Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments

Sandrine Dudoit[1*], Yee Hwa Yang[2*], Matthew J. Callow[3], and Terence P. Speed[2,4]

Technical report # 578, August 2000

1. Department of Biochemistry, Stanford University
2. Department of Statistics, University of California at Berkeley
3. Genome Sciences Department, Lawrence Berkeley National Laboratory
4. Genetics and Bioinformatics Group, The Walter and Eliza Hall Institute

Address for correspondence:
Sandrine Dudoit
Department of Biochemistry
Stanford University School of Medicine
Beckman Center, B400
Stanford, CA 94305-5307
Tel: (650) 736-0076
Fax: (650) 723-7016
E-mail: `sandrine@genome.stanford.edu`

* *These authors contributed equally to this work.*

Abstract

Microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes simultaneously. This 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, we stress the importance of issues such as image processing and normalization. Image processing is required to extract measures of transcript abundance for each gene spotted on the array from the laser scan images. Normalization is needed to identify and remove systematic sources of variation, such as differing dye labeling efficiencies and scanning properties. There can be many systematic sources of variation and their effects can be large relative to the effects of interest. After a brief presentation of our image processing method, we describe a within-slide normalization approach which handles spatial and intensity dependent effects on the measured expression levels.

Given suitably normalized data, our proposed method 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. No specific parametric form is assumed for the distribution of the expression levels and a permutation procedure is used to estimate the joint null distribution of the test statistics for each gene. Several data displays are suggested for the visual identification of genes with altered expression and of important features of these genes. The above methods are applied to microarray data from a study of gene expression in two mouse models with very low HDL cholesterol levels. The genes identified using data from replicated slides are compared to those obtained by applying recently published single-slide methods.

Keywords: replicated cDNA microarrays; differential gene expression; normalization; adjusted p-values; permutation test; graphical display.

1 Introduction

DNA microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes simultaneously (see Section 2 for an introduction to microarrays). Applications of microarrays range from the study of gene expression in yeast under different environmental stress conditions to the comparison of gene expression profiles for tumors from cancer patients. In addition to the enormous scientific potential of microarrays to help in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research. Microarray experiments raise numerous statistical questions, in diverse fields such as image processing, experimental design, and discriminant analysis.

This paper describes statistical methods for the analysis of gene expression data from a study of lipid metabolism in mice (Callow *et al.* [6]). The goal of the cDNA microarray experiments is to identify genes with altered expression in two mouse models with very low HDL cholesterol levels (treatment groups) compared to inbred control mice. The two mouse models considered in this study are the apolipoprotein AI (apo AI) knock-out and the scavenger receptor BI (SR-BI) transgenic mice, where apo AI and SR-BI are two genes known to play pivotal roles in HDL metabolism.

The identification of differentially expressed genes is a question which arises in a broad range of microarray experiments (Callow *et al.* [6], Friddle *et al.* [12], Galitski *et al.* [13], Golub *et al.* [14], and Spellman *et al.* [34], to name a few). The types of experiments include: *single-slide* cDNA microarray experiments, in which one compares transcript abundance (*i.e.*, expression levels) in two mRNA samples, the red and green labeled mRNA samples hybridized to the same slide; *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 experiments which we will not discuss here.

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 expression 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.* [8], Schena *et al.* [30, 31]) 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.* [30] 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.* [8]

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 methods have turned to 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 *et al.* [7] 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 [27] suggest identifying differentially expressed genes using posterior probabilities of change under a mixture model for the log expression ratio $\log R/G$ (after a type of background correction, the orthogonal residuals from the robust regression of $\log R$ *vs.* $\log G$ are essentially normalized log expression ratios). A limitation of these two methods is that they both ignore the information contained in the product RG. Recognizing this problem, Newton *et al.* [23] 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.* [25] is based on assuming that R and G are approximately independently and normally distributed, with variance depending on the mean. It thus 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) falls outside the region between the two curves. The relative merits of the methods depend on their ability to successfully identify differentially expressed genes (*i.e.*, their power or one minus their false negative 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 concern of an elevated chance of committing at least one Type I Error. Finally and most importantly, the gene expression data may be too variable (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 processing methods for segmentation and background correction of the laser scan images can have a very 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 systematic sources of 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. With many different users of the technology and differing protocols, a substantial portion of the variation is likely to reflect a host of systematic effects. Until these 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.

Statistical methods for identifying differentially expressed genes in multiple-slide experiments seem to have received relatively little attention. Instead of considering a testing problem, some investigators have turned to exploratory cluster analysis tools (Alizadeh *et al.* [3], Ross *et al.* [26], Tamayo *et al.* [35]). In this setting, cluster analysis methods, such as hierarchical clustering or self-organizing maps, are used to group genes with correlated expression profiles across experimental conditions. Groups of differentially expressed genes are identified by visual inspection of the resulting clusters, using, for example, red and green images to display the log intensity ratios for each gene in each of the slides (Eisen *et al.* [11]). Such methods are unsupervised in that they do not use the class of the samples hybridized to the slides (*e.g.* mRNA from treatment or control mice). Another approach is to identify single differentially expressed genes by computing for each gene the correlation of its expression profile with a reference expression profile, such as a vector of indicators for class membership (in the case of two classes, this correlation coefficient is a type of t-statistic). Genes are then ranked according to their correlation with the reference profile and permutation methods are used to determine cut-offs for controlling the number of false positives (Galitski *et al.* [13] and Golub *et al.* [14]). In a recent paper, Kerr *et al.* [20] stress the importance of replication in order to assess the variability of estimates of change and suggest applying techniques from the analysis of variance (ANOVA). They 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 identified based on contrasts for the (treatment $\times$ gene) interactions.

