Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 129 (2005) 19–37

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Sequential prediction bounds for identifying differentially expressed genes in replicated microarray experiments

Robert D. Gibbons[a,*], Dulal K. Bhaumik[a], David R. Cox[b], Dennis R. Grayson[c], John M. Davis[c], Rajiv P. Sharma[c]

[a] *Departments of Biostatistics, Psychiatry, and Center for Health Statistics, University of Illinois at Chicago, 1601 W. Taylor, Chicago, IL 60612, USA*

[b] *Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA*

[c] *The Psychiatric Institute, University of Illinois at Chicago,1601 W. Taylor, Chicago, IL 60612, USA*

Available online 21 August 2004

Abstract

Microarrays are new biotechnological devices that permit the simultaneous evaluation of expression levels of thousands of genes in one or more tissue samples. We develop a new method for identifying differentially expressed genes in replicated cDNA and oligonucleotide microarray experiments. The method is based on a nonparametric prediction interval which is computed as an order statistic of n control measurements and is applied sequentially to a series of p replicate sets of experimental measurements, each of size n_i. We illustrate how reasonable experiment-wise false positive and false negative rates can be attained for any practical number of genes based on manipulating the order statistics, n, p and n_i. The method is used to identify gene expression levels that are associated with a pathological condition beyond chance expectations given the large number of genes tested. We illustrate use of the method on replicated gene expression data in tumor and normal colon tissues, and compare it to an alternative approach based on permutation tests.

Keywords: Microarray; Simultaneous prediction intervals; Molecular genetics; Statistical genetics; Multiple comparisons; Nonparametric statistics

[*] Corresponding author. Tel.: +1-312-413-7755; fax: +1-312-996-2113.
E-mail address: rdgib@uic.edu (R.D. Gibbons).

0378-3758/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.jspi.2004.06.037

1. Introduction

With the advent of methods for large-scale gene expression studies, such as microarray technology (Chee et al., 1996), evaluation of large numbers of gene mRNA levels amongst different individuals is now possible. An immediate consequence of this new technology is the ability to differentiate gene expression among normal and diseased states, a problem of central importance to modern biology and medical research and has wide applicability to high throughput pharmaceutical screening. From a statistical perspective, the problem is complex for three primary reasons. First, determining whether an observed difference between two sources (e.g., normal children versus children with Down's syndrome) is due to chance, is an almost insurmountable problem when testing several thousand genes. Second, little is known about the distributional form of intensity data of this type. Often, the data vary over several orders of magnitude with a proportion of the distribution censored below a limit of detection (Audic and Claverie, 1998). In many cases (e.g., cDNA microarrays), the data for each gene are expressed as a ratio of two fluorescence intensity measures, leading to both left and right censoring problems. Third, the number of available measurements is typically small (i.e., 10 or 20 at most), at least in part, due to the high expense of these new technologies, and for some applications due to the limited availability of postmortem human tissue of sufficient quality for the analysis.

Note that the problem addressed in this paper (i.e., identifying differential expression levels in replicated microarray experiments) is only one of several important statistical problems associated with microarrays (see Claverie, 1999 for a review of the earlier work). Some of the earliest approaches to this problem involved the identification of groups of genes with similar function using cluster analysis (Eisen et al., 1998). Such methods are often referred to as "unsupervised" in that they do not consider auxiliary information regarding gene function or their relation to relevant outcomes of interest. Alternative "supervised" methods have been introduced to take advantage of auxiliary information by incorporating disease classes or related outcomes. For example, some investigators have focused on the identification of genes of similar function (e.g., classification of various tumor types) using various forms of discriminant analyses applied to the expression level profiles (Slonim et al., 2000; Dudoit et al., 2000; Brown et al., 1999; Ben-Dor et al., 2000). Hastie et al. (2000) have combined supervised and unsupervised approaches by developed "gene shaving" and "tree harvesting"' algorithms in which hierarchical clustering is used to reduce the dimensionality of the problem and the resulting gene clusters are then related to outcomes of interest (e.g., survival time). A similar approach using linear modeling of "genetic profiles" (i.e., identified using unsupervised hierarchical clustering) has been suggested by van Someren et al. (2000).

Methods for analysis of variance components in a single microarray slide have been developed by Newton et al. (2001); Kerr et al. (2000) and Sapir and Churchill (2000). In such experiments, the data for each gene consist of two fluorescence intensity measures, (R, G), representing the expression level of the gene in the red (Cy5) and green (Cy3) labeled mRNA samples. Typically, one of the two dyes corresponds to a pooled set of control tissues (e.g., normal colon tissue samples) and the other to an experimental tissue of interest (e.g., tissue taken from a malignant tumor in the colon). Dudoit et al. (2002) classify various "single-slide" methods in terms of whether they are based solely on the ratio (R/G) or incorporate overall abundance measured by the product *RG*. Early attempts at identification of differentially expressed genes (where *R* may reflect a control sample and *G* an experimental sample, or the reverse) relied upon an arbitrary cutoff for the ratio (R/G) (DeRisi et al., 1997; Schena et al., 1995, 1996). Chen et al. (1997) proposed a statistical criterion for selecting differentially expressed genes based on assuming normality and constant variance for the R/G ratio. Sapir and Churchill (2000) and Lee et al. (2000) considered a finite mixture model for the raw and log expression ratios and developed classification rules based on the posterior probability of a gene belonging to the differentially expressed component distribution. As pointed out by Newton et al. (2001) and Dudoit et al. (2002), limitations of these approaches include the false assumption of normality, and that the product *RG*, which contains information regarding the variability of R/G, is ignored. Newton and colleagues have proposed a hierarchical regression model based on the gamma distribution that permits variability to be proportional to the magnitude of expression level (i.e., the signal R/G depends on the overall abundance *RG*). Furthermore, Lee et al. (2000) found that there was considerable variability in the probability that a gene will be expressed, even across experimental replicates obtained from the same tissue. This finding clearly supports the need for obtaining a series of independent measurements from both control and experimental tissues.

Dudoit et al. (2002) consider a univariate approach to this problem by constructing a *t*-statistic for comparison of n_1 control hybridizations with n_2 experimental hybridizations, in terms of the log intensity ratios (R/G) for each gene. To adjust the corresponding *p-values* for each gene given the testing of the other $k-1$ genes, they use a permutation algorithm (Westfall and Young, 1993), in which the n columns of the entire $k \times n$ data matrix X are permuted into the $\binom{n}{n_1}$ possible control versus treatment allocations. Permuting the columns of X introduces independence between treatment assignment and gene expression, while retaining the dependence structure of the genes. The permutation distribution of the *t*-statistic for a particular gene is then provided by the empirical distribution of the permuted *t*-statistics. A clear advantage of this approach is that it makes no parametric assumption regarding the distribution of the expression levels, with the exception of a continuity assumption implicit in the use of the sample standard deviation as a measure of variability. It is also one of the first examples of an approach that attempts to rigorously address the multiple comparison problem.

