

# Improved Statistical Inference from DNA Microarray Data Using Analysis of Variance and A Bayesian Statistical Framework

ANALYSIS OF GLOBAL GENE EXPRESSION IN *ESCHERICHIA COLI* K12

Received for publication, November 8, 2000, and in revised form, March 16, 2001  
Published, JBC Papers in Press, March 20, 2001, DOI 10.1074/jbc.M010192200

Anthony D. Long<sup>‡§</sup>, Harry J. Mangalam<sup>¶</sup>, Bob Y. P. Chan<sup>\*\*‡‡</sup>, Lorenzo Toller<sup>§§¶¶</sup>,  
G. Wesley Hatfield<sup>§§¶¶</sup>, and Pierre Baldi<sup>\*\*‡‡</sup>

From the <sup>‡</sup>Department of Ecology and Evolutionary Biology, School of Biological Sciences, the <sup>\*\*</sup>Department of Information and Computer Science, the <sup>§§</sup>Department of Microbiology and Molecular Genetics, the <sup>‡‡</sup>Department of Biological Chemistry, College of Medicine, and the <sup>¶¶</sup>Department of Chemical and Biochemical Engineering, School of Engineering, University of California, Irvine, California 92697, the <sup>¶</sup>National Center for Genome Resources, Santa Fe, New Mexico 97505, and <sup>¶¶</sup>tag Informatics, Irvine, California 92612

We describe statistical methods based on the *t* test that can be conveniently used on high density array data to test for statistically significant differences between treatments. These *t* tests employ either the observed variance among replicates within treatments or a Bayesian estimate of the variance among replicates within treatments based on a prior estimate obtained from a local estimate of the standard deviation. The Bayesian prior allows statistical inference to be made from microarray data even when experiments are only replicated at nominal levels. We apply these new statistical tests to a data set that examined differential gene expression patterns in IHF<sup>+</sup> and IHF<sup>-</sup> *Escherichia coli* cells (Arfin, S. M., Long, A. D., Ito, E. T., Toller, L., Riehle, M. M., Paegle, E. S., and Hatfield, G. W. (2000) *J. Biol. Chem.* 275, 29672–29684). These analyses identify a more biologically reasonable set of candidate genes than those identified using statistical tests not incorporating a Bayesian prior. We also show that statistical tests based on analysis of variance and a Bayesian prior identify genes that are up- or down-regulated following an experimental manipulation more reliably than approaches based only on a *t* test or fold change. All the described tests are implemented in a simple-to-use web interface called Cyber-T that is located on the University of California at Irvine genomics web site.

Fluorescently or isotopically labeled cDNA or RNA probes are hybridized to high density arrays of cDNA clones on glass supports (1–4, 6–8), nylon membranes (9–15), or oligonucleotides directly synthesized on silica wafers (16, 17). Signals are quantified using phosphorimaging, photomultiplier tubes, or CCD imaging, and a data set is created that consists of expression measurements for all of the elements of the array.

Despite rapid technological developments, the statistical tools required to analyze these fundamentally different types of DNA microarray data are not in place. Data often consist of expression measures for thousands of genes, but experimental replication at the level of single genes is often low. This creates problems of statistical inferences because many genes will show fairly large changes in gene expression purely by chance alone. Therefore, to interpret data from DNA microarrays it is necessary to employ statistical methods capable of distinguishing chance occurrences from biologically meaningful data.

The *t* test can be used to determine whether the observed difference between two means is statistically significant (18). The *t* test incorporates a measure of within treatment error into the statistical test; as a result only genes showing a large change in gene expression relative to the within treatment variance are considered to have significantly changed. In a perfect world, all DNA microarray experiments would be highly replicated. Such replication would allow accurate estimates of the variance within experimental treatments to be obtained, and the *t* test would perform well. However, samples may be available in limited supply, and DNA microarray experiments are expensive and time consuming to carry out. As a result, the level of replication within experimental treatments is often low. This results in poor estimates of variance and a correspondingly poor performance of the *t* test itself.

An alternative to the *t* test is to ignore the within treatment variance and only look at fold change as a proxy for statistical significance. Intuition suggests that larger observed fold changes can be more confidently interpreted as a stronger response to the experimental manipulation than smaller observed fold changes. However, an implicit assumption of this reasoning is that the variance among replicates within treatments is the same for every gene. In reality, the variance varies among genes (*e.g.* see Fig. 2), and it is critical to incorporate this information into a statistical test. These different approaches demonstrate the statistical problem faced when analyzing DNA microarray data. Ignoring the sampling variance is incorrect, yet incorporating it in the traditional manner may not be much better because the number of replicate experi-

The recent availability of complete genomic sequences and/or large numbers of cDNA clones from model organisms coupled with technical advances in DNA arraying technology have made it possible to study genome-wide patterns of gene expression. Most high density microarray experiments consist of one of two types: examining changes in gene expression over a temporal or treatment gradient (1) or comparing gene expression between two different cell sample types or genotypes (2–5).

\* This work was supported in part by National Institute of Health Grants GM-58564 (to A. D. L.) and GM55073 (to G. W. H.), National Science Foundation Grant MCB-9743252 (to G. W. H.), Grant 99-15 from the University of California Biotechnology Research and Education Program, a University of California Irvine Laurel Wilkening Faculty Innovation Award (to P. B.), and a predoctoral fellowship from the Programma di Dottorato in Scienze Genetiche, Università degli Studi di Pavia (to L. T.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed.

ments is often quite low.

Although there is no substitute for experimental replication, confidence in the interpretation of DNA microarray data with a low number of replicates can be improved by using a Bayesian statistical approach (19) that incorporates prior information of within treatment measurement. The Bayesian prior assumes that genes of similar expression levels have similar measurement errors, amounting to parallel pseudo-replication of the experiment. For example, the variance of any single gene can be estimated from the variance from a number of genes of similar expression level. More specifically, the variance of any gene within any given treatment can be estimated by the weighted average of a prior estimate of the variance for that gene. This weighting factor, or hyperparameter, is controlled by the experimenter and will depend on how confident the experimenter is that the background variance of a closely related set of genes approximates the variance of the gene under consideration. In the Bayesian approach employed in this study, the weight given to the within gene variance estimate is a function of the number of observations contributing to that value. This leads to the desirable property of the Bayesian approach converging to the  $t$  test as the experimenter carries out additional replications and thus becomes more confident in the observed estimate of within treatment variance.