In this paper, we focus on the identification of single differentially expressed genes in replicated cDNA microarray experiments. After a brief presentation of our image processing method, we describe a within-slide normalization method which handles spatial and intensity dependent effects on the measured expression levels. Next, given suitably normalized data, our basic approach is to consider a univariate testing problem for each gene and correct for multiple testing using adjusted p-values. More specifically, for the lipid metabolism study described above, the genes of interest are found by testing for each gene the null hypothesis of equal mean expression levels in the treatment and control groups. This involves the calculation of a t-statistic for each gene. Various data displays are suggested for the visual identification of genes with altered expression and of important features of these genes. A more precise assessment of the evidence against the null hypothesis of constant expression can be obtained by calculating p-values. However, with a typical microarray dataset comprising thousands of genes, an immediate concern is multiple testing. Adjusted p-values are used to control the family-wise Type I Error rate [18, 32, 33, 37]. Because the joint null distribution of the test statistics is unknown, the adjusted p-values are estimated by permutation (Westfall and Young [37]).

The paper is organized as follows. Section 2 contains a brief introduction to the biology and technology of cDNA microarrays. The datasets are presented in Section 3 along with a summary of our image processing methods for segmentation and background correction. Section 4 describes our proposed normalization method, which allows print-tip and intensity dependent effects. The test statistics, the calculation of adjusted p-values and the data displays are also discussed in Section 4. In Section 5, we present the results of the study and compare the genes identified using replicated slides to those identified by single-slide methods. Finally, Section 6 discusses our findings and outlines open questions.

2 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 many 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. Next, we briefly review basic genetic notions useful for understanding microarray experiments.

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), guanine (G), cytosine (C), and thymine (T). The two chains 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. We refer the reader to Gonick and Wheelis [15] and Griffiths *et al.* [16] for an introduction to the relevant biology.

Different properties 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 focussed primarily on expression at the transcription stage, *i.e.*, on mRNA levels. Although the regulation of protein synthesis in a cell is by no means controlled solely by mRNA levels, mRNA levels sensitively reflect the type and state of the cell. Microarrays derive their power and universality from a key property of DNA molecules described above: *complementary base-pairing*. The term *hybridization* refers to the annealing of nucleic acid strands from different sources according to the base-pairing rules. To utilize the hybridization property of DNA, *complementary DNA* or *cDNA* is obtained from mRNA by reverse transcription. There are different types of microarray systems, including cDNA microarrays [8, 9, 11, 30, 31] and high-density oligonucleotide arrays (proprietary Affymetrix chips) [22]; the 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 (*e.g.* a red-fluorescent dye Cy5 and a green-fluorescent dye Cy3), then mixed and hybridized with the arrayed DNA sequences or *probes* (following the definition of probe and target adopted in the January 1999 supplement to Nature Genetics [1]). 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 fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two nucleic acid samples. The diagram in Figure 1 describes the main steps in a cDNA microarray experiment.

Aside from the enormous scientific potential of microarrays to help in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research. By comparing gene expression in normal and disease cells, microarrays may be used to identify disease genes and targets for therapeutic drugs. The supplement to Nature Genetics [1] and the books *DNA Microarrays : A Practical Approach* [28] and *Microarray Biochip Technology* [29] provide general overviews of microarray technologies and of different areas of application of microarrays.

*** Place Figure 1 about here ***

3 Data

3.1 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.* [6]). 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 8 mice with the apo AI gene knocked-out and the control group consists of 8 "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 8 control mice. The design for the second experiment is similar, but with 8 SR-BI transgenic mice comprising the treatment group and 8 "normal" FVB mice comprising the control group.

In each experiment, target cDNA is hybridized to microarrays containing 5,548 cDNA probes, including 200 related to lipid metabolism. Note that we call the spotted cDNA sequences "genes", whether they are actual genes, ESTs (expressed sequence tags), or DNA sequences from other sources.

3.2 Image processing

The red and green fluorescence intensities (R, G) are already highly processed data. In a cDNA microarray experiment, the hybridized arrays are imaged using a *scanner* (*e.g.* laser scanning confocal microscope) and the output stored as 16-bit image files, one for each dye. We view these image files as "raw" data. Image processing is required to extract measures of transcript abundance for each gene spotted on the array from the laser scan images.

We have developed a new method for extracting information from microarray images. It is implemented in the software `Spot` which was written in collaboration with CSIRO Mathematical and Information Sciences (Buckley [5], Yang *et al.* [38]). The image processing can be divided into three steps which are briefly described next.

1. **Automatic address:** In order to extract spot intensities from microarray images it is necessary to accurately identify the location of each of the spots. This process is called *addressing* and it assigns coordinates to each of the spots. We have developed an automatic *gridding* procedure for addressing the spots; this procedure also includes *registration* (alignment) between the two channels when necessary.

2. **Segmentation:** *Segmentation* allows the classification of pixels as signal (*i.e.*, as corresponding to a spot of interest) or background. Our software implements an adaptive segmentation algorithm known as *seeded-region-growing* [2]. Seeded-region-growing makes no assumptions regarding the size or the shape of the spots, and segmentation is performed using the total pixel intensities for both channels. Other commonly used segmentation methods include fixed circle (*e.g.* `ScanAlyze` software [10]) and adaptive circle (*e.g.* commercial image processing software `GenePix` for the Axon scanner).

3. **Intensity extraction:** After detecting the location, size, and shape of each spot, we calculate signal (spot) intensities, background intensities, and quality measures for each dye at each spot on the array.

    - **Signal:** For each dye, the signal for any given spot is measured by the sum of the pixel intensities within the spot. This sum represents the total amount of cDNA hybridized to the spotted DNA sequence.
    - **Background:** The motivation for background correction is that a spot's measured fluorescence intensity (in each of the two channels) includes a contribution which is not specifically due to the hybridization of the target to the probe (*e.g.* fluorescence of the background due to other chemicals and the glass). Common approaches to background calculation include taking the median of the pixel

intensities in a region surrounding the spot (*e.g.* `ScanAlyze`) and taking the median of the pixel intensities in the local valleys in between spots (*e.g.* `GenePix`). The `Spot` software implements a non-linear filter known as *morphological opening* which provides an estimate of the background drift.

