Biometrics 60, 774–782
September 2004

Multiple-Testing Strategy for Analyzing cDNA Array Data on Gene Expression

Robert R. Delongchamp,[1,*] John F. Bowyer,[2] James J. Chen,[1] and Ralph L. Kodell[1]

[1]Division of Biometry and Risk Assessment, National Center for Toxicological Research, Jefferson, Arkansas 72079, U.S.A.

[2]Division of Neurotoxicology, National Center for Toxicological Research, Jefferson, Arkansas 72079, U.S.A.

[*]*email*: rdelongchamp@nctr.fda.gov

Summary. An objective of many functional genomics studies is to estimate treatment-induced changes in gene expression. cDNA arrays interrogate each tissue sample for the levels of mRNA for hundreds to tens of thousands of genes, and the use of this technology leads to a multitude of treatment contrasts. By-gene hypotheses tests evaluate the evidence supporting no effect, but selecting a significance level requires dealing with the multitude of comparisons. The p-values from these tests order the genes such that a p-value cutoff divides the genes into two sets. Ideally one set would contain the affected genes and the other would contain the unaffected genes. However, the set of genes selected as affected will have false positives, i.e., genes that are not affected by treatment. Likewise, the other set of genes, selected as unaffected, will contain false negatives, i.e., genes that are affected. A plot of the observed p-values $(1 - p)$ versus their expectation under a uniform [0, 1] distribution allows one to estimate the number of true null hypotheses. With this estimate, the false positive rates and false negative rates associated with any p-value cutoff can be estimated. When computed for a range of cutoffs, these rates summarize the ability of the study to resolve effects. In our work, we are more interested in selecting most of the affected genes rather than protecting against a few false positives. An optimum cutoff, i.e., the best set given the data, depends upon the relative cost of falsely classifying a gene as affected versus the cost of falsely classifying a gene as unaffected. We select the cutoff by a decision-theoretic method analogous to methods developed for receiver operating characteristic curves. In addition, we estimate the false discovery rate and the false nondiscovery rate associated with any cutoff value. Two functional genomics studies that were designed to assess a treatment effect are used to illustrate how the methods allowed the investigators to determine a cutoff to suit their research goals.

Key words: Decision theory; False discovery rate; False nondiscovery rate; p-value plot; ROC curves; Subset selection.

1. Introduction

The statistical analysis of gene expression data from cDNA array experiments faces two challenges that arise from characteristics of the technology. First, intensity measurements representing levels of gene expression can be influenced by confounding factors such as spatial heterogeneity and signal saturation that arise from variability in the technically complex steps involved in cDNA array methodology (Nguyen et al., 2002). Second, cDNA array experiments simultaneously interrogate many hundreds or thousands of genes. As a result, the chance for false discovery of genes perceived to be differentially expressed between comparison groups can be unacceptably high (Craig, Black, and Doerge, 2003). In this article, we are concerned with the second challenge, that of dealing with the multitude of statistical tests on individual genes that typically arise with cDNA data. It will be assumed that the first challenge has successfully been met through appropriate normalization of the expression data (Chen et al., 2003).

Statistical analysis of normalized gene expression data often begins with comparisons between two groups (e.g., treated versus control, diseased tissue versus normal tissue) of the multitude of genes by way of individual statistical tests (e.g., Student's t-tests). Frequently, this first analysis is followed by a second analysis that seeks to determine whether particular genetic profiles (groups of coexpressed genes or functional relationships among genes) are predictive of specific biological outcomes. Often, this involves some type of clustering of genes with similar expression patterns. An analysis of the first type can serve as a screen to select a subset of genes for which to conduct an analysis of the second type. This article will consider statistical aspects of the first type of analysis, primarily as a "filtering" exercise preliminary to a profiling analysis.

Because many individual tests assessing differential expression are conducted, the family-wise type I error rate (FWE: the probability of making at least one false positive error) can be very large. Statistical approaches that are customarily used to control the FWE, such as the procedures of Holm

(1979), Hochberg (1988), or Hommel (1988), were not designed for thousands, or even hundreds, of individual comparisons. When applied to cDNA array data on a multitude of genes, they will tend to screen out all but a handful of genes that show extreme differential expression. If the objective is to identify a few genes that differ with treatment(s), such an analysis may well be appropriate. If, on the other hand, the objective is to determine functional relationships among genes that have been affected by treatment, then other selection criteria may be more appropriate. In particular, we consider criteria that do not eliminate as many genes from further analysis. Controlling the false discovery rate (FDR: the expected proportion of incorrectly rejected null hypotheses) has been proposed by Benjamini and Hochberg (1995) as an alternative to controlling the FWE in multicomparison problems. The use of this approach in preliminary genome scans has been discussed (Weller et al., 1998; Zaykin et al., 2000). In addition, somewhat informal approaches for screening the interrogated genes are sometimes used, such as eliminating all but 100 genes with the largest t-statistics (Li and Taylor, 2001).

The need exists for statistical procedures to reliably eliminate most of the unaffected genes from further consideration while keeping essentially all genes whose functions are potentially affected by the biological or toxicological phenomena under study. This article proposes statistical methodology for the analysis and interpretation of cDNA array data based on the use of p-value plots (plots of significance probabilities obtained from multiple individual tests) and receiver operating characteristic (ROC) curves. The strategy is designed primarily as a filter, to screen out many hundreds (or thousands) of genes from consideration prior to genetic profiling; however, it can also be used to identify a relatively small number of truly differentially expressed genes.

