Statistica Sinica **12**(2002), 111-139

STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN REPLICATED cDNA MICROARRAY EXPERIMENTS

Sandrine Dudoit[1], Yee Hwa Yang[1], Matthew J. Callow[2] and Terence P. Speed[1,3]

[1] *University of California, Berkeley,* [2] *Lawrence Berkeley National Laboratory and* [3] *The Walter and Eliza Hall Institute*

Abstract: DNA microarrays are a new and promising biotechnology which allows the monitoring of expression levels in cells for thousands of genes simultaneously. The present paper describes statistical methods for the identification of differentially expressed genes in replicated cDNA microarray experiments. Although it is not the main focus of the paper, new methods for the important pre-processing steps of image analysis and normalization are proposed. Given suitably normalized data, the biological question of differential expression is restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and responses or covariates of interest. Differentially expressed genes are identified based on adjusted p-values for a multiple testing procedure which strongly controls the family-wise Type I error rate and takes into account the dependence structure between the gene expression levels. No specific parametric form is assumed for the distribution of the test statistics and a permutation procedure is used to estimate adjusted p-values. Several data displays are suggested for the visual identification of differentially expressed genes and of important features of these genes. The above methods are applied to microarray data from a study of gene expression in the livers of mice with very low HDL cholesterol levels. The genes identified using data from multiple slides are compared to those identified by recently published single-slide methods.

Key words and phrases: Adjusted p-value, differential gene expression, DNA microarray, image analysis, multiple testing, normalization, permutation test.

1. Introduction

DNA microarrays are a new and promising biotechnology which allows the monitoring of expression levels in cells for thousands of genes simultaneously. Microarrays are being applied increasingly in biological and medical research to address a wide range of problems, such as the classification of tumors or the gene expression response of yeast to different environmental stress conditions (Alizadeh et al. (2000), Alon et al. (1999), Gasch et al. (2000), Golub et al. (1999),

Perou et al. (1999), Pollack et al. (1999), Ross et al. (2000)). An important and common question in microarray experiments is the identification of differentially expressed genes, i.e., genes whose expression levels are associated with a response or covariate of interest. The covariates could be either polytomous (e.g., treatment/control status, cell type, drug type) or continuous (e.g., dose of a drug, time), and the responses could be, for example, censored survival times or other clinical outcomes. The types of experiments include *single-slide* cDNA microarray experiments, in which one compares transcript abundance (i.e., gene expression levels) in two mRNA samples, the red and green labeled mRNA samples hybridized to the same slide, and *multiple-slide* experiments comparing transcript abundance in two or more types of mRNA samples hybridized to different slides. Time-course experiments, in which transcript abundance is monitored over time for processes such as the cell cycle, are a special type of multiple-slide experiment which will not be discussed here.

The present paper describes statistical methods for the analysis of gene expression data from multiple-slide cDNA microarray experiments. The experiments which motivated the development of these approaches are part of a study of lipid metabolism aimed at identifying genes with altered expression in the livers of mice with very low HDL cholesterol levels (treatment group) compared to inbred control mice (Callow et al. (2000)). Although it is not the main focus of the paper, new methods for the important pre-processing steps of image analysis and normalization are proposed. Given suitably normalized data, the biological question of differential expression is restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the treatment/control status. As a typical microarray experiment measures expression levels for several thousands of genes simultaneously, we are faced with an extreme multiple testing situation. Special problems arising from the multiplicity aspect include defining an appropriate Type I error rate and devising powerful multiple testing procedures which control this error rate and take into account the *joint* distribution of the test statistics. Our basic approach for identifying differentially expressed genes consists of two steps: (1) computing a test statistic for each gene, and (2) obtaining adjusted p-values for a multiple testing procedure which strongly controls the family-wise Type I error rate and takes into account the dependence structure between the gene expression levels (Westfall and Young (1993)). No specific parametric form is assumed for the distribution of the test statistics and a permutation procedure is used to estimate adjusted p-values. In addition, various

data displays are suggested for the visual identification of differentially expressed genes and of important features of these genes.

The paper is organized as follows. The remainder of this section contains a brief introduction to the biology and technology of cDNA microarrays (Section 1.1) and a discussion of recent proposals for the identification of differentially expressed genes in single- and multiple-slide experiments (Section 1.2). Section 1.3 presents the microarray experiments which motivated the approaches developed in the paper. After a summary of our proposed image analysis and normalization methods, Section 2 describes a multiple testing procedure for identifying differentially expressed genes. Section 3 presents the results of the study and compares the genes identified using multiple slides to those identified by recently published single-slide methods. Finally, Section 4 discusses our findings and outlines open questions.

1.1. Background on cDNA microarrays

The ever-increasing rate at which genomes are being sequenced has opened a new area of genome research, functional genomics, which is concerned with assigning biological function to DNA sequences. With the complete DNA sequences of many genomes already known (e.g., the yeast *S. cerevisae*, the round worm *C. elegans*, the fruit fly *D. melanogaster*, and numerous bacteria) and the recent release of the first draft of the human genome, an essential and formidable task is to define the role of each gene and understand how the genome functions as a whole. Innovative approaches, such as the cDNA and oligonucleotide microarray technologies, have been developed to exploit DNA sequence data and yield information about gene expression levels for entire genomes. Basic genetic notions useful for understanding microarray experiments are reviewed next.

A *gene* consists of a segment of DNA which codes for a particular *protein*, the ultimate expression of the genetic information. A *deoxyribonucleic acid* or *DNA* molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. Each *nucleotide* comprises a phosphate group, a deoxyribose sugar, and one of *four nitrogen bases*. The four different bases found in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). The two chains of the DNA molecule are held together by hydrogen bonds between nitrogen bases, with base-pairing occurring according to the following rule: G pairs with C, and A pairs with T. While a DNA molecule is built from a four-letter alphabet, proteins are sequences of twenty different types of *amino acids*. The expression of the genetic information stored in the DNA molecule occurs in two stages: (i) *transcription*, during which DNA is transcribed into *messenger ribonucleic*

acid or *mRNA*, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine; (ii) *translation*, during which mRNA is translated to produce a protein. The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the *genetic code*, which relates nucleotide triplets to amino acids.

Different aspects of gene expression can be studied using microarrays, such as expression at the transcription or translation level, and subcellular localization of gene products. To date, attention has focused primarily on expression at the transcription stage, i.e., on mRNA or transcript levels. Microarrays derive their power and universality from a key property of DNA molecules described above, *complementary base-pairing*, and the term *hybridization* is used to refer to the annealing of nucleic acid strands from different sources according to the base-pairing rules. There are several types of microarray systems, including the cDNA microarrays developed in the Brown and Botstein labs at Stanford (DeRisi et al. (1997), Hughes et al. (2001)) and the high-density oligonucleotide chips from the Affymetrix company (Lockhart et al. (1996)); the brief description below focuses on the former.

cDNA microarrays consist of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic *arrayer*. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples or *targets* are reverse-transcribed into cDNA, labeled using different fluorescent dyes (usually a red-fluorescent dye, Cyanine 5 or Cy5, and a green-fluorescent dye, Cyanine 3 or Cy3), then mixed in equal proportions and hybridized with the arrayed DNA sequences or *probes* (following the definition of probe and target adopted in *The Chipping Forecast* (1999)). After this competitive hybridization, the slides are imaged using a *scanner* and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples. The diagram in Figure 1 describes the main steps in a cDNA microarray experiment; see *The Chipping Forecast* (1999) for a more detailed introduction to the biology and technology of cDNA microarrays and oligonucleotide chips.