- **Quality measures:** For each spot, our software computes quality measures such as spot size, shape and relative signal to background intensity. We have yet to make use of these measures in our analyses.

A detailed discussion of our choice of image processing methods and a comparison to popular alternatives is presented in Yang *et al.* [38]. Thus, starting with two 16-bit images, the image processing steps described above produce two main quantities for each spot on the array: R and G, which are measures of the fluorescence intensities (transcript abundance) for the red and green labeled mRNA samples, respectively.

4 Methods

4.1 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.* [23], Sapir and Churchill [27], and papers in Schena [29]) [1]. We find that such plots give an unrealistic sense of concordance and make interesting features of the data harder to see. We prefer to plot the log intensity 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.* [25]). An M *vs.* A plot amounts to a 45^o counterclockwise rotation of the $(\log_2 G, \log_2 R)$-coordinate system, followed by scaling of the coordinates (Figure 3). If M' and A' denote the rotated coordinates, then $A = A'/\sqrt{2}$ and $M = M'\sqrt{2}$.

An M *vs.* A plot is thus another representation of the (R, G) data in terms of the log intensity ratios M which are the quantities of interest to most investigators. We have found M *vs.* A plots to be more revealing than their $\log_2 R$ *vs.* $\log_2 G$ counterparts in terms of identifying spot artifacts and detecting intensity dependent patterns in the log ratios. They are also very useful for normalization as illustrated next.

*** Place Figure 3 about here ***

4.2 Normalization

The purpose of normalization is to identify and remove systematic sources of variation (*e.g.* different labeling efficiencies and scanning properties of the dyes, print-tip or spatial effects)

[1] 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 additive for logged intensities; (iii) taking logs evens out highly skew 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 $2^{16} - 1$.

and allow between-slide comparisons. Imbalance in the red and green intensities can manifest itself when two identical mRNA samples are labeled with different dyes and hybridized to the same slide. In such a situation, the red intensities tend to be lower than the green intensities and the magnitude of the difference may depend on overall intensity A. Reasons for the imbalance in the two channels include properties of the dyes themselves (*e.g.* different labeling efficiencies and scanning properties) and experimental variability resulting, for example, from separate reverse transcription and labeling of the two samples.

The plots in Figures 4 and 9 of $M = \log_2 R/G$ *vs.* $A = \log_2 \sqrt{RG}$ clearly show the dependence of the log ratio M on overall spot intensity A. This suggests that an intensity or A dependent normalization method may be preferable to global methods such as normalization by the mean or median of M values. Furthermore, for the apo AI experiment, the image in Figure 2 and the within print-tip group lowess curves in Figure 4 suggest the existence of spatial or print-tip effects on the fluorescence intensities [2]. In the image, the bottom 4 grids tend to have high red signal, and this is reflected in the M *vs.* A plot, where the corresponding within print-tip group lowess curves clearly stand out from the remaining 12 curves. We thus perform a within print-tip group intensity dependent normalization using the scatter-plot smoother implemented in the `lowess()` function from the Splus software (Venables and Ripley [36]):

$$\log_2 R/G \rightarrow \log_2 R/G - c_j(A) = \log_2 k_j(A) R/G,$$

where $c_j(A)$ is the `lowess()` fit to the M *vs.* A plot for spots printed using the jth print-tip (*i.e.*, data from the jth grid only). In the apo AI and SR-BI experiments $j = 1, \ldots, 16$. The `lowess()` function is a scatter-plot smoother which uses robust locally linear fits. We typically use $f = 20$ to 40% for the parameter specifying the fraction of the data used for smoothing at each point. For the experiments considered here, a small proportion of the genes are expected to vary in expression between the red and green labeled mRNA samples. Thus, normalization is performed using all 5,548 genes. In other circumstances, a number of "housekeeping" genes may be spotted on the slide and used for normalization purposes (Yang *et al.* [39]). Normalization is a challenging question due to the possibly large number of sources of systematic variation and the choice of a suitable gene set.

*** Place Figures 4 and 9 about here ***

4.3 Test statistics

The gene expression data considered here can be summarized by a matrix X of log intensity ratios $\log_2 R/G$, with k rows corresponding to the genes being studied and $n = n_1 + n_2$

[2] 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 and the spots on a given grid are printed using the same print-tip or pin. We say that spots printed using the same print-tip are part of the same print-tip group. 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 be possible to separate print-tip effects from spatial effects.

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). For ease of notation, let the first n_1 columns refer to the n_1 control hybridizations and let the last n_2 columns refer to the treatment hybridizations. In the two experiments considered here $n_1 = n_2 = 8$ and $k = 5,548$.

Let H_j denote the null hypothesis of equal treatment and control mean expression levels for gene j, $j = 1, \ldots, k$ (in the remaining analysis we consider expression levels relative to the green-labeled reference mRNA, *i.e.*, the expression level of a gene is $\log_2 R/G$). Here, we consider only two-sided alternative hypotheses; one-sided alternatives can be handled in a similar manner. For gene j, the *t-statistic* comparing gene expression in the control and treatment groups is

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\sqrt{\frac{s_{1j}^2}{n_1} + \frac{s_{2j}^2}{n_2}}},$$

where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ denote the average expression level of gene j in the n_1 control and n_2 treatment hybridizations, respectively. Similarly, s_{1j}^2 and s_{2j}^2 denote the sample variances of gene j's expression level in the control and treatment hybridizations, respectively. 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 we are not assuming that the t-statistics actually follow a t-distribution, rather we use permutation to estimate their distribution (see Section 4.5.2).