Commonly used software packages are poorly suited for implementing the Bayesian statistical methods we develop in this work. However, we have created a program, Cyber-T, which accommodates this approach. Cyber-T is available for on-line use at the genomics web site at the University of California at Irvine. This program is ideally suited to experimental designs in which replicate control cDNA samples are being compared with replicate experimental cDNA samples.

In this study we use the statistical tools incorporated into Cyber-T to compare and analyze the gene expression profiles obtained from a wild-type strain of *Escherichia coli* and an otherwise isogenic strain lacking the gene for the global regulatory protein, integration host factor (IHF),<sup>1</sup> previously reported by Arfin *et al.* (5). We apply different statistical methods for identifying genes showing changes in expression to this data set and show that a Bayesian approach identifies a stronger set of genes as being significantly up- or down-regulated based on our biological understanding of IHF regulation. We show that commonly used approaches for identifying genes as being up- or down-regulated (*i.e.*, simple  $t$  test or fold change thresholds) require more replication to approach the same level of reliability as Bayesian statistical approaches applied to data sets with more modest levels of replication. We further show that statistical tests identify a different set of genes than those based on fold change and argue that the set of genes identified by fold change is more likely to harbor experimental artifacts.

#### MATERIALS AND METHODS

**Statistical Framework**—As data are commonly background-subtracted, zeros in the input file do not actually represent no expression but instead an expression level that escapes detection. A decision may be made to statistically analyze only genes for which a predetermined number of replicates have above zero expression level estimates. Cyber-T allows the user to define a constant,  $\rho$ , which is the minimum number of replicates required to perform the  $t$  test. Depending on the choice of  $\rho$  and the nature of the data, different statistical tests are carried out. For example, when fewer than  $\rho$  non-zero expression levels are measured for both the control and experimental treatments no statistical test is carried out. The other two cases are detailed below.

In the case of  $\rho$  or greater non-zero observations for both the control and experimental treatments, a two-sample  $t$  test is carried out (see Equation 9.2 in Ref. 18). In the case of equal variances and sample sizes

for each of the control and experimental treatments, the  $t$  statistic reduces to

$$t = \frac{\sqrt{n}(\bar{X}_C - \bar{X}_E)}{\sigma} \quad (\text{Eq. 1})$$

where  $n$  is the number of replicates within each of the control and experimental treatments,  $\sigma$  is the standard deviation in expression level observed within each treatment, and  $\bar{X}_C$  and  $\bar{X}_E$  are the mean expression levels of the control and experimental treatments, respectively. It can be seen from the above expression that large differences in expression level only result in large values of the  $t$  statistic (and hence reach statistical significance) when the within treatment variance is small. This test has intuitive appeal because large differences in expression are unlikely to be of biological significance (or at least less worthy of further investigation) if they are inherently unreplicable.

When the control treatment has less than  $\rho$  non-zero observations, but the experimental treatment has  $\rho$  or greater non-zero observations, a different statistical test is carried out. Data of this form correspond to the biological situation of a gene that is expressed at an undetectable level in the control treatment but induced by the experimental treatment. In this case a  $t$  statistic is constructed that has the following form.

$$t = \frac{\sqrt{n}(\bar{X}_E - \tau)}{\sigma_E} \quad (\text{Eq. 2})$$

where  $n$  is the number of replicates within the experimental treatment,  $\sigma_E$  is the standard deviation in expression level observed within the experimental treatment,  $\bar{X}_E$  is the mean expression level of the experimental treatment, and  $\tau$  is a constant whose definition depends on the number of non-zero observations in the control treatment. If there are any non-zero control observations, then  $\tau$  is equal to the mean of the non-zero control observations. Otherwise  $\tau$  will reflect the smallest expression level reliably detected. That is, we test whether the observed experimental expression level is greater than the smallest expression level that can be experimentally detected. The smallest expression level that can be experimentally detected is defined as the average of the two 0.25% quantiles associated with the average non-zero control and experimental expression levels over all genes on the array. Less precisely,  $\tau$  equals the expression level above which 99.75% of the average expression levels fall. In the case where the experimental treatment appears greatly repressed relative to the control treatment, the statistic is appropriately modified.

Generally high density array experiments are replicated only a few times. In this case the estimate of  $\sigma^2$  obtained may be a poor estimate of the true variation present among replicates within treatment. It is possible to improve upon the poor performance of the  $t$  test through the use of Bayes formula, which defines probability ( $P$ ) of the model ( $M$ ) conditional on the observed data ( $D$ ):

$$P(M|D) = [P(D|M) P(M)]/P(D) \quad (\text{Eq. 3})$$

At its most basic level, the Bayes formula can be used to calculate the posterior probability of any model or hypothesis in light of the data and any prior information one may have. In the case of high density array data it is convenient to model expression levels using a Gaussian probability distribution function of unknown mean and variance. We use the corresponding self-conjugate prior (self-conjugate refers to the case where both the prior and the posterior have the same functional form) of the Gaussian distribution with unknown mean ( $\mu_0$ ) and variance ( $\sigma_0^2$ ). Without going into the mathematical details, which will be presented elsewhere (20), it can be shown that the posterior estimate of the within sample variation is as follows.

$$\sigma_p^2 = \frac{v_0\sigma_0^2 + (n-1)\sigma^2}{v_0 + n - 2} \quad (\text{Eq. 4})$$

where  $\sigma^2$  is the actual estimate of within treatment among replicate variation,  $n$  is the number of replicates, and  $v_0$  is weighting factor which scales with confidence in the prior (hyperparameter). That is, the variance is estimated as if one had, in addition to the sample square deviation, a set of  $v_0$  observations with variance  $\sigma_0^2$  (21). There are a number of reasonable choices for  $\sigma_0^2$ , and we choose to estimate  $\sigma_0^2$  based on a local average of the standard deviations for genes showing similar within treatment expression levels to the gene under consideration. Local averaging is carried out by ordering all genes within a given treatment based on their average expression level and then taking the average standard deviation as the standard deviation observed for any given gene and the  $k$  next higher and lower expressing genes (where  $k$

<sup>1</sup> The abbreviations used are: IHF, integration host factor; ORF, open reading frame.

is a user defined constant). This local averaging is carried out using a C function described by Spector (22) that accounts for “edge” effects. The  $t$  tests employed are the same as those described above with  $\sigma_p^2$  substituted for  $\sigma^2$ .