2. Method

2.1 *Expected Rates of Misclassification in Testing Multiple Genes*

In a study involving a hypothesis test that is repeated for m genes, we assume that the null hypothesis is true (expression is unaffected) for some genes, m_0 genes, and false (expression is affected) for the remainder, $m - m_0$ genes. The p-values that are generated from these tests, $\{p_i : i{=}1, \ldots, m\}$, order the genes, and standard practice is to reject the null hypothesis for the genes where the p-value is less than a specified significance level, i.e., $p_i < \alpha$. Thus, a choice of significance level divides the genes into two sets: those declared to differ significantly, m_α genes, and those declared not to differ significantly, $m - m_\alpha$ genes. Ideally, the m_α genes, which are *selected* because of their small p-value, would correspond to the $m - m_0$ *affected* genes for which the null hypothesis is false. However, it is expected from the definition of a p-value that the m_α selected genes will be contaminated with $\alpha\, m_0$ unaffected genes for which the null hypothesis is true. Table 1 gives expected numbers in a two-way classification of affected and selected genes, where only the number of true hypotheses, m_0, is unknown.

Assuming that m_0 can be estimated, then various misclassification rates defined from the entries in Table 1 can be estimated. The FDR estimates the proportion of the selected genes that are expected to be false positives. Tsai, Hsueh, and Chen (2003) investigated various formal definitions of the FDR and associated statistical properties. We adopt what Tsai et al. (2003) termed the *empirical FDR* for testing at level α, that is,

$$\mathrm{FDR}(\alpha) = \frac{\alpha\, m_0}{m_\alpha}.$$

The expected proportion of affected genes among the genes not selected, the false nondiscovery rate (FNR), is

$$\mathrm{FNR}(\alpha) = \frac{m - m_0 - (m_\alpha - \alpha\, m_0)}{m - m_\alpha}.$$

The fraction of genes not selected (FNS) among the affected genes is

$$\mathrm{FNS}(\alpha) = \frac{m - m_0 - (m_\alpha - \alpha\, m_0)}{m - m_0}.$$

The complement, $1 - \mathrm{FNS}(\alpha) = (m_\alpha - \alpha m_0)/(m - m_0)$, represents the *sensitivity* of the testing procedure, where $1 - \alpha$ is the *specificity*.

For our purposes, these rates are computed as a function of the number of genes selected. Let $p_{(k)}$ denote the kth-order statistic for the p-values, $\{p_i : i{=}1, \ldots, m\}$. Then these misclassification rates can be defined for the set of genes determined by the k smallest p-values. This simply amounts to substituting k for m_α and $p_{(k)}$ for α in the previous formulae, i.e.,

$$\mathrm{FDR}(k) = \frac{p_{(k)}\, m_0}{k}, \tag{1}$$

$$\mathrm{FNR}(k) = \frac{m - m_0 - (k - p_{(k)}\, m_0)}{m - k}, \tag{2}$$

and

$$\mathrm{FNS}(k) = \frac{m - m_0 - (k - p_{(k)}\, m_0)}{m - m_0}. \tag{3}$$

These rates quantify the degree to which the selected genes reflect the affected genes. Hence, they are basic summaries for the study that can be computed from an estimate, say $\hat{m}_0$, of the number of true null hypotheses.

The empirical FDR is a good approximation to the conditional FDR, i.e., conditional on selecting k genes, without the complexities of specifying distribution(s) under the alternative hypothesis (Tsai et al., 2003). It will suffice for most applications and is used in the exposition of the proposed approach. The empirical estimates of FDR, FNR, and FNS only approximate rigorously defined probabilities, so they sometimes take on values outside the interval [0, 1]. Generally this

Table 1
Expected numbers in a crossclassification of selected genes with affected genes

	Selected	Not selected	Total
Affected	$m_\alpha - \alpha m_0$	$m - m_0 - (m_\alpha - \alpha m_0)$	$m - m_0$
Not affected	αm_0	$m_0 - \alpha m_0$	m_0
Total	m_α	$m - m_\alpha$	m

occurs at uninteresting values of α, e.g., $\alpha > 0.5$, and in our work such values are set to the appropriate boundary.

While it is important in genetic profiling to eliminate as many nuisance genes from further consideration as possible (low FDR), it is perhaps more important to ensure that few of the genes that are differentially expressed are removed from consideration prematurely (low FNR). Controlling the FDR (or FNR) in the sense of Benjamini and Hochberg (2000) is not the objective here. Rather, the rejection region is determined first, in the spirit of Storey (2002), and then corresponding estimates of the FDR, FNR, and FNS are made. As will be shown in the examples in Section 3, the success of the proposed strategy depends on having a good estimate of m_0.

2.2 *Estimating the Number of True Null Hypotheses*

Schweder and Spjotvoll (1982) plotted i versus $q_{(i)} = 1 - p_{(m-i+1)}$, which they called a p-value plot. Under the null hypothesis (also assumes independence between tests), the p-values are distributed uniformly over the interval $[0, 1]$. Under this uniform distribution, the expected value of $q_{(i)}$ is i/m. Linearity in a p-value plot indicates agreement between the empirical cumulative distribution of the p-values and the cumulative uniform distribution. In this context one can test whether $m_0 = m$, i.e., that none of the genes are affected by treatment. For example, the Kolmogorov statistic,

$$D_{\max} = \max_i \left\{ \left| q_{(i)} - \frac{i}{m} \right| : i = 1, \ldots, m \right\},$$

could be used for this purpose (Conover, 1971).

More often, observed p-values will be a mixture of m_0 observations from the null distribution, i.e., uniform [0,1], and $m - m_0$ observations from alternative distributions, which provide more support for smaller p-values. As a didactic device suppose that the $m - m_0$ p-values from alternative distributions are always less than $1 - \pi$. Then smaller values in the Schweder and Spjotvoll plot, i.e., $q_{(i)} < \pi$, are from the null distribution, and their distribution on this subinterval is equivalent to a sample that is uniformly distributed on the interval $[0, \pi]$. Hence, the plot will be linear in this range and will have slope m_0. Furthermore, the number of genes with $q_{(i)} < \pi$, say k, has expectation, $\pi\, m_0$, and variance, $m_0\, \pi(1 - \pi)$. This suggests an estimate for the number of true null hypotheses, $\hat{m}_0 = k/\pi$ (Schweder and Spjotvoll, 1982), i.e., the estimated slope of the p-value plot in the linear range.

