 Human Molecular Genetics, 2002, Vol. 11, No. 19 2223–2232

Evaluating test statistics to select interesting genes in microarray experiments†

Charles Kooperberg[1,*], Simonetta Sipione[3], Michael LeBlanc[1], Andrew D. Strand[2], Elena Cattaneo[3] and James M. Olson[2]

[1]Division of Public Health Sciences and [2]Clinical Research Division, Fred Hutchinson Cancer Research Center, PO Box 19024, MP 1002, Seattle, WA 98109-1024, USA and [3]University of Milan, Department of Pharmacological Sciences and Center of Excellence on Neurodegenerative Diseases, Italy

Received March 12, 2002; Revised and Accepted July 11, 2002

A randomization procedure to evaluate the significance level and the false-discovery rate in complex microarray experiments is proposed. A related graph can be used to compare different test statistics that can be used to analyze the same experiment. This graph is closely related to receiver operator characteristic (ROC) curves. The proposed method is applied to a subset of the data from a cell-line experiment related to Huntington's disease. A small simulation study compares the effectiveness of the proposed procedure with the significance analysis of microarrays (SAM) procedure.

INTRODUCTION

DNA microarray experiments make it possible to study the variation of expression for many genes simultaneously. The most common types of microarray experiments are either unstructured, in which a large number of biological samples are compared with each other (1), or situations in which two groups of samples are compared (2). For the former type of experiments, the most common methods used for analysis are clustering and related methods (3,4); for the latter types, common methods involve *t*-statistics and variations thereof (5,6). ANOVA models are discussed in (7).

In this paper, we are concerned with more complicated designs: designs in which more than one experimental factor varies, and some of those factors (including the main one of interest) may have more than two levels. In traditional statistics, the most common way to analyze such data is to build a regression or ANOVA model and to test for the appropriate effects. When there is no clear best model, different models are often examined and compared using (graphical) model diagnostics or summary measures, such as the Akaike Information Criterion (AIC) (8), which compare how different models fit the data.

Many of these techniques would be quite appropriate if genes would be analyzed one at a time. Some techniques (e.g. ANOVA models, *t*-tests, *F*-tests) can easily be carried out for many genes simultaneously, since the design matrix is typically the same for all genes. Model selection and diagnosis, however, translate less easily to microarray experiments, since these methods often involve visual inspection of modeling results, which is impractical when there are more than a few genes.

Even after selecting a model, a complication with *t*-tests and other procedures yielding *P*-values is that multiple comparison corrections need to be made, since many tests are carried out simultaneously. The most common multiple comparisons correction is the Bonferoni correction. Dudoit *et al.* (9) discuss a variety of other multiple comparison corrections. An entirely different approach is advocated by Storey (10). Rather than adjusting *P*-values for individual genes, he suggests to control the false-discovery rate (FDR), which is the fraction of false positives among the genes that are called changed. The relation of the proposed approach to the FDR is discussed further in the Methods section.

In this paper, a graphical tool based on randomization to compare various test statistics (or model summaries) for analyzing complicated microarray experiments is proposed. This tool has similarities to receiver operating characteristic (ROC) curves (11) and is related to the FDR.

RESULTS

Data and questions

The proposed methodology is illustrated on a subset of a microarray study of cell line experiments (12) related to Huntington's disease. In this paper, data on three cell lines that have been engineered with a construct encoding the first

*To whom correspondence should be addressed. Tel: +1 2066677808; Fax: +1 2066674142; Email: clk@fhcrc.org
†This paper is part of the Microarray Report Special Series. See Orr, H.J. (2002) *Hum. Mol. Genet.*, **11**: 1909–1910.

N548 amino acids of the Huntington gene and an expansion in the polyglutamine tract (13) are used. In particular, we use data on the cell lines HD12(Q67), HD40(Q118) and HD43(Q105). [The complete data set used in (12) contains data on several additional cell lines.] For each cell line, all experiments were carried out independently twice. Potentially, activation of the Huntington gene can increase or decrease the expression of other genes (14,15). Let $s=0$ be the time that the gene is induced. For each of the six experiments, there are measurements using gene chips (16) at times $s=-6$, 0, 12, 24, 48 and 72 hours. We are interested in identifying genes that change their expression pattern over time. There are a number of caveats in carrying out a standard analysis. First, as HD12(Q67), HD40(Q118) and HD43(Q105) lines expressing the mutant HD constructs were independently constructed and the experiments of each line were independently carried out, it was expected that some gene expression changes would represent generalizable findings (changes in more than one experiment) whereas others would represent changes that are specific to a single cell line. The changes in expression pattern should, however, be consistent among both experiments carried out with the same cell line. Ideally, we should like to have the same effect each time an exogenous gene is inserted in a cell. However, these clonal cells could respond in different ways to an exogenous stimulus. As such, the requirement that the (small) changes in a gene be observable in all three cell lines may be too stringent. In general, we should like to see a similar pattern in at least two of the cell lines, to prevent identification of changes that are not reproducible. Second, not all changes in the expression pattern necessarily happen at time $s=0$: some genes may only change in expression after, say, 12 or 24 hours; the expression patterns of other genes may change smoothly; potentially the timing may be different between cell lines. For these two reasons, a standard statistical model may not be appropriate, and various modeling options need to be considered.

In the Methods section, we describe an approach to assess the significance of a test statistic that is associated with a model for the expression of a single gene. An example of such a model is discussed at the beginning of the Methods section. As set out in the Introduction, for a microarray experiment with a complicated design, like the one we are considering in this paper, it is not always clear which model to use. Thus, we must choose a model from a set of competing models, after which we may want to choose a cutoff for the test statistic to control the FDR. Model selection using the true-discovery plot is an iterative process. Thus, we shall present models based on *t*-tests and regression-type models for the cell line data, and compare the result as one might in an interactive modeling session.

Randomization

For each of the models that we considered, we made inference using randomization, for which we randomly permuted all time points. We report results using only 100 permutations to evaluate test statistics. While this is a small number for computing extreme *P*-values accurately, it is ample for the comparison of different test statistics; in fact, results based on 25 permutations are virtually indistinguishable from those shown here.

Modeling using *t*-tests

We define two functions that are used to select among a few test-statistics. Let

$$\rho(t_i, i=1,\ldots,I)=\begin{cases}\max_i\{|t_i|\} & \text{if } \max_i\{|t_i|\}=\max_i\{t_i\},\\ -\max_i\{|t_i|\} & \text{otherwise,}\end{cases}$$

and let

$$\phi(x_i, i=1,2,3)=\begin{cases}|\text{med}_i\{x_i\}| & \text{if } |\text{med}_i\{x_i\}|=\text{med}_i\{|x_i|\},\\ 0 & \text{otherwise.}\end{cases}$$

The function $\rho(\cdot)$ selects the absolute largest from a set of test statistics, keeping the original sign, and $\phi(\cdot)$ selects the median of the absolute value of three test statistics, provided that the signs of the test statistics are consistent.