Although there exist methods for choosing an optimal hyperparameter,<sup>2</sup> these complex approaches are not implemented here nor in the current version of the Cyber-T software. Instead, a heuristic approach is taken whereby the user can select the constant  $v_0$  representing the total number of data points deemed necessary to compute reasonable estimates of means and variances. In Cyber-T, the default value for  $v_0$  is 10. If for a given gene  $n < v_0$  values are empirically measured, Cyber-T complements these measurements with an additional  $v_0$  “pseudo-observations” equal to the background value derived by forming an average over neighboring genes. Given the typical levels of replication we have seen in microarray experiments (*i.e.* 2–5), we have found the default value for the hyperparameter of 10 to work well. Analysis of actual data sets (this paper) and simulated data sets<sup>2</sup> shows that neither doubling nor halving the default value of 10 has a great effect on the subsequent ranking of results.

In the case of high density arrays a very large number of  $t$  tests will be carried out on the data. Carrying out such a large number of statistical tests will result in an elevated false positive rate if the  $p$  value used for statistical significance is comparable with the value used when only a single  $t$  test is performed (*e.g.*  $p < 0.05$  or  $p < 0.01$ ). To overcome this problem a statistical threshold must be set on individual  $p$  values so that the experiment-wide false positive rate is held constant at a fixed level. Such a correction applied to individual  $p$  values is often referred to as a Bonferroni correction. If one wishes to hold the experiment-wide false positive rate to  $\alpha_{\text{Bonf}}$  and carries out  $N$  independent statistical tests, then the marginal  $p$  value associated with any individual test must be set as follows.

$$\alpha = 1 - \exp\left[-\frac{\ln(1 - \alpha_{\text{Bonf}})}{N}\right] \quad (\text{Eq. 5})$$

In Cyber-T a slightly more stringent Bonferroni correction is applied that requires  $p$  values to exceed  $\alpha$  for the  $t$  tests carried out on both the log-transformed and raw data. Because a Bonferroni correction assumes  $p$  values for different genes to be independent, it will often be overly conservative. Nonetheless, it has utility when it is important to demonstrate strict statistical significance.

**Analysis of IHF Data**—The IHF data set is modified from that published by Arfin *et al.* (5). The four replicates within a given cell type are each averaged over two independent hybridizations of the same RNA sample and two measures of intensity from each double spotted membrane. Only genes for which all eight resulting measures of gene expression are above background are used (1973 genes in total). We refer to the resulting eight measures of gene expression per gene as 1–8 where 1–4 are from IHF<sup>+</sup> cells and 5–8 are from IHF<sup>−</sup> cells. For the comparisons involving 2-fold replication, all 12 possible 2 by 2 comparisons differing by at least two replicates were carried out (*i.e.*, 12v56, 12v78, 13v57, 13v68, 14v58, 14v67, 23v58, 23v67, 24v57, 24v68, 34v56, and 34v78). A corresponding set of 9 possible 3 by 3 comparisons differing by at least one replicate was also carried out (*i.e.*, 123v567, 123v568, 123v678, 124v567, 124v568, 124v678, 234v567, 234v568, and 234v678). For both the 2 by 2 and 3 by 3 comparisons, the 120 genes with the lowest  $p$  values were identified for seven statistical tests (*i.e.* Bayesian  $t$  test,  $t$  test, ratio of means on the raw data, Bayesian  $t$  test,  $t$  test, ratio of means on the log-transformed data, and a difference of means for the log-transformed data). In the case of the two Bayesian  $t$  tests three different sliding window sizes (*i.e.* 41, 101, and 161) as well as three different hyperparameter values (*i.e.*, 2, 4, and 16) were tested in all possible combinations. The number of genes commonly identified was tallied for each of the 66 possible combinations of the 12 different 2 by 2 comparisons and for each of the 36 possible combinations of the 9 different 3 by 3 comparisons. For each statistical approach the mean number of genes in common is presented in Table I. With the exception of the two statistics based on ratios of means (standard deviation for the ratio of raw means for 2 by 2, ratio of raw means for 3 by 3, ratio of mean of logs for 2 by 2 and ratio of mean of logs for 3 by 3 were 25, 32, 26, and 13, respectively), standard deviations over comparisons were generally small and between 7 and 13. The different combinations of window size and hyperparameter seemed to have little effect on the consistency of the Bayesian approach so only outcomes corresponding to the best and worst parameter combinations are presented in Table I. The compari-

sons among statistical approaches presented in Table II were generated in a manner similar to those described above.

To identify those genes differentially expressed in strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>−</sup>) most likely as a direct consequence of IHF-mediated effects on transcription initiation, 500 base pairs upstream of the ORF for each of these genes were examined for high affinity IHF-binding sites. To identify these sites we demanded a 12 of 13 match with the core consensus sequence (5′-WATCAANNTR-3′, where W = A/T and R = pyrimidine) and required that at least 10 of 15 of the base pairs immediately upstream of the core sequence were AT base pairs (5). These are very stringent criteria that certainly identify high affinity IHF-binding sites; in fact, they exceed the criteria for several documented high affinity IHF-binding sites. Also, IHF-binding sites that perform a DNA loop function located farther than 500 base pairs of an ORF would not be identified in our search. Nevertheless, these stringent criteria have been used in an earlier report to identify IHF-binding sites upstream of 46 operons demonstrated to be regulated by IHF (5).