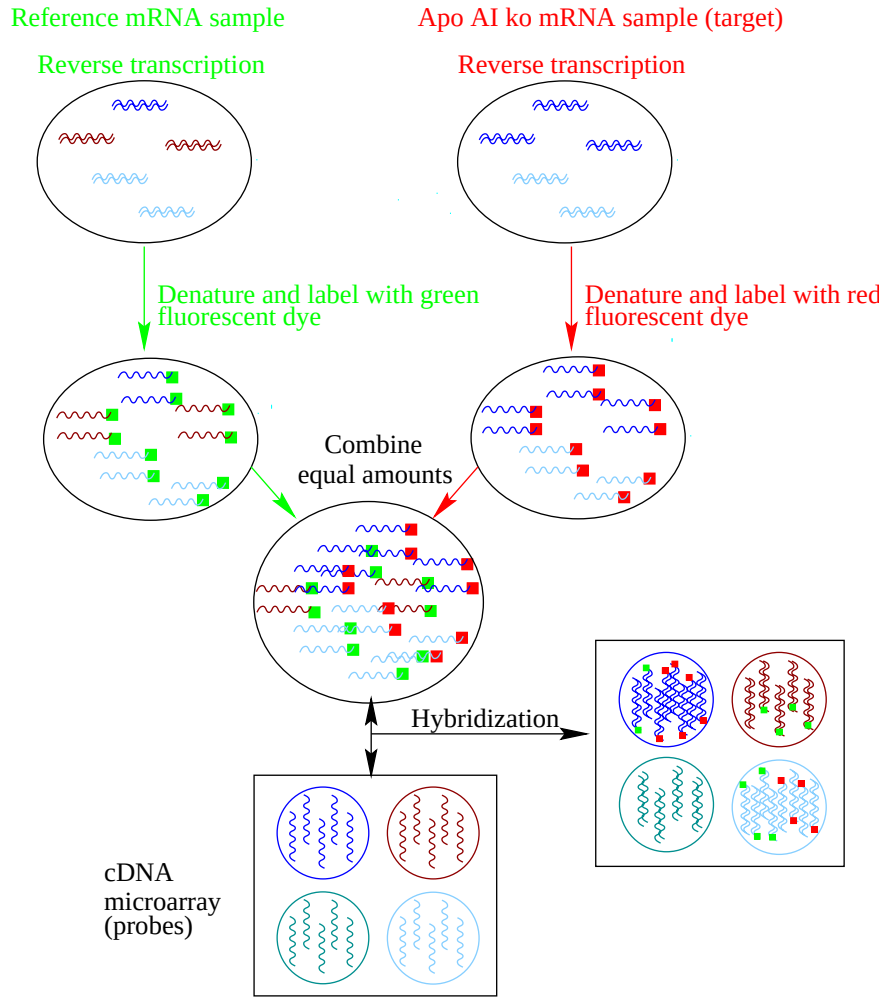


Figure 1. cDNA microarray experiment for apo AI knock-out mice. For each apo AI knock-out mouse, target cDNA is obtained from liver mRNA by reverse transcription and labeled using a red-fluorescent dye (Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations is prepared by pooling cDNA from the 8 C57Bl/6 control mice. The two target samples are mixed and hybridized to a microarray containing 6,384 spots. Following the competitive hybridization, the slides are imaged using a scanner and fluorescence intensity measurements are made separately for each dye at each spot on the array.

1.2. Identification of differentially expressed genes

1.2.1. Single-slide methods

A number of methods have been suggested for the identification of differentially expressed genes in single-slide cDNA microarray experiments. In such

experiments, the data for each gene (spot) consist of two fluorescence intensity measurements, (R, G), representing the expression level of the gene in the red (Cy5) and green (Cy3) labeled mRNA samples, respectively (the most commonly used dyes are the cyanine dyes, Cy3 and Cy5, however, other dyes such as fluorescein and X-rhodamine may be used as well). We distinguish two main types of single-slide methods: those which are based solely on the value of the intensity ratio R/G and those which also take into account overall transcript abundance measured by the product RG. Early analyses of microarray data (DeRisi et al. (1996), Schena et al. (1995), Schena et al. (1996)) relied on fold increase/decrease cut-offs to identify differentially expressed genes. For example, in their study of gene expression in the model plant *Arabidopsis thaliana*, Schena et al. (1995) use spiked controls in the mRNA samples to normalize the signals for the two fluorescent dyes (there, fluorescein and lissamine) and declare a gene differentially expressed if its expression level differs by more than a factor of 5 in the two mRNA samples. DeRisi et al. (1996) identify differentially expressed genes using a ± 3 cut-off for the log-ratios of the fluorescence intensities, standardized with respect to the mean and standard deviation of the log-ratios for a panel of 90 "housekeeping" genes (i.e., genes believed not to be differentially expressed between the two cell types of interest). More recent approaches are based on probabilistic modeling of the (R, G) pairs and differ mainly in the distributional assumptions they make for (R, G) in order to derive a rule for deciding whether a particular gene is differentially expressed. Chen, Dougherty and Bittner (1997) propose a data dependent rule for choosing cut-offs for the red and green intensity ratio R/G. The rule is based on a number of distributional assumptions for the intensities (R, G), including normality and constant coefficient of variation. Sapir and Churchill (2000) suggest identifying differentially expressed genes using posterior probabilities of change under a mixture model for the intensity log-ratio $\log R/G$ (after a form of background correction, the orthogonal residuals from the robust regression of $\log R$ *vs.* $\log G$ are essentially normalized log-ratios). A limitation of these two methods is that they both ignore the information contained in the product RG. Recognizing this problem, Newton, Kendziorski, Richmond, Blattner and Tsui (2001) consider a hierarchical model (Gamma-Gamma-Bernoulli model) for (R, G) and suggest identifying differentially expressed genes based on the posterior odds of change under this hierarchical model. The odds are functions of $R + G$ and RG, and thus produce a rule which takes into account overall transcript abundance. The approach of Roberts et al. (2000) is based on assuming that R and G are approximately independently and normally distributed, with variance depending on the mean, and also produces a rule which takes into account overall transcript abundance.

At the end of the day, each of these methods produces a model dependent rule which amounts to drawing two curves in the $(R, G)-$plane and calling a gene differentially expressed if its (R, G) measured intensities fall outside the region between the two curves. The relative merits of the procedures depend on their ability to successfully identify differentially expressed genes (i.e., their power or one minus their Type II error rate), while avoiding to call unchanged genes differentially expressed (i.e., their false positive or Type I error rate). For any given single-slide experiment, thousands of comparisons are made, raising the issue of multiple testing. Finally, and most importantly, the gene expression data may be too noisy for successful identification of differentially expressed genes without replication, no matter how good the rule.

Note that the fluorescence intensity pairs (R, G) are already highly processed data and the choice of image analysis methods for segmentation and background correction of the laser scanned images can have a large impact on these quantities. Before applying any of the above single-slide methods, or for that matter any inference or cluster analysis method, it is essential to identify and remove sources of systematic variation (e.g., different labeling efficiencies and scanning properties of the Cy3 and Cy5 dyes, print-tip or spatial effects) by an appropriate normalization method. Until these systematic effects are properly accounted for, there can be no question of the system being in statistical control and hence no basis for a statistical model to describe chance variation.

1.2.2. Multiple-slide methods

Until recently, statistical methods for identifying differentially expressed genes in multiple-slide experiments have received relatively little attention. A common approach has been to rely on exploratory cluster analysis methods (hierarchical clustering or partitioning methods such as self-organizing maps) to group genes with correlated expression profiles across experimental conditions (Alizadeh et al. (2000), Ross et al. (2000)). Groups of differentially expressed genes are then identified by visual inspection of the resulting clusters, using, for example, red and green images to display the intensity log-ratios for each gene in each of the slides (Eisen et al. (1998)). Such methods are "unsupervised", in that they ignore the covariates or responses for the samples hybridized to the slides (e.g., treatment or control status of the mice). A more direct and appropriate approach to the problem of differential expression is to exploit this available information by, for example, computing for each gene a test statistic relating its expression levels to the covariates or responses (e.g., t-statistic) and ranking the genes according to this statistic (Galitski et al. (1999), Golub et al. (1999)). Kerr, Martin and Churchill (2000) take this approach and stress

the importance of replication in order to assess the variability of estimates of change. They suggest applying techniques from the analysis of variance and assume a fixed effect linear model for the logged intensities, with terms accounting for dye, slide, treatment, and gene main effects, as well as a few interactions between these effects. Differentially expressed genes are then identified based on contrasts for the treatment $\times$ gene interactions (these contrasts are related to averages of intensity log-ratios).

1.3. Apo AI and SR-BI experiments

The goal of the study is to identify genes with altered expression in the livers of two lines of mice with very low HDL cholesterol levels compared to inbred control mice (Callow et al. (2000)). The two mouse models are the apolipoprotein AI (apo AI) knock-out and the scavenger receptor BI (SR-BI) transgenic mice; apo AI and SR-BI are two genes known to play pivotal roles in HDL metabolism.

In the first experiment, the treatment group consists of eight mice with the apo AI gene knocked-out and the control group consists of eight "normal" C57Bl/6 mice. For each of these 16 mice, target cDNA is obtained from mRNA by reverse transcription and labeled using a red-fluorescent dye (Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations was prepared by pooling cDNA from the eight control mice. The design for the second experiment is similar, but with eight SR-BI transgenic mice comprising the treatment group and eight "normal" FVB mice comprising the control group. In each experiment, target cDNA is hybridized to microarrays consisting of $6,384$ spots, which include 257 genes thought to be related to lipid metabolism. The microarrays were printed using 4×4 print-tips and are thus partitioned into a 4×4 grid matrix. Each grid contains 19×21 spots that were printed with the same print-tip. (cDNA microarrays are spotted using different printing set-ups, such as 4×4 or 4×8 print-tip clusters. The arrays are divided into grids (also called sectors) and the spots on a given grid are printed with the same print-tip or pin. We say that spots printed using the same print-tip are part of the same print-tip group.) Note that the spotted cDNA sequences are usually referred to as "genes", whether they are actual genes, ESTs (expressed sequence tags), or DNA sequences from other sources. The raw data from each of these two experiments consist of 16 pairs of image files, one red and green image pair for each of the slides.

