedom, and $\gamma = 1-(1-\alpha)^{1/N}$ is the adjusted significance level for the test of hypothesis. The estimate of the standard deviation (s) and the degrees of freedom of the reference t-distribution (df) will depend on the assumptions that can be made about the homogeneity of variances of log-transformed differential expressions across different genes. It turns out, as it is shown in the following section, that the issue of variance homogeneity is of crucial importance for the ability of the statistical procedure to detect differentially expressed genes with a limited number experimental replicates.

Estimating Standard Deviation

The optimal approach to calculating the estimate of the standard deviation (s) depends on the whether it can be assumed that the standard deviation is equal for all genes or not. If the standard deviation of log-transformed differential expressions is assumed to vary across different genes, the gene specific estimate of standard deviation can be calculated as

$$s_i = \sqrt{\frac{\sum_{v=1}^{n} (L_{iv} - \bar{L}_{i\bullet})^2}{n-1}}, \text{ for all } i=1,..,N. \quad (1)$$

The degrees of freedom for the reference t-distribution will then be equal to n-1. On the other hand, if the standard deviation of log-transformed differential expressions is assumed to be homogeneous across different genes, the common estimate of standard deviation can be calculated by combining individual estimates given in (1)

$$s = \sqrt{\frac{\sum_{i=1}^{N} (n-1)s_i^2}{N(n-1)}}.$$

The degrees of freedom for the reference t-distribution will in this case be equal to N(n-1).

It turns out, as it is shown in the results section, that the effect of the difference in the degrees of freedom depending on the homogeneity of variance assumption on the probability of detecting differentially expressed genes is tremendous.

Statistical Power

The probability of detecting a certain number of genes that were differentially expressed with a prescribed level of confidence represents statistical power of the procedure. In other words, it is the probability of rejecting the null hypothesis that the gene is not differentially expressed in the situation when the gene is actually differentially expressed. In our model, the t-statistic under the alternative hypothesis has a non-central t-distribution with the noncentrality parameter

$$nc = \frac{\mu_{i1} - \mu_{i2}}{\sigma\sqrt{1/n}} \quad (2)$$

Hence the probability of rejecting the null hypothesis for any differentially expressed gene is given by

$$\text{Power} = \text{Prob}(t_{df,nc} > t_{\gamma/2,df}).$$

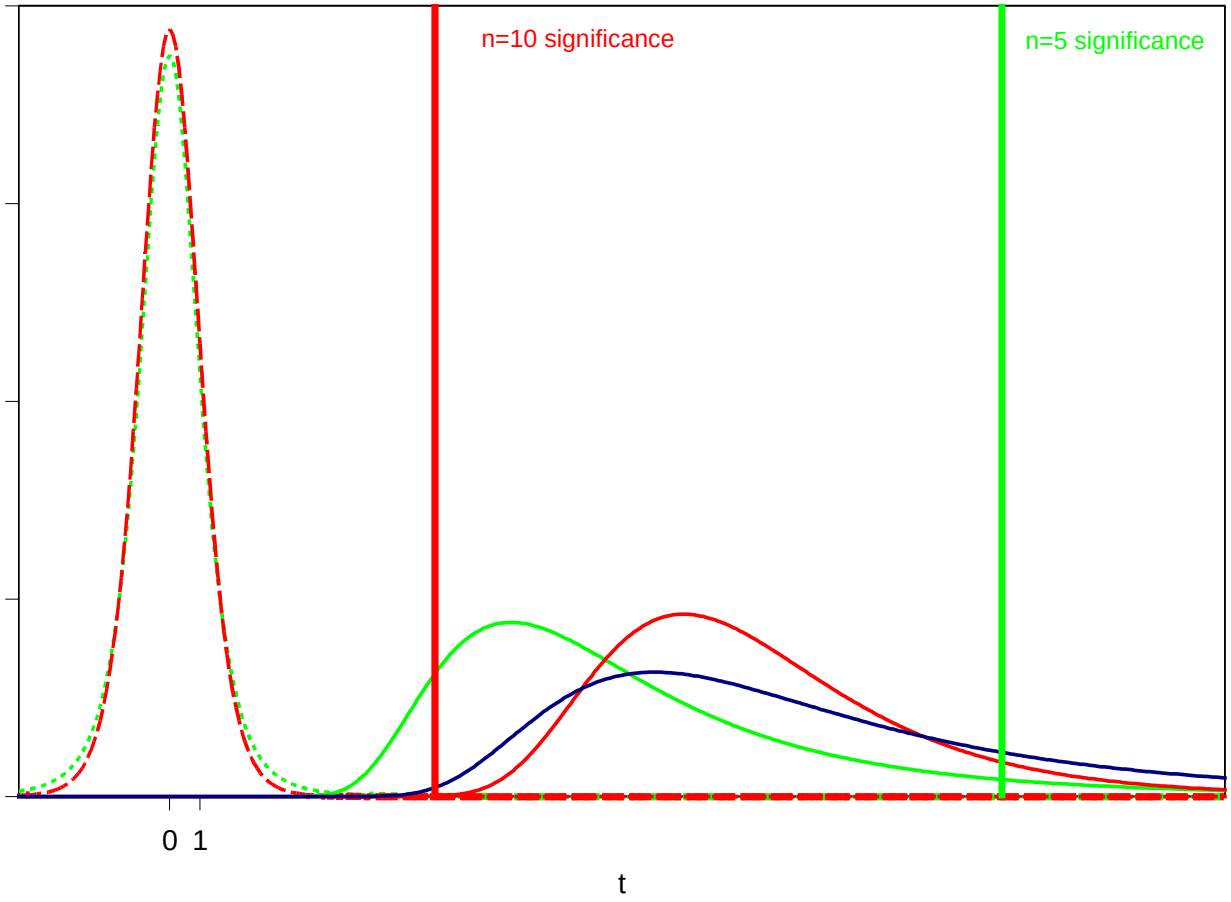
If O<N is the number of differentially expressed genes, than the probability of detecting at least K of them using the paired t-test is