4.4 Data displays for test statistics

4.4.1 Quantile-Quantile plots

Quantile-Quantile plots (Q-Q plots) are a useful display of the test statistics for the thousands of genes being studied in a typical microarray experiment. In a normal Q-Q plot, the quantiles of the data are plotted against the quantiles of a standard normal distribution. In general, Q-Q plots are used to assess whether data have a particular distribution or whether two datasets have the same distribution. If the distributions are the same, then the plot will be approximately a straight line. A plot with a "U" shape means that one distribution is skewed relative to the other. An "S" shape implies that one distribution has heavier tails than the other. In our application, we are not so much interested in testing whether the t-statistics follow a particular distribution, but in using the Q-Q plot as a visual aid for identifying genes with "unusual" t-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. For the two datasets considered here, fewer than 20 genes have t-statistics which deviate markedly from a line going through the first and third quartiles of a normal Q-Q plot.

4.4.2 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.

4.5 Adjusted p-values

4.5.1 Definitions

The normal Q-Q plots for the t-statistics are useful visual aids for identifying genes with altered expression in the treatment mice compared to the controls. A more precise assessment of the evidence against the null hypothesis may be obtained by calculating p-values for each gene. However, with a typical microarray dataset comprising thousands of genes, an immediate concern is multiple testing [18, 32, 33, 37]. 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 (FWE)*, *i.e.*, the probability of at least one error in the family (see Shaffer [33] for a review of such methods). Some procedures provide *strong control* of the FWE, *i.e.*, control this error rate for any combination of true and false hypotheses, while others provide only *weak control*, *i.e.*, control the FWE only when all null hypotheses in the family are true. The procedures described below provide strong control of the error rate.

To account for multiple hypothesis testing, one may calculate adjusted p-values (Shaffer [33] and Westfall and Young [37]). According to Shaffer [33], 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. There are several approaches for computing adjusted p-values and these vary in the severity of the correction for multiplicity. Let p_j and $\tilde{p}_j$ denote respectively the unadjusted and adjusted p-values for hypothesis H_j (gene j), $j = 1, \ldots, k$, and let $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_k}$ denote the ordered unadjusted p-values. Hypothesis H_j is rejected at FWE α if $\tilde{p}_j \leq \alpha$.

The Bonferroni method is perhaps the best known method for dealing with multiple testing. The *Bonferroni single-step adjusted p-values* are given by $\tilde{p}_j = \min(kp_j, 1)$. Closely related to Bonferroni's method is the Šidák method which is exact for protecting the FWE when the unadjusted p-values are independently distributed as $U[0,1]$. The *Šidák single-step adjusted p-values* are given by $\tilde{p}_j = 1 - (1 - p_j)^k$. These methods are called *single-step* because they perform equivalent multiplicity adjustments for all hypotheses, regardless of the ordering of the observed p-values. While single-step adjusted p-values are simple to calculate, they tend to be very conservative. Improvement in power, while preserving strong control of the FWE, may be achieved by considering *step-down* methods which order p-values and make successively smaller adjustments. *Holm's step-down Bonferroni adjusted p-values* are given by

$$\begin{aligned}\tilde{p}_{r_1} &= kp_{r_1} \\ \tilde{p}_{r_j} &= \max\Big(\tilde{p}_{r_{j-1}},\ (k-j+1)p_{r_j}\Big) \qquad \text{for } 2 \le j \le k,\end{aligned} \tag{1}$$

with p-values greater than 1 set to 1. Holm's procedure is less conservative than the standard Bonferroni procedure which would multiply the p-values by k at each step. However, neither Holm's method nor the single-step methods presented above take into account the dependence structure between the variables. In a microarray experiment, groups of genes tend to have highly correlated expression levels for reasons such as co-regulation. A more general and less conservative definition of adjusted p-values, which takes into account the dependence structure between variables, is proposed by Westfall and Young [37]. The *Westfall and Young step-down adjusted p-values* are defined by

$$\begin{aligned}\tilde{p}_{r_1} &= pr(\min_{l\in\{r_1,\ldots,r_k\}} P_l \le p_{r_1} \mid \mathrm{H}_0) \\ \tilde{p}_{r_j} &= \max\Big(\tilde{p}_{r_{j-1}},\ pr(\min_{l\in\{r_j,\ldots,r_k\}} P_l \le p_{r_j} \mid \mathrm{H}_0)\Big) \qquad \text{for } 2 \le j \le k,\end{aligned} \tag{2}$$

where H_0 denotes the intersection of all null hypotheses and P_l the random variable for the unadjusted p-value of the lth hypothesis. Note that computing the quantities in (2) under the assumption that $P_l \sim U[0,1]$ and using the upper bound provided by Boole's inequality yields (1).

4.5.2 Estimation of adjusted p-values by permutation

In practice, the joint null distribution of the t-statistics $T_1, \ldots, T_k$ is unknown, it can however be estimated by permuting the columns of the 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. The permutation distribution of the test statistics T_j, $j = 1, \ldots, k$, is obtained by the following algorithm.

Basic permutation algorithm. For the bth iteration, $b = 1, \ldots, B$:

1. Permute the n columns of the data matrix X. The first (last) n_1 (n_2) columns now refer to the "fake" control (treatment) group.

2. For each gene, compute t-statistics as above: $t_1^{(b)}, \ldots, t_k^{(b)}$.

When computationally feasible, the above steps are applied to each of the $\binom{n}{n_1}$ possible treatment/control allocations. For the knock-out and transgenic mouse datasets, there are $\binom{16}{8} = 12,870$ such permutations. Otherwise, these steps are repeated for thousands of random permutations of the columns of X. The permutation distribution of the t-statistic for gene j is given by the empirical distribution of $t_j^{(1)}, t_j^{(2)}, \ldots, t_j^{(B)}$.