2. Methods

2.1. Image analysis

The red and green fluorescence intensities (R, G), which are inputs to the methods described in Section 1.2, are already highly processed data. We view

the image files produced by the scanner as the "raw" data; these are typically pairs of 16-bit tagged image file format (TIFF) files, one for each fluorescent dye. Image analysis is required to extract foreground and background fluorescence intensity measurements for each spotted DNA sequence. We have developed new addressing, segmentation, and background correction methods for extracting information from microarray scanned images. The addressing method uses the fact that microarrays are generally produced in *batches* and that, within a batch, important characteristics, particularly the print-tip configuration, are very nearly the same. The segmentation component is based on the *seeded region growing algorithm* of Adams and Bischof (1994) and places no restriction on the size or shape of the spots. The background adjustment method relies on a non-linear filter known as *morphological opening* to generate an image of the estimated background intensity for the entire slide. These new image analysis procedures are implemented in a software package named `Spot`, built on the `R` environment for statistical computing (Buckley (2000), Ihaka and Gentleman (1996)). A detailed discussion of the proposed image analysis methods and a comparison to popular alternatives can be found in Yang, Buckley, Dudoit and Speed (2001a). Thus, starting with two images for each slide, the image processing steps outlined above produce two main quantities for each spot on the array: the red and green fluorescence intensities, R and G, which are measures of transcript abundance for the red and green labeled mRNA samples, respectively.

2.2. Single-slide data displays

Single-slide expression data are typically displayed by plotting the log intensity $\log_2 R$ in the red channel vs. the log intensity $\log_2 G$ in the green channel (Newton et al. (2001), Sapir and Churchill (2000), Schena (2000)). (It is preferable to work with logged intensities rather than absolute intensities for a number of reasons, including the facts that: (i) the variation of logged intensities and ratios of intensities is less dependent on absolute magnitude; (ii) normalization is usually additive for logged intensities; (iii) taking logs evens out highly skewed distributions; and (iv) taking logs gives a more realistic sense of variation. Logarithms base 2 are used instead of natural or decimal logarithms as intensities are typically integers between 0 and $2^{16} - 1$.) We find that such $\log_2 R$ vs. $\log_2 G$ plots give an unrealistic sense of concordance between the red and green intensities and can mask interesting features of the data. We thus prefer to plot the intensity log-ratio $M = \log_2 R/G$ vs. the mean log intensity $A = \log_2 \sqrt{RG}$ (a similar display was used in Roberts et al. (2000)). An $MA-$plot amounts to a 45^o counterclockwise rotation of the $(\log_2 G, \log_2 R)-$ coordinate system, followed by scaling of the coordinates, and is thus another representation of the (R, G) data

in terms of the log-ratios M which are the quantities of interest to most investigators. We have found $MA-$plots to be more revealing than their $\log_2 R$ vs. $\log_2 G$ counterparts in terms of identifying spot artifacts and for normalization purposes. Figure 2 displays a $\log_2 R$ vs. $\log_2 G$ plot and an $MA-$plot for a simple self-self microarray experiment in which two identical mRNA samples were labeled with different dyes and hybridized to the same slide.

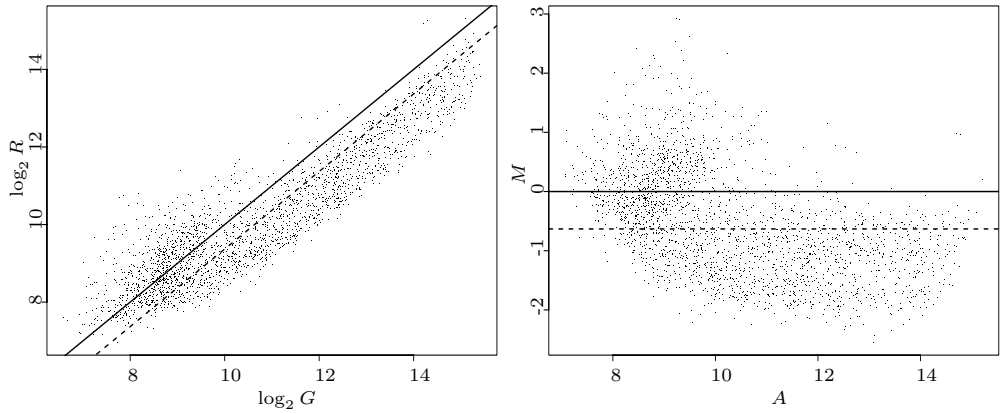


Figure 2. *Self-self hybridization.* $\log_2 R$ vs. $\log_2 G$ plot and $MA-$plot for self-self hybridization. The $M = 0$ (solid) and $M = \text{median}\, M_j$ (dashed) lines are drawn for reference.

2.3. Normalization

The purpose of normalization is to identify and remove sources of systematic variation, other than differential expression, in the measured fluorescence intensities (e.g., different labeling efficiencies and scanning properties of the Cy3 and Cy5 dyes; different scanning parameters, such as PMT settings; print-tip, spatial, or plate effects). It is necessary to normalize the fluorescence intensities before any analysis which involves comparing expression levels within or between slides (e.g., clustering, discriminant analysis). The need for normalization can be seen most clearly in self-self experiments (Figure 2). Although there is no differential expression and one expects the red and green intensities to be equal, the red intensities often tend to be lower than the green intensities. Furthermore, the imbalance in the red and green intensities is usually not constant across the spots within and between arrays, and can vary according to overall spot intensity A, location on the array, plate origin, and possibly other variables.

The simplest approach to within-slide normalization is to subtract a constant from all intensity log-ratios, typically their mean or median. Such global normalization methods are still the most widely used in spite of the evidence of spatial or intensity dependent dye biases in numerous experiments. We favor more flexible normalization methods which allow the normalization function to depend on a number of predictor variables, such as spot intensity A, location, and plate origin. The normalization function can be obtained by *robust locally weighted regression* of the log-ratios M on the predictor variables. For the apo AI and SR-BI experiments, the print-tip group was used as a proxy for the location of the spots on the slide. (Systematic differences may exist between the print-tips, such as differences in length or in the opening of the tip. There may also be spatial effects due, for example, to the placement of the cover-slip. Note that it may not always be possible to separate print-tip effects from spatial effects.) Within print-tip group intensity dependent normalization was performed using the lowess scatterplot smoother implemented in the `lowess` function from the `R` software package (Cleveland (1979), Ihaka and Gentleman (1996)): $\log_2 R/G \leftarrow \log_2 R/G - l(A, j)$, where $l(A, j)$ is the lowess fit to the $MA-$plot for spots printed using the jth print-tip (i.e., data from the jth grid only), $j = 1, \ldots, 16$. The parameter f, specifying the fraction of the data used for smoothing at each point, was set between 20 and 40%. For the experiments considered here, only a small proportion of the genes were expected to vary in expression between the red and green labeled mRNA samples; normalization was thus performed using all $6,384$ probes. In other circumstances, a subset of control sequences may be spotted on the slide and used for normalization purposes (see Yang et al. (2001) for greater detail on normalization).

2.4. Graphical displays for test statistics

2.4.1. Test statistics

For the purpose of identifying differentially expressed genes in the treatment and control mice, the normalized gene expression data can be summarized by a matrix X of intensity log-ratios $M = \log_2 R/G$, with m rows corresponding to the genes being studied and $n = n_1 + n_2$ columns corresponding to the n_1 control hybridizations (C57Bl/6 or FVB) and n_2 treatment hybridizations (apo AI knock-out or SR-BI transgenic). The fluorescence intensity log-ratio x_{ji} thus represents the expression response of gene j in either a control or treatment mouse. In the two experiments considered here $n_1 = n_2 = 8$ and $m = 6,384$.

Let H_j denote the null hypothesis of no association between the expression level of gene j and the treatment, $j = 1, \ldots, m$. Only two-sided alternative

hypotheses are considered here; one-sided alternatives can be handled in a similar manner. Differentially expressed genes are identified by computing a two-sample *Welch t-statistic* for each gene j; the random variable and realization of the t-statistic for gene j are denoted by T_j and t_j, respectively. Large absolute t-statistics suggest that the corresponding genes have different expression levels in the control and treatment groups. Note that replication is essential for such an analysis, as it is required for assessing the variability of the gene expression levels in the treatment and control groups. Also note that the t-statistics are not assumed to actually follow a t-distribution; a permutation procedure is used to estimate their joint distribution (see Section 2.5.2).