Tusher et al. (2001) propose a method known as significance analysis of microarrays (SAM) that assigns a score to each gene. The numerical value of the score depends on the change of its' expression relative to the standard deviation. SAM also adjusts for multiple comparisons over genes by estimating the false discovery rate (FDR, Benjamini and Hochberg, 1995), which identifies a threshold, below which, group differences in expression levels are considered to be consistent with chance expectations. More recently, both parametric and nonparametric Bayesian methods have been proposed for analysis of microarray data. Efron et al. (2001), proposed a nonparametric empirical Bayesian mixture model to identify differentially expressed genes. Their method adjusts for multiple comparisons by relying on joint inference to determine significant differences in expression levels. They show that their method is statistically similar to FDR. Ibrahim et al. (2002) proposed a parametric Bayesian model for simultaneous analysis of multiple genes in microarray experiments. To identify a subset of genes that are differentially expressed in experimental and control conditions, they derived a Bayesian model selection rule.

An alternative strategy for the case of replicated microarray experiments has been proposed by Kerr et al. (2000), who proposed a linear model for log intensities. The basic idea is to control for differences between the dyes, genes and multiple arrays, using main effects and interactions in a fixed-effects ANOVA model. In contrast to Dudoit and colleagues who consider each gene individually, Kerr and colleagues consider genes to be a factor in the design with potentially thousands of levels. Dudoit et al. (2002) point out that for main effects in the model (e.g., the dye main effect), this implies a single error term for all genes, resulting in a potential loss of sensitivity relative to single gene approaches. Furthermore, gene-specific effects are estimated by interactions in the model, which lead to quite different standard errors than single gene approaches and no attempt is made to adjust for multiple comparisons, although presumably a form of permutation test could be applied here as well. In this and a related paper (Kerr and Churchill, 2001) are the first to seriously consider experimental design issues in replicated microarray experiments. This is a very important area for future research and Kerr and Churchill (2001), and Yang and Speed (2002) provide excellent foundations for further study. Zien et al. (2002) extend this work to consider sample size determination.

Finally, Dudoit et al. (2002), Newton et al. (2001) and Schuchhardt et al. (2000) consider experimental and statistical aspects of normalization of intensity measurements that identify and eliminate slide-specific sources of variation and enable a more meaningful synthesis of intensity measurements across different microarrays and experimental conditions. Such normalization methods range from use of internal standards or "houskeeping" genes, to dividing each gene intensity by the mean intensity over all genes on a given microarray slide, to robust local linear fits based on scatter-plot smoothing (Venables and Ripley, 1999; Dudoit et al., 2002). Depending on the application, any or all of these adjustments may prove useful for between-slide comparisons.

With the exception of the paper by Zien et al. (2002), a feature that is noticeably absent from this emerging area of statistical work is the consideration of statistical power. As the number of genes considered increase, control of the overall experiment-wise false positive rate must, of course, have a profound effect on the power to detect a real difference should one exist. Here we focus on replicated microarray experiments where power is the ability to reject the null hypothesis of no difference between experimental and control conditions when, in fact, a real difference exists.

In the following sections we propose an alternative approach to screening replicated genetic expression data for differentially expressed genes using sequentially applied simultaneous nonparametric prediction limits. These sequential prediction limits can be used to identify differentially expressed genes in replicated cDNA and oligonucleotide microarray experiments. Sequential prediction limits provide a useful addition to the growing arsenal of statistical tools for analysis of microarray data in general, and are particularly useful in those cases in which the number of control samples is large and the number of experimental samples is small. For example, assume that a large database of control expression levels is available (e.g., 50 independent samples). As an initial screening test, we may be interested in obtaining microarray data from a small number of experimental subjects (e.g., colon tumors from five patients with colon cancer). Alternatively, we may wish to sequentially collect experimental samples one at a time (or in small groups), until we have reasonable confidence that the identified genes are inconsistent with chance expectations, given the large number of genes being simultaneously tested. Of course, we must also be satisfied that we have sufficient statistical power to detect a differentially expressed gene should it actually be present. On the basis of these limited data, we can then determine which genes are good candidates for further study. Note that this type of comparison is different from the traditional two-sample approach in several ways. First, there is generally a large imbalance between the number of available control and experimental samples. Second, the number of experimental samples may be too small (e.g., 5 or 10 experimental subjects) to perform a valid two-sample comparison, particularly since we are routinely screening thousands of genes. Third, when using prediction limits, the comparison is specific to which of the two groups is considered to be the reference group. A prediction limit computed on the control data and used to screen experimental data, may not identify the same genes, had the group labeling been reversed. For this reason, the statistical methodology introduced here, may have limited utility where two experimental conditions are being compared. Furthermore, these differences should make it clear that the approach based on prediction limits is not intended to be a replacement for traditional two-sample tests, either parametric or nonparametric.

In the following sections, issues for both design and analysis of microarray experiments are addressed and comparison to traditional two-sample comparison methods is presented. The methodology is illustrated using data on colon cancer and results are compared to a permutation test.

2. Statistical development

2.1. cDNA versus oligonucleotide arrays

While the basic foundation of both cDNA and oligonucleotide microarrays are quite similar, their implementations are typically quite different. In oligonucleotide arrays, the analysis produces intensities for a single tissue sample (e.g., a colon tumor), whereas the cDNA array involves the within-array comparison of two tissue samples, typically, an experimental sample and pooled control. In the discussion that follows we assume that for a cDNA microarray, both the individual control samples and the experimental samples are both expressed as a ratio to the pooled control intensity for each gene. In this way, the result of both cDNA and oligonucleotide microarrays each yield a single intensity measurement for each gene and tissue sample. Of course, the intensity measurements for the two types of microarrays are not measured on a common scale and therefore, only replicated microarrays of a single type (i.e., cDNA or oligonucleotide) should be included in a particular analysis.

2.2. Normalization

Although intensity measurements are already the result of considerable numerical processing (e.g., methods for extracting signals from background, “signal segmentation,” and background correction), the scale of intensity measurements can vary from slide to slide for numerous systematic reasons. For example, differences in the dyes used can produce changes in the red and green intensity measurements and their ratio. Furthermore, the overall intensity for a given slide can change gene-specific variability given the observed dependence of the intensity ratio R/G on the overall intensity RG (Dudoit et al., 2002). As such, the intensity dependent normalization method described by Dudoit et al. (2002) may be preferable to global adjustments such as mean or median normalization for analysis of cDNA microarrays. Several authors (Dudoit et al., 2002; Kerr et al., 2000) have suggested that log-transformed intensity ratios be analyzed to decrease the dependence of variability on overall intensity. At a minimum, to normalize for differences between slides, the intensity level for each gene on a given slide should be adjusted for the overall intensity level of all genes (mean or median) on that slide. Clearly, normalization is a challenging area in need of further research. In our example involving oligonucleotide arrays (which do not involve intensity ratios) in normal colon tissue and colon tumors, we use a global adjustment (i.e., divide the overall mean intensity from each gene's intensity measurement). Since we use nonparametric prediction limits, the data were not transformed prior to analysis.