Unadjusted permutation p-values. For two-sided alternative hypotheses, permutation p-values for the t-statistics are

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

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

Adjusted p-values for the methods of Bonferroni, Šidák, and Holm can be estimated by replacing p_j by p_j^*. However, for the adjusted p-values of Westfall and Young, the joint null distribution of $P_1, \ldots, P_k$ needs to be estimated. When the unadjusted p-values themselves are unknown and estimated using resampling methods, additional resampling for estimating adjusted p-values can be computationally intractable. Rather than obtaining unadjusted p-values by permutation, Westfall and Young suggest approximating these p-values using asymptotic theory. Adjusted p-values can then be estimated by permutation (p. 114 in Westfall and Young [37]). In our setting, all the hypotheses are tested by t-statistics. In principle, the t-statistics for individual genes could have a different null distribution. However, asymptotically, these statistics have the same null distribution and the p-values should be monotone in the observed t-statistics across genes. We follow Algorithm 4.1 in Westfall and Young [37] and assume that $pr(\min_{l\in\{r_j,\ldots,r_k\}} P_l \leq p_{r_j} \mid \mathrm{H}_0) = pr(\max_{l\in\{r_j,\ldots,r_k\}} |T_l| \geq |t_{r_j}| \mid \mathrm{H}_0)$.

Permutation algorithm for Westfall and Young step-down adjusted p-values

For the bth permutation

1. Permute the n columns of the data matrix X. The first (last) n_1 (n_2) columns now refer to the "fake" control (treatment) group.

2. For each gene, compute t-statistics: $t_1^{(b)}, \ldots, t_k^{(b)}$.

3. Next, compute

$$\begin{aligned} u_k^{(b)} &= |t_{r_k}^{(b)}| \\ u_j^{(b)} &= \max\Big(u_{j+1}^{(b)}, |t_{r_j}^{(b)}|\Big) \qquad \text{for } 1 \leq j \leq k-1, \end{aligned}$$

   where r_j are such that $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_k}|$ for the original data.

The above steps are repeated B times and 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

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

5 Results

5.1 Normalization

Figures 4 and 9 display M *vs.* A plots for individual slides from the apo AI and SR-BI experiments. These plots illustrate the non-linear dependence of the log ratio M on the overall intensity A; the dependence is especially strong in the SR-BI experiment. This suggests that an intensity dependent normalization method may be preferable to a global one. Also, for the apo AI hybridizations, 4 within print-tip group lowess curves clearly stand out from the remaining 12 curves, suggesting strong print-tip or spatial effects. The 4 curves correspond to the last row of pins in the 4×4 print-tip cluster (pins 13, 14, 15, and 16). This pattern is visible in the images, where the bottom 4 grids tend to have high red signal (Figure 2). In these two experiments, relatively few genes are expected to vary in expression between the treatment and control mice. We thus normalize the data using `lowess` fits to the within print-tip group M *vs.* A plots.

*** Place Figure 2 about here ***

5.2 Identification of differentially expressed genes with replicated experiments

Q-Q plots. For the apo AI experiment, the normal Q-Q plot in Figure 5 indicates that 8 genes have t-statistics that deviate markedly from an otherwise linear relationship. All 8 genes have negative t-statistics, suggesting down-regulation in the knock-out mice compared to the controls. For the SR-BI experiment (Figure 10), the deviations from linearity are more subtle and gradual. There are about a dozen genes with "unusual" t-statistics; these seem like possible candidates for differential expression (both up and down-regulation). 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 adjusted p-values.

*** Place Figures 5 and 10 about here ***

Adjusted p-values. Figures 6 and 11 display plots of the unadjusted and adjusted p-values for the 50 genes with the largest absolute t-statistics. For both experiments, unadjusted p-values are very close to zero while adjusted p-values can be quite large. For the apo AI experiment, 8 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-values is much more gradual than in the apo A1 knock-out experiment. Thus, adjusted p-values sensitively reflect the pattern seen in the Q-Q plots, while unadjusted p-values are as expected much too small and lack the sensitivity of adjusted p-values.

*** Place Figures 6 and 11 about here ***

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 (Figures 7 and 12, and Tables 1 and 2). For both experiments, the genes with large absolute t-statistics tended to have high total expression levels, as measured by $\bar{A}$. They typically had large differences in their relative expression levels (M) between the two groups (numerator) as well as low standard errors (SEs in denominator).

*** Place Figures 7 and 12 and Tables 1 and 2 about here ***

Identity of differentially expressed genes. Many of the genes 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. This resulted in several of the genes in Tables 1 and 2 appearing multiple times after re-sequencing.

For the apo AI knock-out experiment, apo AI appeared 3 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. 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 identified 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 biological 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 were quantitated by RT-PCR (real-time quantitative polymerase 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.* [6]).

Note that in Callow *et al.* [6], the gene expression data were analyzed using different image processing and normalization methods than presented here. In Callow *et al.*, the scan images were processed using the `ScanAlyze` [10] software and the data were normalized within-slide by subtracting the median of all log intensity ratios from individual log intensity ratios. For the apo AI experiment, the same 8 genes clearly stood out from the rest, but had slightly larger adjusted p-values than here. The gap between the 8 genes and the other genes was also smaller. For the SR-BI experiment, our new image processing and normalization methods produced a longer list of genes with small adjusted p-values; we have not yet confirmed the identity of all the new genes.