Figure 1 is a Schweder and Spjotvoll plot that was computer generated under the scenario outlined above with $\pi = 0.8$, $m_0 = 500$, and $m = 1000$. In particular, $\{q_i : i = 1, \ldots, m\}$ is a mixture of 500 random samples from a uniform [0, 1], i.e., representing the null distribution, and 500 samples from a uniform [0.8, 1], i.e., representing an alternative distribution albeit somewhat artificial. This figure illustrates the basic ideas in a setting that is unencumbered by complications that are usually present in actual studies. The empirical cumulative distribution of the $\{q_i : i = 1, \ldots, m\}$, henceforth denoted by $\hat{F}(q)$, is the irregular curve. The expected "null hypothesis" component of $\hat{F}(q)$ is the dashed line that connects the origin to the point ($q = 1$, $i = 500$). For $q < 0.8$, $\hat{F}(q)$ is consistent with this line. Hence, a linear approximation to $\hat{F}(q)$ provides

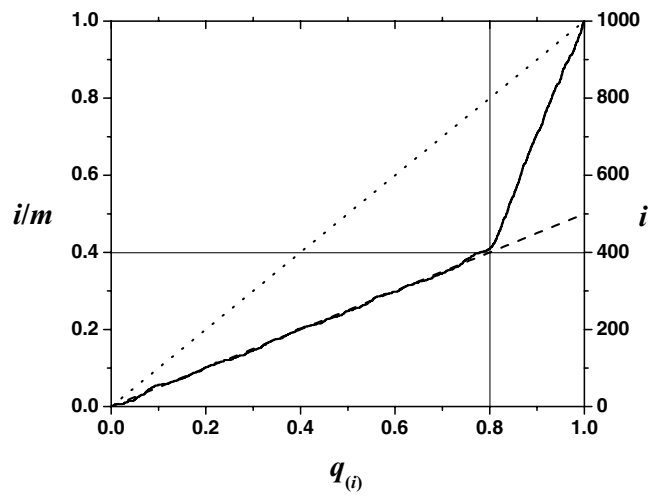

Figure 1. Schweder and Spjotvoll plot of simulated data, 500 observations with a uniform distribution on [0, 1] and 500 observations with a uniform distribution on [0.8, 1]. Expected values for a uniform distribution on [0, 1] (dashed line) for 500 observations.

an estimate of m_0, where it intersects the right-hand vertical axis.

The estimate of m_0 that was suggested above, namely $\hat{m}_0 = k/\pi$, is such an estimate when π is known. It corresponds to a line (not shown) connecting the origin to $\hat{m}_0$ on the right-hand vertical axis. Although somewhat informal, a visual inspection of a p-value plot to assess the region of linearity, i.e., a visual assessment of π, produces reasonable estimates of m_0. This simple procedure works well with the simulated data (Figure 1). Serious departures from an expected Schweder and Spjotvoll plot will diagnose data sets where estimating m_0 is problematic.

Recently, there has been a revival of interest in algorithms for estimating m_0. Hsueh, Chen, and Kodell (2003) investigated several methods (Schweder and Spjotvoll, 1982; Benjamini and Hochberg, 2000; Storey, 2002). They showed via Monte Carlo simulation that their own least squares (LS) method that uses information from both null and alternative distributions was the least biased method. For the sake of computational clarity, the LS method of Hsueh et al. (2003) will be adopted and used for further development and illustration of the proposed testing strategy. However, we caution against an uncritical use of any procedure. The Schweder and Spjotvoll plot should always be evaluated, and the estimated m_0 should be reasonable in the context of the plot. During the review of this article, we became aware of a mixture-model approach to the estimation of m_0 (Allison et al., 2002), which also gives reasonable estimates in many cases.

2.3 *Using ROC Curves to Determine an Optimal p-Value Cutoff*

ROC curves, which originated in the context of electronic signal detection, are commonly used in medicine to assess the quality of clinical diagnostic tests (Zweig and Campbell, 1993). ROC curves are derived from large groups of patients whose disease status (positive or negative) is known, by

plotting the true positive rate (sensitivity) of a diagnostic test against the test's false positive rate (1 − specificity) for various decision thresholds, or cutoff values, of the test. An ROC-like strategy can be adapted to the present context by substituting gene-specific hypotheses for patients and p-values for diagnostic test results. False null hypotheses (affected genes) correspond to patients having the disease and true null hypotheses (unaffected genes) to patients without the disease. Unlike the diagnostic testing setting, the truth or falseness of each hypothesis is unknown; however, having an estimate of the number of true hypotheses enables the construction and use of ROC curves in the present context.

One way to construct an ROC plot for a cDNA array experiment is to plot the estimated sensitivity,

$$1 - \mathrm{FNS}(k) = \frac{k - p_{(k)}\,\hat{m}_0}{m - \hat{m}_0},$$

against 1 − specificity, i.e., $p_{(k)}$. This gives a *nonparametric* ROC curve. The closer an ROC curve is to the top-left corner of the unit square the better the quality of the experiment in differentiating false from true hypotheses. The closer the curve is to the 45° diagonal the less able the experimental data to resolve differences. The area under an ROC plot can be used as a measure of test quality and provides a basis for comparison of different diagnostic procedures (Zweig and Campbell, 1993) or, in this case, different hypothesis tests about the genes. Factors that may contribute to differences in areas under the curves include effect sizes, sample sizes, and underlying variability.

