Nonparametric Modeling Approach for Discovering Differentially Expressed Genes in Replicated Microarray Experiments

Ali Gannoun[1,2] Jérome Saracco[2] Wolfgang Urfer[3]

George E. Bonney[1]

(1) Statistical Genetics and Bioinformatics Unit

National Human Genome Center

2216 6th street, suite 206

Washington D.C. 20059, USA

(2) Université Montpellier II

Laboratoire de Probabilités et Statistique, cc 051

34095 Montpellier Cedex 05, France

(3) Department of Statistics

University of Dortmund

D-44221 Dortmund, Germany

August 28, 2002

Summary

Microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels of thousands of genes simultaneously. In microarray data analysis, the comparison of gene-expression profiles with respect to different conditions and the selection of biologically interesting genes are crucial tasks. Multivariate statistical methods have been applied to analyze these large data sets. In particular, Dudoit *et al.* (2002) developed methods using t-statistics with p-values calculated through permutations, and with the Westfall and Young step-down approach to correct for multiple testing. Thomas *et al.* (2001) developed a regression modelling approach. Following the idea of Efron *et al.* (2000) and Tusher *et al.* (2001), Pan (2002) proposed mixture modelling approach that relaxes many strong assumptions on the null distributions of the test statistics. In this paper, we replace the based Normal mixture density estimators proposed by Pan *et al.* (2002), with less restrictive nonparametric ones.

Keywords: Kernel estimator; Microarray; Mixture modeling; Regression modelling, t-test.

1 Introduction

Gene expression regulates the production of protein, the ultimate expression of the genetic information, which in turn governs many cellular processes in biological systems. The knowledge of gene expression has applications ranging from basic research on the mechanism of protein production to applications such as diagnosing, staging, treating and preventing of diseases. Microarray techniques provide a way of studying the RNA expression levels of thousands of genes simultaneously; see for example Brown and Botstein (1999), Lander (1999), Quackenbush (2001). The identification of differentially expressed genes is a question which arises in a broad range of microarray experiments which produce enormous amounts of data, see Spellman *et al.* (1998), Galitski *et al.* (1999), Golub *et al.* (1999), Callow *et al.* (2000), Friddle *et al.* (2000), and Guimaraes and Urfer (2002), to name a few. The expression level can refer to summary measure of relative red to green channel intensities in a fluorescence-labeled complementary DNA or cDNA array, a radioactive intensity of a radiolabeled cDNA array, or summary difference of the perfect match (PM) and mis-match (MM) scores from an oligonucleotide array, see Li and Wong (2001)). Microarray experiments involve a number of distinct stages which are discussed in Smyth *et al.* (2002). The expression levels may have been suitably preprocessed to acquire red and green foreground and background intensities for each spot of the arrays, including dimension reduction, data normalization and data transformation; see for example Dudoit *et al.* (2002), Efron *et al.* (2000), Li and Wong (2001). We suppose here that all stages to get data are satisfied. For the purpose of the paper, let Rf and Gf (resp. Rb and Gb) be the foreground (resp. the background) red and green intensities for each spot. The *log*-differential expression ratio will be $Y = log_2(R/G)$ where usually $R = Rf - Rb$ and $G = Gf - Gb$, where $G > 0$. One of the core goals of microarray data is to compare, for example, the expression levels of genes in samples drawn from two different cell types, such as healthy versus diseased cells, and to identify which of the genes show good evidence of being differentially expressed. Statistical methods are very helpful to reach this goal. In the early days, many data analysis programs sort the genes according to the absolute level of $\overline{Y}$, where $\overline{Y}$ is Y -values for any particular gene across the replicate arrays, see Smyth *et al.* (2002) for more details. This is known to be unreliable (see Chen *et al.* (1997)) because statistical variability of the expression levels for each genes was not taken account. It has also been noticed that data based on a single array may not be reliable and may contain high noises (see Lee *et al.* (2000)). Moreover, the need for independent replicates has been recognized (see Lee *et al.* (2000)), and several methods from combining information from several arrays have been proposed. These methods assign a test score to each of the genes and then select those that are "significant". The test statistics included the t-test (Zhang *et al.* (1997), Alon *et al.* (1999)), the ANOVA F-statistics (Kerr *et al.* 2000)) and the information theoric measure known as InfoScore (Hadenfalk *et al.* (2001)). Recently, Chilingaryan *et al.* (2002) used a multivariate approach based on Mahalanobis distance between vectors of gene expressions as a criterion for simultaneously comparing a set of genes, and developed an algorithm for maximizing it. A Similar vectorial approach, including principal components analysis, is also given by Kuruvilla *et al.* (2002). Bayesian probabilistic framework for microarray data analysis are also developed by Friedman *et al.* (2000), Baldi *et al.* (2001), Imato *et al.* (2002) among others. In this paper we consider the detection of differentially expressed genes with replicated measurements of expression levels using Bayesian inference with the mixture model approach of Pan *et al.* (2002). It is one of the three methods reviewed by Pan (2002). In particular, we introduce a kernel estimator of density functions in order to form the test statistic in the Bayesian techniques.