Let w_{si}, with $s\in\{12, 24, 48, 72\}$, $i=1, 2, 3$ referring to cell line, be the regular *t*-statistics comparing log(Average Differences) y_{si1} and y_{si2} with y_{0i1}, y_{-6i1}, y_{0i2} and y_{-6i2}. Let

$$T_a=\min_i|\rho(w_{si}, s\in\{12,24,48,72\})|.$$

Thus, $\rho(w_{si}, s\in\{12, 24, 48, 72\})$ is the largest *t*-statistic for cell line *i*, and T_a takes the smallest of these statistics. The true-discovery plot for T_a (Fig. 1) shows that out of the 100 genes with the largest T_a, only ~10 are expected to be true positives. Alternative test statistics constructed from *t*-statistics gave similar results. The reason for the bad performance of regular *t*-statistics is likely that these *t*-statistics have very few degrees of freedom, and thus have variance estimates that are very noisy.

Therefore, we modeled the variance of log(Average Difference) from replicates at the same time point as a function of the mean of the log(Average Difference) for these two experiments. This relation (Fig. 2) shows that the variance decreases with expression level.

Let t_{sij}, with $s\in\{12, 24, 48, 72\}$, $i=1, 2, 3$ referring to cell line and $j=1, 2$ referring to experiment, be the *t*-statistics comparing log(Average Differences) y_{sij} with y_{0ij} and y_{-6ij} using the smoothed estimate of the variance (Fig. 2) instead of the usual estimate of the variance. Let $v_i=\rho(\min_j\{t_{sij}\}, s\in\{12, 24, 48, 72\})$ for each *i*. The statistic v_i will be large if there is a times $s\geq 12$ for which both repeats *j* of cell line *i* are significantly different from the baseline measurements at times $s=0$ and -6.

The first two test statistics that we consider are

$$T_b=\min_i|v_i|$$

and

$$T_c=\phi(v_i, i=1,\ldots,3);$$

thus, for T_b all three cell lines must show a significant result, where for T_c we like two of the three cell lines to show a significant result, as measured using v_i.

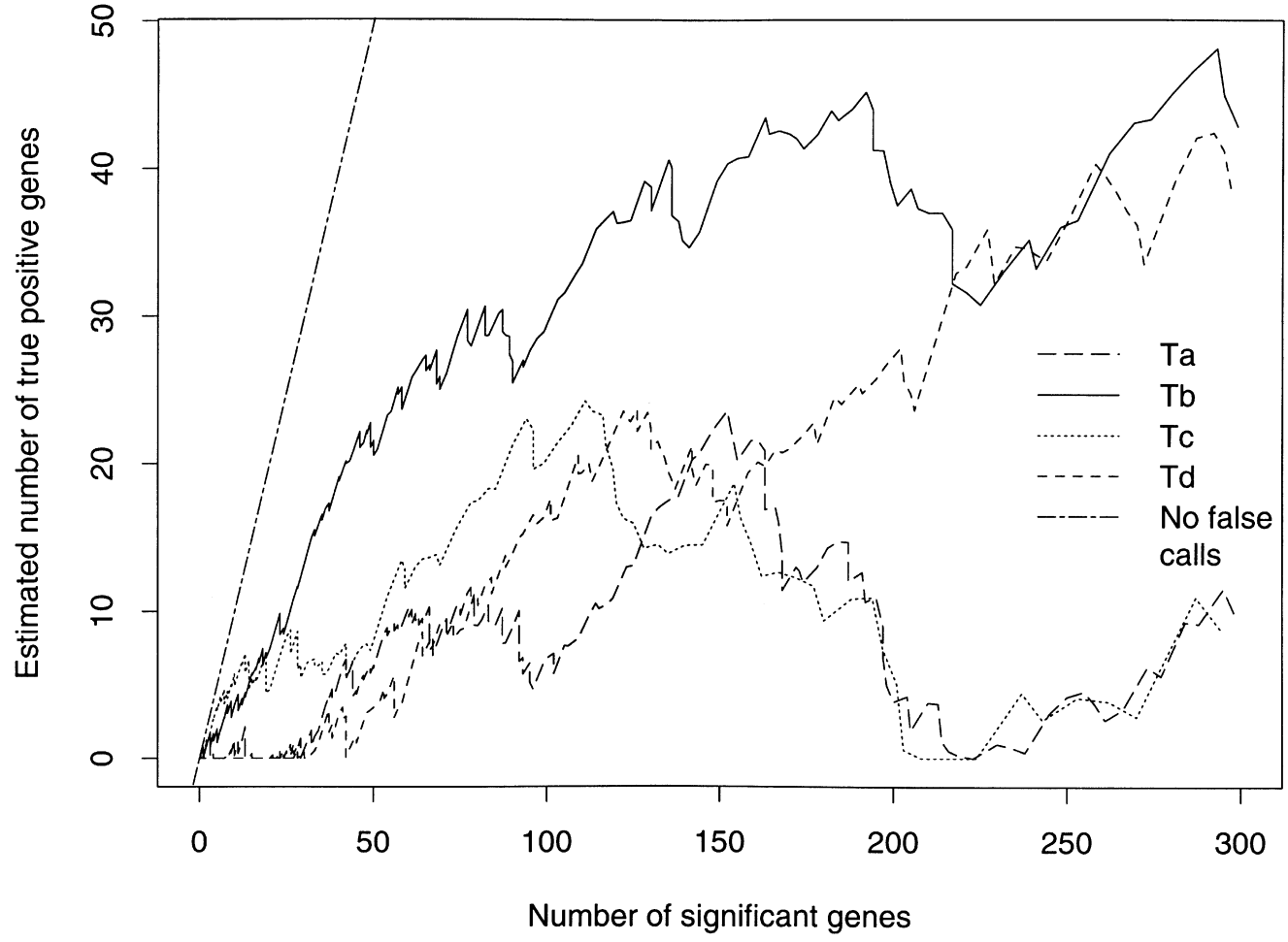

Figure 1. True-discovery plot for three test statistics based on the *t*-test. This plot shows the estimated number of true-positive genes when a particular number of the most significant genes are selected. Large estimated number of true-positive genes are preferred, and the best possible procedure would be one for which the true-discovery plot coincides with the 'no false calls' line. Of the four procedures shown in this plot, T_b performs best. The number of true-positive genes is quite low even for T_b, though.