2.4.2. Quantile-Quantile plots

Quantile-Quantile plots or *Q-Q plots* are a useful display of the test statistics for the thousands of genes being studied in a typical microarray experiment. In general, Q-Q plots are used to assess whether data have a particular distribution or whether two datasets have the same distribution. In our application, we are not so much interested in testing whether the test statistics follow a particular distribution, but in using the Q-Q plot as a visual aid for identifying genes with "unusual" test statistics. Q-Q plots informally correct for the large number of comparisons and the points which deviate markedly from an otherwise linear relationship are likely to correspond to those genes whose expression levels differ between the control and treatment groups. In a normal Q-Q plot, the ordered test statistics are plotted against the quantiles of a standard normal distribution. Alternately, Q-Q plots may be obtained by plotting the ordered test statistics against quantiles estimated from the permutation distribution of these test statistics. For the microarray datasets we have encountered so far, the normal and permutation Q-Q plots were virtually identical.

2.4.3. Plots vs. absolute expression levels

Important features of the genes with large absolute t-statistics can be identified by examining plots of the t-statistics, their numerators and denominators, against absolute expression levels. The absolute expression level for a particular gene is measured by $\bar{A}$, the average of $A = \log_2 \sqrt{RG}$ over the 16 hybridizations for the apo AI or SR-BI experiments.

2.5. Multiple hypothesis testing

2.5.1. Adjusted p-values

The Q-Q plots for the t-statistics are useful visual aids for identifying genes with altered expression in the treatment mice compared to the control mice. A

more precise assessment of the evidence against the null hypothesis of no differential expression may be obtained by calculating a p-value, $p_j = pr(|T_j| \geq |t_j| \mid \mathrm{H}_j)$, for each gene j, $j = 1, \ldots, m$. However, with a typical microarray dataset comprising thousands of genes, an immediate concern is multiple testing. When many hypotheses are tested, as is the case here, the probability that at least one Type I error is committed can increase sharply with the number of hypotheses. Numerous methods have been suggested for controlling the *family-wise Type I error rate (FWER)*, i.e., the probability of at least one Type I error in the family (see Shaffer (1995) for a review of such methods). Some procedures provide *strong control* of the FWER, i.e., control this error rate for any combination of true and false hypotheses, while others provide only *weak control* of the FWER, i.e., control the FWER only under the *complete null hypothesis* $\mathrm{H}_0^C = \cap_{j=1}^m \mathrm{H}_j$ that all hypotheses in the family are true. The procedures described below provide strong control of the FWER.

The concept of p-value can be extended to multiple testing procedures. Given any test procedure, the *adjusted p-value* corresponding to the test of a single hypothesis H_j can be defined as the level of the entire test procedure at which H_j would just be rejected, given the values of all test statistics involved (Shaffer (1995), Westfall and Young (1993)). If interest is in controlling the FWER, the adjusted p-value for hypothesis H_j is $\tilde{p}_j = \inf\{\alpha : \mathrm{H}_j \text{ is rejected at } FWER = \alpha\}$, and hypothesis H_j is rejected at FWER α if $\tilde{p}_j \leq \alpha$. There are several approaches to p-value adjustment and these vary in the severity of the correction for multiplicity.

The Bonferroni procedure is perhaps the best known method for dealing with multiple testing. The *Bonferroni single-step adjusted p-values* are given by $\tilde{p}_j = \min(mp_j, 1)$, $j = 1, \ldots, m$. Closely related to the Bonferroni method is the Šidák procedure which is exact for protecting the FWER when the unadjusted p-values are independently distributed as $U[0, 1]$ under the complete null. The *Šidák single-step adjusted p-values* are given by $\tilde{p}_j = 1 - (1 - p_j)^m$. These two procedures are called *single-step* because they perform equivalent multiplicity adjustments for all hypotheses, regardless of the ordering of the unadjusted p-values. While single-step adjusted p-values are simple to calculate, they tend to be conservative. Improvement in power, while preserving strong control of the FWER, may be achieved by considering *step-down* procedures which order p-values and make successively smaller adjustments. Let $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$ denote the *ordered unadjusted p-values*. The *Holm step-down adjusted p-values* are given by

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \left\{ \min((m - k + 1)\, p_{r_k}, 1) \right\}. \tag{1}$$

Holm's procedure is less conservative than the standard Bonferroni procedure which would multiply the unadjusted p-values by m at each step. Note that taking successive maxima of the quantities $\min((m-k+1)\,p_{r_k}, 1)$ enforces monotonicity of the adjusted p-values. That is, $\tilde{p}_{r_1} \leq \tilde{p}_{r_2} \leq \cdots \leq \tilde{p}_{r_m}$, and one can only reject a particular hypothesis provided all hypotheses with smaller unadjusted p-values were rejected beforehand. However, neither Holm's method nor the single-step methods presented above take into account the dependence structure between the test statistics. In microarray experiments (and many other situations), the test statistics and hence the p-values are correlated due, for example, to co-regulation of the genes. Westfall and Young (1993) propose adjusted p-values for less conservative multiple testing procedures which take into account the dependence structure between the test statistics. The Westfall and Young (1993) *step-down minP adjusted p-values* are defined by

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \left\{ pr\Big(\min_{l \in \{r_k,\ldots,r_m\}} P_l \leq p_{r_k} \mid \mathrm{H}_0^C \Big) \right\}, \tag{2}$$

where H_0^C denotes the complete null hypothesis and P_l the random variable for the unadjusted p-value of the lth hypothesis H_l. Alternately, one may consider procedures based on the *step-down maxT adjusted p-values* which are defined in terms of the test statistics T_j themselves as

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \left\{ pr\Big(\max_{l \in \{r_k,\ldots,r_m\}} |T_l| \geq |t_{r_k}| \mid \mathrm{H}_0^C \Big) \right\}, \tag{3}$$

where $|t_{r_1}| \geq |t_{r_2}| \geq \cdots \geq |t_{r_m}|$ denote the ordered observed test statistics. Note that computing the quantities in (2) using the upper bound provided by Boole's inequality yields Holm's p-values. Procedures based on the step-down minP adjusted p-values are thus less conservative than Holm's procedure. The maxT p-values are easier to compute than the minP p-values and are equal to the minP p-values when the test statistics T_j are identically distributed. However, the two procedures generally produce different adjusted p-values, and considerations of balance, power, and computational feasibility should dictate the choice between the two approaches. In the case of non-identically distributed test statistics T_j (e.g., t-statistics with different degrees of freedom), not all tests contribute equally to the maxT adjusted p-values and this can lead to unbalanced adjustments (Westfall and Young (1993, p.50)). When adjusted p-values are estimated by permutation (Section 2.5.2), the minP p-values require more computations than the maxT p-values, because the unadjusted p-values must be estimated before considering the distribution of their successive minima.

2.5.2. Estimation of adjusted p-values by permutation

In many situations, the joint (and marginal) distribution of the test statistics is unknown. Resampling methods (bootstrap or permutation) can be used to estimate unadjusted and adjusted p-values while avoiding parametric assumptions about the joint distribution of the test statistics. In the microarray setting, the joint distribution under the complete null hypothesis of the test statistics $T_1, \ldots, T_m$ can be estimated by permuting the columns of the gene expression data matrix X. Permuting entire columns of this matrix creates a situation in which membership to the control or treatment group is independent of gene expression, while attempting to preserve the dependence structure between the genes. When computationally feasible, all possible permutations of the columns are considered, otherwise, a random subset of B permutations (including the observed) may be considered. For the knock-out and transgenic mouse datasets, there are $B = \binom{16}{8} = 12,870$ possible permutations of the treatment/control labels and p-values are estimated using the full set of permutations.

Box 1. Permutation algorithm for unadjusted p-values

For the bth permutation, $b = 1, \ldots, B$,

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.

The permutation distribution of the test statistic T_j for hypothesis H_j, $j = 1, \ldots, m$, is given by the empirical distribution of $t_{j,1}, \ldots, t_{j,B}$. For two-sided alternative hypotheses, the permutation p-value for hypothesis H_j is

$$p_j^* = \frac{\sum_{b=1}^{B} I(|t_{j,b}| \geq |t_j|)}{B},$$

where $I(\cdot)$ is the indicator function, equaling 1 if the condition in parentheses is true, and 0 otherwise.

