 Human Molecular Genetics, 2002, Vol. 11, No. 17 1977–1985

A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model

Xie L. Xu[1], James M. Olson[2] and Lue Ping Zhao[1,*]

[1]Division of Public Health Sciences and [2]Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Received March 12, 2002; Revised and Accepted June 7, 2002

Time-course studies with microarray technologies provide enormous potential for exploring underlying mechanisms of biological phenomena in many areas of biomedical research, but the large amount of gene expression data generated by such studies also presents great challenges to data analysis. Here we introduce a regression-based statistical modeling approach that identifies differentially expressed genes in microarray time-course studies. To illustrate this method, we applied it to data generated from an inducible Huntington's disease transgenic model. The regression method accounts for the induction process, incorporates relevant experimental information, and includes parameters that specifically address the research interest: the temporal differences in gene expression profiles between the mutant and control mice over the time course, in addition to heterogeneities that commonly exist in microarray data. Least-squares and estimating equation techniques were used to estimate parameters and variances, and inferences were made based on efficient and robust *Z*-statistics under a set of well-defined assumptions. A permutation test was also used to estimate the number of false-positives, providing an alternative measurement of statistical significance useful for investigators to make decisions on follow-up studies.

INTRODUCTION

It is a common practice in biomedical research to observe a biological phenomenon and to measure one or more responses over a period of time. Such time-course studies are essential in biomedical research to understand biological phenomena that evolve in a temporal fashion. The recent advance in microarray technologies allows researchers to simultaneously measure the expression levels of thousands of transcripts over time, and promises to be a powerful tool for studying the underlying mechanisms of biological processes.

Microarray time course (MTC) studies can be largely categorized into four common types: single-series and multiple-series, each with and without time-varying cofactors. A typical single-series MTC study uses one microarray slide (and its replicates) to measure expression profiles at each time point. The primary objective in these single-series MTC studies is to characterize the temporal patterns of gene expression (changes). Examples of such a single-series MTC experiment include the yeast cell cycle study by Cho *et al.* (1), which measured the expression profiles of one sample (the synchronous yeast cultures) over time on high-density oligonucleotide arrays, and the fibroblast serum stimulation response study by Iyer *et al.* (2), which monitored the changes in expression profiles of one pair of samples—the serum-stimulated fibroblasts (treatment group) and the quiescent fibroblasts (control group) that were competitively hybridized onto the same cDNA array—and measured the ratios of their fluorescence intensities over time.

In contrast, a typical multiple-series MTC study utilizes multiple microarray slides at each time point, hybridized with different samples—for example, one or more mutants versus a wild-type control, or different drug treatment groups versus a placebo- or mock-treated control. Replicates may also be included at each time point. The primary interest in such multiple-series MTC studies is to discover temporal differences in gene expression profiles among groups. In either single-series or multiple-series studies, researchers may be also interested in the relationships of gene expression with other time-varying cofactors, such as drug concentrations, phenotype

*To whom correspondence should be addressed at: Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, MW-805, Seattle, WA 98109, USA. Tel: +1 2066676927; Fax: +1 2066672437; Email: lzhao@fhcrc.org

progression or cellular functions. The primary objective of such studies is to correlate time-varying cofactors with gene expression over time. While these MTC studies are clearly very powerful in many areas of biomedical research, they also generate vast amounts of data with substantial variations. How to extract relevant information presents an immense challenge to data analysis.

Currently, cluster analysis, which is commonly used to analyze data from various studies using microarray technologies (3–7), is also most frequently used to analyze MTC data (1,2,8–11). In MTC studies, a typical cluster analysis computes pairwise correlations (or any other distance measures) among genes, then groups genes into clusters (12,13) or hierarchical trees (14) without utilizing timing information. Once the clusters are formed, the average of the expression values of the genes within the cluster is computed and visually displayed to assess patterns associated with timing. The key strength of cluster analysis is to group genes so that one can visualize meaningful patterns. However, cluster analysis has some weaknesses worth noting. First, the timing information or any time-varying cofactors are not incorporated in the analysis. Second, a typical cluster analysis does not measure statistical significance, in that it will always find clusters regardless of whether or not meaningful clusters exist. Third, cluster memberships are easily influenced by choice of scales, transformations or filtering processes.

To overcome these challenges with cluster analysis, we propose a regression-based statistical modeling approach to analyze MTC data. The general framework of this approach has been described and applied to identify cell cycle-dependent genes in budding yeast (15) and in a two-group comparison study to identify genes differentially expressed between acute lymphoblastic leukemia and acute myeloid leukemia samples (16). The principle of this approach is to treat all gene expression values as multiple responses and regress them on specific experimental variables, such as time, cell or tissue type, drug dose, etc. The regression model is indexed with biologically meaningful gene-specific parameters. Furthermore, it also includes normalization factors that adjust for heterogeneity among arrays. Applying both least-squares and estimating-equation techniques, the model yields estimates of all relevant gene-specific parameters and their statistical significance. In this paper, we describe an application of this regression-based modeling approach to a multiple-series MTC study, using data generated from an inducible Huntington's disease animal model. Similar analytic strategies can also be applied to the three other types of MTC studies.