Let t_{si} be the *t*-statistics, using the variance estimate from Figure 2, comparing log(Average Differences) y_{si1} and y_{si2} with y_{0i1}, y_{-6i1}, y_{0i2} and y_{-6i2}. Let

$$T_d = \min_i |\rho(t_{si}, s \in \{12, 24, 48, 72\})|.$$

The difference between T_b and T_d is that for T_b we compute the significance between baseline and follow-up times separately for both repeats, requiring consistency in combining them, while for T_d we combine both repeats in one two-sample test. Otherwise, both are using *t*-statistics with smoothed variances and require consistency between all three cell lines.

In Figure 1, we plot the true-discovery plot for T_b, T_c and T_d. The straight line labelled 'no false calls' in this figure corresponds to $m-m^* = m$. The best test statistic is the one for which the true-discovery plot is closest to this line. We note that all three test statistics do in fact have a high rate of suspected false positives. For the best of these three statistics, T_b, ~40% of the 100 most significant genes may be true positives. Considering more genes as significant does not appear to yield many more true positive genes.

It may be counterintuitive that the true-discovery plots (Fig. 1) can be decreasing. However, this is partly due to the inequalities in Equations 5 and 6 (see the Methods section). As α becomes larger, the number of significant genes increases while the inequality becomes less sharp, and this is also because we estimate α using simulation. A more complicated randomization scheme, comparing order statistics (see e.g. 9, 18), can circumvent this. We shall consider the advantages and disadvantages of both approaches in the Discussion.

Regression models

We concluded from Figure 1 that the use of *t*-tests with smooth estimates of the variance, while yielding better results than regular *t*-tests, failed to select differentially expressed genes, as the number of true discoveries was too low. Therefore, we now

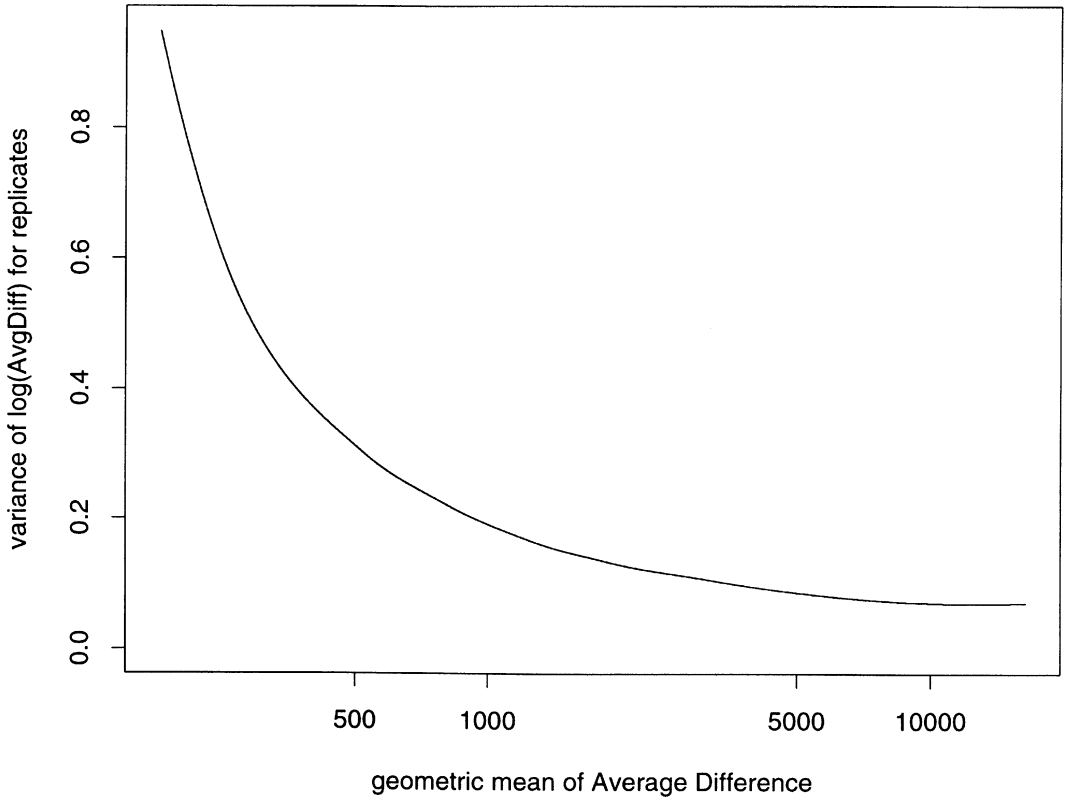

Figure 2. Smooth estimate of the residual variance as a function of expression level. From each pair of replicate observations of the expression for each gene at a particular time point for a particular cell line, we obtain a one-degree-of-freedom estimate of the residual variance σ^2. We smooth those estimates using a `loess` smoother (17) as a function of the expression level, as measured by the mean average difference.

discuss test statistics using regression models. Consider models of the form

$$y_{sij} = \beta_0^{(i)} + \beta_1^{(i)}\mathrm{Ind}(j=2) + \beta_2^{(i)}Z(s) + \epsilon_{sij}, \qquad i = 1, 2, 3. \quad \mathbf{1}$$

For each cell line i, this model combines data from both replicates, allowing different intercepts $\beta_0^{(i)}$ when $j=1$ and $\beta_0^{(i)}+\beta_1^{(i)}$ when $j=2$. Depending on the choice of $Z(s)$, we can search for different gene expression patterns in the data.

For test statistic T_e, we use the model in Equation 1, with $Z(s)=\mathrm{Ind}(s\geq 12)$. Let t_i be the (t-)test statistic corresponding to $\beta_2^{(i)}$ in Equation 1 for cell line i. This choice of $Z(s)$ looks for a difference between times $s\leq 0$ and $s\geq 12$, and, in fact, the t-statistic of $\beta_2^{(i)}$ would be the same as the usual t-statistic comparing $s\leq 0$ with $s\geq 12$ if the term $\beta_1^{(i)}\mathrm{Ind}(j=2)$ was not in Equation 1. We set

$$T_e = \min_i\{|t_i|\};$$

that is, we look for consistency between all three lines.

The test statistics T_f and T_g are defined similarly to T_e, but with $Z(s)=\mathrm{Ind}(s\geq 24)$ and $Z(s)=\mathrm{Ind}(s\geq 48)$, respectively. Thus, these two statistics also look for jumps in the expression level, but between times $s=12$ and $s=24$ for T_f and between times $s=24$ and $s=48$ for T_g. Test statistic T_h uses $Z(s)=s\times\mathrm{Ind}(s\geq 0)$, so that it looks for a linear trend in the log(Average Difference) after time $s=0$.