2.3. Sequential sampling

To provide a foundation, consider the case in which we have n control slides and m experimental slides. Assume that the slides represent $n + m$ independent tissue samples and if we are using a cDNA microarray, intensity measurements are expressed as a ratio to the pooled control sample (i.e., a composite of the n control tissue samples). This is the standard replicated microarray comparison strategy described by Dudoit et al. (2002). The goal is to identify only those genes that are differentially expressed in the experimental and control populations. Using replicated data, Dudoit et al. (2002) compare the mean concentrations of the n control and m experimental samples respectively using a t-statistic, or alternatively, the n control samples can be used to construct a prediction interval for the mean of the m experimental samples (Guttman, 1970; Hahn and Meeker, 1991). If the observed experimental mean lies within the prediction interval then we conclude with $100(1 - \alpha)\%$ confidence that the experimental data could have been drawn from the same distribution as the control data. In fact, at this point, the two strategies are actually quite similar and have relatively similar power characteristics, depending on the relative sizes of n and m. As an alternative approach, however, we might proceed sequentially by again, obtaining n control samples, but in this case we collect n_i experimental samples in each of up to $i=1, \ldots, p$ subsets. After each subset is obtained, we determine which, if any, genes have a mean intensity that is outside of the prediction interval in that subset and all preceding subsets. We continue until (a) all genes are within the prediction interval in at least one subset, or (b) we reach p subsets. Those genes that have mean concentration consistently above(below) the prediction interval in all p subsets are considered to be differentially expressed. Note that this sequential strategy is, in fact, quite different from the traditional two-sample comparison approach. For example, if $n_i = 1$ then we can refine our list of candidate differentially expressed genes after each new subject is obtained.

For example, suppose that we have tissue samples available from $n=50$ postmortem brains from non-psychiatric patients and we wish to screen individual schizophrenic patients until we have reasonable confidence, say 95%, that one or more genes out of the 12,000 genes tested are differentially expressed. Note that 12,000 genes is the approximate number of genes that can be simultaneously evaluated given current microarray technology. Numerous possible decision rules exists, however, consider the simple decision rule that the gene is differentially expressed if the expression levels for all n_i schizophrenic samples are above the maximum control sample. Note that this decision rule is equivalent to selecting p sequential replicates each of size $n_i = 1$. For some applications, requiring that all n_i samples exceed the control maximum may be too stringent a decision rule. As a less stringent alternative, we may wish to include the median of say $n_i = 3$ samples (denoted $s_i = 2$). The question now becomes, how many sequential replicates (p) of size $n_i = 3$ must be obtained such that we can have 95% confidence that one or more genes out of the 12,000 genes tested are truly differentially expressed. Note that by replicates, we are referring to independent sets of samples, obtained sequentially in time.

In the general case, our objective is to use the n control measurements to derive an upper (lower) intensity bound for some or all of the n_i experimental samples in at least one of the p experimental subsets. Although previously not considered in this way, this is a classical problem in statistical prediction (Guttman, 1970; Hahn and Meeker, 1991) and the sequential case has been studied in considerable detail in environmental statistics (see Gibbons, 1994 for an overview). The primary advantages of the prediction limit approach over the traditional two sample comparison approach are (1) that it provides an approach to screening large numbers of genes when only small numbers of experimental subjects are available, and (2) the experimental subjects may be tested sequentially, leading to important intermediate information on potential differentially expressed genes. The question is whether or not adequate statistical power can be obtained.

2.4. A sequential nonparametric prediction interval

Given the previous characterization of the problem and the questionable distributional form of the intensity measurements, a natural approach to the solution of this problem is to proceed nonparametrically. To fix ideas, for a particular gene (i.e., cDNA or expressed sequence tag (EST) intensity on a microarray), let us define an upper “prediction limit” as the uth largest source intensity out of the n control sources. If $u = n$ then our prediction limit is the largest control intensity for that gene. If $u = n - 1$ then our prediction limit is the second largest control intensity for that gene. A natural advantage of using $u < n$ is that it provides an automatic adjustment for outliers in that the largest $n - u$ values are removed. Note, however, that the larger the difference between u and n the lower the overall confidence keeping everything else equal.

Now consider the experimental subjects. Assume that we have n_i experimental subjects obtained sequentially in up to p experimental replicates or subsets, and let s_i be the number of experimental measurements required to be contained within the interval. For example, if $n_i = 5$ and we wish to have the median experimental value in subset i below the prediction limit, then $s_i = 3$. In contrast, if $n_i = 1$, then $s_i = 1$ and subset i contains a single measurement. Note that as an example, $s_i = n_i = 5$ and $p = 1$ is equivalent to $s_i = n_i = 1$ and $p = 5$. A gene is selected only if the s_ith experimental sample exceeds the prediction limit in all p subsets.

The questions of interest are: (1) What is the probability of a chance exceedance in all p experimental subsets for different values of n, u, n_i, s_i and p? (2) How is this probability affected by varying numbers of genes (i.e., k)? (3) What is the power to detect a real difference between expressed genes in two populations for a given statistical strategy ?

To address these questions, let $y_{(s_i,n_i)}$ denote the s_ith largest value (i.e., order statistic) from the n_i sources in subset i ($i = 1, \ldots, p$) and $x_{(u,n)}$ denote the uth order statistic from a control sample of size n. We can now express the previous discussion mathematically as

$$Pr\{y_{(s_1,n_1)} > x_{(u,n)}, y_{(s_2,n_2)} > x_{(u,n)}, \ldots, y_{(s_p,n_p)} > x_{(u,n)}\} \leqslant \alpha^*/k, \tag{1}$$

where α^* is the experiment-wise false positive rate (e.g., $\alpha^* = 0.05$). In order to evaluate the joint probability in Eq. (1) note that the probability density function of the uth order statistic from a sample of size n (i.e., $x_{(u,n)}$) is

$$g(x; n, u) = \frac{n!}{(u-1)!(n-u)!}[F(x)]^{u-1}[1 - F(x)]^{n-u} \cdot f(x), \tag{2}$$

where

$$\begin{aligned}\int_{-\infty}^{\infty}[F(x)]^{u-1}[1 - F(x)]^{n-u} \cdot f(x)\,\mathrm{d}(x) &= \left[\frac{n!}{(u-1)!(n-u)!}\right]^{-1}\\ &= \left[\frac{n(n-1)!}{(u-1)!(n-u)!}\right]^{-1}\\ &= \left[n\binom{n-1}{u-1}\right]^{-1},\end{aligned} \tag{3}$$