Huntington's disease (HD) is caused by an expansion of CAG repeats at the 5′ end of the *IT15* gene (17). The CAG repeats are translated into a polyglutamine [poly(Q)] sequence in the N-terminal portion of huntingtin (htt), the protein products of the *IT15* gene. While normal individuals have a poly(Q) length of 6–34 repeats, individuals with more than 40 repeats develop HD with virtually 100% penetrance. Several transgenic mouse models have been generated to gain insights into the pathogenesis of HD (18–23). Mice that express all or part of the htt protein with an abnormal poly(Q) length exhibit a progressive HD-like phenotype with various levels of severity. The inducible HD model (HD94) utilized a tetracycline-inducible system (tet-off) to control the expression of a chimeric mouse/human exon 1 containing 94 repeats of CAG (24). When expression of the mutant htt was induced at birth, the mice showed progressive limb clasping and other signs of motor dysfunction, as well as an HD-like neuropathology.

In an attempt to understand the early events during HD pathogenesis, striatal RNAs were extracted from the brains of these mice at various timepoints, before and after the induction of the mutant htt, and then hybridized to Affymetrix Mu11K oligonucleotide arrays. Initial evaluation of the data revealed that gene expression changes in this model were subtler than those in a previous HD mouse model (25) (R. Luthi-Carter and J. Olson, unpublished observation). Using the statistical modeling approach described here, we attempt to identify best-candidate genes that are differentially expressed between the mutant and control mice, and evaluate their statistical significance. This paper focuses on statistical methods for MTC studies. The complete dataset, confirmation studies and biological interpretation will be presented in another article.

RESULTS

Modeling

In order to gain insights into the pathogenesis mechanisms of poly(Q) extension in huntingtin, RNA were prepared from the striata of HD94 mice (24) or the single transgenic control mice carrying the tTA molecule (tTA control) at the following weeks: −2, −1.5, −1, 0, 2, 4, 6 and 8, with week 0 as the time of induction (Table 1). These were hybridized to Affymetrix Mu11K Sub B oligonucleotide arrays (details are described in Materials and Methods). The primary goal of this study is to identify genes that are expressed differentially between the control and the mutant mice over time after the induction of mutant htt. To achieve this goal, the regression model should incorporate parameters that specifically address the research interest, while at the same time utilizing all available relevant information. The expression profile generated from each hybridization can be conceptualized as a vector of J responses. Let $Y_k = (Y_{1k}, Y_{2k}, \ldots, Y_{Jk})'$ denote the expression profile generated from the kth hybridization, where Y_{jk} denotes the expression of the jth gene in the kth profile ($j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$). Two pieces of information regarding the sample used in the kth hybridization are relevant to the research goal: the genotype of the mouse from which the sample was taken, and the timepoint during the course of induction when the sample was collected (listed in Table 1). The genotype was incorporated using the covariate X_k: $X_k = 1$ for HD94 mice, and $X_k = 0$ for the tTA controls. To incorporate the induction process into the model, the timepoints were coded into three variables: t_{bk}, t_{ak} and I_k. Their values corresponding to the

Table 1. Variables t_{bk}, I_k, t_{ak} correspond to the timepoints in the induction course

Time (weeks)	−2	−1.5	−1	0	2	4	6	8
t_{bk}	−2	−1.5	−1	0	0	0	0	0
I_k	0	0	0	0	1	1	1	1
t_{ak}	0	0	0	0	0	2	4	6

timepoints in the induction course are as described in Table 1. Specifically, the binary variable I_k indicates the time of induction of the mutant htt transgene: before induction, $I_k = 0$; after induction, $I_k = 1$. t_{bk} and t_{ak} correspond to the timepoints before (t_{bk}) and after (t_{ak}) induction. Statistically, to capture the essence of temporal trends and their differences between the mutant and control mice, we propose the following regression model:

$$Y_{jk} = \delta_k + \lambda_k(\alpha_j + \beta_j t_{bk} + \gamma_j I_k + \vartheta_j t_{ak} + \theta_j t_{ak}^2 + \chi_j I_k X_k + \kappa_j t_{ak} X_k + \eta_j t_{ak}^2 X_k) + \varepsilon_{jk}, \quad \mathbf{1}$$

in which (δ_k, λ_k) are profile-specific additive and multiplicative heterogeneity factors. These are used to adjust systematic errors associated with each hybridization that are identical for all genes in the same profile, but vary from profile to profile (16). ε_{jk} is the random variation associated with each gene in each hybridization due to all sources other than those that have already been incorporated into the model. The parameters (α_j, β_j, γ_j, ϑ_j, θ_j, χ_j, κ_j, η_j) are gene-specific regression coefficients, and their biological interpretations are summarized in Table 2.

Figure 1 illustrates the possible patterns following the regression. The model conceptualizes the relationship between the expression level of a gene and the time variable before induction as a linear function regardless of the genotype of the mice, since before induction of the mutant htt transgene (weeks −2, −1.5, −1 and 0), the HD94 mice should not differ from the tTA controls. (α_j, β_j) measure the basal expression level (intercept) in both HD94 and tTA mice at week 0 and the slope of the linear function before induction, respectively (Fig. 1 and Table 2). After induction, the model approximates the relationship between the expression level of a gene and the time variable as a quadratic curve (Fig. 1). γ_j measures the shift in the basal expression level in the tTA control mice 2 weeks after induction and χ_j measures the difference at basal expression level between the HD94 and the tTA control mice at week 2 (referred to as intercept difference). After week 2, ϑ_j measures the slope in the tTA controls and κ_j measures the slope difference between the HD94 and the tTA control mice (referred to as slope difference). The quadratic coefficient θ_j indicates the curved trend of the expression level of a gene in the tTA controls after week 2 and the parameter η_j measures the difference in the curved trend between the HD94 and the tTA control mice (referred to as quadratic coefficient difference). The model is aimed at systematically identifying genes that follow a quadratic curve trend and show differences between the control and the mutant mice after induction of the transgene. Using a quadratic model, the direction of the temporal trend is not fixed, thus allowing detection of both consistent and temporary changes in gene expression profiles over the time course (Fig. 1).