The paper is organized as follows. After describing the statistical model and tests in Section 2, we discuss the kernel estimation procedure. Finally, Section 3 summarizes some concluding remarks and gives an outlook for further activities.

2 Statistical model and methods

In this section, we give a general statistical model from which we make the comparative studies. Then, we recall the construction of the t-test and the mixture modeling approach.

2.1 The model

Various models are proposed to summarize multiple measurements of gene expression. A general survey is given by, for example, Thomas *et al.* (2001), Li and Wong (2001) and Sebastiani and Romani (2002). We will focus on a simple model studied in particular by Pan et *al.* (2002).

Suppose that Y_{ij} is the expression level of gene i in array j, $i = 1, ..., n$ and $j = 1, ..., J$. We suppose that $J = J_1 + J_2$ and that the first J_1 and last J_2 arrays are obtained under the different conditions, say treatment and control, respectively.

We consider the following general statistical model:

$$Y_{ij} = \beta_i + \mu_i x_j + \varepsilon_{ij} \tag{1}$$

where $x_j = 1$ for $1 \leq j \leq J_1$ and $x_j = 0$ for $J_1 + 1 \leq j \leq J_1 + J_2$, and ε_{ij} are independent random errors with mean 0.

Hence $\beta_i + \mu_i$ and β_i are the mean expression levels of gene i under the two conditions respectively.

Let H_{0i} denote the null hypothesis of equal treatment and control mean expression levels for gene i, $i = 1, ..., n$. Here we consider only two-sided alternative hypotheses; one-sided alternatives can be handled in similar manner. Then, determining whether a gene has differential expression is equivalent to testing the null hypothesis

$$H_{0i} : \mu_i = 0 \text{ against } H_{1i} : \mu_i \neq 0.$$

A statistical test consists of two parts. The first is to construct a summary test statistic which will rank the genes in order of evidence for differential expression,

from strongest to weakest evidence. The second is to choose a critical-value, or the significance level or p-value associated with the test statistic above which any value is considered to be significant. In many microarray studies the aim is to identify a number of candidate genes for confirmation and further study.
To focus on the main issue, we use $\alpha = 0.01$ as the genome-wide significance level. To account for multiple hypothesis testing, one may calculate adjusted p-values, see Shaffer ((1995) and Westfall and Young (1993). According to Shaffer (1995), given any procedure, the adjusted p-value corresponding to the test of single hypothesis H_{0i} can be defined as the level of the entire test procedure at which H_{0i} would just be rejected, given the values of all test statistics involved. The Bonferroni method is perhaps the best known method with multiple testing (see Dudoit *et al.* (2002) and Thomas *et al.* (2001)). This method should be used here. Hence the gene-specific significance level (for a two-sided test) is $\alpha^* = \alpha/(2n)$.
In the following, we review two methods along the line.

2.2 The t-test

Let us recall that H_{0i} denote the null hypothesis of equal expression levels under the two different conditions (e.g. control and treatment) for gene i, $i = 1, ..., n$. We consider only two-sided alternative hypotheses. For gene i , the t-statistic comparing gene expression is

$$Z_i = \frac{\overline{Y}_{i(1)} - \overline{Y}_{i(2)}}{\sqrt{\frac{s^2_{i(1)}}{J_1} + \frac{s^2_{i(2)}}{J_2}}}, \tag{2}$$

where $\overline{Y}_{i(1)}$ and $\overline{Y}_{i(2)}$ denote the average expression level of gene i in the J_1 treatment and J_2 control hybridizations, respectively. Similarly, $s^2_{i(1)}$ and $s^2_{i(2)}$ denote the sample variances of gene i's expression level in the treatment and control hybridizations, respectively.
Large absolute t-statistics suggest that the corresponding genes have different expression levels in the control and treatment groups. Note that the replication

is essential for such an analysis because of the need for assessing the variability of gene expression levels in the treatment and control groups.