Permutation adjusted p-values for the Bonferroni, Šidák, and Holm procedures can be obtained by replacing p_j by p_j^* in the equations defining $\tilde{p}_j$. However, for the Westfall and Young minP adjusted p-values, the *joint* null distribution of $P_1, \ldots, P_m$ needs to be estimated. When the unadjusted p-values themselves are unknown, additional resampling for estimating these p-values can be computationally intensive. For ease of computation, a multiple testing procedure based on the maxT adjusted p-values is used to identify differentially expressed genes in the apo AI and SR-BI experiments.

Box 2. Permutation algorithm for step-down maxT adjusted p-values - based on Algorithms 2.8 and 4.1 in Westfall and Young (1993)

For the bth permutation, $b = 1, \ldots, B$,

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.
3. Next, compute successive maxima of the test statistics

$$u_{m,b} = |t_{r_m,b}|,$$
$$u_{j,b} = \max\Big(u_{j+1,b}, |t_{r_j,b}|\Big) \qquad \text{for } j = m-1, \ldots, 1,$$

where r_j denotes the ordering of the *observed* test statistics such that $|t_{r_1}| \geq |t_{r_2}| \geq \cdots \geq |t_{r_m}|$.

The adjusted p-values are estimated by

$$\tilde{p}^*_{r_j} = \frac{\sum_{b=1}^{B} I(u_{j,b} \geq |t_{r_j}|)}{B},$$

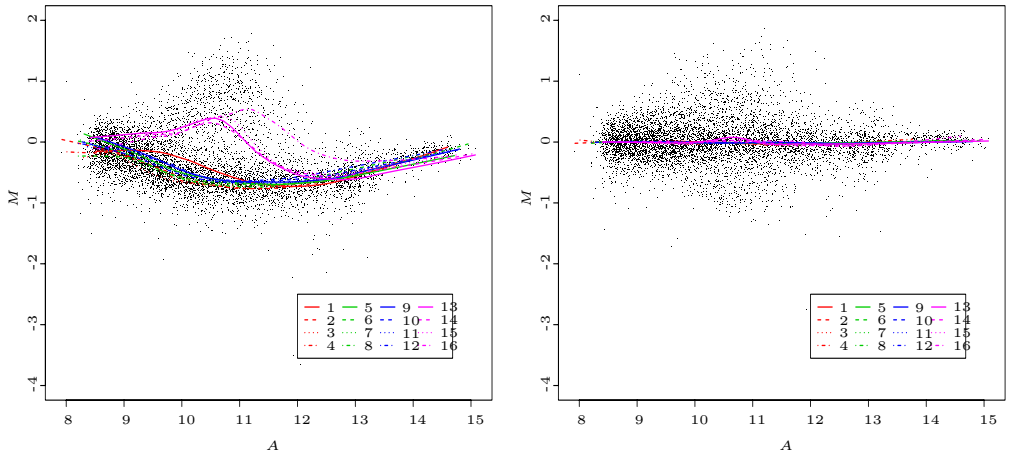
with the monotonicity constraints enforced by setting

$$\tilde{p}^*_{r_1} \leftarrow \tilde{p}^*_{r_1}, \qquad \tilde{p}^*_{r_j} \leftarrow \max(\tilde{p}^*_{r_j}, \tilde{p}^*_{r_{j-1}}) \qquad \text{for } j = 2, \ldots, m.$$

3. Results

3.1. Normalization

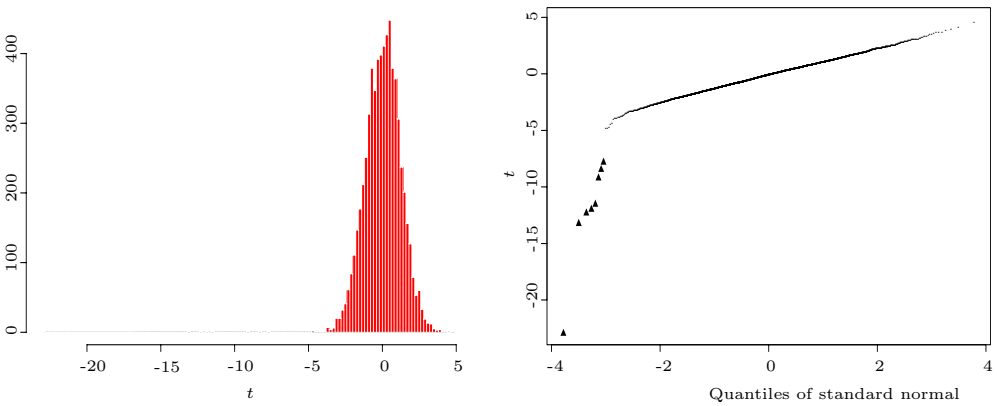
Figure 3 displays the $MA-$plot for a single slide from the apo AI experiment before (panel (a)) and after (panel (b)) within print-tip group intensity dependent normalization. Panel (a) illustrates the non-linear dependence of the log-ratio M on the overall spot intensity A and suggests that an intensity or A-dependent normalization method is preferable to a global one. Also, for the apo AI arrays four print-tip group lowess curves clearly stand out from the remaining twelve curves, suggesting strong print-tip or spatial effects. The four curves correspond to the last row in the 4×4 print-tip cluster, i.e., to print-tips 13, 14, 15, and 16. This pattern is visible in the images, where the bottom four grids tend to have high red signal. The normalized log-ratios in the $MA-$plot of panel (b) are evenly distributed about zero across the range of intensities. Similar effects were observed for the other apo AI slides and for the SR-BI experiment (see web supplement http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html).



(a) (b)

Figure 3. *Apo AI.* (a) $MA-$plot illustrating the need for within print-tip group intensity dependent normalization. (b) $MA-$plot after within print-tip group intensity dependent normalization. Both panels display the lowess curves ($f = 40\%$) for each of the 16 print-tips (data from apo AI knock-out mouse #8). Different colors are used to represent lowess curves for print-tips from different rows and different line types are used to represent lowess curves for print-tips from different columns.

3.2. Identification of differentially expressed genes with replicated slides



(a) (b)

Figure 4. *Apo AI.* Histogram and normal Q-Q plot for two-sample t-statistics. The points corresponding to genes with maxT adjusted p-values less than 0.01 are indicated by filled triangles.

Q-Q plots. For the apo AI experiment, the Q-Q plot in Figure 4 indicates that eight genes (filled triangles) have t-statistics that deviate markedly from an

otherwise linear relationship. All eight genes have negative t-statistics, suggesting down-regulation in the knock-out mice compared to the controls. For the SR-BI experiment, the deviations from linearity are more subtle and gradual (see web supplement for figure). There are about a dozen genes with "unusual" positive and negative t-statistics and these seem like possible candidates for differential expression. In order to determine whether the extreme t-statistics do indeed reflect significant differences between the control and transgenic or knock-out mice we turn to p-value adjustment procedures.

Adjusted p-values. For the apo AI experiment, Figure 5 displays a plot of the unadjusted p-values and Westfall and Young maxT adjusted p-values for the 50 genes with the largest absolute t-statistic. As expected, adjusted p-values are much larger than the corresponding unadjusted p-values. For this experiment, eight genes have very small ($\tilde{p}^* \leq 0.01$) adjusted p-values and the remaining genes have markedly higher p-values ($\tilde{p}^* \geq 0.60$). In the SR-B1 experiment, 13 genes have adjusted p-values lower than 5% and the increase in p-value is much more gradual than in the apo AI knock-out experiment (see web supplement). Thus, adjusted p-values reflect the patterns seen in the Q-Q plots, while unadjusted p-values are, as expected, much too small and lack the specificity of adjusted p-values.

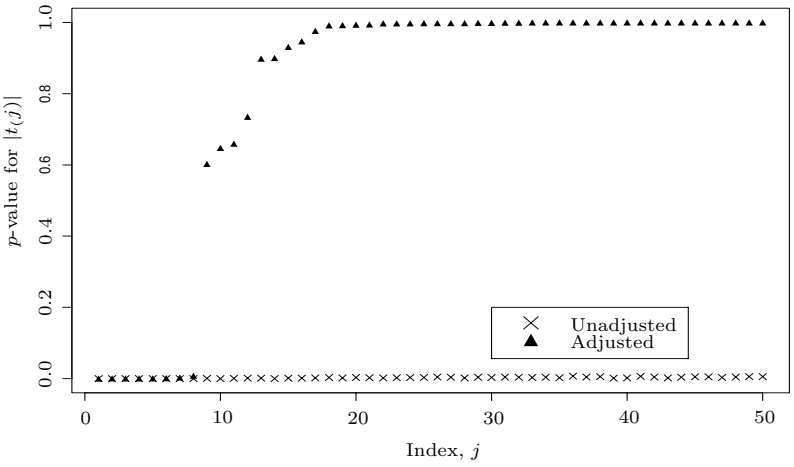


Figure 5. *Apo AI.* Westfall and Young maxT adjusted p-values (filled triangles) and unadjusted p-values (crosses) for the 50 genes with the largest absolute t-statistic.