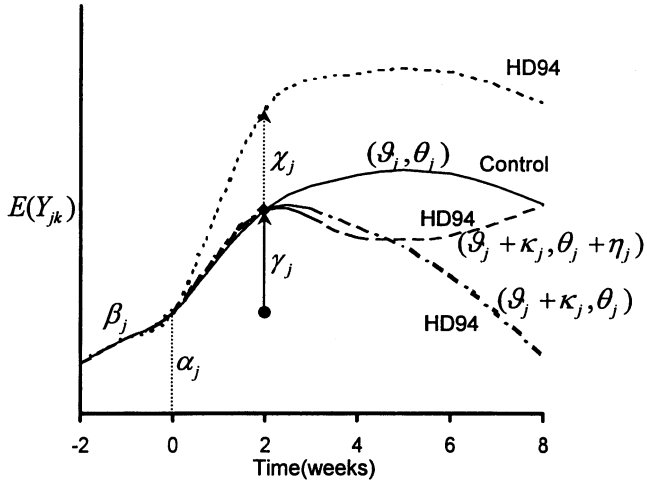

Figure 1. Schematic representation of the model. The horizontal axis indicates the timepoints during the time course; the vertical axis indicates the predicted gene expression level $E(Y_{jk})$. α_j (as indicated by a vertical dotted line) measures the intercept in both the HD94 and the tTA control mice at week 0 and β_j indicates the slope of the function before induction. After induction, for the tTA control mice (solid line), the intercept at week 2 is $\alpha_j + \gamma_j$; thus, γ_j (as indicated by a solid upward-pointing arrow) measures the shift at the basal level in the tTA control mice after induction. (ϑ_j, θ_j) is the slope and quadratic coefficient for the tTA controls (solid line) after induction; both of them determine the curve of the function for the tTA controls. If the expression of the gene in HD94 mice only differs from that in the tTA controls by intercept (dotted line), its slope and quadratic coefficient remain (ϑ_j, θ_j), and the curve remains the same as for the tTA controls. For HD94 mice, the intercept at week 2 is $\alpha_j + \gamma_j + \chi_j$; thus, χ_j (as indicated by a dotted upword-pointing arrow) measures the intercept difference between HD94 and the tTA controls. If the expression of the gene in HD94 mice differs from that in the tTA controls only by the slope (dash–dotted line), it shares the same intercept as the tTA controls ($\alpha_j + \gamma_j$); however, its slope becomes $\vartheta_j + \kappa_j$, and the curve is determined by ($\vartheta_j + \kappa_j$, θ_j). Thus, κ_j measures the slope difference between HD94 mice and the tTA controls. If the expression of the gene in HD94 differs from that in the tTA controls only by slope and quadratic coefficient (dashed line), its intercept is same as that of the tTA controls ($\alpha_j + \gamma_j$); however, its slope becomes $\vartheta_j + \kappa_j$, its quadratic coefficient becomes $\theta_j + \eta_j$, and the curve is determined by ($\theta_j + \kappa_j$, $\theta_j + \eta_j$); thus, η_j measures the quadratic coefficient difference between the HD94 and the tTA control mice.

Table 2. The biological interpretations of the regression coefficients in the model

Coefficient	Biological interpretation
α_j	Basal expression level at week 0 in both HD94 and tTA control mice
β_j	Slope of changing expression levels prior to induction in both HD94 and tTA control mice
γ_j	Shift of basal expression level in the tTA control mice after induction
χ_j	Difference at basal expression level between HD94 and tTA control mice at week 2
ϑ_j	Slope of changing expression levels in the tTA control mice after week 2
κ_j	Slope difference between HD94 and tTA control mice after week 2
θ_j	Curved trend of the expression level of a gene in the tTA control mice after week 2
η_j	Difference of the curved trend between HD94 and tTA control mice after week 2

Parameter estimation and inference

We used the least-squares technique to estimate parameters and the estimating-equation technique to assess their corresponding standard errors as described in (15). When the sample size is large, estimates derived from estimating equations have an

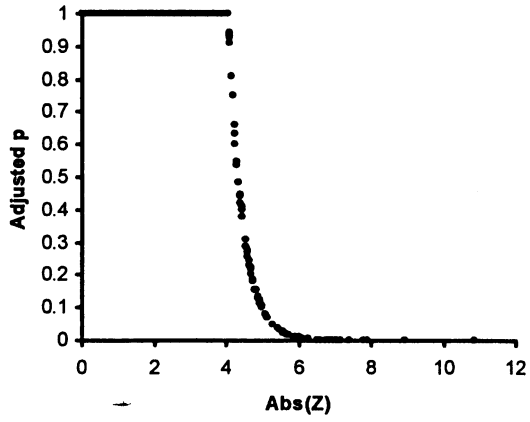

Figure 2. A plot of adjusted *P*-values versus *Z*-statistics. The horizontal axis shows the absolute value $|Z|$ of the *Z*-statistics; the vertical axis is the *P*-value adjusted by a modified Bonferroni correction (28). An adjusted *P*-value of 0.1 corresponds to $|Z| = 4.99$.