**Computer Software**—Cyber-T is designed to accept data in the large data spreadsheet format that is generated as output by the software typically used to analyze array experiment images. An element may correspond to a single spot on the array (typical of membrane- or glass slide-based arrays) or a set of spots (typical of GeneChips (16, 17) designed to query labeled RNA). This data file is uploaded to Cyber-T using the “Browse” button in the Cyber-T browser window. Fig. 1 shows a typical Cyber-T window. This window will vary somewhat depending on the exact analysis that the user wishes to carry out. Detailed instructions for using Cyber-T are accessed from this web page. We briefly describe the use of Cyber-T below.

After uploading the data file, the user defines the columns on which analysis will be performed. One button allows the user to define a Bonferroni significance level: the probability of a single gene being identified as significantly different between the control and experimental treatments by chance alone given the number of genes examined. A second button allows the user to define an integer “confidence” given to the Bayesian prior estimate of within treatment variance (referred to as the Bayesian prior). Larger confidence gives greater weight to the Bayesian prior and smaller confidence gives greater weight to the experimentally observed variance for the particular gene being tested. We have observed satisfactory performance when the actual number of observations and the confidence value total approximately ten. A third button allows the user to define the size of the window over which the Bayesian prior is calculated. If a window of size  $w$  is selected, the Bayesian prior for a given gene is estimated as the average standard deviation of a window of  $w$  genes ranked by expression level, centered on the gene of interest. Window sizes of larger than 100 genes often perform well. Additional buttons exist to control data formatting (Fig. 1).

Cyber-T generates three output files, two of which (allgenes.txt and siggenes.txt) can either be viewed in the browser window or downloaded and imported into a spreadsheet application for user-specific formatting. The file siggenes.txt contains a set of columns similar to those of allgenes.txt except only those genes, if any, that pass a Bonferroni test are listed. These files return the original data and a number of additional columns containing summary statistics such as the mean and standard deviation of both raw and log-transformed data, estimates of the standard deviations employing the Bayesian prior,  $t$  tests incorporating the Bayesian prior on both the raw and log-transformed data,  $p$  values associated with  $t$  tests, and “signed fold change” associated with the experiment. The exact content of these files is detailed on line. In addition to allgenes.txt and siggenes.txt, an additional postscript file, CyberT.ps, is generated. CyberT.ps contains a self-explanatory set of six graphs useful in visualizing the data submitted to the program (A–F of Fig. 2 contain examples of these plots). In some cases, two additional graphs are generated (not shown) that plot the  $p$  values of genes that pass the Bonferroni test against their absolute fold change in expression. These plots are suppressed if only a few genes pass the Bonferroni test, which will often be the case.

All statistical analysis is carried out using the *hdarray* library of functions written in R. R is a freely available statistical analysis environment adhering to the Open Source development model. The *hdarray* functions are normally invoked through the Cyber-T web-based interface but can also be used directly in an X-Window session. A brief tutorial on how to analyze data directly in R is available at the University of California at Irvine genomics web site. This tutorial also lists the functions available as part of the *hdarray* library and R resources. The library and Cyber-T web interface also includes routines for analyzing paired samples, which would be produced from two-dye glass slide DNA microarray experiments (1–4, 6–8).

<sup>2</sup> P. Baldi and A. D. Long, unpublished observations.

## Gene Expression Array Analysis for C + E Data using Cyber-T

<a href="#">General Help</a>	<a href="#">Example C+E Data Set</a>
------------------------------	--------------------------------------

---

Data File to Upload: ( [Format expected](#), [Data Coding](#) )

Data fields delimited by:  \* whitespace = TABS & spaces

Delete lines which have NULL Labels.

If your file won't upload, or Cyber-T won't process it, [check these possible reasons.](#)

---

Please enter any text that you would like to have as a header for the analysis output.

Columns start at 0, not 1; leading / lagging spaces are bad.

Label Columns (as # # #...):

Control Data Replicate Columns (as # # #...):

Experimental Data Replicate Columns (as # # #...):

Minimum non-zero Replicates Required (#):

If left blank, the number of values of Experimental Data will be used.

Values less than  will be set to 0 and ignored in calculations. (Leave blank to include all values).

---

In the Analytical selections below, this LIGHT GREEN SECTION refers to the Bayesian analysis and is optional (it will not be done unless there is a value in the "Confidence Value" space).	
Choose a sliding window size for approximating the variance of your values. (ie. How many total samples around the point of interest will give you a satisfactory estimate of the local variance?)	101
Enter a confidence value here that applies to the Bayesian Variance Estimate that you set immediately above. A decent default would be '10' or about 3 times the number of replicates per treatment. <i>If left blank, the whole Bayesian Estimate Analysis (light green) will be skipped.</i>	<input type="text"/>
Bonferroni Correction - Experiment-wide false positive rate (the probability of a single gene significant by chance alone)	0.25
Repeat Label Line every ~50 lines to tell you what the columns are.	<input type="checkbox"/>
In the Graphics Output, this many plots should be placed on 1 page.	2
Convert default postscript output to PDF (can view the results with Acrobat).	<input type="checkbox"/>
View input/output from top N results (by 'p' value) in 2/3D using <a href="#">xgobi</a> interactively. If you know your X DISPLAY value, you can type it in the following window. Otherwise, we'll assume that it's the same machine that your Netscape is running on (usually a good bet). Your X DISPLAY: <input type="text"/> [machine.net.domain.edu:0.0 or 128.200.34.145:0.0 ] Leave entries blank to skip and DON'T do this if you're more than 5 network hops from the server as this generates a lot of traffic and is sensitive to latency. (Requires a running X Window Server allowing X access from this server)	Where N = <input type="text"/>

---

This analysis takes SEVERAL MINUTES to run.