$$\text{Power}(K) = 1 - \sum_{k=0}^{K-1} \binom{O}{k} \text{Power}^k (1 - \text{Power})^{D-k}.$$

The relationship between the Power of detecting a differentially expressed gene and the magnitude of the noncentrality parameter is illustrated in the Figure 1. An increase in the noncentrality parameter will generally result in an increase in the power of detecting the corresponding gene. The effect of the magnitude of the mean log-transformed differential expression (μ_{i1}- μ_{i2}) and the level of the noise in the system (σ) on the Power through their effect on the noncentrality parameter is obvious from the equation (2).

The effect of the number of experimental replicates (n) is two-fold: first, the increase in the number of replicates will result in an increase in the noncentrality parameter (2), and second, an increase in the number of replicates will result in an increase in degrees of freedom of the reference t-distribution resulting in an additional increase in Power. Both of these effects are demonstrated in the Figure 1. The increase in the sample size from 5 to 10 replicates will first reduce the cut-off point for the t-statistic to reach in order for the corresponding null hypothesis to be rejected (green vertical line vs the red vertical line) due to the difference in shapes of t-distribution with 4 and 9 degrees of freedom (dashed green and red line respectively). Second, the probability of t-statistic reaching this threshold (Power), which corresponds to the area under the curve representing the distribution of t-statistic under the alternative hypothesis (solid curves) to the right of the corresponding threshold line is affected by the change in degrees of freedom, and the size of the noncentrality parameter of the corresponding noncentral t-distribution.

Figure 1. Effect of the sample size on the Power



Green Dashed Line – Distribution of t-statistic under the null hypothesis with n=5 (t-distribution with 4 degrees of freedom)
Red Dashed Line – Distribution of t-statistic under the null hypothesis with n=10 (t-distribution with 9 degrees of freedom)
Green Solid Line – Distribution of t-statistic under the alternative hypothesis with n=5 (noncentral t-distribution with 4 degrees of freedom and the noncentrality parameter of 6.1)
Red Solid Line – Distribution of t-statistic under the alternative hypothesis with n=10 (noncentral t-distribution with 9 degrees of freedom and the noncentrality parameter of 8.6)
Blue Solid Line – Distribution of t-statistic under the alternative hypothesis with n=5 (noncentral t-distribution with 4 degrees of freedom and the noncentrality parameter of 8.6)

Results

The magnitude of the mean log-transformed differential expression (μ_{i1}-μ_{i2}) that will result in 80% chance (Power) of detecting the corresponding differential expression as a function of the number of experimental replicates (n) and the standard deviation (σ) at the significance level of α=.05 is shown in Figure 2.