Features of differentially expressed genes. Important features of the genes with large absolute t-statistics can be identified by examining plots of the numerator and denominator of the t-statistics against absolute expression levels

(see Figure 6 and web supplement for SR-BI). For both experiments, the genes with large absolute t-statistics tended to have high absolute expression levels, as measured by $\bar{A}$. They typically had large differences in their expression levels between the two groups (numerator) as well as fairly low standard errors (SEs in denominator).

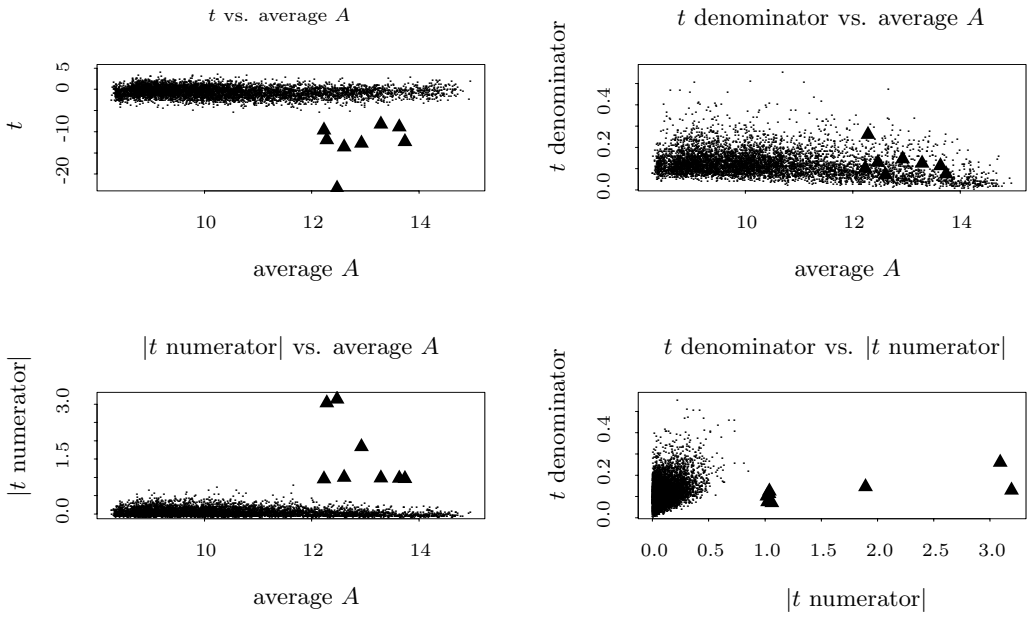


Figure 6. *Apo AI.* Plots of t-statistics, numerator, and denominator, against overall intensity $\bar{A}$. The points corresponding to genes with maxT adjusted p-values less than 0.01 are highlighted with filled triangle.

Identity of differentially expressed genes. Many of the spotted cDNA probes with large absolute t-statistics were re-sequenced because of the known possibility of mixed populations of clones, chimeric clones, or errors in plate arraying of the bacterial clones. As a result, several of the probes in Tables 1 and 2 were found to correspond to the same gene after re-sequencing.

For the apo AI knock-out experiment, apo AI appeared three times and apo CIII, a gene physically very close to apo AI and also associated with lipoprotein metabolism, appeared twice. Sterol C5 desaturase, an enzyme involved in the later stages of cholesterol synthesis, also appeared twice.

As expected, SR-B1 was the most significantly altered gene in the SR-B1 transgenic experiment, as the two probes with the smallest adjusted p-values corresponded to this gene. Glutathione s-transferase and Cytochrome p450 2B10

both appeared twice along with the hemoglobin alpha and beta chains. Although there is no obvious link between the latter two genes and cholesterol metabolism, the known functions of these genes may suggest altered oxidative and steroid metabolism associated with over-expression of SR-B1. SR-B1 is believed to not only facilitate the uptake of cholesterol by cells but also other molecules such as phospholipids. Several other genes were identified but have not yet been confirmed by re-sequencing.

In an alternative method of analysis, expression levels of some of the genes in Tables 1 and 2 were quantitated by RT-PCR (real-time quantitative polymerase

Table 1. *Apo AI.* Genes with maxT adjusted p-values ≤ 0.01. For each gene, the table lists the gene name, the permutation adjusted p-value ($\tilde{p}^*$), the two-sample t-statistic (t), the numerator (Num) and denominator (Den) of the t-statistic.

Gene name	$\tilde{p}^*$	t	Num	Den
Apo AI	0.00	-22.85	-3.19	0.14
Sterol C5 desaturase	0.00	-13.14	-1.06	0.08
Apo AI	0.00	-12.21	-1.90	0.16
Apo CIII	0.00	-11.88	-1.02	0.09
Apo AI	0.00	-11.44	-3.09	0.27
EST AA080005	0.00	-9.11	-1.02	0.11
Apo CIII	0.00	-8.36	-1.04	0.12
Sterol C5 desaturase	0.01	-7.72	-1.04	0.13

Table 2. *SR-BI.* Genes with maxT adjusted p-values ≤ 0.05. For each gene, the table lists the gene name, the permutation adjusted p-value ($\tilde{p}^*$), the two-sample t-statistic (t), the numerator (Num) and denominator (Den) of the t-statistic.

Gene name	$\tilde{p}^*$	t	Num	Den
SR-BI	0.00	13.70	3.05	0.22
SR-BI	0.00	12.13	3.30	0.27
Glutathione s-transferase	0.00	9.66	1.25	0.13
Un-identified	0.00	9.46	1.22	0.13
Glutathione s-transferase	0.00	8.79	1.11	0.13
Un-confirmed	0.02	6.97	0.60	0.09
Un-confirmed	0.02	6.96	0.13	0.02
Cytochrome P450 2B10	0.03	-6.90	-0.74	0.11
Hemoglobin alpha chain	0.03	6.85	0.74	0.11
Cytochrome P450 2B10	0.03	-6.83	-1.46	0.21
Un-confirmed	0.03	6.80	0.50	0.07
Un-confirmed	0.03	-6.77	-0.32	0.05
Hemoglobin beta chain	0.04	6.69	0.55	0.08

chain reaction). In this method, cDNA was first synthesized from the mRNA by random priming and gene specific DNA primers were then used to amplify DNA specific for the gene of interest. Production of DNA was quantitated during the cycles of amplification with SYBR green dye in a 7700 sequence detector (Perkin Elmer). This alternative method of quantitation confirmed changes observed by microarray analysis (Callow et al. (2000)).

Note that (Callow et al. (2000)) applied different image analysis and normalization methods than presented here. The scanned images were processed using the ScanAlyze (Eisen (1999)) software package and a global median normalization was performed. For the apo AI experiment, the same eight genes clearly stood out from the rest, but had slightly larger adjusted p-values than here. The gap between the eight genes and the other genes was also smaller. For the SR-BI experiment, our new image analysis and normalization methods produced a longer list of genes with small adjusted p-values; the identity of all the new genes has not yet been confirmed.

3.3. Comparison with single-slide methods

The single-slide methods of Chen, Dougherty and Bittner (1997), Newton et al. (2001), and Sapir and Churchill (2000) were applied to individual slides from the apo AI and SR-BI experiments. (Note that for the Sapir and Churchill method we are not performing the orthogonal regression for the log-transformed intensities (Part I of the poster). The orthogonal residuals are essentially normalized intensity log-ratios. We have simply implemented Part II of the poster and are applying the mixture model to our already normalized log-ratios.) The above three methods were used to identify genes with differential expression in mRNA samples from individual treatment mice compared to pooled mRNA samples from control mice. Using an $MA-$plot, Figure 7 shows the contours for the posterior odds of change in the Newton et al. (2001) method, the upper and lower limits of the Chen, Dougherty and Bittner (1997) 95% and 99% "confidence intervals" for M, and the contours for the Sapir and Churchill 90%, 95%, and 99% posterior probabilities of differential expression. The regions between the contours for the Newton et al. (2001) method are wider for low and high intensities A; this is a property of the Gamma distribution which is used in the hierarchical model.

In general, the genes identified as differentially expressed seem to vary more between methods than within method for different significance thresholds (e.g., different posterior probability cut-offs). Furthermore, the gene lists varied from slide to slide. For the eighth knock-out mouse in the apo AI experiment (Figure 7), the Chen, Dougherty and Bittner (1997) 95% and 99% rules both single out the eight clones identified using replicated slides (green points). However, the rule makes a large number of Type I errors, especially in the positive M region.