---

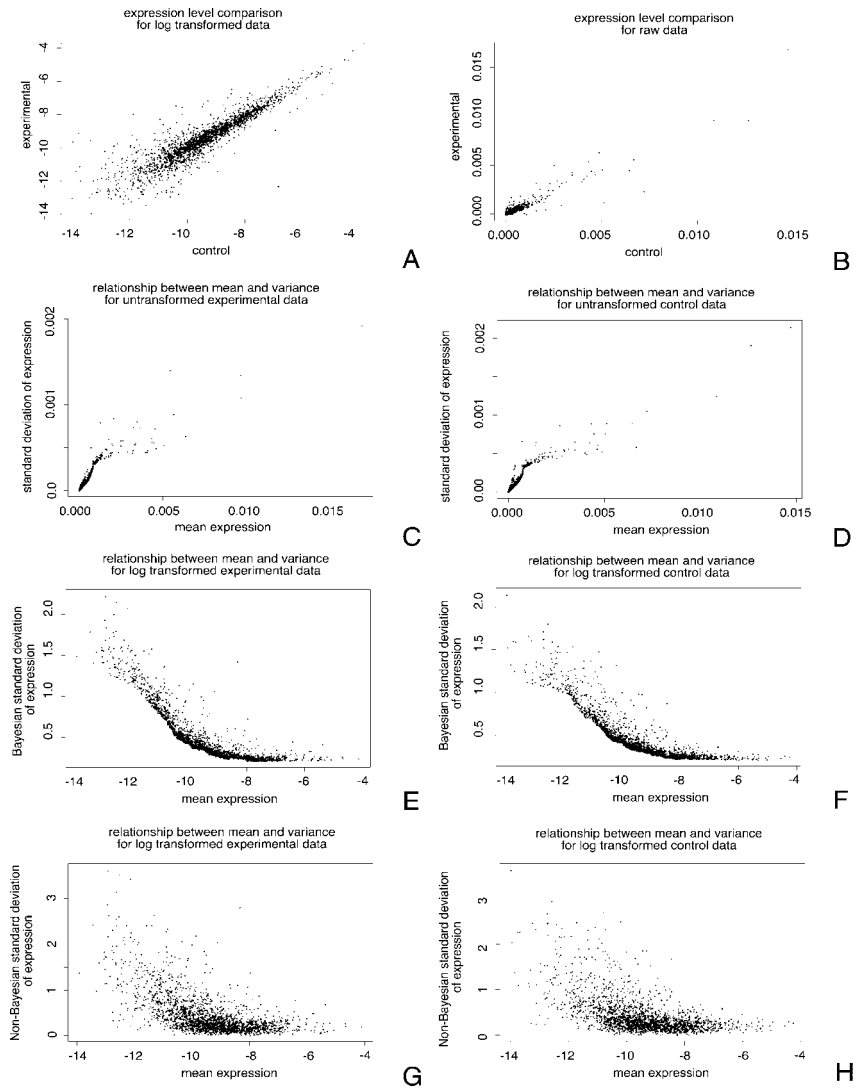
FIG. 1. An example of the Cyber-T interface. See text for description.

## RESULTS

*An Application of Cyber-T to Microarray Data*—Cyber-T was used to analyze a data set that examined gene expression levels in IHF mutant and wild-type strains of *E. coli* (5). This data set was obtained from <sup>33</sup>P-labeled cDNA reverse transcribed from total RNA with random hexamer oligonucleotides and hybridized to commercially available nylon arrays obtained from Sigma Genosys. These arrays are spotted in duplicate with each of the 4,290 predicted *E. coli* ORFs. This is an excellent data set for examining some of the statistical inferences that can be drawn from high density arrays, because four separate RNA preparations from each of the IHF<sup>+</sup> and IHF<sup>-</sup> strains were independently hybridized to two separate arrays for each independent experiment. Cyber-T was used to analyze an ed-

ited IHF data set consisting of the 1973 genes for which an above background signal was detected in all eight replicates of the experiment.

We first applied the simple *t* test to the IHF data set and identified 128 genes significant at  $p < 0.01$  (properties of the gene identified are discussed below). We then employed a Bayesian analysis with a sliding window size of 101 and a Bayesian confidence value of 10. The Bayesian analysis resulted in the detection of 68 genes that showed a significant change in expression at  $p < 0.01$  (not corrected for multiple tests). Of these genes the smallest fold change detected (1.9-fold) was larger than the smallest fold change detected by the simple *t* test (1.3-fold), indicating that the Bayesian approach is slightly more conservative. It is interesting to note that genes



**FIG. 2. An example of the graphical output generated by Cyber-T.** The data presented are described in the text. In all cases *experimental* refers to IHF<sup>-</sup> *E. coli* cells and *control* refers to IHF<sup>+</sup> *E. coli* cells. *B–D* are analyses on raw data, whereas the remainder of the panels are analyses carried out on log-transformed data. Log transformations of raw data generally change both average expression level and the standard deviation in expression over replicates. *E* and *F* incorporate the Bayesian approach described in the text to stabilize estimates of the within gene standard deviation, whereas panels *G* and *H* do not incorporate a Bayesian prior. The output is essentially that generated automatically by the Cyber-T program, except the axis labels and figure titles have been edited and/or rescaled.

showing very high fold changes in expression often showed little statistical support associated with their expression change when using the Bayesian approach. For example, one gene (b2734) showed an 8-fold reduction in expression but only had a *p* value of 0.48. Large nonsignificant fold changes in expression often result from a single extreme observation likely to be an experimental artifact. The 128 genes with the smallest *p* values identified using the two methods have 86 genes in common. This indicates that although the two statistical approaches are consistent in identifying a large number of genes, there are also many differences in the set of genes identified (~33%). We have found many instances where the simple *t* test identifies genes as showing significant changes in expression when the fold change is quite modest. Generally these instances are associated with “too small to be true” within treatment variance estimates, reflecting the large sampling variance on the estimate of the within treatment variance when the number of observations is small. Bayesian estimates of the within treatment variation tend to reduce the rate of this source of false positives.