Under the Normality assumption of Y_{ij}, Z_i approximately has a t-distribution with degree of freedom

$$d_i = \frac{(s^2_{i(1)}/J_1 + s^2_{i(2)}/J_2)^2}{(s^2_{i(1)}/J_1)^2/(J_1-1) + (s^2_{i(2)}/J_2)^2/(J_2-1)} \tag{3}$$

A classical approximation of d_i is given by $J_1 + J_2 - 1$, see for example Scheffé (1970) and Best and Rayner (1987).

If we do not assume the t-distribution, we use permutation to estimate their distribution, see Dudoit *et al.* (2002) for more details. Wastfall and Young (1993) suggest approximating the p-values using asymptotic theory, see also Dudoit *et al.* (2002) for a computational algorithm.

2.3 The mixture model

The ordinary t-statistic is not ideal because of its restrictive assumptions. Strong assumptions (e.g. normality, equality of variances) are needed for the null distribution of the test statistics. To estimate the null distribution, Pan (2002) and Pan *et al.* (2002) constructed the following null statistics

$$z_i = \frac{Y_{i(1)}u_i/J_1 - Y_{i(2)}v_i/J_2}{\sqrt{\frac{s^2_{i(1)}}{J_1} + \frac{s^2_{i(2)}}{J_2}}} \tag{4}$$

where $Y_{i(1)} = (Y_{i1}, Y_{i2}, ..., Y_{iJ1}), Y_{i(2)} = (Y_{iJ_1+1}, Y_{iJ_1+2}, ..., Y_{iJ_1+J_2}), u_i$ is a random permutation of column vector containing $J_1/2$ 1's and -1's respectively, and v_i is a random permutation of column vector containing $J_2/2$ 1's and -1's respectively. Let f and f_0 be the distribution densities of Z_i and z_i. If there is no expression change for gene i , then Z_i should have the same distribution as that of z_i. Under the weak assumption that the random variable ε_{ij} in (1) has a distribution symmetric about its mean 0, then under H_{0i}, $f = f_0$.

If we assume that the distribution of Z_i's for genes that are differentially expressed is f_1, f can be expressed as a mixture of f_0 and f_1,

$$f = p_0 f_0 + p_1 f_1 \tag{5}$$

where p_1 is an unknown proportion of the genes that are differentially expressed and $p_0 = 1 - p_1$.

For any given Z, if we know f_0 and f, we use the likelihood ratio test statistic

$$LR(Z) = f_0(Z)/f(Z) \tag{6}$$

to test for H_0.

Then, by the optimal Neyman-Pearson test, a small value of $LR(Z)$, say $LR(Z) < c$, provides evidence to reject H_0. The cut-off point c is determined such that the type I error kind is

$$\frac{\alpha}{n} = \int_{LR(z)<c} f_0(z)dz \tag{7}$$

where α is the genome wide significance level.

In the absence of strong parametric assumptions, the parameters p_0, f_0 and f_1 are not identifiable, see Efron *et al.* (2000). By assuming a normal distribution for Z_i, for each i, one can estimate the components of the mixture by using for example the EM algorithm (see Dempster *et al.* (1977)). Lee *et al.* (2000) and Newton *et al.* (2001) considered parametric approaches by assuming Normal or Gamma distributions for f_0 and f_1 respectively. Efron *et al.* (2000) avoided such parametric assumptions and considered a nonparametric empirical Bayesian approach.

Using z_i's and Z_i's we will estimate f_0 and f by a kernel method and develop a procedure to determine the type I error.

2.4 Kernel estimation of f_0 and f

The construction of a kernel estimator requires a choice of a density function K, and a bandwidth h_n which is a a sequence of positive numbers tending to 0 as n tends to infinity. From Z_i and z_i, $i = 1, ..., n$, f and f_0 can be estimated nonparametrically by

$$f_n(z) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{z - Z_i}{h_n}) \tag{8}$$

and

$$f_{0n}(z) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{z - z_i}{h_n}). \tag{9}$$

In order to get smoother estimation, one can use a kernel K which is bounded, symmetric and satisfying $|z| K(z) \to 0$ as $|z| \to \infty$ and $\int z^2 K(z) dz < \infty$. Some special kernel functions are given by:

Kernel	$K(z)$
Uniform	$\frac{1}{2} 1(\|z\| \leq 1)$
Triangle	$(1 - \|z\|) 1(\|z\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - z^2) 1(\|z\| \leq 1)$
Quartic	$\frac{15}{16}(1 - z^2)^2 1(\|z\| \leq 1)$
Triweight	$\frac{35}{36}(1 - z^2)^3 1(\|z\| \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2}$
Cosines	$\frac{\pi}{4} \cos(\frac{\pi}{4} z) 1(\|z\| \leq 1)$

Well known theoretical results show that the choice of reasonable K does not seriously affect the quality of the estimators (8) and (9). On the contrary the choice of h_n turns to be crucial for the accuracy of the estimator. Some indications about this choice are given in Bosq and Lecoutre (1987). We will use in practice

$$h_n = \widehat{\sigma}_n n^{-1/5}$$

where $\widehat{\sigma}_n$ denotes the empirical standard deviation. This choice minimizes some asymptotic mean square error (see Deheuvels (1977)).