Note that β_2 can be either positive or negative in each of the models, so both genes whose expression increases and genes whose expression decreases are captured. In Figure 3, we show the true-discovery plot for T_b (the best of the previously examined statistics), T_e, T_f, T_g and T_h. From this figure, we note that, except for T_e, all regression models perform much better than T_b. The test statistics using a linear relation between time and log(Average Difference), T_h, seems to perform best. Out of the 100 most significant genes, ~75% are expected to be true positives, while out of the first 250, ~68% are expected to be true positives for T_h. This suggests that the changes in expression at time $s=12$ are still modest, and that most changes occur later in time.

We proceeded with examining alternative ways to use the regression model with $Z(s)=s\times\mathrm{Ind}(s\geq 0)$. In particular, set $T_i=\phi(t_i,\ i=1,\ 2,\ 3)$ using the same t-statistics used to create T_h, and let T_j be the absolute value of the test statistic corresponding to the model in Equation 2 (see the Methods section). Thus, for T_h, we required all three cell lines to show a (linear) effect in log (Average Difference); for T_i, we required two of the three cell lines to show such an effect; and for T_j, we combine all data, and therefore model the change for each cell line with the same slope. In Figure 4, we show the true-discovery plot for T_h, T_i and T_j. From this figure, we notice that

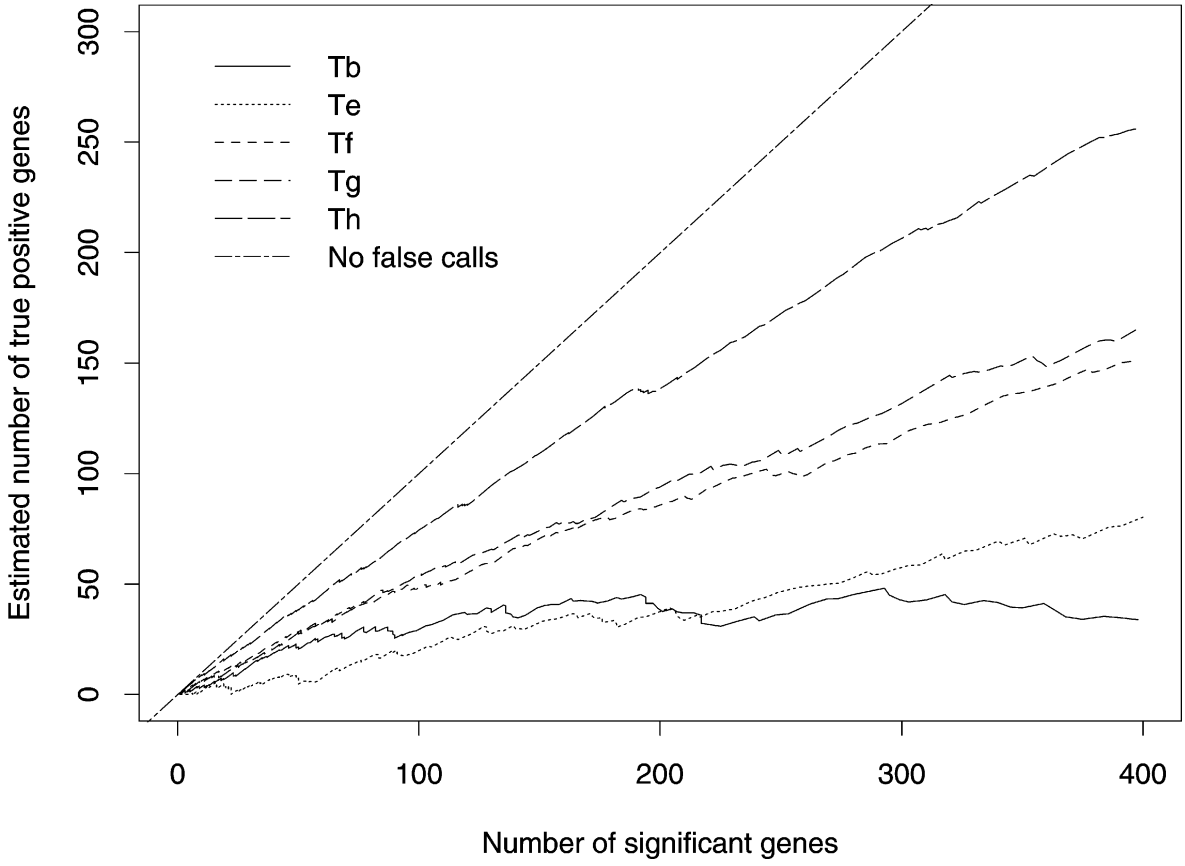

Figure 3. True-discovery plot for four test statistics based on a regression model and the best test statistic based on a *t*-test. The test statistics $T_e, \ldots, T_h$ are all based on the regression model in Equation 1 but use different forms of $Z(s)$. The test statistic T_b was the best test statistic from Figure 1. The best test statistic in this plot is clearly T_h, which uses $Z(s) = s \times \text{Ind}(s \geq 0)$. The plot suggests that for T_h, ~75% of the significant genes are true positives.

using the ϕ-function (essentially requiring two of the three cell lines to be in agreement), as is done for T_i, yields a slightly larger fraction of positives than combining all three cell lines. For the other functions $Z(s)$, used for T_e, T_f and T_g, we also found that using the ϕ-function yielded a larger estimate of true-positive genes found than either using one regression model or requiring all cell lines to be consistent.

For each design, we shall have to select a randomization scheme. For our experiment, we randomized the times. As an alternative, in Figure 5, we show true-discovery plots for T_i using the randomization scheme where the times were randomized and an alternative randomization scheme in which all 36 arrays were randomized. From this plot, we notice that the complete randomization scheme seems to suggest a larger fraction of true positives. However, this is an artifact of using complete randomization: as there is experiment-to-experiment and cell-line-to-cell-line variation, randomization using complete randomization yields larger variance estimates for the randomized results, and smaller test statistics. Thus, by using the incorrect randomization scheme, we would incorrectly assume that there were more true-positive genes than are actually present.

Simulation

We carried out a small simulation study to validate the procedure of computing α-levels using randomization and to investigate the power for detecting genes that change expression level. Let Y_{icj} be the log expression level for the *j*th replicate of the *i*th gene for class *c*. For our simulation, we generated data for 250 genes, with a two-class design and four replicates for each class. For 230 of the 250 genes, the simulated log expression ratio was independent standard normal data for both classes. For the remaining 20 genes, $Y_{i1j} = Z_{i1j}$ and $Y_{i2j} = M_i + Z_{i2j}$, for $i = 231, \ldots, 250$, $j = 1, \ldots, 4$, with all Z independent standard normal pseudo-random numbers and M_i different for each simulation set-up. The values for M_i for the different simulations are summarized in Table 1.