Tremendous effect of homogeneity of the variability across different genes is best illustrated by comparing Figure 2A and 2B. In the situation when the standard deviation in log-transformed differential expressions varies across different genes (Figure 2A), minimum of 5 replicates are needed to detect 10 fold differential expressions at even minimal levels of variability. On the other hand, in the case of homogenous standard deviation (Figure 2B), there is a reasonable chance of detecting a single 2-fold differentially expressed genes at low but achievable levels of variability in the system. Furthermore, chance of detecting all 10 differentially expressed genes with 2 replicates in the homogenous variability case (Figure 2D), for a fixed level of variability, is similar to the 10 replicates situation in the non-homogeneous case (Figure 2C).

Figure 2A:

Detecting any single differentially expressed gene when standard deviation varies across different genes

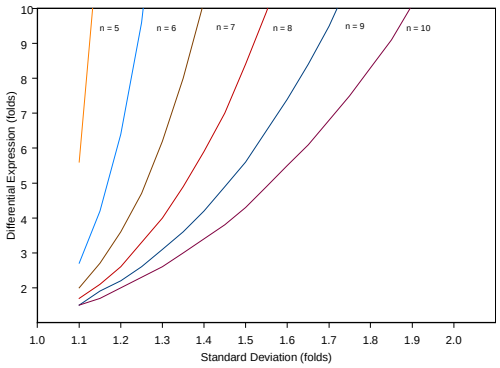


Figure 2B:

Detecting any single differentially expressed gene when standard deviation is equal for for all genes

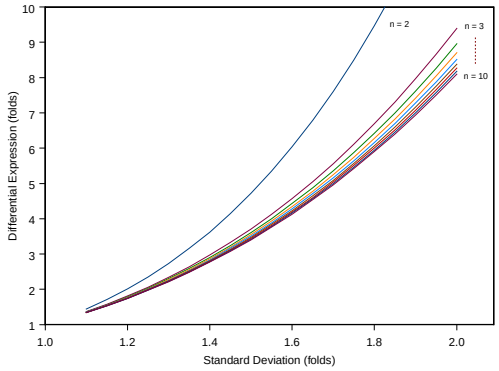


Figure 2C:

Detecting all differentially expressed genes when standard deviation varies across different genes

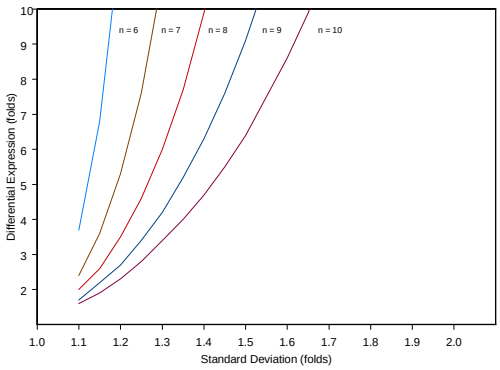
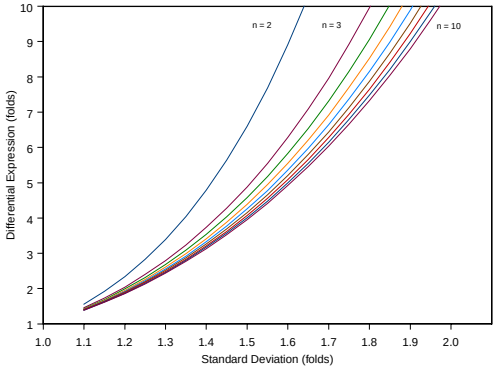


Figure 2D:

Detecting all differentially expressed genes when standard deviation is equal for for all genes



Conclusions

One of the biggest promises held by microarray technology is that it will allow for dissection of genetic regulatory networks (Lander, 1996, Lander, 1999). Karp et al. (1999) proposed a method for designing optimal microarray experiments with this purpose in mind. In order for their method to be applicable one has to be able to identify "red" (induced), "green" (repressed) and "yellow" (unchanged) genes. False interpretation of differences in fluorescence intensities observed in microarray experiments is likely to make the whole analysis much more difficult. On the other hand the reliable identification of "red", "green", and "yellow" genes by statistical analysis requires a certain number of experimental replicates to be performed. In this article, we described statistical criteria for determining the needed number of experimental replicates.