2.5 Determination of the cut-off point c

From Efron and Tibshirani (1993), Pan (2002) and Pan et *al.* (2002), Efron *et al.* (2000, 2001), a parametric bootstrap approach proceeds as follows.
We draw B random samples from f_0, $z^{(1)}, z^{(2)}, ..., z^{(B)}$, where $z^{(b)} = \{z_1^{(b)}, z_2^{(b)}, ..., z_N^{(b)}\}$ for $b = 1, ..., B$. Then for a possible cut-off point c, we calculate the average of false rejections:

$$False(c) = \frac{1}{B}\sum_{b=1}^{B} \#\{i : LR(z_i^{(b)}) < c\}. \tag{10}$$

Based on a desired false rejection number, we can choose the corresponding c. Note that with the normal mixture model in Pan *et al.* (2002), it is possible to numerically determine the c using the bisection method (Press *et al.*(1992)) to solve equation (7).
Here we propose an empirical method to solve (7) based on nonparametric estimation of f_0 and f. We start by replacing f_0 and f by their kernel estimators f_{0n} and f_n proposed in (8) and (9), and solve the modified equation

$$\frac{\alpha}{n} = \int_{\widehat{LR}(z)<c} f_{0n}(z)dz, \tag{11}$$

where $\widehat{LR}(z) = f_{0n}(z)/f_n(z)$.
Let $T = LR(z)$ and $\widehat{T}_i = \widehat{LR}(z_i)$ and $\widehat{T}_{(1)}, \widehat{T}_{(2)}, ..., \widehat{T}_{(n)}$ be the ordered values of $\widehat{T}_1, \widehat{T}_2, ..., \widehat{T}_n$. For a fixed value $c \in \left[\widehat{T}_{(1)}, \widehat{T}_{(n)}\right]$, let $A_c = \{z : T < c\}$, $\widehat{A}_c = \left\{z_i : \widehat{T}_i < c, i = 1, ..., n\right\}$ and $z_{c,(1)}, z_{c,(2)}, ..., z_{c,(q)}$ be the ordered values of $\widehat{A}_c$. Then

$$\int_{A_c} f_0(z)dz \approx \int_{\widehat{A}_c} f_{0n}(z)dz \approx \int_{z_{c,(1)}}^{z_{c,(q)}} f_{0n}(z)dz \tag{12}$$

Now, the approximate cut-off point is the value $c_j = \widehat{T}_{(1)} + \frac{j}{m}(\widehat{T}_{(n)} - \widehat{T}_{(1)})$, where m is as large as possible, such that

$$\frac{\alpha}{n} \approx \int\limits_{\widehat{A}_{c_j}} f_{0n}(z)dz. \tag{13}$$

3 Discussion and concluding remarks

We have reviewed and extented methods for the analysis of microarray experiments. Following the principle of "letting the data speak about themselves", we have introduced a nonparametric kernel estimation into mixture models. Our method has three principal advantages:

1) An assumption of normality is not needed.

2) The estimation of the degrees of freedom in the existing t-test is avoided.

3) We need not use bootstrap to estimate the cut-off point.

Hence, our method can be implemented unambiguously and efficiently. In a forthcoming paper, extensive simulation results and applications to real data will be reported.

Acknowledgment *The work was supported by the United States Public Service Grant No AG 16996 from the National Institute of Health, and the German Research Council through the Collaborative Research Center (SFB 475) at the University of Dortmund (Germany).*

The authors thank Dr. Wei Pan for helpful discussions.

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci.,* USA, **96**, 6745–6750.

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.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.

Best, D.J., and Rayner, J.C.W. (1987). Welsh approximate solution for the Behrens-Fisher problem. *Technometrics*, **29**,205–210.

Bosq, D., and Lecoutre, J.P. (1987). *Théorie de l'estimation fonctionnelle.* Economica.

Brown, P.O. and Botstein D. (1999). Exploring the New World of the genome with DNA microarrays. *Nature Genetics,* **21**, 33–37.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, **10**, 2022–2029.