Fig. 2 shows the plots that result from the Bayesian analysis of the IHF data. *A* and *B* show plots of expression level in control *versus* experimental for the raw and log-transformed data. *C–F* show plots of the standard deviation in expression level *versus* its mean over genes for raw and log-transformed data for both the IHF<sup>+</sup> and IHF<sup>-</sup> cells. It can be seen that the variance in expression level is a strong function of the mean.

This is a justification for carrying out statistical analyses on log-transformed data. After a log transformation the relationship between the mean and the standard deviation is somewhat uncoupled. The coefficient of variation, which is equal to the standard deviation divided by the mean expressed as a percentage, is a measure of the relationship between the mean and the standard deviation, and a log transformation is often considered desirable before statistical analyses when the coefficient of variation is large. In the case of the data set of this paper, the coefficient of variation is smaller for every gene for log-transformed (average coefficient of variation = 4.5%) than for untransformed data (average coefficient of variation = 38%). Once the data are log-transformed the standard deviation in gene expression is larger for lowly expressed genes than for highly expressed genes. This relationship makes biological sense because it is difficult to accurately measure and quantify genes showing very low expression levels using high density arrays. Based on our experience with other data sets, the strong relationship between the variance and the mean expression levels are not peculiar to the radioisotope/filter array technology employed here but appear to be a general feature of DNA microarray experiments.

*G* and *H* of Fig. 2 show plots of the standard deviation as a function of the mean on the log-transformed data without incorporating the Bayesian prior. In comparing *G* and *H* with *E* and *F* it can be seen that Bayesian conditioning reduces the large stochastic variance in within treatment estimates of the

TABLE I

Average number of genes of the 120 with the smallest  $p$  values identified in common based on analyzing subsamples of the IHF data set

For the Bayesian analysis the number of commonly identified genes is given only for the best and worst window size and hyperparameter combination.

Statistical approach	2 by 2 comparisons	3 by 3 comparisons
Bayesian of raw data (worst)	59	89
Bayesian of raw data (best)	62	93
$t$ test of raw data	33	80
Ratio of means of raw data	35	51
Bayesian of log data (worst)	63	89
Bayesian of log data (best)	67	91
$t$ test of log data	31	75
Ratio of means of log data	39	74
Difference of means of log data	59	84

standard deviation as evidenced by the spread of the points.

*A Comparison of Statistical Approaches Used to Identify Genes Showing Changes in Gene Expression*—The primary objective of hypothesis testing is to strike a balance between minimizing the likelihood of detecting false positives while maximizing the likelihood of detecting true positives. If high density array data sets existed in which the true positives and false positives were known, then the evaluation of different statistical methods for detecting these genes would be straightforward. Such data sets do not exist. An alternate means of assessing the appropriateness of a statistical method is to measure its consistency over independent realizations of an experiment. In the case of the IHF data set we are able to compare the ability of the different statistical methods to consistently identify the same set of genes as being up- or down-regulated in different subsets of the data. For the IHF data set there are 12 possible IHF<sup>-</sup> versus IHF<sup>+</sup> comparisons of size two replicates and 9 possible comparisons of size three replicates. For each comparison we sorted all genes by the  $p$  value associated with the statistical test (or ratio score in the case of the ratio tests) and examined the top 120 genes. For each comparison we were able to measure the consistency in identifying the same set of genes as being up- or down-regulated over a number of quasi-independent comparisons. In the case of the comparisons of sample size two, 66 measures of consistency were obtained from comparison pairs having no more than two replicates in common. Similarly, in the case of the comparisons of size three, 36 measures of consistency were made from comparisons having one replicate in common.

Table I summarizes the consistency of the different statistical approaches. As might be expected additional replications of an experiment result in greater consistency at identifying the same set of genes as being up- or down-regulated. Log transformations of expression data result in more consistent statistical inferences, except for statistical tests that were already consistent, in which case a log transformation of the data makes little difference. In the case of little experimental replication (the comparisons of size two) both the simple  $t$  test and commonly used ratio tests performed poorly; only identifying the same set of genes 25% of the time. In the case of little replication, Bayesian statistical approaches performed the best, although a modified ratio approach based on the difference of the logs was almost as good as the Bayesian approach. With additional replication, all tests performed better with the exception of ratio tests based on raw data (which performed poorly). The disparity between the poorly performing tests (the  $t$  test and ratio of logs) and the better performing tests (Bayesian and difference of logs) was lessened by increasing replication. It is important to note that even in the best case the average consistency was only about 75%, adding credence to the idea that high density array experiments with modest

levels of replications are still largely an exploratory tool.

Although the consistency of different statistical approaches appears fair with modest levels of replication, it should be noted that the different statistical approaches are not necessarily consistently identifying the same set of genes. Table II lists the number of genes of the 120 co-identified using the seven different statistical approaches, when all four replicates of the IHF data set are considered. It is apparent from this table that the different statistical approaches based on the  $t$  test (*i.e.* the  $t$  test and  $t$  test with a Bayesian prior on the variance with both log-transformed and raw data) converge to a similar set of genes as the amount of replication increases. Statistical approaches based on the  $t$  test have 73% of their genes in common on average. For the  $t$  test-based tests the decision to log transform the data has less of an effect on the number of genes shared than the decision to incorporate a Bayesian prior. Similarly, statistical tests based on ratios of the means (ratio of the means of raw and log-transformed data and the difference of the means of log-transformed data) also converge on a similar set of up- or down-regulated genes, having 77% of their genes in common on average. Of the three ratio-based tests, the two tests based on a log transformation of the data appear to identify more genes in common than the test based on the raw data. It is important to note that the tests based on the  $t$  test and those based on ratios do not necessarily converge on the same set of genes. Comparisons among approaches yield only 54% of their genes in common on average. It seems likely that the tests based on ratios are consistently identifying a set of genes with a large fold change on average but a great deal of variation among within treatment replicates.