asymptotic normal distribution, and therefore do not require any further distributional assumptions when making inferences on these estimates (26,27). These techniques have been successfully used in two other microarray studies (15,16). Among all the parameters in the models, χ_j (intercept difference), κ_j (slope difference) and η_j (quadratic coefficient difference) are of particular interest. If any of these is significantly different from zero, it implies that the expression level of the *j*th gene is significantly different between the HD94 and the tTA control mice at some point after induction. To determine if they are significantly different from zero, we used the following *Z*-statistic:

$$Z_{\chi_j} = \frac{\hat{\chi}_j}{\mathrm{SE}_{\chi_j}}, \qquad Z_{\kappa_j} = \frac{\hat{\kappa}_j}{\mathrm{SE}_{\kappa_j}}, \qquad Z_{\eta_j} = \frac{\hat{\eta}_j}{\mathrm{SE}_{\eta_j}}, \qquad \mathbf{2}$$

where the hatted letters are estimates of the corresponding parameters, and the SEs in the denominators are their standard errors. As stated earlier, when the sample size k is sufficiently large, the above *Z*-statistics follow a normal distribution, on the basis of which *P*-values can then be calculated. However, since the sample size in this dataset is small, we used a *t*-distribution to compute approximated *P*-values. To control false-positives on the global scale of all genes tested, it is essential to adjust for multiple comparisons. In this application, we used a stepdown modified Bonferroni correction method (28). An example of adjusted *P*-values versus *Z*-statistics is shown in Figure 2. For *Z*-statistics with absolute values <4, the adjusted *P*-values approach 1.0. A *P*-value of 0.1 corresponds to an absolute value of *Z*, $|Z|$, equal to 4.994.

The analytical procedures

To focus on genes that are readily detected by the arrays, we filtered the genes using the criterion that the gene must have been called 'Present' in at least 12 samples out of the total 30 samples from week −2 to week 8 by Affymetrix's Absolute Analysis Algorithms. This resulted in 2143 genes being selected. The 'absolute difference' values of the 2143 genes (referred to as 'expression values' or 'the raw data' in this paper) were then subjected to a normalization procedure using a regression model to adjust for systematic heterogeneity among the samples (16). As noted above, the parameters χ_j, κ_j and η_j in the model are of particular importance. To identify genes that differ on a specific parameter (intercept χ_j, slope κ_j or quadratic coefficient η_j), we adopted a forward stepwise model-fitting procedure: sequentially adding one parameter a time, χ_j, κ_j, then η_j, computing Z_{χ_j}, Z_{κ_j} or Z_{η_j} and their corresponding *P*-values (p_{χ_j}, p_{κ_j}, p_{η_j}) in each step. Figure 3 shows six genes with the most positive or negative Z_{χ_j}, Z_{κ_j} or Z_{η_j}, calculated from the raw data of this dataset. At $P < 0.1$ ($|Z| > 4.994$), seven candidate genes were significantly different at slope κ_j between HD94 mice and the tTA controls. No gene was found to be significantly different at intercept χ_j or quadratic coefficient η_j (Table 3).

The rank transformation

In addition to using the raw expression values, we also transformed them to rank scores, that is, we replaced the actual expression values by their ranks (from 1 to 2143) as described by C. Cheng, R. Kimmel, P. Neiman and L.P. Zhao (manuscript in preparation). The advantages of using rank scores are (i) the transformation automatically adjusts for the systematic heterogeneity among samples, thus making results more stable, and (ii) the results will not be significantly influenced by genes with exceptionally large expression values. The disadvantage is that the rank transformation loses some degree of quantitative information. Figure 4 shows a typical monotonic relationship between rank scores and raw expression values. As shown, the large expression values were scaled down after rank transformation. Since the use of ranks automatically adjusts for heterogeneity, the factors (δ_k, λ_k) were set to (0, 1) in the model. Using the same estimation and inference technique, we obtained a set of test statistics using Equation 2 and computed their corresponding *P*-values. As examples, Figure 5 shows the six genes with the most positive or negative Z_{χ_j}, Z_{κ_j} or Z_{η_j}, calculated from the rank scores of the dataset. Comparing with the raw data, the use of rank scores generated more candidate genes at $P < 0.1$ ($|Z| > 4.994$): 1 significantly different at intercept χ_j, 32 at slope κ_j (2 overlap with the 7 candidate genes derived from the raw data) and 3 at quadratic coefficient η_j. A total of 36 unique genes were identified at this significance level (Table 3). As also shown in another study (C. Cheng, R. Kimmel, P. Neiman and L.P. Zhao, manuscript in preparation), rank transformation appeared to increase the power of analysis due to reducing variations in the raw data, when heterogeneities among data are large.