5.3 Comparison with single-slide methods

We have applied the Chen *et al.* [7], Newton *et al.* [23], and Sapir and Churchill [27] [3] single-slide methods to individual slides from the apo AI and SR-BI experiments. These methods are used to identify genes with differential expression in mRNA samples from individual treatment mice compared to pooled mRNA samples from control mice. Using an M *vs.* A representation, Figures 8 and 13 show the contours for the posterior odds of change in the Newton *et al.* method, the upper and lower limits of the Chen *et al.* 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.* method are wider for low and high intensities A; this is a property of the Gamma distribution which is used in the hierarchical model.

For the *8th* knock-out mouse in the apo AI experiment (Figure 8), the Chen *et al.* 95% and 99% rules both pick out the 8 genes identified using replicated data (green points). However, it also picks out a large number of false positives, especially in the positive M region. The Newton *et al.* rule with 1:1 posterior odds identifies all but one of the 8 genes and selects a large number of false positives. With posterior odds of 100:1, the method now only identifies 4 out of the 8 genes, with still a fairly large number of false positives, especially in the positive M region. The Sapir and Churchill method is a lot more conservative than the Chen *et al.* method and yields contours similar to the Newton *et al.* method. 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). Similar patterns were observed for the other slides (not shown) and for the SR-BI experiment (Figure 13).

*** Place Figures 8 and 13 about here ***

6 Discussion

In this paper, we have presented statistical methods for the identification of single differentially expressed genes in replicated microarray experiments. Although it is not the main focus of the paper, we have stressed the importance of issues such as imaging (*e.g.* effect of laser power and gain), image processing (segmentation and background adjustment), and normalization (Yang *et al.* [38, 39]). 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

[3]Note that we are not performing the orthogonal regression for the log transformed intensities (Part I of the poster). The orthogonal residuals of Sapir and Churchill are essentially normalized log expression ratios. We have simply implemented Part II of the poster and are applying the mixture model to our already normalized log ratios.

testing or clustering.

Our first recommendation is to examine single-slide data using M *vs.* A plots. Such a representation is useful for the identification of spot artifacts and specific features of the slide (*e.g.* print-tip effects). The intensity or A dependent normalization method proposed here deals with the intensity dependence also often observed in M *vs.* A plots for experiments in which the same mRNA is labeled in both channels (data not shown). The usefulness of a print-tip group normalization for the apo AI experiment is clearly illustrated by its impact on the results from single-slide methods: without a print-tip dependent normalization, the single-slide methods are essentially picking genes from only four of the print-tips and are thus making a large number of false positives (Figure 4). "Global" methods such as mean, median, or ANOVA normalization do not deal with these features. Recently, Sapir and Churchill [27] 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 our `lowess` based normalization method it only allows a *linear* relationship between the log intensities in the two channels. 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 important systematic sources of 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 systematic sources of 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 data are acquired and processed. However, given our current limited knowledge of the possible sources of systematic variation, normalization remains a challenging question which cannot always be addressed in a simple generic manner or by relying on unverified modeling assumptions.

For suitably normalized data, our 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, we used a t-statistic to test the null hypothesis of equal mean expression levels in the treatment and control groups. One could have also used a non-parametric rank statistic 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 is used to estimate the joint null distribution of the test statistics for each gene. 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 good correspondence between the patterns seen in the Q-Q plots and the adjusted p-values. In the SR-BI experiment, there was no clear discontinuity in the t-statistics or their corresponding 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 FWE acceptable for their purposes.

A common criticism for procedures that control the FWE in the strong sense, such as the Westfall and Young adjusted p-values, is that they are too conservative and that strong control is not always needed. Recently, Benjamini and Hochberg [4] have proposed a less conservative approach to multiple testing which calls for controlling the expected proportion of falsely rejected hypotheses or *false discovery rate (FDR)*. Control of the FDR implies weak control of the FWE. Benjamini and Hochberg give a simple Bonferroni-type procedure which controls the FDR for independent test statistics. This procedure is not applicable to the gene expression data for which genes tend to be highly correlated.

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 (Figures 8 and 13). Single-slide methods tend to produce a large number of false positives and at the same time miss a few of the confirmed genes. There is no easy way to tell which genes are differentially expressed on the basis of data from a single microarray experiment. Recently proposed methods are based on assumed parametric models (*e.g.* Gamma or Gaussian) for the (R, G) intensities and at this point, we do not know enough about the systematic and random variation within a microarray experiment to justify such strong 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 we examined. Most importantly, gene expression data may be too noisy for successful identification without replication, no matter how good the rule.

The importance of replication was also stressed by Kerr *et al.* [20] who proposed a linear model for the log intensities. 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 array 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. We are currently exploring the use of smoothed variance estimators for situations in which only a few replicates are available. 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. Firstly, the reference sample used in all 16 hybridizations (for treatment and control mice) consists of a mix of mRNA from the 8 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 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. Secondly, 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 assignment 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 [19].

A natural question arising with the design of this study is whether there is any need to make use of comparisons involving mRNA from individual control mice and pooled control mRNA, rather than simply comparing mRNA from individual treatment mice to pooled mRNA from control mice. In an obvious sense, using 8 treatment mice and 8 control mice leads to a more symmetric experimental design, and one which admits a permutation analysis, but is it necessary? We can get a partial answer to this question by examining our two data sets, but this time using only the data from the 8 experiments comparing treatment mouse mRNA to pooled control mouse mRNA. By analogy with our initial analysis, we can compare the mean relative expression levels to zero by computing one-sample rather than two-sample t-statistics. We can then make normal Q-Q plots and plots of t-numerator, t-denominator and t against overall intensity $\bar{A}$, all as before. However, in such an analysis, we can no longer determine the statistical significance of outlying points in the Q-Q plots by permutation. It seems to us that we now have no choice but to assume the normality of the t-statistics, or to carry out a more extensive bootstrapping approximation based on resampling the 8 treatment mice. We do not present the results of this here, but simply comment that for the knock-out experiment, 7 out of the 8 genes identified with the 16 slides were among the 20 genes with the largest absolute one sample t-statistics. The remaining gene (apo CIII) had a large t-numerator, but also a fairly large SE. The other 13 (out of 20) genes tended to have fairly low standard error and $\bar{A}$. For the SR-BI experiment, only 4 out of the 13 genes identified with the 16 slides were among the 20 genes with the largest absolute one sample t-statistics. We do not yet have a good explanation for this discrepancy, but hope to find one, as the design issue is an important one.