We employed a regular *t*-test as our test statistic. We carried out all $\binom{8}{4} = 70$ possible randomizations to select a cutoff value that has the correct α-level among the randomizations. For each gene, we then checked whether or not it was called significant. We repeated this procedure 500 times. In Table 2, we present the fraction of times that genes were called significant. We

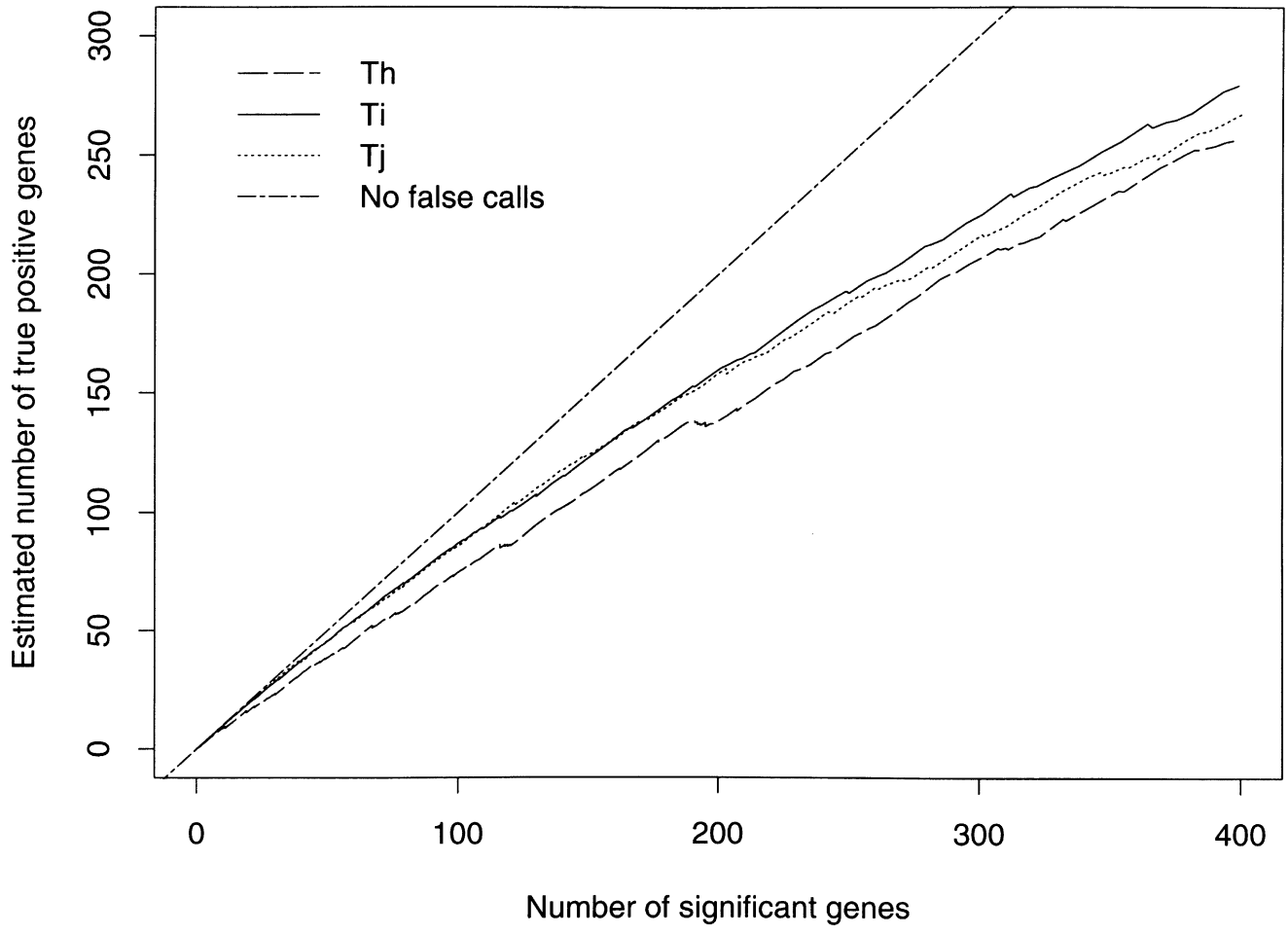

Figure 4. True-discovery plot for three test statistics based on a regression model that uses time linearly. The three test statistics in this plot all use $Z(s) = s \times \text{Ind}$ $(s \geq 0)$, but they combine the three cell lines in different ways. There is not much difference in performance between the three test statistics. The test statistic T_i, which uses the second largest t-statistic among the three for the different cell lines, is marginally better than T_h and T_j.

distinguish between the first 230 genes (we should like a fraction α of those to be significant) and the last 20 genes (except for set-up A, we should like as many genes as possible to be significant).

As a comparison, we carried out the same simulation using our own implementation of significance analysis of microarrays (SAM) (18). (The Excel format of the official SAM software did not allow us to use it for a large simulation.) The significance calls in SAM are based on a quantity Δ, the difference between the sorted t-statistics (the variance of these statistics are slightly inflated in SAM) and the sorted t-statistics of the randomizations. Further details can be found in (18) and the online website for SAM.

While SAM is intended to control the FDR, we can also control α by selecting the parameter Δ, such that among the simulations at most a fraction α of the genes is declared significant. We did not impose a cutoff on the expression ratio, as is suggested as a secondary threshold in the SAM paper. Again, we used all 70 possible randomizations. The results for our implementation of SAM can be found in Table 3.

To investigate the FDR, we counted how many out of the k most significant genes were among the last 20 genes for which there was a signal (except for set-up A, where there was no signal at all). The results are shown in Table 4. Note that when $k = 40$, the best possible FDR is 0.5.

DISCUSSION

There are many different (regression) models possible for the Huntington's disease data; we reported on some of the ones we explored. The model in Equation 2 (see the Methods section), but fitted separately for each cell line, resulted in the best true-discovery plot. The reason that a linear model fits the data best may be that changes occur linearly, or, alternatively, that changes happen at different times for different genes, where the linear model effectively averages these times. It is quite possible that some other models (e.g. random-effects models) yield equally good results. An advantage of the models that we considered is that all models can be fit simultaneously to all genes, so that the randomization procedure is fast.