*The Bayesian Approach Allows the Identification of More True Positives with Fewer Replicates*—The data in Table I show that additional experimental replications result in the identification of a more consistent set of up- or down-regulated genes and that the Bayesian statistical approach identifies a more consistent set than a simple  $t$  test. The natural question that arises is whether these genes are true positives. That is, whether these are IHF-regulated genes. This question is addressed by the data shown in Fig. 3. Here we define genes whose differential expression is likely to be due to a direct effect of IHF, and therefore a true positive, as those genes that possess a documented or predicted high affinity IHF binding site within 500 base pairs upstream of the ORF for each gene, or the operon containing the gene (see “Materials and Methods”). Of the 120 genes differentially expressed between IHF<sup>+</sup> and IHF<sup>-</sup> strains with the lowest  $p$  values identified by a simple  $t$  test based on two experiments (2 by 2  $t$  test), 51 genes contain an upstream IHF site, whereas the Bayesian analysis of these same two data sets (2 by 2 Bayesian) identifies 59, or 15% more genes with upstream IHF sites. Furthermore, comparison of the differentially expressed genes identified by the 2 by 2  $t$  test or the 2 by 2 Bayesian to the differentially expressed genes identified by a simple  $t$  test performed on four experimental data sets (4 by 4  $t$  test) shows that the 2 by 2 Bayesian analysis also identifies 15% more genes (38 versus 33) in common with the genes identified with the 4 by 4  $t$  test. The same results are observed when the log-transformed data are compared in this way. These data illustrate the utility of the Bayesian approach; replicating an experiment 2-fold and performing a Bayesian analysis is comparable in inference to replicating an experiment 4-fold and using a traditional  $t$  test.

#### DISCUSSION

At present, the significance of high density array data is often judged solely on the basis of observed fold change in expression. An arbitrary fold change threshold is created, with

TABLE II  
The number of genes identified in common by the different statistical approaches for the entire IHF data set

	Bayesian log data	Bayesian raw data	<i>t</i> test log data	<i>t</i> test raw data	Difference in means of log data	Ratio of means of log data
Bayesian raw data	103					
<i>t</i> test log data	77	78				
<i>t</i> test raw data	82	82	102			
Difference in means of log data	58	63	48	57		
Ratio of means of log data	74	78	61	69	104	
Ratio of means of raw data	68	74	58	65	85	87

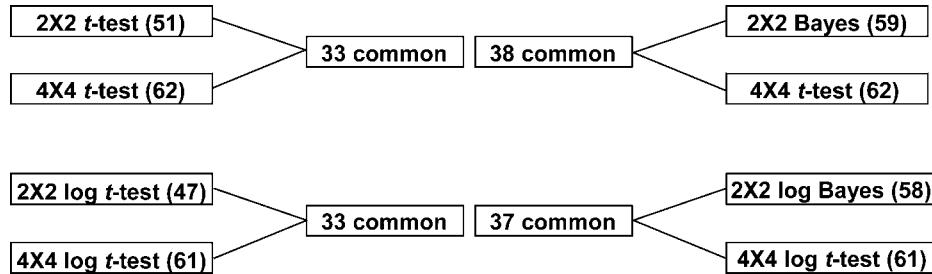


FIG. 3. Analysis of IHF data with and without Bayesian treatment. The numbers in parentheses represent the number of genes of the 120 genes with the lowest *p* values that contain a documented or predicted IHF binding site less than 500 base pairs upstream of each ORF. In each case the raw or log-transformed data from two (2 × 2) or four (4 × 4) independent experiments were analyzed either with or without Bayesian treatment.

genes showing greater change than that threshold declared significant and all others declared nonsignificant. It is apparent from Fig. 2 that fold change in expression may not be a good proxy for statistical significance. If analyses are carried out on nontransformed data, any given threshold of fold change in expression will be liberal for genes expressed at a high level and conservative for genes expressed at a low level. Conversely, if analyses are carried out on log-transformed data, the threshold will be conservative for genes expressed at a high level and liberal for genes expressed at a low level. In cases where either the control or experimental observations are replicated, it is possible to assess the significance of the difference between the control and experimental data relative to the observed level of within class variation. This results in smaller fold changes being significant for genes whose expression levels are measured with great accuracy and large fold changes being nonsignificant for genes whose expression levels cannot be measured very accurately.

Ignoring the variation among replicates or not carrying out replication and determining significance based solely on fold change do not negate the problems discussed above; it amounts to assuming the variance among replicates is equal for all genes. The support for an expression level difference being meaningful relative to the observed variation within treatments is conveniently represented by the *t* statistic (see “Materials and Methods”).

Observed differences between a single replication of a control versus experimental treatment can be due to inadequately controlled experimental factors as opposed to the experimental condition itself. Examples of such variables may include small differences in the time or method of harvesting cells, differences between tissue samples not related to the experiment, and variation induced by the RNA isolation or labeling protocol. For this reason, replicates of high density array experiments are particularly useful. Replication will increase the likelihood of detecting subtle changes in expression between treatments while decreasing the likelihood of false positives. In an ideal world, high density array experiments would be replicated a “large” number of times (*e.g.* >10), and the *t* statistic would measure the relative support for a difference between control and experimental treatments being due to chance alone. In practice, high density array experiments are rarely

replicated this many times, and the sampling variance on the *t* statistic is therefore quite large. In this context, using only the *t* statistic (or a corresponding *p* value) as a measure of whether or not a gene is significant can be misleading. Nonetheless, we have found it useful to sort genes on *p* values as an exploratory tool for identifying potentially interesting genes (5).

A Bayesian approach to estimating the within treatment variation among replicates has been implemented within Cyber-T. The use of the weighted average of the “local” standard deviation for genes with similar expression levels and the observed gene-specific standard deviation stabilizes within treatment variance estimates. Increasing the precision of variance estimates in both the control and experimental treatments results in more stable *t* statistics. This allows inferences to be drawn from high density array experiments that have been carried out with nominal levels of replication. This is demonstrated in Fig. 3 where statistical inference using the Bayesian approach with only two replicates approaches that normally achieved with more replication. There remains a possibility that different genes of similar expression levels have widely differing true variances. Under this possibility and the current prior, poorly replicable genes will be falsely declared significant, and intrinsically highly replicable genes will be falsely declared not significant. Although this would represent an undesirable outcome, when experiments show little replication the relative error in inference introduced by an incorrect prior is likely to be less than the error in inference introduced from very poor estimates of within treatment variance. Ultimately, it will be important to derive empirical guidelines for the determination of the correct hyperparameter to use in weighting the prior information (*i.e.* the local average standard deviation) relative to the observed within treatment estimate of variation. It is possible that the best weighting will depend on factors such as the biological system being studied, experimental conditions employed, and high density array technology used.