(see Sarhan and Greenberg, 1962). Since

$$Pr\{y_{(j,m)} \geqslant x\} = \sum_{i=0}^{j-1}\binom{m}{i}[F(x)]^{i}[1 - F(x)]^{m-i}, \tag{4}$$

the joint probability in (1) becomes

$$\frac{n}{\sum_{i=1}^{p} n_i + n}\sum_{j_1=0}^{s_1-1}\sum_{j_2=0}^{s_2-1}\cdots\sum_{j_p=0}^{s_p-1}\frac{\binom{n_1}{j_1}\binom{n_2}{j_2}\cdots\binom{n_p}{j_p}\binom{n-1}{u-1}}{\binom{\sum_{i=1}^{p} n_i + n - 1}{\sum_{i=1}^{p} j_i + u - 1}} = \alpha, \tag{5}$$

(Chou and Owen, 1986; Gibbons, 1990, 1991, 1994; Davis and McNichols, 1999). A lower bound on the probability of the s_ith largest experimental measurement (e.g., the median) in all p subsets exceeding the uth largest control measurement for any of the k expressed genes is given by $\alpha^* = 1-(1-\alpha)^k$. One minus this probability provides the corresponding confidence level. Ultimately, for practical applications we would typically like the overall confidence level to be approximately 95% (i.e., $\alpha^* \leqslant 0.05$). Note that this simple Bonferroni/Sidak-type adjustment for the effects of multiple genes is overly conservative in that it assumes that the genes are uncorrelated. Nevertheless, for design purposes (i.e., determining provisional values of n, u, n_i, s_i and p) it is a quite useful lower bound on the experiment-wise Type I error rate, which can then be verified, or improved upon, using randomization or permutation tests (Dudoit et al., 2002; Westfall and Young, 1993) once the data have been collected.

To determine if a gene's intensity level is significantly decreased in experimental relative to control conditions, let $x_{(l,n)}$ denote the lth smallest value from a control subset of size n. Then Eq. (1) becomes

$$\Pr\left\{y_{(s_1,n_1)} < x_{(l,n)},\, y_{(s_2,n_2)} < x_{(l,n)},\, \ldots,\, y_{(s_p,n_p)} < x_{(l,n)}\right\} \leqslant \alpha^*/k, \tag{6}$$

which leads to

$$\frac{n}{\sum_{i=1}^{p} n_i + n} \sum_{j_1=s_1}^{n_1} \sum_{j_2=s_2}^{n_2} \cdots \sum_{j_p=s_p}^{n_p} \frac{\binom{n_1}{j_1}\binom{n_2}{j_2}\cdots\binom{n_p}{j_p}\binom{n-1}{l-1}}{\binom{\sum_{i=1}^{p} n_i + n - 1}{\sum_{i=1}^{p} j_i + l - 1}} = \alpha. \tag{7}$$

The probability of the s_ith largest experimental measurement (e.g., the median) in all p subsets being simultaneously less than the lth smallest control measurement for any of the k expressed genes is given by $\alpha^* = 1 - (1-\alpha)^k$, and one minus this probability provides the corresponding confidence level. For a two-sided interval, we compute upper and lower limits each with probability $\alpha^*/2$. If s_i is the median of the n_i experimental measurements, and the n_i are odd, then the upper prediction limit can be computed with probability approximately $\alpha^*/2$ and the lower limit is simply the expression level of the $l = (n - u + 1)$th ordered measurement.

The reader should note that selection of the tuning parameters (i.e., n, u, n_i, s_i and p) is very much dependent on the particular problem to which the method is applied, and the effect of changing the tuning parameters on the false positive and false negative rates for the microarray study as a whole. As such, it is difficult to provide definitive guidelines of what should be done in general. In the following illustrations and example, we outline the process by which the tuning parameters are selected and illustrate the effect of changing these parameters on statistical power and gene-wide false positive rates.

3. Illustration

To illustrate the method, consider the previously described problem of using the maximum of $n = 50$ control expression measurements as an upper prediction limit (i.e., $u = n = 50$) for a series of n_i sequentially obtained experimental samples for each of $k = 12,000$ cDNA intensities, where k is the number of cDNAs on the gene chip or microarray. From Eq. (5), we obtain an overall confidence level of less than 0.01 for both $n_i = 1$ and $n_i = 2$, indicating an extremely high likelihood (i.e., 99%) of at least one significant association by chance alone. For $n_i = 3$, the overall confidence level increases to 0.60 or a 40% chance of a significant result by chance alone. However, for $n_i = 4$, the overall confidence level increases to 0.96 or only a 4% chance of a significant result by chance alone. This rather astounding result indicates that with a background set of $n = 50$ control samples, but only $n_i = 4$ experimental samples, we can have 95% confidence that a gene that was expressed above the maximum control sample in all four experimental samples is not due to chance alone, despite simultaneous testing of 12,000 genes. In fact, this is a lower bound on the true confidence level to the extent that the genes are correlated. This simple procedure would then permit rapid screening of huge numbers of genes using microarrays and the resulting differentially expressed genes could then be validated using, for example RT-PCR (reverse transcription-polymerase chain reaction). RT-PCR is a way to amplify RNA and/or DNA, is quite sensitive, and can be used to validate microarray assays. RNase protection assays, which are also fairly sensitive, can also be used to validate differentially expressed candidate genes from microarray assays.

3.1. Statistical power

One might be tempted to stop here and design the experiment accordingly, since we have shown that a reasonable experiment-wise false positive rate can be achieved with 50 control and only four experimental subjects. Note, however, that we have said nothing about the likelihood that such a strategy will, in fact, have sufficient power to detect a true difference. To this end, we can evaluate the power of the test for detecting a true control versus experimental group difference via simulation. As a conservative approach, we can simulate control and experimental measurements from a standard normal distribution with control and experimental mean values separated by differences of 0–5 standard deviation units. Different strategies can then be compared on the basis of their sampling requirements and their power to detect a real difference of a given magnitude. For example, given the previous example of $n = u = 50$, $n_i = s_i = 4$, $p = 1$, and $k = 12,000$, power of 80% is achieved for a control versus experimental difference of 4.02σ. While this is a large effect size, we are asking a great deal from only four experimental measurements.