Permutation test and number of false discoveries (NFD)

For small sample sizes such as in this dataset, the distributions of *Z*-statistics (Z_{χ_j}, Z_{κ_j} or Z_{η_j}) may not be well approximated by either the normal distribution or the *t*-distribution. Therefore, the *P*-values calculated based on the normal or *t*-distribution are only an approximate indication of significance. The permutation test requires no assumption regarding the underlying distribution; thus, inferences made by this test are more reliable when the sample size is small. We can use

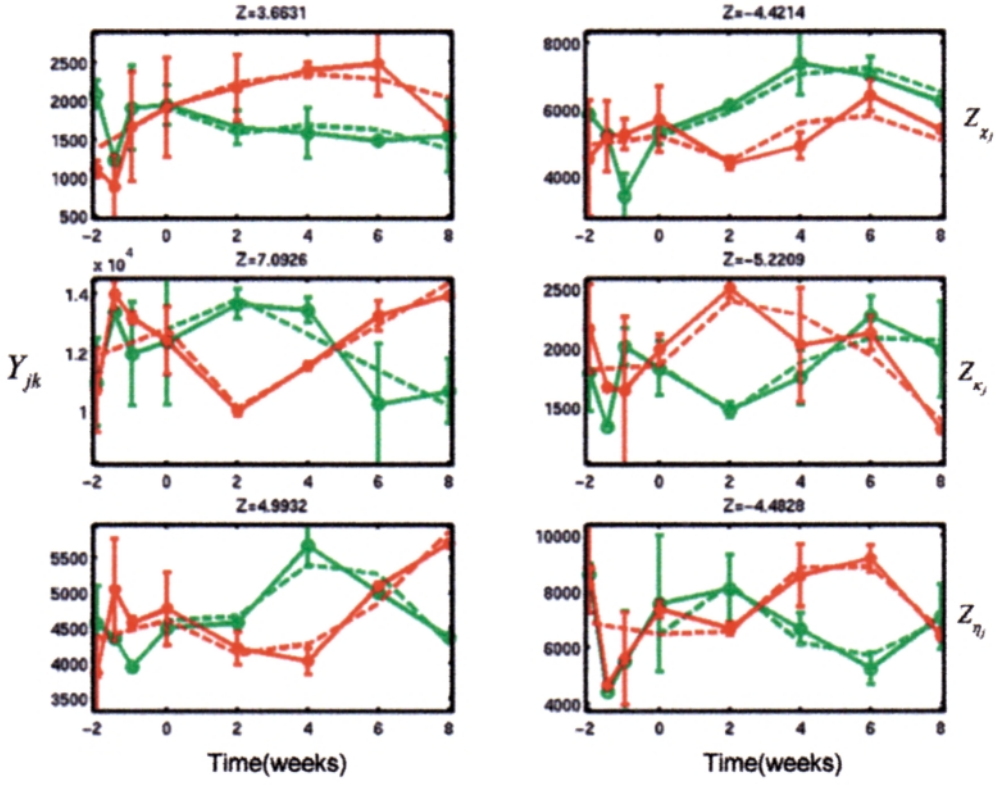

Figure 3. The expression of the genes with the six most positive or negative Z_{χ_j}, Z_{κ_j} or Z_{η_j}, calculated from the raw data. The corresponding Z-statistic (Z_{χ_j}, Z_{κ_j} or Z_{η_j}) is indicated on the right, next to the plots, and its value is shown on the top of each plot. The plots on the left show the genes with the most positive Z_{χ_j}, Z_{κ_j} or Z_{η_j}, while those on the right show the genes with the most negative Z_{χ_j}, Z_{κ_j} or Z_{η_j}. For each plot, the average of the normalized expression level of the gene in tTA controls (solid line, green) or in HD94 mice (solid line, red) at each time point (vertical axis) was plotted against the timepoints in the induction course (horizontal axis). The ranges of the gene expression level among the tTA controls or HD94 mice at each timepoint are indicated by the green (tTA controls) or red (HD94 mice) error bars. The dashed lines indicate the model-predicted expression level for the gene in the tTA controls (green) or in HD94 mice (red).

permutations to compute exact P-values (C. Cheng, R. Kimmel, P. Neiman and L.P. Zhao, manuscript in preparation); however, permutation-based P-values that also adjust for multiple comparisons could be very conservative. A P-value of 0.05 indicates that on all the genes tested, the chance of falsely identifying one gene as positive is 5%. In biomedical research, microarray experiments are often used as screening methods to identify a list of potential candidate genes, which are then subjected to additional verification experiments. With the numerous verification methods currently available, a higher number of false-positives is generally tolerable, and a criterion such as $P < 0.05$ after adjusting for multiple testing may be too stringent. Lowering the threshold will lead to a higher power of discovery; thus, more candidate genes can be identified, at a risk of an increased number of false-positives. However, reliable assessment of the risk of increased false-positives is essential.

In the current context, instead of computing exact P-values using permutations, we are more interested in estimating the number of falsely significant genes when a list of candidate genes are identified at a specific significant level, such as $P < 0.1$ ($|Z| > 4.994$). This has been formally defined as the number of false discoveries (NFD) (L.P. Zhao and C. Cheng, manuscript in preparation). We designed a specific permutation scheme that retained the timing structure but permutated the genotypes. More specifically, given all k samples, we retained the original relationship between the timing variables and the corresponding array. At each timepoint after induction, we

Table 3. The number of candidate genes at $P < 0.1$ ($|Z| > 4.994$) and the expected numbers of false discovers (NFD)

	Candidates from raw data	Expected NFD (from raw data)	Candidates from rank scores	Expected NFD (from rank scores)	Overlaps
By intercept	0	2.58	1	2.94	0
By slope	7	8.61	32	17.60	2
By quadratic coefficient	0	11.62	3	11.42	0
Total	7	22.77	36	31.90	2