Chen, Y., Dougherty, E. R., and Bittner, M. (1997). Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Biomedical Optics*, **2**, 364–374.

Chilingaryan, A., Gevorgyan, N., Vardanyan, A., Jones, D. , and Szabo, A. (2002). Multivariate approach for selecting sets of differentially expressed genes. *Mathematical Biosciences*, **176**, 59–69.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood estimation from incomplete data, via the EM algorithm (with discussion). *J.R. Statist Soc. B*, **39**, 1–38.

Deheuvels, P. (1977). Estimation non parametrique de la densité par histogramme generalise. *Rev. Stat. Apppliquée.* 35F42.

Dudoit, S., Yang, Y. H. , Speed, T. P. and Callow M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Efron, B. and Tibshirani, R.J. (1993). *An introduction to the Bootsrap.* Chapman & Hall: London.

Efron, B., Tibshirani, R. , Goss, V. and Chu, G. (2000). Microarrays and Their Use in a Comparative . Technical report, Stanford University.

Efron, B., Storey, J. and Tibshirani, R. (2001). Microarrays, Empirical Bayes Methods, and False Discovery Rates. Technical report, Univ. California, Berkeley.

Friddle C., Koga T., Rubin, E. and Bristow, J. (2000). Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Nat. Acad. Sci.,* USA, **97**, 6745–50.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology* , **7**, 601–620.

Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S., and Fink, G.R. (1999). Ploidy regulation of gene expression. *Science,* **285** , 251–254.

Golub,T.R. , Slonim, D. K., Tamayo, P., Huard, C ., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grant, G., Manduchi, E., and Stoeckert, C. (2002). *Using non-parametric methods in the context of multiple testing to identify differentially expressed genes. Methods of microarray data analysis.* Editors S.M. Lin and K.F. Johnson, Kluwer Academic Publishers, 37–55.

Guimaraes, G., Urfer, W. (2002). Self-organizing maps and its applications in sleep apnea research and molecular genetics. In: O.Opitz and M. Schwaiger (eds) Exploratory data analysis in empirical research. Studies in classification, data analysis and knowledge organization. Springer Verlag, Heidelberg, 332–345.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O. P., Borg, Å., and Jeffrey Trent. (2001). Gene Expression Profiles in Hereditary Breast Cancer. *New England Journal of Medicine*, **244**, 539–548.

Imoto, S., Goto, T., and Miyano, S. (2002). Estimation of Genetic Networks and Functional Structures between Genes by using Bayesian Network and Non-parametric Regression. *Proc. Pacific Symposium on Biocomputing*, **7**, 175–186

Kerr, M.K., Martin, M., and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data.*Journal of Computational Biology* , **7**, 819–837.

Kuruvilla, F.G., Park, P.J., and Schreiber, S.L. (2002). Vector algebra in the analysis of genome-wide expression data. *Genome Biology*, **3**, 1–11.

Lander, E. S. (1999). Array of hope. *Nature Genetics*, **21**, 3–4.

Lee, M-L.T., Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000). Importance of microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, *Proc. Nat. Acad. Sci.,* USA, **97**, **18**, 9834–9839.

Lee, M-L, Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Nat. Acad. Sci.,* USA, **97**, 9834–9839.

Li, C., and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *PNAS*. **98**, 31–36.

Newton, M.A., Kendziorski, C.M. Richmond, C.S. Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37–52.

Pan, W. (2002). A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. *Bioinformatics*, **12**, 546–554.

Pan, W., Lin, J. and Le, C.T. (2002). A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data. *Genome Biology* (To appear).

Press, W.H., Teukolsky, C.M., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical recepes in C, The Art of Scientific Computing.* 2nd ed. Combridge: New York.

Quackenbush J. (2001). Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–27.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Sociaty*, **65**, 1501–1508.

Sebastiani, P., and Ramoni, M. (2002). Statistical Challenges in Functional Genomics. Technical report, Massachusetts University.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.*, **46**, 561–584.

Smyth, G. K., Yang, Y.-H., Speed, T. P. (2002). Statistical issues in microarray data analysis. In: *Functional Genomics*, Methods and Protocols, M. J. Brownstein and A. B. Khodursky (eds.), Methods in Molecular Biology series, Humana Press, Totowa, NJ.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9** , 3273–3297.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, LP. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

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

Westfall, P.H., Young, S.S. (1993). *Resampling-based multiple testing: Examples and Methods for p-value adjustment.* Wiley series in probability and mathematical statistics. Wiley, 1993.

Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler, K.W. (1997). Gene Expression Profiles in Normal and Cancer Cells. *Science*, **276**, 1268–1272.