There are, however, several ways to increase statistical power of the procedure. For example, what if we decreased u to the second largest control expression level? Doing so decreases our overall confidence to 83%, however, 80% power is now achieved for a 3.58σ difference between the true means of control and experimental populations. If 83% confidence is unacceptable, we can increase n_i and s_i to five and confidence increases to 98%. Of course, the increase in confidence is not without a price. With $n_i = s_i = 5$, 80% power is achieved at 3.70σ.

Additional gains in statistical power can be achieved by not requiring that all n_i experimental samples be above(below) the upper(lower) prediction limit. For example, consider the case in which we require the median of n_i experimental samples to be below the upper prediction limit. As an illustration, let us return to our original example of $n = 50$ control samples and $k = 12,000$ genes, and consider a single ($p = 1$) experimental group with $n_i = 5$ subjects for which we want the median expression level $s_i = 3$ to be below the largest control measurement ($u = 50$). As one might expect, the overall confidence level is poor, in this case approximately 1%. However, if we collect a second set of $n_i = 5$ experimental measurements (i.e., $p = 2$), and require that both medians be above the prediction limit for a gene to be differentially expressed, then confidence is now 97.5% and 80% power is now achieved at only 3.14σ. Further increases in power can be obtained by adding replicates and decreasing u. For example, with $u = 47$ i.e., the prediction limit is the expression level of the fourth largest control measurement, and $p = 3$ replicates each of size n_i, confidence is 93% and power of 80% is obtained at 2.46σ. In this case, we would collect $p = 3$ sets of $n_i = 5$ sequentially, where after each set of five experimental measurements, we could decrease the list of genes that are potential candidates for differential expression to only those whose median was above(below) the upper(lower) prediction limit up to that replicate.

These examples are just a tiny subset of the possible comparative strategies that could have been developed for this illustration. The choice among them should be based on the strategy that achieves maximal statistical power while achieving the nominal experiment-wise false positive rate, given available resources.

4. Bootstrap and permutation tests

As previously noted, a potential limitation of the simple Bonferroni/Sidak type adjustment for multiple genes, is that it does not incorporate the effect of inter-gene communication, that is, correlation of expression levels among subsets of genes. While the simple adjustment may be useful for design purposes prior to obtaining the data, as the correlation among the gene expression levels increases, the simple uniform adjustment for multiple genes can be overly conservative. Once the data have been collected, better estimates of the true experiment-wise Type I error rate and adjusted gene-specific confidence levels can be obtained via bootstrap or permutation tests as suggested by Dudoit et al. (2002), following the work of Westfall and Young (1993). The primary difference between bootstrap and permutation resampling is that the former is done with replacement whereas the latter is done without replacement (Westfall and Young, 1993). The disadvantage of the permutation approach is that we must enumerate all of the possible patterns or maintain an index of which patterns have already been sampled. As the number of samples and replicates increase, the number of permutations becomes too large for practical purposes. In the current context, the following simple randomization algorithm can be used. Following Westfall and Young (1993) and Manly (1997), we can either fully or randomly assign the $N = n + \sum_i^p n_i$ measurements to each of the $p+1$ groups (i.e., n controls and p experimental subsets each of size n_i) without regard to their actual group membership. Over all permutations or a large number of randomizations, we can then compute the actual probability that the s_ith largest measurement in all p randomized experimental groups is above(below) the randomized prediction interval for one or more genes. The probability is obtained by simply enumerating the number of instances in which all p of the randomized experimental medians are above(below) the upper(lower) randomized prediction limit for at least one gene and dividing by the number of randomizations. The same algorithm can be used for all possible permutations. If randomization is used, several thousand resamples should be obtained (Manly, 1997).

5. Comparison to traditional two-sample tests

At what point is a traditional two-sample nonparametric test more powerful than a sequential prediction limit? First, the sequential nature of the prediction limits is in many instances a distinct advantage over the traditional two-sample test (e.g., the Mann–Whitney U-test). For the traditional test, we must wait until all of the data have been collected before deriving any inference regarding differentially expressed genes, whereas as previously shown for the sequential prediction limits, information on differential expression is available after each individual sample or subset of samples. Second, for small numbers of experimental samples (e.g., less than 10) and large numbers of genes (e.g., 10,000 or more), traditional nonparametric two-sample tests are typically not even an option, in that they cannot provide an overall experiment-wise false positive rate of 5% or less. For our original example of $n=50$, $n_i=4$, and $k=12,000$ genes, the traditional nonparametric test is powerless to detect an experimental versus control difference when adjusted for the multiple comparisons using a Bonferroni-type correction. Nevertheless, there are many situations in which the traditional two-sample test provides increased statistical power over the sequential prediction limit described here. In the following, we explore these conditions using simulation.

For the two-sample nonparametric test, we use Bonferroni adjustment of the individual gene Type I error rate to control for the simultaneous testing of all k genes. This is similar to the sequential prediction limits, and both methods can be improved using bootstrap and permutation tests as described in the previous section. For example, assume that we have $n=20$ control samples and $m=20$ experimental samples. For the sequential prediction limit, we can divide the $m=20$ experimental samples into $p=4$ sequential replicates of $n_i=5$ samples each. Setting the prediction limit for the $p=4$ medians ($s_i=3$) to the largest control measurement $u=20$ and considering a microarray with $k=12,000$ genes, the prediction limit achieves a 96% confidence level and 80% power is achieved for a difference of 2.92σ. By contrast, 80% power of the traditional two-sample Mann–Whitney U-test is achieved for a difference of only 2.20σ. As such, the additional power of the traditional two-sample test for the case of a balanced design and larger numbers of experimental samples, may in many cases overshadow the benefits of sequential testing.

6. Application to colon cancer data

We now apply the sequential prediction limits to a previously collected dataset to illustrate their use. Note that this not a perfect example in that the number of control samples is, in fact, smaller than the number of experimental samples. This is the case for most, if not all, datasets that have been collected in this emerging field. Furthermore, these data were not collected sequentially, however, we have randomly assigned the 40 experimental samples to $p=8$ subsets of $n_i=5$ samples each to illustrate the sequential process.

Alon et al. (1999) collected gene expression data from 40 tumor and 22 normal colon tissues using Affymetrix oligonucleotide microarrays (Mack et al., 1998), complementary to more than 6500 human genes. In analysis of these data, Alon and co-workers focused on identifying genes that regulate each other or have similar cellular function, using cluster analysis. Their analysis was based on a subset of 2000 genes with highest minimal intensity across the tissue samples. The treatment of raw data from the Affymetrix oligonucleotide arrays is described in detail in the original paper (Alon et al., 1999). In both the original analysis and in our analysis, the data for each array were normalized by dividing the intensity of each EST on an array by the mean intensity of all ESTs on the array, to compensate for possible systematic intensity variations between arrays. Results of the original cluster analysis revealed that tumor and normal tissues were separated into two distinct clusters

Table 1
Number of differentially expressed genes and experiment-wise confidence level after each sequential replicate sample of size $n_i = 5$—colon cancer data

No. of sequential samples (p)	No. of identified genes	Overall confidence independence	Overall confidence randomized
1	446	<0.001	<0.001
2	136	<0.001	0.032
3	78	0.004	0.286
4	57	0.211	0.616
5	47	0.603	0.817
6	33	0.830	0.914
7	23	0.923	0.960
8	18	0.962	0.979

based on tissue composition (i.e., genes that are related to the development of smooth muscles, where tumors generally had a low muscle content).