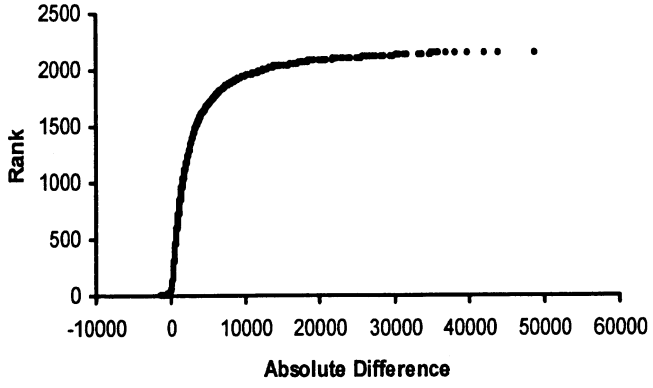

Figure 4. Rank transformation of the raw data: plot of rank versus absolute difference. If any absolute difference values are tied, their average rank was computed and used in transformation. A total of 2143 selected genes were ranked.

randomly permutated the tTA control and HD94 labels, and generated 648 permutated datasets for all possible combinations of the genotypes X_k at timepoints after the induction of the mutant htt transgene. The 648 permutated datasets represent all possible realizations under the null hypothesis that there is no difference between the mutant and control mice after induction, and any genes called significant in the permutated datasets are false-positives.

To calculate NFD at a certain threshold of the test statistic, we followed the same computing procedures as described above on each permutated dataset: sequentially fitting the model in a forward stepwise fashion, using either the raw data or the rank scores, and computing permutated Z_{χ_j}, Z_{κ_j} or Z_{η_j} statistics. At each step of model-fitting, we counted the number of genes with absolute values of Z_{χ_j}, Z_{κ_j} or Z_{η_j} exceeding a certain threshold Z_0. The expected NFD for each step is the average of such counts over all the permutated datasets. In addition, to estimate the total number of genes that were significantly different between the HD94 mice and the tTA controls, regardless of which parameter was considered, we counted the number of genes with any Z_{χ_j}, Z_{κ_j} or Z_{η_j} exceeding the threshold Z_0 in each permuted dataset, and, as above, the expected total NFD is the average of such counts over all the permutated datasets. As an example, Table 4 shows the expected NFD for Z_{κ_j} and the expected total NFD at different thresholds Z_0, compared with the number of candidate genes identified in the real dataset (both of which were calculated from the rank scores).

We used the permutation procedures described above to estimate NFDs with $Z_0 = 4.994$ ($P \sim 0.1$), and the results are listed in Table 3. When the raw data were used, the numbers of candidate genes identified in the real dataset were consistently smaller than the expected NFD under permutations, implying that the confidence in the identified genes is rather low. In contrast, when the rank scores were used in the analysis, 32 genes were found to be significantly different at slope κ_j, whereas the expected NFD was 17.6, predicting that about half of the candidates are likely to be verified in further studies.

The NFD can also serve as an alternative measure of statistical significance. They can be used to set the threshold of the test statistics. As shown in Table 4, with different values of the threshold Z_0, the number of candidate genes called significant and the expected NFD were different. For example, at the threshold $Z_0 = 5.0$, 31 candidate genes were called significant at Z_{κ_j}, while the expected NFD was 17.51, implying that for every 1.77 (31/17.51) candidate genes identified, one false-positive could be expected. However, if the threshold Z_0 were raised to 6.5, then 12 candidate genes would be called significant at Z_{κ_j}, while the expected NFD would be 4.4, suggesting that for each 2.73 (12/4.4) candidate genes identified, one false-positive would be expected. If the investigator has limited resources for verification work, a threshold of 6.5 would be a reasonable criterion; however, if the investigator is interested in finding more candidate genes and a larger amount of verification work is feasible, a threshold of 5.0 instead of 6.5 might be selected. Although the false-positive ratio (expected NFD/number of identified genes) is higher at $Z_0 = 5.0$, with 31 candidate genes identified instead of 12, it is likely that additional truly significant genes could be found from the candidates. Investigators can use such information as a guideline, combining it with other biological information and available resources, to choose a Z_0 that seems most appropriate for their experimental aims.

DISCUSSION

In this paper, we present a regression-based modeling approach to analyze multiple-series MTC data. A typical application of this modeling approach includes three steps: first, formulate a model that approximates the relationship between gene expression and experimental factors, with parameters incorporated to address the research interest; second, use least-squares and estimating-equation techniques to estimate parameters and their corresponding standard errors; third, compute test statistics, P-values and NFD as measures of statistical significance. The advantages of this approach are as follows. First, it addresses the research interest in a specific, systematic way, and maximally utilizes all the data and other relevant information. Second, it accounts for both systematic and random variations associated with the data, and the results of such analysis give not only gene-specific information relevant to the research goal, but also its reliability, thus helping investigators to make better decisions for follow-up studies. Third, this approach is very flexible, and can easily be extended to other types of MTC studies or other microarray experiments by formulating different models based on the experimental design of the studies.

In the current application, we established an order of genes that might express differentially between HD94 and control mice during the time course under study, based on the scores of Z-statistics (Z_{χ_j}, Z_{κ_j} or Z_{η_j}). These can serve as candidates for further studies. Consistent with the initial observation (R. Luthi-Carter and J. Olson, unpublished result), the changes detected on these genes were mild. However, a fraction of them were statistically significant, and, from a statistical point of view, it is likely that with additional replicates and a larger sample size, or focusing on a smaller set of genes, more candidate genes with modest changes can be identified as statistically significant.