The present paper focuses on 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 are being monitored (*e.g.* study of ploidy and mating types in Galitski *et al.* [13]), one could perform an ANOVA for each gene. It is implicit in our approach that there are only a modest number of differentially expressed genes in the experiments we consider, 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 approach makes sense, there are clearly situations such as tissue or organ comparisons in which it may not. When comparing gene expression between whole mouse brain and olfactory bulb 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 [7, 20, 23, 27], is the identification of *single* differentially expressed genes only, *i.e.*, we consider a large scale single gene screen in which the null hypothesis of equal expression is tested for one gene at a time. We recognize that 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.* [17] and Lazzeroni and Owen [21].

Finally, 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 testing approach, diagnostic plots for the test statistics and adjusted p-value calculation extend directly. For example, K. Vranizan and B. R. Conklin (private communication) have used the method outlined in Section 4.5.2 above to adjust p-values for Affymetrix chip data on 6,320 genes from an experiment involving 8 control mice and 9 mice expressing Ro1 at 8 weeks, see Redfern *et al.* [24] and the supplemental material at `http://www.pnas.org` for fuller details. In this comparison, many hundreds of genes had small unadjusted p-values, but just 55 had adjusted p-values less than 0.05, 26 involving a relative over-expression and 29 a relative under-expression at the 8-week time point compared to the control. Our approach to normalization is not directly applicable, however, our general discussion on the identification of systematic sources of variation is equally relevant to this other type of technology.

Acknowledgments

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.

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

[1] *The Chipping Forecast*, volume 21, 1999. Supplement to Nature Genetics.

[2] R. Adam and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:641–647, 1994.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Different types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

[5] M. J. Buckley. *The Spot user's guide*. CSIRO Mathematical and Information Sciences, August 2000. `http://www.cmis.csiro.au/IAP/spotinfo.htm`.

[6] M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in hdl deficient mice. *Genome Research*, Submitted.

[7] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 2:364–374, 1997.

[8] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, 1996.

[9] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–685, 1997.

[10] M. B. Eisen. `ScanAlyze`. `http://rana.Stanford.EDU/software/` for software and documentation.

[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95:14863–14868, 1998.

[12] C. J. Friddle, T. Koga, E. M. Rubin, and J. Bristow. Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Natl. Acad. Sci.*, 97:6745–6750, 2000.

[13] T. Galitski, A. J. Saldanha, C. A. Styles, E. S. Lander, and G. R. Fink. Ploidy regulation of gene expression. *Science*, 285:251–254, 1999.

[14] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[15] L. Gonick and M. Wheelis. *The cartoon guide to genetics.* Harper Perennial, updated edition, 1991.

[16] A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. *An Introduction to Genetic Analysis.* W. H. Freeman and Company, New York, 6th edition, 1996.

[17] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, D. Botstein, and P. Brown. Gene shaving: a new class of clustering methods for expression arrays. Technical report, Department of Health Research and Policy, Stanford University, 2000.
`http://www-stat.stanford.edu/ hastie/Papers/`.

[18] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6:65–70, 1979.

[19] M. K. Kerr and G. A. Churchill. Experimental design for gene expression microarrays. Technical report, The Jackson Laboratory, 2000.
`http://www.jax.org/research/churchill/pubs/index.html`.

[20] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray. Technical report, The Jackson Laboratory, 2000.
`http://www.jax.org/research/churchill/pubs/index.html`.

[21] L. Lazzeroni and A.B. Owen. Plaid models for gene expression data. Technical report, Department of Statistics, Stanford Unisersity, 2000.
`http://www-stat.stanford.edu/research/list.html`.

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

[23] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Technical Report 139, Department of Biostatistics and Medical Informatics, UW Madison, 1999.
`http://www.stat.wisc.edu/ newton/papers/abstracts/btr139a.html`.

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

[25] C. J. Roberts, B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennet, Y. D. He, H. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend.

Singnaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression. *Science*, 287:873–880, 2000. Web supplement.

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

[27] M. Sapir and G. A. Churchill. *Estimating the posterior probability of differential gene expression from microarray data.* Poster, The Jackson Laboratory, 2000. `http://www.jax.org/research/churchill/`.

[28] M. Schena, editor. *DNA Microarrays : A Practical Approach.* Oxford University Press, 1999.

[29] M. Schena, editor. *Microarray Biochip Technology.* Eaton, 2000.

[30] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[31] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human genome analysis: microarray-based expression monotoring of 1000 genes. *Proc. Natl. Acad. Sci.*, 93:10614–10619, 1996.

[32] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *JASA*, 81:826–831, 1986.

[33] J. P. Shaffer. Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46:561–584, 1995.

[34] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[35] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.*, 96:2907–2912, 1999.

[36] W. N. Venables and B. D. Ripley. *Modern applied statistics with S-PLUS.* Springer, 3rd edition, 1999.

[37] P. H. Westfall and S. S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley series in probability and mathematical statistics. Wiley, 1993.

[38] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Image processing on cDNA microarray data. (manuscript in preparation).

[39] Y. H. Yang, S. Dudoit, and T. P. Speed. Normalization for cDNA microarray data. (manuscript in preparation).