The number of replicates that need to be used in experiments utilizing DNA Arrays will ultimately be affected by numerous factors not mentioned in this paper (cost of performing the experiments and availability of biological samples being two potential factors). In this respect, results presented here can be used as a guideline of what kind of results we can expect from our experimentation. On the other hand if a general advice should be given based on our results, in the case when we can expect

homogeneous variability, there is a substantial benefit in using 3 vs 2 microarrays, but not much benefit can be seen in increasing the sample size from 3 to 10 microarrays. The approach of using the minimum of three experimental replicates is in line with conclusions made by Ting Lee et al. 2000. Furthermore, three having three replicates would allow us to factor out the dye effect that might be an important source of bias in cDNA microarrays (Dudiot et al. 2000). In the situation when the variability is not homogenous, the substantial improvement can be seen in each additional experimental replicate.

To use presented curves to precisely determine the needed number of replicates, an estimate of the total variability in the system is needed. The standard deviation of the log-transformed differential expressions (σ) quantifies the total variability in the experimental system relevant to our conclusions. If σ can not be predicted a priori, a pilot experiment will be needed to estimate sigma. Ideally, the pilot experiment consists of two replicates of the experiment in which mRNA obtained from two independent samples of the biologic system of interest is independently extracted and differentially labeled and co-hybridized on a DNA array. From such a pilot experiment, we can estimate σ as well as investigate if the assumption of the variance homogeneity is reasonable. Based on such analysis, and other considerations (cost-benefit, anticipated level of differential expression of interest, etc) we can identify the optimal number experimental replicates for our specific experiment.

Strictly speaking, results presented here apply only to the model with the Gaussian distribution of the random variations in observed log-transformed differential expressions. While this is a reasonable assumption on its own, even if this model is not correct, due to the central limit theorem, with the increase in the number of replicates the distributions of t-statistics used in our analysis will approach their distributions under the Gaussian model regardless of the true underlying distribution of the data. This allows us to use results presented in a more general setting than is required by the model used in obtaining these results.

Reference:

(1) Chen, Y., Dougherty, E.R., Bittner M.L. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2:** 364-374.

(2) Cheung, V.G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., Childs, G. 1999. Making and reading microarrays. *Nature Genetics Supplement* **21:** 15-19.

(3) Claverie, J.-M. 1999. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **8:** 1821-1832.

(4) Dudoit, S., Yang, Y.H., Callow, M. J., Speed, T.P. 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report #578, University of California at Berkely.

(5) Eisen, M.B., Brown, P. O. 1999. DNA Arrays for Analysis of Gene Expression. *Methods in Enzimology* 303: 179-205.

(6) Karp, R.M., Stoughton, R., Yeung, K.Y. 1999. Algorithms for Choosing Differential Gene Expression Experiments. *RECOMB '99, Lyon France*: 208-217.

(7) Kerr, M. K., Churchill, G. A. 2000A. Experimental Design for Gene Expression Microarrays. *Manuscript available at* http://www.jax.org/research/churchill/research/expression/kerr-design.pdf.

(8) Kerr, M. K., Churchill, G. A. 2000A. Analysis of Variance of Gene Expression Microarray Data. *Manuscript available at* http://www.jax.org/research/churchill/research/expression/kerr-synteni.pdf.

(9) Lander, E. S. 1996. The new genomics: Global views of biology. *Science* **274:** 536-539.

(10) Lander, E. S. 1999. Array of Hope. *Nature Genetics Supplement* **21:** 3-4.

(11) Newton M.A., Kendziorski C.M., Richmond C.S., Blattner F.R., Tsui K.W. 2000. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Technical Report 139, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison. *(To appear in Journal of Computational Biology).*

(12) Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P., Davis, R. 1996. Parallel human genome analysis: microarray-based expression of 1000 genes. *Proceedings of the National Academy of Science USA,***94:** 1359-1367.

(13) Ting Lee, M.-L., Kuo, F.C., Whitmore, G.A., Sklar, J. 2000. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Science USA,***97:** 9834 - 9839.

(14) Westfall, P.H., Young S.S. (1993). *Resampling-Based Multiple Testing*. New York, John Wiley & Sons, Inc.