In our reanalysis of these data, to illustrate the sequential process, we randomly split the 40 tumor tissues into $p = 8$ discrete subsets, each with $n_i = 5$ tissues, and developed a prediction bound for the median measurement ($s_i = 3$) in each subset, based on the interval defined by the 4th ($l = 4$) and 19th ($u = 19$) ordered measurements out of the set of $n = 22$ control measurements, simultaneously for all $k = 2000$ genes. $p = 8$ was selected because it provided approximately 95% overall confidence, equally divided the tissues into subsets with an odd number of measurements, and maximized statistical power relative to all of the alternatives. The simple independence based estimated confidence associated with this interval is 96.2%. The adjusted overall confidence level based on the randomization algorithm is 97.9%, which is remarkably close to the independence based estimate. Table 1 displays the number of identified differentially expressed genes and associated independence based and randomization based overall confidence levels as a function of the number of replicates (p) obtained. As can clearly be seen in Table 1, poor overall confidence and large numbers of false positives are identified with insufficient numbers of replicates. Table 1 also reveals that the independence based and randomization based overall confidence levels are quite similar for high and low levels of confidence, but considerably different for intermediate values in the .2 to .8 range. In general, however, we are only interested in high levels of confidence, for example 95% or more. Note that for $p = 7$ replicates, the independence based confidence level is .923, but the adjusted confidence level is .960. With seven replicates, an additional five differentially expressed genes were identified.

It is of interest to note that if we had simply pooled all 40 tumor tissue samples into a single group, the prediction interval defined by the 4th and 19th largest control levels would only have a 13.3% adjusted overall confidence level of containing the median tumor measurement for all 2000 genes. This contrasts with a 97.9% adjusted overall confidence level when the data are split into $p = 8$ replicates each of size $n_i = 5$.

Table 2 displays the 18 genes (out of 2000 genes) that had a median value outside the prediction interval for all $p = 8$ subsets, and the additional 5 genes identified had we stopped after $p = 7$ replicates. As in the original analysis of these data performed by Alon and co-workers, we observed a decrease in muscle specific gene products, including

Table 2
Genes that significantly differentiate tumor from normal colon tissue

Gene number	Sequence	Name	PTB**
Gene expression levels high in tumor tissue			
M26697	Gene	Human nucleolar protein (B23) mRNA, complete code	0.00080
M36981	Gene	Human putative NDP kinase (nm23-H2S) mRNA, complete cds	0.22820
X14958	Gene	Human hmgI mRNA for high mobility group protein Y	0.03040
M22382	Gene	Mitochondrial matrix protein P1 precursor (Human)	0.00001
X12671	Gene	Human gene for hnRNP core protein A1	0.00001
H40095	3’ UTR	Macrophage migration inhibitory factor (Human)	0.00460
T86473	3’ UTR	Nucleoside diphosphate kinase A (Human)	0.00160
R36977	3’ UTR	Transcription factor IIIA	0.00160
T40454	3’ UTR	Antigenic surface determinant protein OA3 precursor (Homo Sapiens)	0.04100
X67155	Gene	Mitotic kinesin-like protein-1 with Alu repetetive element	0.38700
J05032	Gene	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds	0.00080
H08393	3’ UTR	Collagen alpha 2(XI) chain (Homo Sapiens)	0.00001
H38185	3’ UTR	CYL-COA-Binding Protien (Homo sapiens)*	0.94200
R64115	3’ UTR	Adenosylhomocysteinase (Homo sapiens)*	0.01560
X63629	Gene	H.sapiens mRNA for p cadherin*	0.01560
Gene expression levels low in tumor tissue			
M76378	Gene	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.00600
M63391	Gene	Human desmin gene, complete cds	0.00140
Z50753	Gene	H.sapiens mRNA for GCAP-II/uroguanylin precursor	0.01000
R87126	3’ UTR	Myosin heavy chain, nonmuscle (gallus gallus)	0.00001
J02854	Gene	Myosin regulatory light chain 2, smooth muscle isoform	0.00140
M36634	Gene	Human vasoactive intestinal peptide (VIP) mRNA, complete cds	0.00440
H43887	3’ UTR	Complement factor D precursor (Homo sapiens)*	0.00560
T94350	3’ UTR	Peripheral myelin protein 22(Homo sapiens)*	0.86160

*Identified after $p = 7$ replicates,
**PTB = Permutation test probability.

myosin regulatory light chain and myosin heavy chain in tumor tissues compared to control tissues. However, in addition, our analysis revealed several significant differences in gene expression between tumors and controls that were not identified in the initial analysis and which are of direct relevance to colon cancer (Table 2). The first is significant overexpression of nm23-H2S mRNA. Overexpression of this gene product has previously been linked to early stages of colorectal carcinoma (Martinez et al., 1995). The second is overexpression of mRNA for the cell surface antigenic determinant OA3. This surface antigen has previously been reported to be highly specific to ovarian carcinomas, and has been used as a target for immunotherapy of those tumors (Campbell et al., 1992). Our analysis suggests that OA3 surface antigen is frequently overexpressed in colonic carcinomas, and may provide an important target for immunotherapy of colon cancer. The third, cDNA that we found increased corresponded to macrophage migration inhibitory factor (MIF). MIF is a cytokine that has been linked to an inhibition of natural killer cell-mediated lytic activity and is increased in certain human uveal melanomas (Perou et al., 1999). Repp et al. (2000) suggest that MIF is produced as a means to allow uveal melanoma cells to escape NK-cell mediated lysis. While the relevance of this activity remains to be confirmed with respect to colon carcinomas, it raises the intriguing possibility that other tumor cells may possess similar mechanisms to suppress certain immune responses.