The Newton et al. (2001) rule with 1:1 posterior odds identifies all but one of the eight genes and selects a large number of false positives. With posterior odds of 100:1, the method now only identifies four out of the eight probe sequences, with still a fairly large number of false positives, especially in the positive M region. The Sapir and Churchill method is more conservative than the Chen, Dougherty and Bittner (1997) method and yields contours similar to the Newton et al. (2001) method. Similar patterns were observed for the SR-BI experiment (see web supplement).

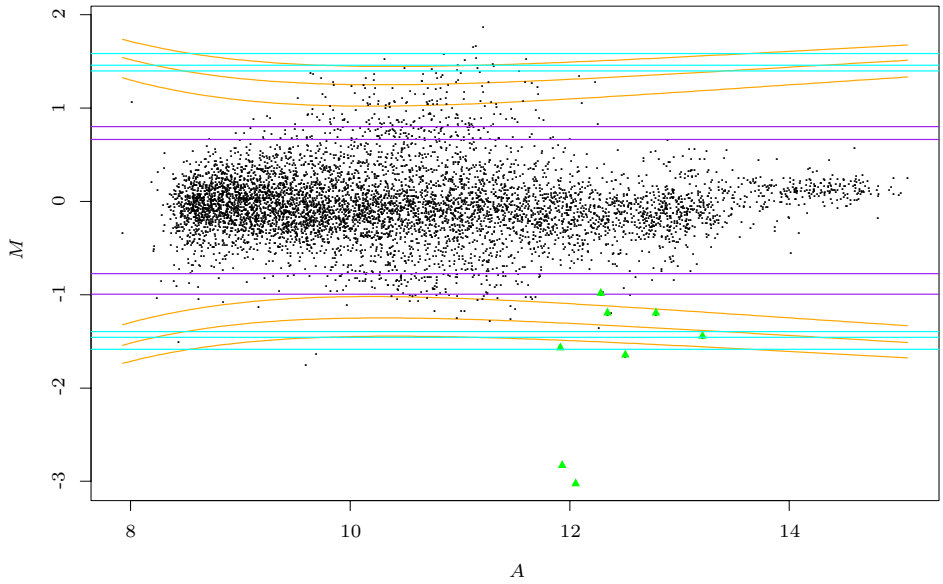


Figure 7. *Apo AI.* Single-slide methods: $MA-$plot with contours for the methods of Newton et al. (orange, odds of change of 1:1, 10:1, and 100:1), Chen et al. (purple, 95% and 99% "confidence"), and Sapir and Churchill (cyan, 90%, 95%, and 99% posterior probability of differential expression). The points corresponding to genes with maxT adjusted p-values less than 0.01 (based on data from 16 slides) are colored in green. The data are from knock-out mouse # 8.

4. Discussion

We have described statistical methods for the identification of differentially expressed genes in replicated microarray experiments. Although it is not the main focus of the paper, we stress the importance of issues such as imaging (e.g., effect of laser power and gain), image analysis (segmentation and background

adjustment), and normalization (Yang et al. (Accepted), Yang et al. (2001)). Each of these pre-processing steps can have a potentially large impact on the (R, G) intensity pairs used in further analyses, such as hypothesis testing or clustering.

Our first recommendation is to examine single-slide intensity data using $MA-$plots. In addition to aiding in the identification of spot artifacts and specific features of the slide (e.g., spatial or print-tip effects), such a representation is useful for normalization purposes. We are proposing normalization methods based on robust local regression to deal with spatial and intensity dependent dye biases observed in numerous experiments. The importance of including the print-tip group (a proxy for the location of the spots on the slide) in the normalization function for the apo AI experiment is clearly illustrated by the results from single-slide methods: without a print-tip dependent normalization, the single-slide methods are essentially calling genes from only four of the print-tip groups and are thus making a large number of false positives (Figure 7). "Global" methods such as mean, median, or ANOVA normalization (Kerr et al. (2000)) do not deal with spatial, plate, or intensity dependent effects. Recently, Sapir and Churchill (2000) have proposed a normalization method based on orthogonal linear regression of log intensities $\log_2 R$ vs. $\log_2 G$ (after a type of background correction of R and G). This is an intensity dependent normalization but, unlike the lowess based normalization method, it only allows a *linear* relationship between the log intensities in the two channels and lacks robustness. We have worked with a number of datasets from different labs and most exhibit *non-linear* relationships between $\log_2 R$ and $\log_2 G$. We do not claim by any means to have identified all relevant sources of systematic variation in a cDNA microarray experiment. Rather, we believe that different systematic features could arise in different types of experiments and that these should be investigated carefully before proceeding to any inference. Until sources of systematic variation are identified and properly accounted for, there can be no question of the system being in statistical control and so no basis for a statistical model to describe chance variation. With many different users of this technology and a variety of experimental protocols, a substantial proportion of the variation is likely to remain systematic and possibly more important than random variation. The situation should improve with a deeper understanding of how the intensity data are acquired and processed. However, given our current limited knowledge of the possible sources of systematic variation, normalization remains an important and challenging question which cannot always be addressed in a simple generic manner or by relying on unverified modeling assumptions. Local regression procedures, such as lowess or loess, are promising tools for devising robust and flexible normalization methods.

For suitably normalized data, the proposed approach for the identification of single differentially expressed genes is to consider a univariate testing problem for each gene and then correct for multiple testing using adjusted p-values. In the lipid metabolism study described above, t-statistics were used to test the null hypothesis of no differential expression between the treatment and control groups. One could have also used non-parametric rank statistics such as the Wilcoxon rank sum statistic. Unlike single-slide methods, no specific parametric form is assumed for the distribution of the (R, G) intensity pairs and a permutation procedure was used to estimate the joint null distribution of the test statistics. We found Q-Q plots and plots of different components of the test statistics against overall intensity $\bar{A}$ particularly useful for the visual identification of genes with altered expression and of important features of these genes. There was a close correspondence between the patterns seen in the Q-Q plots and the adjusted p-values. For the apo AI experiment, 8 cDNA probes (including three copies of the knocked-out gene) clearly stood out from the rest and had very small adjusted p-values ($\tilde{p}^* \leq 0.01$). In the SR-BI experiment, there was no clear discontinuity in the t-statistics or their corresponding adjusted p-values. For brevity, we chose to list only the genes with adjusted p-values less than 5%. However, this cut-off is somewhat arbitrary and biologists may find a higher FWER acceptable for their purposes.

A different approach to multiple testing was proposed in 1995 by Benjamini and Hochberg. These authors argue that in many situations control of the FWER can lead to unduly conservative procedures and that one may be prepared to tolerate some Type I errors, provided their number is small in comparison to the number of rejected hypotheses. These considerations led to a less conservative approach which calls for controlling the expected proportion of Type I errors among the rejected hypotheses — the *false discovery rate* (FDR). The development of FDR controlling procedures is an active area of research. In the microarray setting, where thousands of comparisons are performed simultaneously and a fairly large number of genes are expected to be differentially expressed, FDR controlling procedures present a promising alternative to more conservative FWER approaches.

A comparison of the genes identified with replicated slides and confirmed by RT-PCR to those identified using single-slide methods highlights the importance of replication and a careful study of systematic effects (Figure 7). The genes called by single-slide methods varied across replicated slides, and these methods tended to produce a large number of false positives while missing a few of the confirmed genes. In general, there is no easy way to tell which genes are differentially expressed on the basis of data from a single slide. Recently proposed methods are based on assumed parametric models (e.g., Gamma or Gaussian) for

the (R, G) intensity pairs, and we do not know enough at this point about systematic and random variation in microarray experiments to justify such strong modeling assumptions. In addition, existing single-slide methods do not as yet cope with replicated spots within slides or with between slide variation. The claimed significance levels are thus dubious and it is not clear what progress has been made over the early fold increase/decrease cut-off rules. For the two experiments presented here, "eye-balling" would have worked at least as well as any of the single-slide methods examined in Section 3.3. Most importantly, gene expression data may be too noisy for successful identification without replication, no matter how good the rule.

The need for replication was also stressed by Kerr et al. (2000). These authors assume a linear model for the log intensities, with terms accounting for dye, slide, treatment, and gene main effects, as well as a few interactions between these effects. However, such a "global" model tries to do too much in one step and may lose some of the sensitivity of the experiment: only one main effect for normalization (the dye main effect D_j amounts to a normalization by the mean of log intensities across genes and arrays), only one error term for all genes. Furthermore, interactions are included or not included somewhat arbitrarily and the issue of multiple testing is not addressed. Our approach can also be cast in an ANOVA setting: instead of having one "big" ANOVA for all genes, we consider a "small" ANOVA for each gene, with only treatment and slide effects for already normalized data. The "big" and "small" ANOVAs produce the same contrast estimates, but different SEs for these estimates. The relative merits of these two approaches for the calculation of standard errors deserve further study. For the use of smoothed variance estimators the reader is referred to Lönnstedt and Speed (2001) ; these smoothed estimators represent intermediate ground between the "big" and "small" ANOVA SEs.