We also tried a linear model that approximated the relationship between the expression level of a gene and the time after

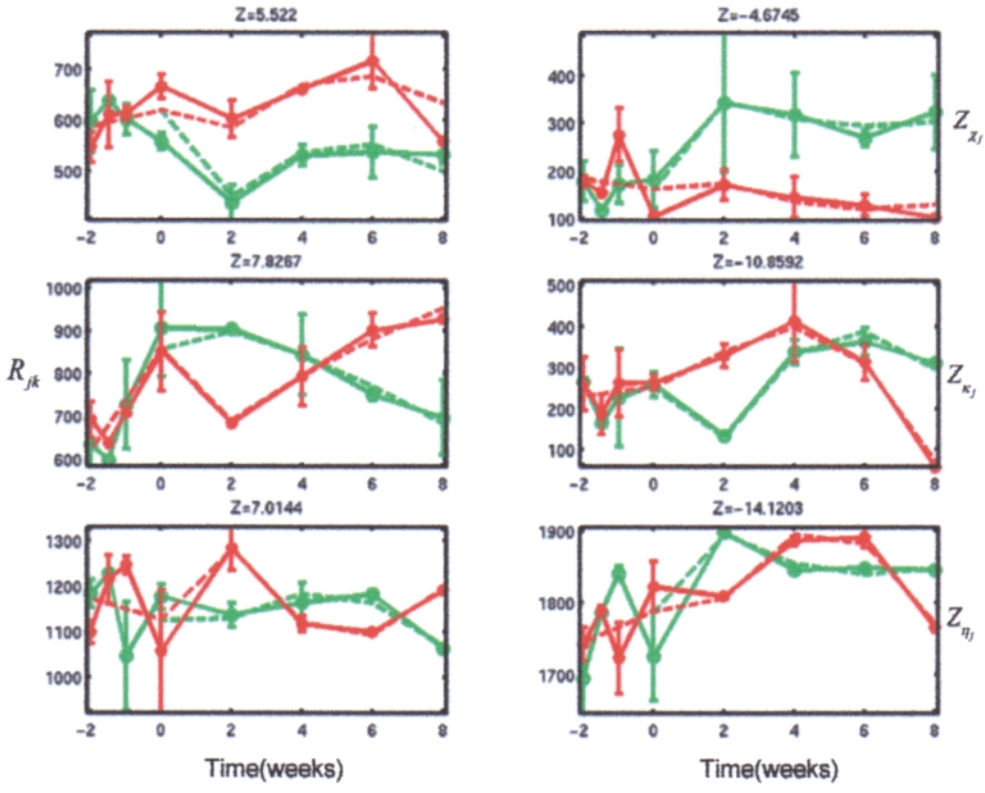

Figure 5. Genes with the most positive or negative Z_{χ_j}, Z_{κ_j} or Z_{η_j}, calculated from the rank scores. The corresponding Z-statistic: (Z_{χ_j}, Z_{κ_j} or Z_{η_j}) is indicated on the right, next to the plots, and its value is shown on the top of each plot. The plots on the left show the genes with the most positive Z_{χ_j}, Z_{κ_j} or Z_{η_j}, while those on the right show the genes with the most negative Z_{χ_j}, Z_{κ_j} or Z_{η_j}. For each plot, the average of the rank scores R_{jk} of the gene in the tTA controls (solid line, green) or in HD94 mice (solid line, red) at each timepoint (vertical axis) was plotted against the timepoints in the induction course (horizontal axis). The ranges of the rank scores among the tTA controls or HD94 mice at each timepoint are indicated by the green (tTA controls) or red (HD94 mice) error bars. The dashed lines indicate the model-predicted rank scores for the tTA controls (green) or HD94 mice (red).

induction as a straight line, as an alternative to the quadratic model described above. The candidate genes that it identified were largely a subset of those identified by the quadratic model presented above (data not shown). The mechanism underlying the temporal trend of the expression of a gene is complicated, and largely remains unknown. The linear model might not be a good approximation for many genes, and a higher-order model or other data transformation may be more appropriate. For example, in this application, we used a quadratic model to approximate various types of temporal trends, including no temporal trend with zero slope coefficients, a linear trend with zero quadratic coefficients, and a quadratic trend. With a larger sample size and more timepoints, one can also expand such a model to include higher orders, resulting in a rich class of polynomial models. These models are closely related to approximations of any complex function via Taylor expansion. The goodness of fit of the models is likely to affect the sensitivity and specificity of the detection. How to choose the most appropriate model is currently under study.

Table 4. NFD and the number of significant genes at different values of threshold Z_0

Z_0	Number of genes with $Z_{\kappa_j} > Z_0$[a]	Expected NFD for Z_{κ_j}[a]
4.5	55	28.17
5.0	31	17.51
5.5	25	10.84
6.0	16	6.80
6.5	12	4.40
7.0	7	2.79
7.5	4	1.84
8.0	2	1.17

[a]Computed from rank scores.

Logarithmic transformation is often performed in microarray data analysis. For datasets generated using arrays with an Affymetrix platform such as this one, it is necessary to remove or recode a large fraction of data with negative values before logarithmic transformation, and the statistical consequences of such recoding of original data are not clear. The rank transformation procedure described above can serve as an alternative transformation, which also simultaneously adjusts for the heterogeneity. In this dataset, results derived from logarithmically transformed data largely overlapped with those from those using rank-transformed data (data not shown). Both analyses appeared to be more powerful than use of the raw data, probably owing to reduced heterogeneity and outlier effects.