Which type of models are considered clearly depends on the biological context of the problem. In the Huntington's disease example used in this paper, changes in gene expression are

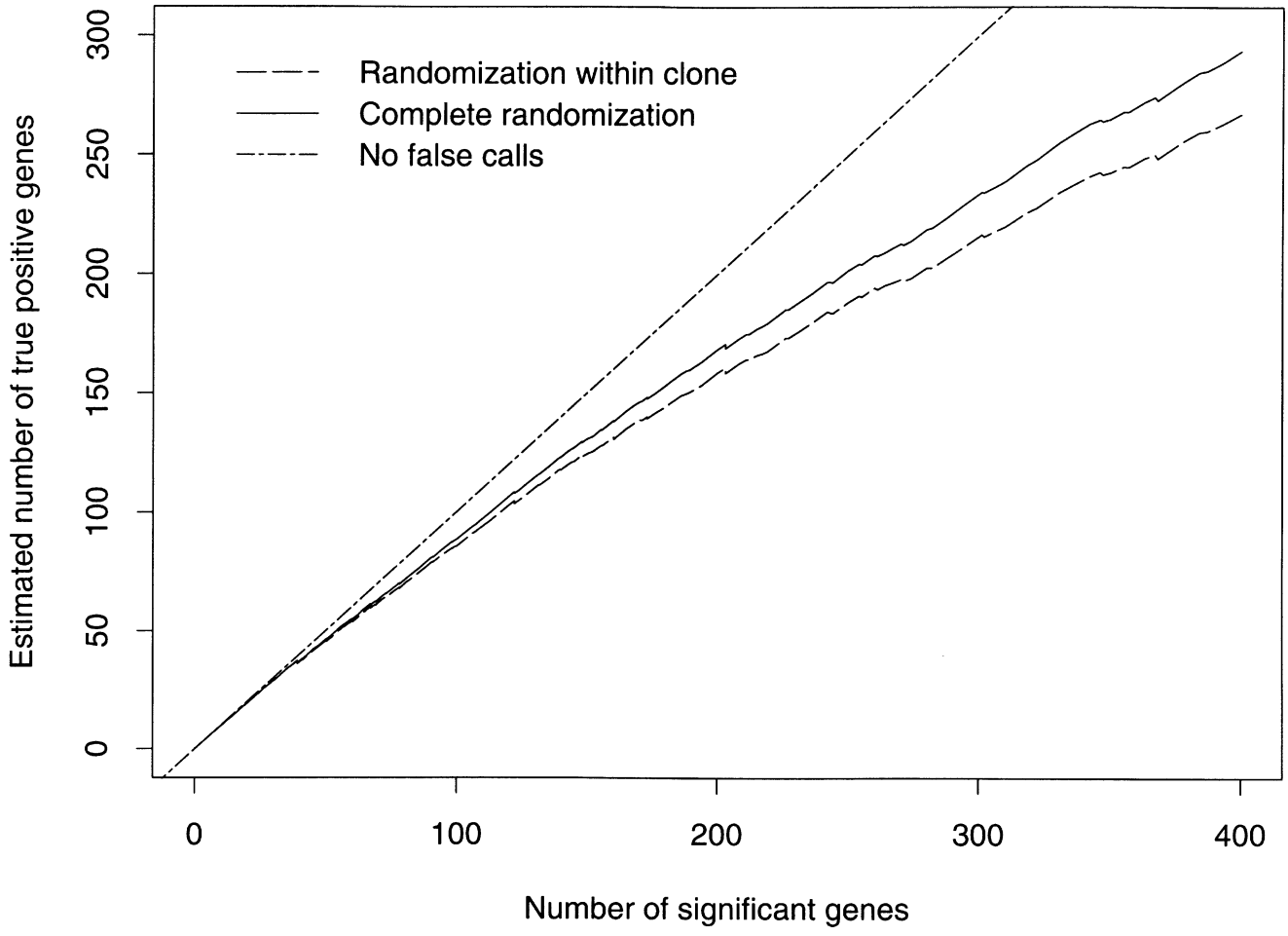

Figure 5. True-discovery plot for two different randomization schemes for the test statistic T_i. Randomization within a clone is the 'correct' way to randomize the Huntington's disease experiment; a complete randomization ignores part of the dependence structure. Using the incorrect randomization, we would get too large an estimate of the number of true discoveries.

expected after the gene is induced at time $s = 0$. Those changes could be gentle, or more abrupt—hence the linear model that averages-out effects yields the best results. Clearly, for other types of experiments, other types of models may be more appropriate; for example, for cell cycle data, we could imagine using the single-pulse model (19).

The approach that we present in the Methods section to evaluate test statistics is simple and requires no additional software. Since it only uses a randomization procedure, and no comparison of order statistics, as do Dudoit *et al.* (9) and Tushner *et al.* (18), it may be intuitive for non-statisticians. From the simulation (Table 2) we note that the fraction of genes indicated to be significant is almost exactly α and that when $|M_i| \geq 2$ the *t*-test appears quite powerful.

The fact that we do not use order statistics is the main difference between our procedure to select significant genes and SAM. It appears clear from Table 3 that for SAM it is much harder to control α. We believe that this is caused by the procedure in SAM to identify which genes are called significant, described in the last two paragraphs on page 5117 of the SAM paper (18). This procedure can sometimes cause large numbers of genes to be called significant at the same α (or FDR) level when Δ is changed. Because the SAM procedure does not match the intended α-levels, it is hard to compare the power. We do note that for most set-ups the actual α is close to the 0.01 when the nominal $\alpha = 0.001$ and that the actual α is close to the 0.05 when the nominal $\alpha = 0.01$. As such, we can roughly compare the 0.001 and 0.01 lines for the 20 genes with signal for SAM (Table 3) with the 0.01 and 0.05 lines, respectively, for the 20 genes with signal in Table 2, suggesting that both procedures have comparable power. We note from Table 4 that SAM has a slightly lower FDR than the procedure proposed here, but the differences are very small.

Since the difference between the ordered test statistics on the actual data and the (average) ordered test statistics on the randomized data is not necessarily monotone, the SAM procedure ends up being very granular, calling genes significant in groups. This is a main reason why it is hard to control the α level using SAM.

The true-discovery plot proposed in this paper is intended as a diagnostic tool for evaluating test-statistics. It is not intended as an alternate estimate of the FDR. In fact, because of the

Table 1. Simulation set-ups

Set-up	Definition of M_i
A	$M_i = 0$, $i = 231, \ldots, 250$
B	$M_i = -1$, $i = 231, \ldots, 240$; $M_i = 1$, $i = 241, \ldots, 250$
C	$M_i = -2$, $i = 231, \ldots, 240$; $M_i = 2$, $i = 241, \ldots, 250$
D	$M_i = -5$, $i = 231, \ldots, 240$; $M_i = 5$, $i = 241, \ldots, 250$
E	$M_i = 5$, $i = 231, \ldots, 250$
F	$M_i = i - 230$, $i = 231, \ldots, 240$; $M_i = 251 - i$, $i = 241, \ldots, 250$

The M_i are the differences in log(gene expression) between the two groups in the simulation study for gene i. For genes 1, ..., 230, there is no difference between the control and experimental group. Standard normal noise is being added to the gene expression level for both the control and experimental group.