In addition to the FDR and FNR, the presently defined ROC curve is a tool derived from the estimate of m_0 that can be used in deciding which genes to declare significantly affected, that is, which cutoff to use. Let FP\$ and FN\$ denote, respectively, the costs of false positive and false negative errors, which may be combinations of dollar costs and other factors. An "optimal" cutoff, $p_{(k)}$, can be found by identifying the combination of sensitivity and specificity that minimizes the total expected cost of making errors, i.e.,

$$\begin{aligned}
&\mathrm{Prob(accept\ a\ false\ null\ hypothesis)} \cdot \mathrm{FN\$} \\
&\quad + \mathrm{Prob(reject\ a\ true\ null\ hypothesis)} \cdot \mathrm{FP\$} \\
&= \mathrm{Prob(false)} \cdot \mathrm{Prob(accept\,|\,false)} \cdot \mathrm{FN\$} \\
&\quad + \mathrm{Prob(true)(Prob(reject\,|\,true)} \cdot \mathrm{FP\$} \\
&= P \cdot (1 - \mathrm{Se}) \cdot \mathrm{FN\$} + (1 - P) \cdot (1 - \mathrm{Sp}) \cdot \mathrm{FP\$},
\end{aligned}$$

where P is the proportion of p-values that correspond to false hypotheses, Se stands for sensitivity, and Sp for specificity. One can show that the minimum of the cost function is defined by

$$\mathrm{Se} = \frac{\mathrm{FP\$}}{\mathrm{FN\$}} \frac{1 - P}{P} (1 - \mathrm{Sp}) + C,$$

that is,

$$\mathrm{Se} = \frac{\mathrm{FP\$}}{\mathrm{FN\$}} \frac{m_0}{m - m_0} (1 - \mathrm{Sp}) + C,$$

where C is the maximum Se intercept among lines that intersect the ROC curve and have positive slope, (FP\$/FN\$) × $[m_0/(m - m_0)]$. Hence, the point (1 − Sp, Se) corresponding to the optimal cutoff value, $p_{(k)}$, may be found by moving a line with the above slope down from above and from the left of the ROC plot until it intersects the plot (Zweig and Campbell, 1993). Using the ratio of the costs of false positive and false negative errors, along with the odds of a true hypothesis, $m_0/(m - m_0)$, enables an investigator to choose a decision threshold in an alternative, but complementary, manner to choosing one based on the FDR and FNR. Clearly, the slope of the line defining the optimal cutoff point will approach unity as the weights on (1 − Se) and (1 − Sp) in the above cost function, i.e., $P \cdot \mathrm{FN\$} = [(m - m_0)/m] \cdot \mathrm{FN\$}$ and $(1 - P) \cdot \mathrm{FP\$} = (m_0/m) \cdot \mathrm{FP\$}$, approach equality.

If the profiling analysis that is to be conducted on the selected genes will focus less on individual genes and more on combinations of genes (e.g., data mining), then the FP\$/FN\$ ratio might be substantially less than unity because little preliminary filtering may be necessary or desirable. The corresponding FDR (FNR) will tend to be large (small). If, on the other hand, follow-up verification of observed expression changes is planned (e.g., quantitative reverse transcription polymerase chain reaction), then the filtering will need to be heavy, i.e., the FP\$/FN\$ ratio could be substantially greater than unity. The corresponding FDR (FNR) will tend to be small (large). In the first case, simply controlling the FWE to select genes for a profiling analysis would not be useful and in the second case it might be too restrictive. Without an estimate of m_0, little insight into the consequences of an FWE-controlling strategy could be gained. The decision-theoretic ROC strategy, which is enabled by having an estimate of m_0 derived from a p-value plot, allows an investigator to strike an appropriate balance between the FP\$ and FN\$. Corresponding estimates of the FDR and FNR inform the investigator of the expected consequences of selecting particular cutoff thresholds for filtering.

3. Examples

Our estimate of m_0 and its use in selecting affected genes is illustrated with data from a study that interrogated expression of 1185 genes in several regions of the rat brain following amphetamine or saline treatment (Jakab and Bowyer, 2002; Delongchamp, Harris, and Bower, 2003; Bowyer et al., 2004). Figure 2a plots $q_{(i)}$ versus i/m for the hypothesis test of a treatment difference in gene expression in the posteriolateral cortical amygdaloid nucleus (PLCo). The plot has values for 752 genes, which represents an "expressed" subset of the 1185 interrogated genes. That is, this subset represents genes that were judged as being expressed above background in at least one of studied brain tissues (Delongchamp et al., 2003). In our original analyses of these data, we estimated m_0 using the least slope estimate of Hochberg and Benjamini (1990). Their admittedly conservative procedure, which was devised mainly to sharpen sequentially rejective procedures (Hochberg and Benjamini, 1990), gave an estimate of 739. This is shown by the dashed line in Figure 2a. The LS estimate of Hsueh et al. (2003), $\hat{m}_0 = 645$, is depicted by the right-hand intercept of the solid straight line in Figure 2a. It is evident in the graph that this estimate, 645, fits the lower $q_{(i)}$ values better than 739. Curvature over the range of this graph suggests that the