Analyses in Cyber-T are performed on both log-transformed and nontransformed data. Log transformations are carried out for three reasons. First, in plots of raw data (Fig. 2A) many of the data points are clustered at the low end of the values. Plots of log-transformed data tend to expand these low values and make them easier to examine visually (Fig. 2B). Second, an assumption of the *t* test is that the variances of the two groups

being tested are equal. Although the  $t$  test is fairly robust with respect to violations of this assumption (especially when the sample sizes of the two groups are equal), if the variances of the two treatments are widely different the statistical test for a difference between means may not be valid. Often, unequal variances between treatments result from the variance in a set of observations scaling with their mean. Log transformations often reduce or eliminate this dependence. It can be seen from  $C$  and  $D$  of Fig. 2 that the variance in raw expression level is a function of the mean, and in this case a log transformation may be appropriate.  $E$  and  $F$  of Fig. 2 show that the dependence of the variance on the mean is somewhat uncoupled following a log transformation. Interestingly, in these plots it appears that the variance in log-transformed expression levels is higher for genes expressed at lower rather than at higher levels. These plots suggest that genes expressed at low or near background levels may be good candidates for ignoring in expression analyses. The variance in the measurement of genes expressed at a low level is large enough that in many cases it is difficult to detect significant changes in expression for this class of loci. Third, statistical tests of log-transformed data have an intuitive appeal. The difference between the log of two numbers raised to the base of the log is equivalent to the ratio of the two numbers (*i.e.*  $a/b = e^{\ln a - \ln b}$ ). Thus a test of the significance of the difference between the log expression levels of two genes is equivalent to a test of whether or not their fold change is significantly different.

We have shown that statistical tests for changes in genes that incorporate the within treatment variance and a Bayesian prior on the estimate of the within treatment variance have a number of desirable properties. They are generally more consistent than tests not employing a Bayesian prior, implying that they give similar results when high density array experiments are replicated. Tests incorporating a measure of experimental error into the test statistic do not identify genes showing large fold changes in expression that also show little correspondence over within treatment replicates. The IHF data presented in Fig. 3 suggest that a Bayesian statistical framework facilitates the identification of more true positives and fewer false positives with fewer replications. A primary deterrent to a more widespread adoption of statistical approaches incorporating a Bayesian prior for the analysis of high density array data is the lack of software that can easily be used to carry out such analyses. We have implemented the approaches described in this work and have created a simple to use web interface that make these tools widely available and accessible.

In summary, Cyber-T provides an easily accessible interface

that allows routine assessment of high density array data for statistical significance. The incorporation of a Bayesian prior into the commonly accepted  $t$  test allows statistical inferences to be drawn from high density array data that is not highly replicated. Although it is often difficult to achieve the levels of statistical significance necessary to satisfy a stringent criterion for experiment-wide significance, the  $p$  values generated in Cyber-T can be used to rank genes and determine those differences most likely to be real.

*Acknowledgments*—Suzanne Sandmeyer and the members of the Functional Genomics group of the University of California at Irvine Institute of Genomics and Bioinformatics provided helpful input and data during the development of the programs described here.

#### REFERENCES

- DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) *Science* **278**, 680–686
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) *Science* **270**, 467–470
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996) *Proc. Natl. Acad. Sci. U. S. A.* **93**, 10614–10619
- Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94**, 13057–13062
- Arfin, S. M., Long, A. D., Ito, E. T., Toller, L., Riehle, M. M., Paegle, E. S., and Hatfield, G. W. (2000) *J. Biol. Chem.* **275**, 29672–29684
- DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996) *Nat. Genet.* **14**, 457–460
- Shalon, D., Smith, S. J., and Brown, P. O. (1996) *Genome Res.* **6**, 639–645
- Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedillon, T., Gilmore, J., Woolley, D. E., and Davis, R. W. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94**, 2150–2155
- Lennon, G. G., and Lehrach, H. (1991) *Trends Genet.* **7**, 314–317
- Gress, T. M., Hoheisel, J. D., Lennon, G. G., Zehetner, G., and Lehrach, H. (1992) *Mamm. Genome* **3**, 609–619
- Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P., and Jordan, B. R. (1995) *Genomics* **29**, 207–216
- Takahashi, N., Hashida, H., Zhao, N., Misumi, Y., and Sakaki, Y. (1995) *Gene (Amst.)* **164**, 219–227
- Zhao, N., Hashida, H., Takahashi, N., Misumi, Y., and Sakaki, Y. (1995) *Gene (Amst.)* **156**, 207–213
- Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Samson, R., Houlgatte, R., Soularue, P., and Auffray, C. (1996) *Genome Res.* **6**, 492–503
- Rovere, P., Trucy, J., Zimmerman, V. S., Granjeaud, S., Rocha, D., Nguyen, C., Ricciardi-Castagnoli, P., Jordan, B. R., and Davoust, J. (1997) *Adv. Exp. Med. Biol.* **417**, 467–473
- Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas D. (1991) *Science* **251**, 767–773
- Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., Lockhart, D. J. (1999) *Nat. Genet.* **21**, 20–24
- Sokal, R. R., and Rohlf, F. J. (1995) *Biometry*, pp. 219–227, W. H. Freeman and Co., New York
- Baldi, P., and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, MA
- Baldi P., and Long A. D. (2001) *Bioinformatics*. **17**, 509–519
- Box, G. E. P., and Tiao, G. C. (1992) *Bayesian Inference in Statistical Analysis*, pp. 92–112, John Wiley & Sons, Inc., New York
- Spector, P. (1994) *An Introduction to S and S-Plus*, pp. 131–134, Duxbury Press, Belmont, CA