Table 2. Fraction of times that genes were called significant during the simulations

α	Set-up					
	A	B	C	D	E	F
230 noise genes:						
0.001	0.0012	0.0012	0.0011	0.0002	0.0017	0.0001
0.01	0.0104	0.0103	0.0100	0.0083	0.0127	0.0086
0.05	0.0509	0.0508	0.0497	0.0516	0.0549	0.0513
20 genes with signal:						
0.001	0.0012	0.0095	0.0604	0.3805	0.3598	0.4172
0.01	0.0091	0.0657	0.3060	0.9864	0.9364	0.7918
0.05	0.0495	0.2284	0.6488	1.0000	0.9520	0.8832

For the set-ups described in Table 1, we carried out 500 simulations of four control and four experimental microarrays with 250 genes. To determine significance, a t-statistic, with significance levels set using the randomization procedure described in the Methods section, was employed. Since for genes 1, ..., 230 the M_i were 0, a fraction α of these genes should be significant, while, except for set-up A, for the 20 genes with signal ($i = 231, \ldots, 250$) as many genes as possible should be significant.

inequalities appearing in Equations 5 and 6 (see the Methods section) an estimate of the FDR obtained using these equations would be biased downwards. See (20) for a better estimate of the FDR. In fact, the true-discovery plot is very similar to the ROC curves used to evaluate different testing procedures (11). In particular, the true-discovery plot plots $m - m^*$ versus m, while an ROC curve would plot $(m - m^*)/n$ versus m^*/n, the estimated fraction of true positives versus the estimated fraction of true negatives. Using ROC curves, we should prefer the test statistic with the largest 'area under the curve'. Clearly, test statistics that are 'good' on the true-discovery plot will be good on the ROC curve, and *vice versa.* (In fact, we could formalize the selection of the test statistic using the true-discovery plot, by selecting the test statistic that has the largest area under the true discovery plot over the area $0 < m < k$, for some maximum number k of genes of interest.) We feel that for DNA microarrays, the total number of significant genes, m, is the quantity that needs to be controlled, since this is the number of follow-up experiments, such as northern blots, that need to be carried out.

A key assumption for our approach to evaluating test statistics, as well as those of Dudoit *et al.* (9) and Tushner *et al.* (18) is that, after both explicit normalization and the implicit normalization that the computation of a test statistic adds to this, the test statistic for any one gene can be compared with the test statistics for other genes. This exchangability condition is somewhat suspect when the variability of the test statistic differs from the expression level—for example because there are more outliers when expression levels are low, causing more large test statistics. One approach to remedy this would be to 'bin' genes by expression level, and only use the test statistics for genes that are in the same bin. Another, complementary, possibility is to use robust test statistics, such as those proposed by Lönstedt and Speed (21), that inflate the variance estimate for t-statistics.

METHODS

Randomization

Assume that we want to use a particular model to identify genes for which the expression pattern may have changed and that this model yields some sort of test statistic. If this test statistic is extreme (say large), then the gene is called 'significant'. An example of such a test statistic for one gene for the Huntington's disease data described above would be the t-statistic for β_7 in the regression model

$$y_{si} = \sum_{j=1}^{6} \beta_j \, \mathrm{Ind}(i = j) + \beta_7 [s \times \mathrm{Ind}(s > 0)] + \epsilon_{si}, \qquad \mathbf{2}$$

where $i = 1, \ldots, 6$ refers to the six experiments that were carried out at each time point (three cell lines times two replicates) and $s \in \{-6, 0, 12, 24, 48, 72\}$, the times at which experiments were carried out, y_{si} is the expression level, as measured by the logarithm of the Average Difference, provided by the GeneChip software, and $\mathrm{Ind}(x)$ is the usual indicator function [i.e. $\mathrm{Ind}(x) = 1$ is x is true, and 0 otherwise]. The interpretation of the model in Equation 2 is that for each cell line and each experiment, there is a different baseline level β_j, and that after times $s = 0$ there is a linear trend in the expression with the same slope β_7 for all cell lines. Statistical significance of β_7, as measured by its t-statistic, would measure whether there is evidence of a linear slope.

The model in Equation 2 is an example of one that may be of interest when analyzing the experiment; in the Results section many more possible models are considered. Typically, the significance of each such model would be summarized by a single test statistic. An important question is to determine a cutoff for such a test statistic. In traditional (parametric) statistical models, this cutoff is determined such that the probability of identifying a gene as having changed, given that it actually has not changed, to be a prespecified level α by referring to a known reference distribution (e.g. $Z > 1.96$). In the context of a microarray experiment, with many thousands of genes being examined at the same time, it is more useful to identify a cutoff value that controls the fraction of the genes that are significant while they are in fact unchanged. This was identified as the FDR (10).

Let t_i be the test statistic for the ith gene ($i = 1, \ldots, n$) using the original data. Assume that r times the arrays

Table 3. Fraction of times that genes were called significant during the simulations for SAM

α	Set-up					
	A	B	C	D	E	F
230 noise genes:						
0.001	0.0011	0.0014	0.0021	0.0000	0.0010	0.0000
0.01	0.0110	0.0198	0.0449	0.0670	0.0650	0.0674
0.05	0.0543	0.1037	0.1604	0.1604	0.1354	0.1590
20 genes with signal:						
0.001	0.0010	0.0156	0.1641	0.4396	0.4318	0.4454
0.01	0.0110	0.1112	0.6328	1.0000	0.9524	0.8993
0.05	0.0536	0.2978	0.8450	1.0000	0.9562	0.9276

This table provides the same information as Table 2, when the significance levels for the test statistic are determined using the SAM (18) procedure, which uses randomizations and order statistics, instead of the randomization procedure described in this paper.

Table 4. False-discovery rate for various numbers of selected genes during the simulations

Number k selected	Set-up				
	B	C	D	E	F
Procedure proposed in this paper:					
10	0.686	0.295	0.008	0.053	0.004
20	0.736	0.420	0.044	0.094	0.176
40	0.789	0.596	0.500	0.523	0.545
SAM:					
10	0.658	0.227	0.001	0.026	0.000
20	0.713	0.382	0.014	0.058	0.160
40	0.779	0.586	0.500	0.519	0.537

Instead of tuning test statistics to select the significance level α, the k genes with the largest test statistic are selected. This table displays the average fraction of these k that have in fact no changes (i.e. $i=1, \ldots, 230$). Any gene i with $i=1, \ldots, 230$ is in fact a false discovery. For $k=40$, it is not possible to have a false-discovery rate of under 50%, since there are only 20 genes with true changes in the simulation. The top half of the table uses t-statistics; the bottom half of the table uses the statistic Δ from SAM (18).

are randomized yielding test statistics t_{ij}^* $(i=1, \ldots, n, j=1, \ldots, r)$. If the experiment is large enough, we can, for a gene i, compare t_i with the t_{ij}^* for the same gene, and the fraction of times that $t_{ij}^* > t_i$ is the P-value for gene i. However, many complicated microarray experiments will not have enough replicates to allow for an appropriate randomization scheme that allows r to be large enough to estimate extreme P-values separately for each gene. For example, in the experiment used in this paper, we randomize over the time s, so that only $6!/2 = 360$ unique randomizations are possible (the division by 2 is necessary since $s=-6$ and $s=0$ are interchangeable in this experiment).