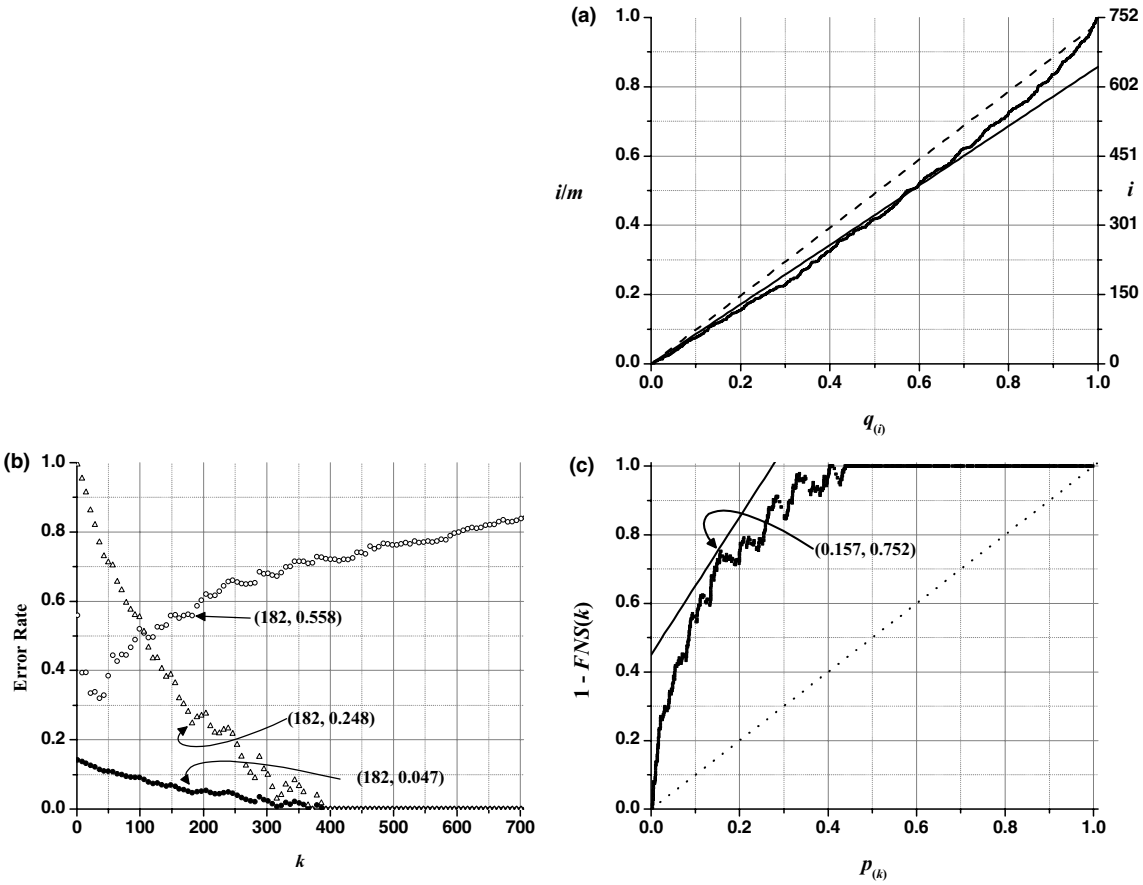

Figure 2. (a) Schweder and Spjotvoll plot for the 752 hypothesis tests that amphetamine treatment affects gene expression in the rat PLCo. Expected values for a uniform distribution (solid line) using the least squares method of Hsueh et al., $\hat{m}_0 = 645$, and expected values for a uniform distribution (dashed line) using the least slope estimate of Benjamini and Hochberg, $\hat{m}_0 = 739$. (b) The false discovery rate (○), false nondiscovery rate (●), and proportion of missed genes (Δ) using $\hat{m}_0 = 645$. (c) ROC curves for the hypotheses that amphetamine treatment affects gene expression in the rat PLCo. The optimum rejection level assuming FN\$ = 3 FP\$ is the intersection of the solid line with the ROC curve.

conceptual uniform distribution is contaminated throughout. Nevertheless, the observed p-values are reasonably consistent with a mixture of 645 uniform deviates contaminated with 107 p-values from a distribution with stochastically smaller p-values.

The practical issue is how well can we identify the 100+ affected genes that are implied by Figure 2a. Assuming 645 true null hypotheses, Figure 2b plots estimates of the FDR, FNR, and FNS as functions of the number of selected genes, equations (1), (2), and (3), respectively. Since the FDR never dips below 0.30, any p-value-based selection of "affected" genes will include many false positives. In Delongchamp et al. (2003), we exploited the dip in the FDR that occurs in these data and chose $k = 50$. This choice minimized the FDR for $k > 0$ and yields a somewhat conventional p-value, $p_{(50)} = 0.025$. Typically, the minimum for the FDR occurs at the boundary, so this criterion is not generally applicable. We also considered sets that select "all affected genes." Figure 2b indicates that selecting half of the genes ($k = 376$) would capture 95% of the affected genes, $1 - \text{FNS}(376) = 0.952$. This list implies a p-value cutoff of $p_{(376)} = 0.425$. Although this list is mostly false positives, $\text{FDR}(376) = 0.729$, it halves the number of selected genes with few of the affected genes being misclassified, and it may be a reasonable starting place for ascertaining affected biological processes that involve multiple genes.

A more formal approach is to use the ROC curve, as described in Section 2. Figure 2c gives the nonparametric ROC plot for the 752 expressed genes in the PLCo region of the brain. Because we place a higher cost on prematurely discarding potentially important genes than on falsely selecting unimportant genes, we take FN\$ = 3 FP\$. These relative costs give the optimal slope of the Se versus $(1 - \text{Sp})$ line as $(1/3)$ $[645/(752 - 645)] = 2.01$. The desired intersection of the ROC and this line will be around $(1 - \text{Sp}, \text{Se}) = (0.157, 0.752)$, i.e., a cutoff of $p_{(k)} = 0.157$. The FNS is the complement of the sensitivity. So, $\text{Se} = 0.752$ corresponds to $\text{FNS}(k) = 0.248$. From Figure 2b, this corresponds to selecting $k = 182$ of the 752 genes as significantly affected by treatment. In the sense of FP\$ and FN\$ trade-offs, this would be optimal.

From Figure 2b, the FDR and FNR values at $k = 182$ would be 0.558 and 0.047, respectively. Thus, it is estimated that 102