Another interesting overexpressed gene product, encoding P-cadherin, is a member of the cadherin superfamily of adhesion proteins. P-cadherin is aberrantly expressed in many adenocarcinomas and appears to be preferentially expressed in invasive tumors and has been implicated as a useful marker of aggressive clinical behavior in certain cervical cancers (Han et al., 2000). Fatty acyl-CoA binding protein (ACBP) has been shown to be elevated in transformed colon cells. (Gossett et al., 1997). The authors indicate that the overexpression of ACBP may be a consequence of the transformation process. Finally, we have shown that S-adenosylhomocysteinase (SAM-hydroloase) is overexpressed in this study. This enzyme has been shown to inactivate certain anticancer nucleoside analogues and may be increased in response to specific therapeutic strategies. While it remains to be established whether there is a correlation between individual patient therapies and the elevation of this gene product, the clinical relevance of this observation may have important implications with respect to treatment strategy.

As a point of comparison, we reanalyzed the colon cancer data using the permutation based method of Westfall and Young (1993), as described by Dudoit et al. (2002). The permutation based probabilities for the 23 genes identified by the prediction limit method are displayed in Table 2. Of these, 19 are less than 0.05, indicating, in general, good agreement between the two methods. Of the four genes that were identified by the prediction limits but not by the permutation tests (i.e., $p > 0.05$ in Table 2), they have overall means between the two groups that are similar due to the presence of a few elevated values in the control group. As an illustration, Table 3 presents the actual ordered data for genes X14958 which had a permutation-based probability of 0.00001 and X67155 which had a permutation-based probability of 0.38700. For X14958, there is large separation between the two groups and in all cases, the median of the 8 subsets of 5 colon cancer samples is well above the nonparametric upper prediction limit (i.e., the 19th order statistic). In contrast, for X67155, the means are quite similar and there is considerable overlap between the two distributions. However, in all cases, the median of the 8 subsamples of 5 colon cancer patients also exceeds the upper nonparametric prediction limit, indicating a significant difference. This is not surprising given the nonparametric nature of the prediction limit test.

Note however, that the permutation-based method identified an additional 36 genes that were not identified by the prediction limit method. Without independent confirmation, it is impossible to determine if these additional genes represent false positive results for the permutation test or false negatives for the prediction limit methodology. However, we have evaluated these genes from a functional perspective to determine their relevance to colon cancer.

The 36 genes represented several categories of proteins. Five genes are involved in protein synthesis, two genes are involved in cell cycle control, two genes are involved in DNA binding or replication or repair, five genes are involved in chromatin or transcription, and four genes are involved in signal transduction. It is difficult to conclude that these genes

Table 3
Comparison of permutation test and nonparametric sequential prediction limit for one gene in which they agree (X12671) and for one gene in which they disagree (X67155)

X12671 gene Human gene for hnRNP core protein A1
Permutation based probability = 0.00001

Controls ($n = 22$, $\bar{x} = 0.8406$, $s = 0.4636$, LPL = 0.442, UPL = 1.242)

0.261	0.303	0.414	**0.442**	0.463	0.483	0.546	0.550	0.558	0.616	0.728
0.731	0.752	0.781	0.917	1.229	1.237	1.238	**1.242**	1.283	1.731	1.989

Patients ($n = 40$, $\bar{x} = 1.9688$, $s = 0.8756$)

		Median		
0.956	0.979	**1.584**	2.067	2.214
0.618	1.489	**1.558**	2.361	2.428
1.154	1.311	**1.829**	2.362	3.117
1.495	2.124	**2.524**	2.538	3.964
0.664	1.760	**2.343**	2.355	3.392
0.830	1.712	**1.929**	2.287	2.714
0.719	0.990	**1.459**	2.289	3.300

X67155 gene Mitotic kinesin-like protein-1 with Alu repetitive element
Permutation based probability = 0.38700

Controls ($n = 22$, $\bar{x} = 0.2438$, $s = 0.0837$, LPL = 0.152, UPL = 0.291)

0.117	0.134	0.138	**0.152**	0.174	0.203	0.207	0.215	0.216	0.225	0.232
0.239	0.249	0.256	0.263	0.263	0.274	0.290	**0.291**	0.375	0.412	0.438

Patients ($n = 40$, $\bar{x} = 0.3316$, $s = 0.1108$)

		Median		
0.224	0.247	**0.329**	0.448	0.584
0.248	0.295	**0.361**	0.390	0.427
0.075	0.262	**0.319**	0.350	0.387
0.236	0.284	**0.353**	0.404	0.461
0.192	0.295	**0.321**	0.350	0.366
0.190	0.192	**0.346**	0.398	0.586
0.195	0.283	**0.358**	0.368	0.410
0.144	0.237	**0.394**	0.453	0.503

are not related to colon cancer in that there is literature suggesting that each of these processes either has a relationship with some form of cancer or they might be anticipated to be upregulated in rapidly dividing cells. In addition, one gene, Cyctatin, is of potential interest in colon cancer and would be hypothesized to be downregulated which, in fact, it was. Four of the additional genes identified are less likely to be associated with cancerous cells. They may be false positives but it is difficult to know until the data are confirmed independently. The remaining genes, indicated as OPRF (open reading frame), are not as yet identified.

These results indicate that many different statistical methods may be quite relevant in analysis of gene expression levels. The prediction limits presented here have the advantage of (a) being nonparametric and therefore robust to outliers, and (b) being conservative in that they only identify genes with consistent increases and decreases across multiple subsamples of the population.

7. Discussion

The goal of the statistical methodology developed here is to identify one or more genes that are associated with a particular biological or experimental condition of interest. The primary contribution of this work is to provide such an identification that is beyond chance expectations due to the large number of genes that are routinely screened on a microarray, and to understand the statistical power associated with such a decision rule. The primary advantage of the sequential approach over a traditional two-sample comparison is that intermediate information regarding differential expression is available after each new experimental sample (or subset of samples) is obtained. A second advantage is that when a large number of control samples are available, statistical inference with small numbers (i.e., < 10) of experimental samples is possible, even when large numbers of genes (e.g., $> 10,000$) are simultaneously investigated. By contrast, when larger and comparable numbers of measurements are available, traditional two-sample nonparametric tests generally have increased power relative to sequential nonparametric prediction limits. As such, the sequential prediction limits presented here are most useful as a screening tool for large numbers of genes and small numbers of experimental observations. RT-PCR, for example, can then be used to validate the differential expression of these new candidate genes. In addition, a linear combination of the candidate genes in this more manageable subset may then be identified that maximally differentiate control and experimental tissues (e.g., a discriminant function analysis) and that can be used for the purpose of classification, risk assessment and perhaps even as a diagnostic tool.

Application of this statistical methodology to the colon carcinoma data collected by Alon and co-workers clearly illustrated the utility of the approach. Several of the differentially expressed genes identified in our analysis have been previously identified in various forms of cancer, but not previously in connection with colon cancer. These results go beyond those previously reported from the cluster analysis described in the original report by Alon and co-workers. Of course, these differentially expressed candidate genes must now be validated by another method, such as RT-PCR.