For very small P-values, many more randomizations are needed: for example, for a Bonferoni correction, the $0.05/n$ quantile of the randomization distribution needs to be estimated. To estimate this with any accuracy, at least $\sim 5n/0.05 = 100n$ randomizations are needed. Even when enough randomizations are possible, computation time may limit the number of randomizations. Thus, with few replicates and without making a parametric assumption, data for different genes will have to be combined in order to estimate extreme P-values.

Instead, we assume that under the null hypothesis of no gene changes, the distribution of the test-statistic is the same for each gene. Now an estimate of the P-value for each gene using all t_{ij}^* is

$$\hat{\alpha}_i = \alpha(t_i) = \frac{\sum_{i',j} \mathrm{Ind}(t_i > t_{i'j}^*)}{rn}. \qquad \mathbf{3}$$

This approach allows computation of more-extreme P-values using as few as 100 randomizations. The idea of using test statistics of one gene for evaluating other genes was also used to compute the Westfall–Young step-down P-values (9), and for SAM (18). Note that the assumption of using Equation 3 (namely, that under the null hypothesis of no change, the distribution of the test-statistics is the same for each gene) is a considerably weaker assumption than the assumption that the distribution of the expression of all genes is the same.

P-values and the false-discovery rate

If there are no changes in expression ratio, then approximately a fraction α of the genes have a P-value $<\alpha$. If a fraction p_0 of the genes have changes that are large enough that they can be detected with the experiment, then approximately a fraction $p_0+(1-p_0)\alpha = \alpha+p_0(1-\alpha)$ of the genes have a P-value $<\alpha$.

Suppose that out of n genes, m are significant at a level of α. This yields

$$\frac{m}{n} \approx \alpha + p_0(1-\alpha), \qquad \mathbf{4}$$

so that

$$p_0 \approx \frac{m/n - \alpha}{1-\alpha} > \frac{m}{n} - \alpha. \qquad \mathbf{5}$$

Thus, out of the m significant genes, at least about $(m-\alpha n)/(1-\alpha)$ should be true positives.

Using the randomization procedure described above, let T be a particular cutoff level, and let $\hat{\alpha} = \alpha(T)$ be the corresponding significance level (see Equation 3). Let m^* be the average number of genes among the randomized data sets that exceeds a particular the same cutoff level T for the test statistic. Note that $m^* = \hat{\alpha} n$ Then an estimate of the number of true positives that we expect among the m genes having test statistics exceeding T is

$$np_0 \approx n\frac{m/n - \hat{\alpha}}{1-\hat{\alpha}} = \frac{m-m^*}{1-m^*/n} > m - m^*. \qquad \mathbf{6}$$

In many situation, the interest is not in selecting one (or maybe a few) genes for which we are convinced that there is a change in the expression level, but rather we want to identify a longer list of genes that have potential changes in expression level. In making a decision how many genes to select, a critical ingredient is the expected number of true-positive genes among this list. Therefore, we examine a graph of the expected number of true-positive genes, $m-m^*$, versus the total number of

significant genes, m, as the cutoff level of the test statistic T is varied. Such a graph is also a tool to select among test statistics, since test statistics for which the graph is higher for a particular value of m are more powerful in selecting a list of m genes. We refer to this plot as the 'true-discovery plot'.

Storey and Tibshirani (20) point out that $(1 - p_0)m^*/m$ is an estimate of the FDR, and discuss a way to obtain an estimate for p_0. In particular, they point out that it is advantageous to estimate p_0 using a smaller cutoff value for the test statistic than T, which is the cutoff used to determine significance. Their argument does consider all changes—not just those that are large enough to be detected with the actual experiment, which is the starting point for Equation 4. Our approach, in fact, will lead to a slightly conservative analysis, in which we may slightly overestimate the number of false positives.

ACKNOWLEDGEMENTS

C.K. was supported in part by NIH Grant CA 74841. C.K. and M.L. were supported in part by a pilot grant from the Fred Hutchinson Cancer Research Center. J.M.O. was supported in part by NIH Grant NS 42157. S.S., A.S., E.C. and J.M.O. were supported in part by The Hereditary Disease Foundation Cure HD Initiative. The Fred Hutchinson Cancer Research Center array facility were instrumental in generating the data used in this manuscript.

REFERENCES

1. Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.
2. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
3. Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
4. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, O.M., Lau, C. *et al.* (2002) Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature*, **414**, 436–442.
5. Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.
6. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J.R. and Nevins, J.R. (2001) Predicting the clinical status of human breast cancer using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.
7. Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
8. Akaike, H. (1974) A new look at statistical model identification. *IEEE Trans. Autom. Control*, **AC-19**, 716–722.
9. Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12**, 111–139.
10. Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.
11. Begg, C.B. (1991) Advances in statistical methodology for diagnostic medicine in the 1980's. *Statist. Med.*, **10**, 1887–1895.
12. Sipione, S., Rigamonti, D., Valenza, M., Zucato, C., Pritchard, J.I., Kooperbertg, C., Olson, J.M. and Cattaneo, E. (2002) Early transcriptional profiles in huntingtin-inducible stratial cells by microchip analyses. *Hum. Mol. Gen.*, **11**, 1953–1965.
13. Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosome. *Cell*, **72**, 971–983.
14. Luthi-Carter, R., Strand, A., Peters, N.L., Solano, S.M., Hollingsworth, Z.R., Menon, A.S., Frey, A.S., Spektor, B.S., Penney, E.B., Schilling, G. *et al.* (2000) Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. *Hum. Mol. Gen.*, **9**, 1259–1271.
15. Cha, J.-H.J. (2000) Transcriptional dysregulation on Huntington's disease. *Trends Neurosci.*, **23**, 387–392.
16. Lipschutz, R.J., Fodor, S.P.A., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**(Suppl.), 20–24.
17. Cleveland, W.S. and Devlin, S.J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Assoc.*, **83**, 596–610.
18. Tushner, V., Tibshirani, R.J. and Chu, C. (2001) Significance analysis of microarrays applied to ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
19. Zhao, L.P., Prentice R.L. and Breeden, L. (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.
20. Storey, J.D. and Tibshirani, R.J. (2001) Estimating the false discovery rate under dependence, with applications to DNA microarrays. Technical Report 2001-28, Department of Statistics, Stanford University.
21. Lönnstedt, I. and Speed, T.P. (2002) Replicated microarray data. *Statist. Sinica*, **12**, 31–46.