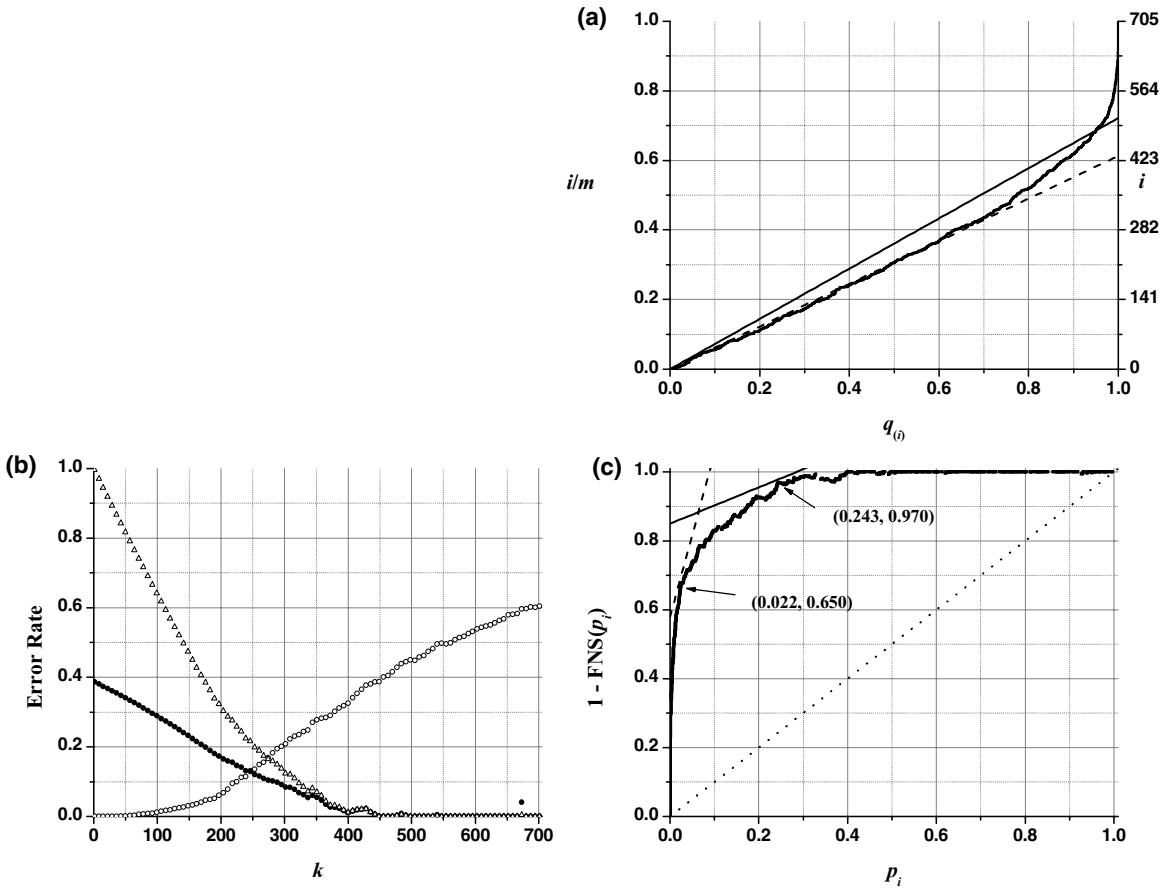

Figure 3. (a) Schweder and Spjotvoll plot for the 705 hypothesis tests that cisplatin affects gene expression in rat kidney. Expected values for a uniform distribution (solid line) using the least squares method of Hsueh et al., $\hat{m}_0 = 509$, and expected values for a uniform distribution (dashed line) using a version of an estimate of Storey, $\hat{m}_0 = 432$. (b) The false discovery rate (○), false nondiscovery rate (●), and proportion of missed genes (Δ) using $\hat{m}_0 = 432$. (c) ROC curves for the hypothesis that cisplatin affects gene expression in rat kidney. The optimum rejection level assuming FN\$ = 3 FP\$ is the intersection of the solid line with the ROC curve and the optimum rejection level assuming 3 FN\$ = FP\$ is the intersection of the dashed line with the ROC curve.

of the 182 selected genes may be false positives, which means that approximately 80 of the estimated 107 affected genes are included. On the other hand, of the 570 genes not selected, it is estimated that 27 are truly affected genes that are not selected.

The second example is a toxicogenomic study of gene expression levels in kidney samples from rats dosed with cisplatin (Thompson, Mirsky, and Sistare, 2002; Chen et al., 2003; Hsueh et al., 2003). This study interrogated 700 rat genes plus five genes from other species to monitor nonspecific background binding of labeled DNA. Figure 3a plots p-values from the two-sample t-test of treatment versus control. The LS estimate, $\hat{m}_0 = 509$, is depicted as the right-hand intercept of the solid line. The computation of this estimate and several alternatives are discussed in the context of these data elsewhere (Hsueh et al., 2003). Visual inspection of Figure 3a suggests that a smaller value of m_0, such as the dashed line representing $\hat{m}_0 = 432$, is more consistent with a hypothesized uniform distribution among lower values of $q_{(i)}$. The estimate, $\hat{m}_0 = 432$, is a version of an estimate of Storey (2002) that is discussed in Hsueh et al. (2003). The point is that a given method may not produce the "best" estimate with all data sets. Here, "best" means agreement with a uniform distribution in the lower values of $q_{(i)}$. So for this illustration, we will use $\hat{m}_0 = 432$.

Figure 3b plots the FDR, FNR, and FNS as functions of the number of selected genes. In this case, the differences are resolved well enough that most genes can be identified. The ROC curve in Figure 3c reflects the higher resolving power of the test compared to the first example, in that the curve in the upper left-hand corner of the plot is well off the diagonal. Using the same ratio of cost functions as above, the optimal slope is estimated to be $(1/3)[432/(705 - 432)] = 0.527$. This slope is shallow enough that intersection of the line with the ROC plot occurs near maximum sensitivity, at a p-value cutoff of 0.243. This corresponds to selecting 370 genes as being affected, which can be approximated using Figure 3b. From Figure 3b, the corresponding FDR is 0.284, and the FNR and FNS rates are small, 0.024 and 0.030, reflecting the higher cost assigned to false negatives. Inverting the cost ratio gives the

optimal slope, $3[432/(705-432)] = 4.747$, which corresponds to a p-value cutoff of 0.022 and the selection of 187 genes. This choice assigns higher costs to false positives and gives FDR = 0.051, FNR = 0.184, and FNS = 0.350.

4. Discussion

The methodology proposed herein is predicated on estimating the number of true null hypotheses, m_0, among a multitude of tests for individual genes in a cDNA array experiment. The p-value plot from these tests is a basic tool for estimating m_0 and helps to turn a multiple-comparison liability into an asset in that the large number of tests (the liability) gives the p-value plot good resolution (the asset). The large number of comparisons also gives good resolution to ROC-like curves used to estimate optimal cutoff points for declaring statistical significance. The utility of these methods was illustrated with expression data from two studies, one using radiolabeled macroarrays (Figure 2) and the other labeling microarrays using two fluorescent dyes (Figure 3).