Multiple testing presents a large problem for microarray data analysis, since most microarray datasets contain thousands of

genes observed on few samples. A calculation of *P*-values that accounts for multiple comparisons could be too conservative for studies with a screening purpose. Alternative measures such as NFD described above may be more appropriate and more relevant in interpreting microarray results, as well as for designing further experiments. A similar approach has been reported in (29).

MATERIALS AND METHODS

HD94 datasets

Animal maintenance and doxycycline (dox) treatment were as described in (24). Doxycycline was removed from the drinking water at postnatal week 4 to induce the expression of mutant htt in HD94 mice. The expression of mutant htt reached peak level in HD94 central nervous system 2 weeks after induction (A. Yamamoto, unpublished observation). RNA was harvested from the striata of HD94 mice or the single transgenic control mice carrying the tTA molecule at weeks −2, −1.5, −1, 0, 2, 4, 6 and 8, with week 0 as the time of induction. Two replicates of HD94 or tTA samples were prepared at each timepoint except in week −1.5 (one tTA controls and two HD94 samples) and week 8 (two controls and one HD94 sample). Each RNA sample was labeled and hybridized to a Mu11K Sub B oligonucleotide array (Affymetrix), and the fluorescence intensity was measured using GeneChip3.1 software (Affymetrix). The absolute difference of each dataset was used for further analysis.

Data preprocessing

Among the 6500 elements on the array, many of them were called 'Absent' by Affymetrix's Absolute Analysis Algorithms. To focus on the genes that were readily detected by the array, we used the following filter: the genes must have been called 'Present' in at least 12 samples out of the total of 30 samples from week −2 to week 8, which resulted in 2143 genes being selected. The 'absolute difference' values of the 2143 genes were subjected to a normalization procedure using a regression model to adjust for systematic heterogeneity among the samples (16). For the rank transformations, ranks were computed for the absolute difference values of the selected genes in each array. In the case of ties, the average rank was assigned. For logarithmic transformation, any absolute difference value <20 was replaced with 20.

Analysis implementation

Methods and algorithms were implemented using MATLAB (The MathWorks, Inc). A user-friendly software GenePlus is available from Enodar Biologic Corporation (http://www.enodar.com).

ACKNOWLEDGEMENTS

We are grateful to René Hen, Ruth Luthi-Carter, Andrew Strand, Ai Yamamoto and the Hereditary Disease Foundation for making the dataset available to us. We also thank Ai Yamamoto, Ruth Luthi-Carter and Najma Khalid for helpful discussions and critical evaluation of the manuscript. This work was supported in part by NIH Grants RO1 NS42157-01, RO1 GM58897-02 and RO1 HG02283-02.

REFERENCES

1. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
2. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J., Jr, Boguski, M. S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
3. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.
4. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.
5. 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.
6. Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.
7. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
8. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
9. Cho, R.J., Huang, M., Campbell, M.J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S.J., Davis, R. W. and Lockhart, D. J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.
10. Lukashin, A.V. and Fuchs, R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.
11. Walker, J. and Rigley, K. (2000) Gene expression profiling in human peripheral blood mononuclear cells using high-density filter-based cDNA microarrays. *J. Immunol. Meth.*, **239**, 167–179.
12. Toronen, P., Kolehmainen, M., Wong, G. and Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451**, 142–146.
13. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
14. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
15. Zhao, L.P., Prentice, R. and Breeden, L. (2001) Statistical modeling of large microarray data sets to identify stimulus–response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.
16. Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.
17. The Huntington's Disease Collaborative Research Group. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.
18. Usdin, M.T., Shelbourne, P.F., Myers, R.M. and Madison, D.V. (1999) Impaired synaptic plasticity in mice carrying the Huntington's disease mutation. *Hum. Mol. Genet.*, **8**, 839–846.

19. Hodgson, J.G., Agopyan, N., Gutekunst, C.A., Leavitt, B.R., LePiane, F., Singaraja, R., Smith, D.J., Bissada, N., McCutcheon, K., Nasir, J. *et al.* (1999) A YAC mouse model for Huntington's disease with full-length mutant huntingtin, cytoplasmic toxicity, and selective striatal neurodegeneration. *Neuron*, **23**, 181–192.
20. White, J.K., Auerbach, W., Duyao, M.P., Vonsattel, J.P., Gusella, J.F., Joyner, A.L. and MacDonald, M.E. (1997) Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nat. Genet.*, **17**, 404–410.
21. Schilling, G., Jinnah, H.A., Gonzales, V., Coonfield, M.L., Kim, Y., Wood, J.D., Price, D.L., Li, X.J., Jenkins, N., Copeland, N. *et al.* (2001) Distinct behavioral and neuropathological abnormalities in transgenic mouse models of HD and DRPLA. *Neurobiol. Dis.*, **8**, 405–418.
22. Reddy, P.H., Williams, M., Charles, V., Garrett, L., Pike-Buchanan, L., Whetsell, W.O., Jr, Miller, G. and Tagle, D.A. (1998) Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA. *Nat. Genet.*, **20**, 198–202.
23. Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S.W. *et al.* (1996) Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, **87**, 493–506.
24. Yamamoto, A., Lucas, J.J. and Hen, R. (2000) Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell*, **101**, 57–66.
25. 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. Genet.*, **9**, 1259–1271.
26. Prentice, R.L. and Zhao, L.P. (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825–839.
27. Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.
28. Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple test of significance. *Biometrika*, **75**, 800–802.
29. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.