As a final note, the methodology described here is applicable to any large scale automated screening system (e.g., high throughput receptor binding screening) where the outputs are measured on a continuous scale. As pointed out for the case of gene expression data, the methodology is uniquely suited to those cases in which the number of potential indicators is large and the number of available subjects is small. The robustness shown with respect to correlation among subsets of the multiple endpoints further insures the widespread utility of the general approach as a tool for experimental design of screening studies and the use of permutation tests provide more realistic estimates of adjusted gene-specific confidence levels which take the correlation between genes into account. To provide ease of application, we have developed a “probability calculator” which computes single gene and experiment-wise confidence levels and corresponding statistical power for any set of values of $N,\ n_i,\ p,\ u,\ l,\ s_i$, and k. The probability calculator is freely available at www.uic.edu/labs/biostat/.

References

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

Audic, S., Claverie, J., 1998. Visualizing the competitive recognition of TATA-boxes in vertebrate promoters. Trends Gen. 14, 10–11.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z., 2000. Tissue classification with gene expression profiles. J. Comput. Biol. 7, 559–583.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57, 289–300.

Brown, M.P.S., Grundy, W.N., Lin, D., Sugnet, C., Ares, M., Haussler, D., 1999. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Nat. Acad. Sci. 97, 262–267.

Campbell, I., Freemont, P., Foulkes, W., Trowsdale, J., 1992. An ovarian tumor marker with homology to vaccinia virus contains an IgV-like region and multiple transmembrane domains. Cancer Res. 52, 5416–5420.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high density DNA microarrays. Science 274, 610–614.

Chen, Y., Dougherty, E.R., Bittner, M.L., 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Opt. 2, 364–374.

Chou, Y., Owen, D., 1986. One-sided distribution-free simultaneous prediction limits for *p* future samples. J. Quality Technol. 18, 96–98.

Claverie, J., 1999. Computational methods for the identification of differential and coordinated gene expression. Human Mol. Gen. 8, 1821–1832.

DeRisi, J.L., Iyer, V.R., Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686.

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

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96, 1151–1160.

Eisen, M., Spellman, P., Brown, P., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. National Acad. Sci. 95, 14863–14868.

Gibbons, R., 1990. A general statistical procedure for Ground-Water Detection Monitoring at waste disposal facilities. Ground Water 28, 235–243.

Gibbons, R., 1991. Some additional nonparametric prediction limits for ground-water monitoring at waste disposal facilities. Ground Water 29, 729–736.

Gibbons, R., 1994. Statistical Methods for Groundwater Monitoring, Wiley, New York.

Gossett, R.E., Schroeder, F., Gunn, J.M., Kier, A.B., 1997. Expression of fatty acyl-CoA binding proteins in colon cells: response to butyrate and transformation. Lipids 32, 577–585.

Guttman, I., 1970. Statistical Tolerance Regions: Classical and Bayesian, Hafner, Darien Connecticut.

Hahn, G., Meeker, W., 1991. Statistical Intervals: A Guide for Practitioners, Wiley, New York.

Han, A.C., Edelson, M.I., Soler, A.P., Knudsen, K.A., Lifschitz-Mercer, B., Czernobilsky, B., Rosenblum, N.G., Salazar, H., 2000. Cadherin expression in glandular tumors of the cervix. Cancer 89, 2053–2058.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., Botstein, D., 2000. Gene shaving: a new class of clustering methods for expression arrays. Technical Report, Stanford University.

Ibrahim, J.G., Chen, M.H., Gray, R., 2002. Bayesian models for gene expression with DNA microarray data. J. Amer. Statist. Assoc. 97, 88–99.

Kerr, M.K., Churchill, G.A., 2001. Experimental design for gene expression microarrays. Biostatistics 2, 183–201.

Kerr, M.K., Martin, M., Churchill, G.A., 2000. Analysis of variance for gene expression microarray data. J. of Comput. Biol., in press.

Lee, M., Kuo, F., Whitmore, G., Sklar, J., 2000. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. Proc. Nat. Acad. Sci. 97, 9834–9839.

Mack, D., Tom, E., Mahadev, M., Dong, H., Mittman, M., Dee, S., Levine, A., Gingeras, T., Lockhart, D., 1998. Deciphering molecular circuitry using high-density DNA arrays. In: Mihich, K., Croce, C. (Eds.), Biology of Tumors. Plenum, New York, pp. 123–131.

Manly, B.F.J., 1997. Randomization Bootstrap and Monte Carlo Methods in Biology, Chapman and Hall, London.

Martinez, J., Prevot, S., Nordlinger, B., Nguyen, T., Lacarriere, Y., Munier, A., Lascu, I., Vaillant, J., Capeau, J., Lacombe, M., 1995. Overexpression of nm23-H1 and nm23-H2 genes in colorectal carcinomas and loss of nm23-H1 expression in advanced tumor stages. Gut 37, 712–720.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui, K.W., 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J. Comput. Biol. 8, 37–52.

Perou, C., Jeffrey, S., vandeRijn, M., Rees, C., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., Lashkari, D., Shalon, D., Brown, P., Botstein, D., 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Science 96, 9212–9217.

Repp, A., Mayhew, E., Apte, S., Niederkorn, J., 2000. Human uveal melanoma cells produce macrophage migration-inhibitory factor to prevent lysis by NK cells. J. Immunol. 165, 710–715.

Sapir, M., Churchill, G.A., 2000. Estimating the posterior probability of gene expression from microarray data. Unpublished manuscript, The Jackson Laboratory. (http://www.jax.org/research/churchill).

Sarhan, A., Greenberg, B., 1962. Contributions to Order Statistics, Wiley, New York.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davis, R.W., 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Nat. Acad. Sci. 93, 10614–10619.

Slonim, D., Tamayo, P., Mesirov, J., Golub, T.R., Lander, E., 2000. Class prediction and discovery using gene expression data. In: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, April 8–11, 2000, Tokyo, Japan.

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci. 98, 5116–5121.

van Someren, E., Wessels, L.F.A., Reinders, M.J.T., 2000. Linear modeling of genetic networks from experimental data. In: Proceedings of the Intelligent Systems in Molecular Biology Conference, August 19–23, 2000, La Jolla, California.

Venables, W.N., Ripley, B.D., 1999. Modern Applied Statistics with S-Plus, Springer, New York.

Westfall, P.H., Young, S.S., 1993. Resampling-based Multiple Testing: Examples and Methods for *p*-Value Adjustment, Wiley, New York.

Yang, Y.H., Speed, T., 2002. Design issues for cDNA microarray experiments. Nature Rev. 3, 579–588.

Zien, A., Fluck, J., Lengauer, T., 2002. Microarrays: how many do you need ? Assoc. Comput. Mach.