No variance estimates have been proposed for the estimates of m_0 discussed in Hsueh et al. (2003). To our knowledge, none have been published. In some of our work, we have employed a somewhat ad hoc estimate based on the binomial variance, $\mathrm{var}(k) = m_0\pi(1-\pi)$, where π is estimated from the "linear range." This approach results in $\mathrm{var}(\hat{m}_0) \approx k(1-\hat{\pi})/\hat{\pi}^2$. When the m genes are independent, simulations indicate that Wald-type confidence intervals for m_0 are reasonably accurate (results not shown). Since this calculation does not account for the variance inflation associated with correlations among genes, it would be too small in general. Simulated values of the variance were reported for each of the estimates of m_0 studied by Hsueh et al. (2003), including the methods used in the two examples of Section 3. The effect of this variation on the estimation of variously defined FDRs was simulated in a follow-up study by Tsai et al. (2003). Their results indicated that estimates of m_0 and the FDR are sufficiently reproducible to be useful in a filtering analysis. Nevertheless, because estimates of not only the FDR but also the FNR and FNS are affected by the error in estimation of m_0, this is an area where further research is needed.

In addition to being used to estimate the number of true null hypotheses, determine appropriate cutoffs, and estimate associated misclassification error rates, p-value plots and ROC curves also give information on the general quality of the data for resolving differences. The ability to distinguish differences in gene expression in the second example was better than in the first example. This appears as a departure from linearity that is larger for the p-value plot in Figure 3a than the plot in Figure 2a. Similarly, the ROC curve in Figure 3c is farther from the diagonal and has more area underneath it than the ROC curve in Figure 2c, indicating the higher "diagnostic" quality of the curve in Figure 3c. This difference in resolution would be expected from the toxicity associated with the treatments used in these two studies and it is unlikely to be due to inherent differences in macroarrays versus microarrays. In the macroarray experiment (Figure 2), the gene expression in brain tissue of the control rats was compared to the same tissue in rats treated with a moderate neurotoxin. In the microarray experiment (Figure 3), control kidney tissue was compared to exposed tissue 7 days after 5 mg/kg cisplatin exposure, known to be toxic to the kidney. In general, p-value plots and ROC curves derived from cDNA array data become more informative as the number of replications increases and the average treatment effect size increases.

Real data sets are more complicated than the simulated data in Figure 1. Three complications occur with some regularity: (1) The statistical test will have limited power to distinguish the treatment effects for some genes, (2) the statistical tests are correlated, and (3) the null distribution of the p-values is not uniform. These complications muddy the interpretation of the resultant plots, but the plot generally remains a useful summary of the observed data.

The first complication implies that there will be some large p-values from the alternative distributions. In more extreme cases, there may not be an obvious "linear range" interval that is not contaminated by p-values from the alternative distribution. This contamination is seen as upward curvature throughout the interval [0, 1]. The plot in Figure 2a illustrates this behavior for an experiment which was interpreted further in Section 3. Upward curvature is the trademark symptom that the null hypothesis is not true for some of the genes. Curvature throughout the range of p-values implies that the study lacks sufficient power to resolve many of the effects. For most estimation methods, this type of curvature tends to bias $\hat{m}_0$ such that it overestimates the number of null hypotheses. Hence, the plot indicates that there are at least $m - \hat{m}_0$ affected genes, and the rates, FDR(k), FNR(k), and FNS(k), quantify the ability of the observed p-values to identify these genes. For a given effect size, this bias becomes negligible as the sample sizes increase.

Correlations among the genes arise because all of the genes are measured on the same array and in the same sample. Hence, gene-by-gene tests to assess differential expression are correlated to some degree. Correlation is also induced by the normalization that is typically applied to array data. Hsueh et al. (2003) simulated correlated data and they observed that moderate correlations among tests can substantially increase the variance of $\hat{m}_0$ but they do not affect its expectation. Hence, lack of independence does not bias estimates of m_0. In our experience, the correlations observed among genes rarely exceed 0.25, which is at or below the simulated correlations (Hsueh et al., 2003; Tsai et al., 2003).

A nonuniform null distribution, the third complication, is observed as an unusual shape in the p-value plot, e.g., inverted u-shapes (convex) or inverted s-shapes. Essentially, any p-value plot where the empirical cumulative distribution rises significantly above the diagonal indicates problems with the assumption that the null distribution is uniform. A nonuniform null distribution indicates that the test statistic does not have the assumed distribution, and one should review test assumptions. An incorrectly specified model is one way that this problem arises and this should be ruled out. In our work, the experimental design may include replicate measurements on each experimental unit and it deals with components of variation associated with normalization, array, dye, and batch, as well as treatment effects. Under more ad hoc designs, there may be sources of variance that are not accounted for, and quite possibly cannot be accounted for by any attempt at normalization or modeling. Even in a "correctly" specified model, convex shapes can occur. Often, the test statistic is

based upon a few degrees of freedom, and asymptotic normality may not be adequately realized. Bootstrapping or nonparametric alternatives may correct this although the more egregious cases often have too few degrees of freedom for such methods to have any power. In addition, measurements vary in precision from array to array and the assumption of constant residual variance will generally be problematic. We have attempted several variance-weighting schemes in a few cases where we presumed that this was the source of the problem with the p-value plot, but these attempted fixes were not very effective. When the plot cannot be "fixed," it still alerts the analyst to a problem with the null distribution, which has implications for virtually any procedure that addresses multiple comparisons. If the departure from a uniform is concave instead of convex, it is unlikely to be recognized as a problem with the null distribution unless there is an a priori reason to expect most genes to have null p-values. It is too common among the published microarray studies to find tests based upon inappropriate estimates of the within-treatment variation. Usually, these estimates underestimate the relevant variation and this wreaks havoc on p-value interpretations that presume uniformity under the null.