The design of the apo AI knock-out and SR-BI transgenic experiments has a number of deficiencies. First, the reference sample used in all 16 hybridizations (for treatment and control mice) consists of a mix of mRNA from the eight control mice. This creates an asymmetry between the treatment and control groups, even in the absence of differential expression. The use of a common reference sample for all hybridizations is favored by biologists in order to compare gene expression levels across slides. In that case, it may have been better to use a more general reference sample, not directly related to the mRNA samples being probed. Second, the reference mRNA was always labeled with the green dye, and the treatment and control mRNA with the red dye. It may be more efficient to have the treatment and control mRNA hybridized to the same slide and reverse the dye assignments in different slides (dye-swap experiment). Clearly more research is needed on the design of microarray experiments; preliminary work on this subject can be found in (Kerr and Churchill (2001) and Lin et al. (2001)).

A natural question arising with the design of this study is whether there is any need to make use of hybridizations involving mRNA from individual control mice and pooled control mRNA. In an obvious sense, using eight treatment mice and eight control mice leads to a more symmetric experimental design, but is it necessary? A partial answer to this question can be found by examining the two datasets, this time using only the eight hybridizations comparing treatment mouse mRNA to pooled control mouse mRNA. By analogy with the initial analysis, one can compare the mean relative expression levels to zero by computing one-sample rather than two-sample t-statistics. For the knock-out experiment, seven out of the eight genes identified with the 16 slides were among the 20 genes with the largest absolute one-sample t-statistic. The remaining gene (apo CIII) had a large t-numerator, but also a fairly large SE. For the SR-BI experiment, only four out of the thirteen genes identified with the 16 slides were among the 20 genes with the largest absolute one-sample t-statistic. We do not yet have a good explanation for this discrepancy and are further exploring the important design issue.

The present paper considered only two types of mRNA samples (treatment and control), but three or more types can be handled in a similar fashion with different test statistics. For factorial experiments, in which several factors such as time and treatment are being monitored simultaneously (Galitski et al. (1999), Lin et al. (2001), Lönnstedt, Grant, Begley and Speed (n.d.)), one could perform an ANOVA for each gene. It is implicit in this approach that there are only a modest number of differentially expressed genes, rather than a continuum, and that it is reasonable to attempt to identify them all. While it is perhaps too early to say in general when this makes sense, there are clearly situations in which it may not. When comparing gene expression between whole mouse brain and cerebellum cells, for example, a large proportion of the genes seem to be differentially expressed, and it seems futile to seek a clear cut-off between the genes which are and which are not. Also, note that the question addressed in this paper, as well as in Chen, Dougherty and Bittner (1997), Kerr et al. (2000), Newton et al. (2001), Sapir and Churchill (2000), is the identification of *single* differentially expressed genes, i.e., the null hypothesis of equal expression is tested for one gene at a time. Having data on many arrays gives us the potential for learning about the *joint* behavior of genes and the next step would be to seek clusters of genes which change in a coordinate manner. However, statistical methods for doing so are still in their infancy; recent efforts include the work of Hastie et al. (2000) and Lazzeroni and Owen (2000).

Although the methods described in the present paper were illustrated on data from a cDNA microarray study, some apply to oligonucleotide arrays (Affymetrix chips) as well. The diagnostic plots for the test statistics and multiple testing procedure extend directly. For example, K. Vranizan and B. R. Conklin (private

communication) have used the method outlined in Section 2.5 above to adjust p-values for Affymetrix chip data on 6,320 genes from an experiment involving eight control mice and nine mice expressing Ro1 at eight weeks (see Redfern et al. (2000) and the supplemental material at http://www.pnas.org for greater detail). In this comparison, many hundreds of genes had small unadjusted p-values, but only 55 had adjusted p-values less than 0.05, 26 involving a relative over-expression and 29 a relative under-expression at the eight-week timepoint compared to the control. The normalization method of Section 2.3 is not directly applicable, however, the general discussion on the identification of systematic sources of variation is equally relevant to this other type of technology.

Finally, the methods described in this paper are implemented in an R package (Ihaka and Gentleman (1996)), SMA (Statistics for Microarray Analysis), which may be downloaded from http://www.R-project.org. Supplementary analyses, figures, and datasets are available at http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html.

Acknowledgements

We would like to acknowledge Juliet Shaffer from the Statistics Department at UC Berkeley for valuable discussions on multiple testing, and for guiding us through the literature on this topic. We would also like to thank Ben Bolstad from the Statistics Department at UC Berkeley for his assistance with the single-slide methods. Members of the Brown and Botstein labs at Stanford University, Ngai lab at UC Berkeley, and David Bowtell and Chuang Fong Kong from the Peter MacCallum Cancer Institute in Melbourne have been most helpful in introducing us to the many statistical questions arising in microarray experiments. We are also grateful to Bruce Conklin and Karen Vranizan, from the Gladstone Institute of Cardiovascular Disease at the University of California at San Francisco, for stimulating discussions on the analysis of Affymetrix chip data. Finally, we thank anonymous referees for valuable comments on an earlier version of this paper.

This work was supported in part by an MSRI and a PMMB postdoctoral fellowship (SD), and by the NIH through grants 5R01MH61665-02 (YHY) and 8R1GM59506A (TPS).

References

Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Trans. Pattern Analysis and Machine Intelligence* **16**, 641-647.

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., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000). Different types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

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.* **96**, 6745-6750.

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

Buckley, M. J. (2000). *The Spot user's guide.* CSIRO Mathematical and Information Sciences. http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* **10**, 2022-2029.

Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomedical Optics* **2**, 364-374.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-685.

DeRisi, J., Penland, L., Brown, P. O. and Bittner, M. L. et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* **14**, 457-460.

Eisen, M. B. (1999). *ScanAlyze.* http://rana.Stanford.EDU/software/ for software and documentation.

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.* **95**, 14863-14868.

Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S. and Fink, G. R. (1999). Ploidy regulation of gene expression. *Science* **285**, 251-254.

Gasch, A. P., Spellman, P. T., Kao, C. M., rel, O. C.-H., Eisen, M. B.,Storz, G., Botstein, D. and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell.* **11**, 4241-4257.

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., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D. and Brown, P. O. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**, 1-21.

Hughes, T. R., Mao, M., Jones, A. R.,Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H. and Linsley, P. S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **19**, 342-347.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *J. Computat. Graph. Statist.* **5**, 299-314.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183-201.

Kerr, M. K., Martin, M. and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J. Computat. Biology* **7**, 819-837.

Lazzeroni, L. and Owen, A. B. (2000). Plaid models for gene expression data. *Technical report*, Department of Statistics, Stanford University. http://www-stat.stanford.edu/research/list.html.

Lin, D. M., Yang, Y. H., Scolnick, J. A., Brunet, L. B., Peng, V., Speed, T. P. and Ngai, J. (2001). A spatial map of gene expression in the olfactory bulb. Submitted.

Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M. and Horton, H. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675-1680.

Lönnstedt, I., Grant, S., Begley, G. and Speed, T. P. (2001). Microarray analysis of two interacting treatments: a linear model and trends in expression over time. In preparation.

Lönnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statist. Sinica* **12**, 31-46.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Computat. Biology* **8**, 37-52.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O. and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci.* **96**, 9212-9217.

Pollack. J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41-46.

Redfern, C. H., Degtyarev, M. Y., Kwa, A. T., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E. K., Coward, P., Shah, N., Warrington, J. A., Fishman, G. I., Bernstein, D., Baker, A. J. and Conklin, B. R. (2000) Conditional expression of a G_i-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proc. Natl. Acad. Sci.* **97**, 4826-4831.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennet, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C. and Friend, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression, *Science* **287**, 873-880. Web supplement.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M. and Spellman, P., Iyer, V., Jeffrey, S. S., van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**, 227-234.

Division of Biostatistics, School of Public Health, University of California, Berkeley, 140 Earl Warren Hall, #7360, Berkeley, CA 94720-7360, U.S.A.

E-mail: sandrine@stat.berkeley.edu

Department of Statistics, University of California, Berkeley, 367 Evans Hall #3860, Berkeley, CA 94720-3860, U.S.A.

E-mail: yeehwa@stat.berkeley.edu

Genome Sciences Department, Lawrence Berkeley National Laboratory.

E-mail: mjcallow@1b1.gov

Genetics and Bioinformatics Group, The Walter and Eliza Hall Institute.

E-mail: terry@stat.berkeley.edu

(Received October 2001; accepted October 2001)