The multiple-testing strategy proposed herein provides a summary of hypothesis-testing results in a way that neither sacrifices statistical power needlessly nor inflates the false positive error rate haphazardly. Although the strategy may be used even if the desire is to identify relatively few highly differentially expressed genes, the analysis is mainly intended for applications in which the desire is to identify a pool of potentially important genes for follow-up testing or for genetic profiling. A decision-theoretic method is provided for taking into account the relative costs of false positive and false negative misclassifications of genes in determining cutoff points. Relevant misclassification rates can be estimated as a way to characterize the reliability of the findings. Importantly, selection of statistical significance levels need not be made a priori, in that the objective is not to control error rates at some prespecified levels. Rather, the objective is to allow decision criteria to be chosen after the data are observed, and then to estimate corresponding error rates specifically for the observed data. This approach allows flexibility in analyzing expression data, guiding follow-up analyses or additional studies to retain genes with changes, albeit less statistically evident.

Résumé

Un objectif de la plupart des études de génomique fonctionnelle consiste à estimer les modifications dans l'expression des gènes induites par le traitement. Les puces à cDNA étudient la quantité de mRNA pour des centaines ou même des dizaines de milliers de gènes dans chaque échantillon tissulaire, et cette technologie conduit à une multitude de contrastes thérapeutiques. Les tests d'hypothèses pour chaque gène évaluent l'évidence en faveur d'une absence d'effet mais le choix d'un niveau de signification est délicat en raison du nombre de comparaisons effectuées. Les degrés de signification de ces tests permettent de classer les gènes de telle façon qu'un seuil donné divise les gènes en deux groupes. Idéalement, un groupe devrait contenir les gènes affectés par le traitement et l'autre les gènes non affectés. Cependant, le groupe de gènes sélectionnés comme affectés contiendra un certain nombre de faux positifs, i.e. des gènes qui en réalité ne sont pas affectés par le traitement. Symétriquement, l'autre groupe de gènes, sélectionnés comme non affectés par le traitement, contiendra des faux négatifs, i.e. des gènes en réalité affectés. Un graphe des p-values (1-p) vs. leur espérance sous une distribution uniforme [0; 1] permet d'estimer le nombre d'hypothèses nulles vraies. Avec cette estimation, le taux de faux positifs et le taux de faux négatifs associés avec n'importe quel seuil de sélection, i.e. une p-value donnée séparant les gènes ne deux groupes, peut être estimé. Si l'on répète cette classification pour une large étendue de seuils, ces taux résument la capacité de l'étude à identifier les effets réels. Dans ce travail, nous sommes plus intéressés par sélectionner les gènes affectés que par se protéger des quelques faux positifs. Un seuil optimal dépend du coût relatif de classer à tort un gène comme affecté versus le coût de classer à tort un gène comme non affecté. Nous proposons de sélectionner ce seuil par une méthode analogue aux méthodes développées pour les courbes ROC. De plus, nous estimons le taux de faux positifs et le taux de faux négatifs associés pour n'importe quel valeur du seuil. Deux études de génomique fonctionnelle dont l'objectif était d'évaluer l'effet d'un traitement sont utilisées pour illustrer comment les différentes méthodes permettent aux investigateurs de choisir un seuil qui correspond le mieux à leur objectif de recherche.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis* **39,** 1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57,** 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25,** 60–83.

Bowyer, J. F., Harris, A. J., Delongchamp, R. R., Jakab, R., Miller, D. B., Little, A. R., and O'Callaghan, J. P. (2004). Selective changes in gene expression in cortical regions sensitive to amphetamine during the neurodegenerative process. *Neurotoxicology* **25,** 555–572.

Chen, Y., Kodell, R. L., Sistare, F., Thompson, K. L., Morris, S. M., and Chen, J. J. (2003). Normalization methods for analysis of microarray gene-expression data. *Journal of Biopharmaceutical Statistics* **13,** 57–74.

Conover, W. J. (1971). *Practical Nonparametric Statistics.* New York: Wiley.

Craig, B. A., Black, M. A., and Doerge, R. W. (2003). Gene expression data: The technology and statistical analysis. *Journal of Agricultural, Biological and Environmental Statistics* **8,** 1–28.

Delongchamp, R. R., Harris, A. J., and Bowyer, J. F. (2003). A statistical approach in using cDNA array analysis to finding modest, 2-fold or less, changes in several brain regions after neurotoxic insult. *Annals of the New York Academy of Sciences* **993,** 363–376.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–803.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine* **9,** 811–818.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75,** 383–386.

Hsueh, H.-M., Chen, J. J., and Kodell, R. L. (2003). Comparison of methods for estimating the number of true hypotheses in multiplicity testing. *Journal of Biopharmaceutical Statistics* **13,** 675–689.

Jakab, R. L. and Bowyer, J. F. (2002). Parvalbumin neuron circuits and microglia in three dopamine-poor cortical regions remain sensitive to amphetamine exposure in the absence of hyperthermia, seizure and stroke. *Brain Research* **958,** 52–69.

Li, C. S. and Taylor, J. M. G. (2001). Ridge regression for the classification of tumors using gene expression data. In *International Biometric Society Eastern North American Region Spring Meeting*. Charlotte, North Carolina.

Nguyen, D. V., Arpat, A. B., Wang, N., and Carroll, R. J. (2002). DNA microarray experiments: Biological and technological aspects. *Biometrics* **58,** 701–717.

Schweder, T. and Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69,** 493–502.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64,** 479–498.

Thompson, K. L., Mirsky, M. L., and Sistare, F. (2002). Concordance of degree of renal injury with gene expression in individual animals treated with the nephrotoxicant cisplatin. *The Toxicologist* **66,** 297–298.

Tsai, C.-A., Hsueh, H.-M., and Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics* **59,** 1073–1083.

Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A., and Ron, M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150,** 1699–1706.

Zaykin, D. V., Young, S. S., and Westfall, P. H. (2000). Using the false discovery rate approach in the genetic dissection of complex traits: A response to Weller et al. *Genetics* **154,** 1917–1919.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39,** 561–577.

Received May 2003. *Revised October* 2003.
Accepted January 2004.