Table 1: Apo AI. Genes with adjusted p-value ≤ 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	$\tilde{p}^*$	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 adjusted p-value ≤ 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	$\tilde{p}^*$	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

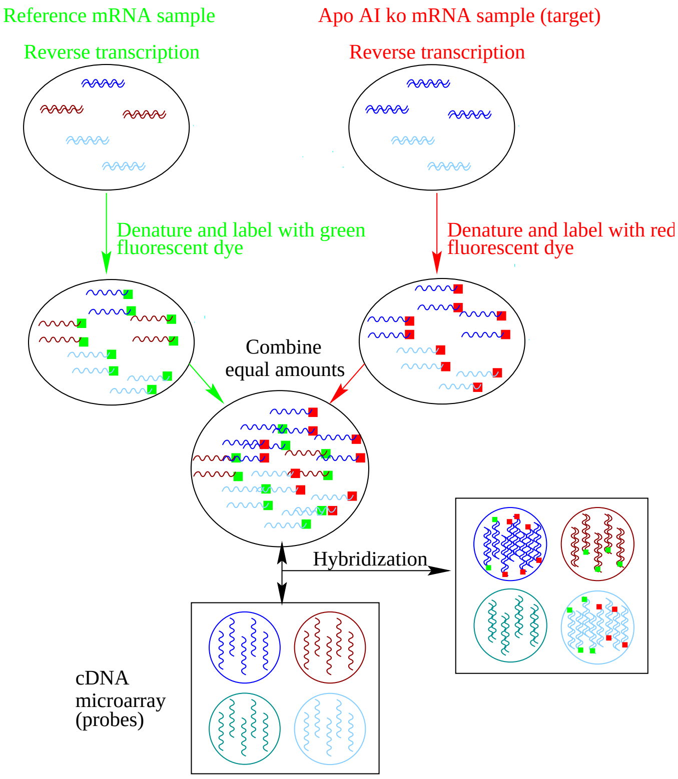

Figure 1: cDNA microarray experiment for apo AI knock-out mice. For each apo AI knock-out mouse, target cDNA is obtained from liver mRNA by reverse transcription and labeled using a red-fluorescent dye (Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations is prepared by pooling cDNA from the 8 C57Bl/6 control mice. The two target samples are mixed and hybridized to a microarray containing 5,548 cDNA probes. Following 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.

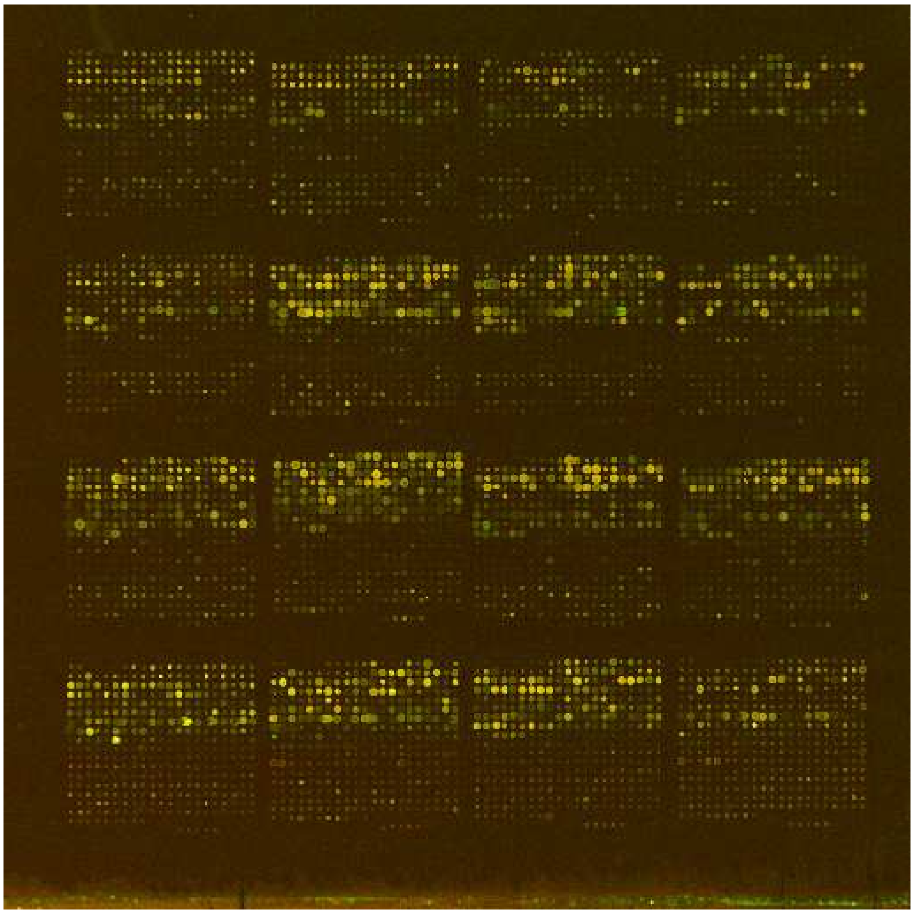

Figure 2: Apo AI. RGB image for visualizing the results from the microarray experiment for knock-out mouse #8. For display purposes, the two 16-bit TIFF images (scan output from the Cy3 and Cy5 channels) were compressed into 8-bit images using a square root transformation. This transformation is required in order to display the fluorescence intensities for both wavelengths using a 24-bit composite RGB overlay image. In this RGB image, blue values (B) are set to zero, red values (R) are used for the Cy5 intensities, and green values (G) are used for the Cy3 intensities. Bright green spots represent genes under-expressed in the knock-out mouse, bright red spots represent genes over-expressed in the knock-out mouse, and yellow spots represent genes with similar expression in the knock-out mouse and the reference sample. The coordinates of the three apo AI clones are (2,2,8,7), (4,1,8,6), and (3,3,8,5), where, for example, (2,2,8,7) is the spot in row 8 and column 7 of the spot matrix which is in row 2 and column 2 of the grid matrix.

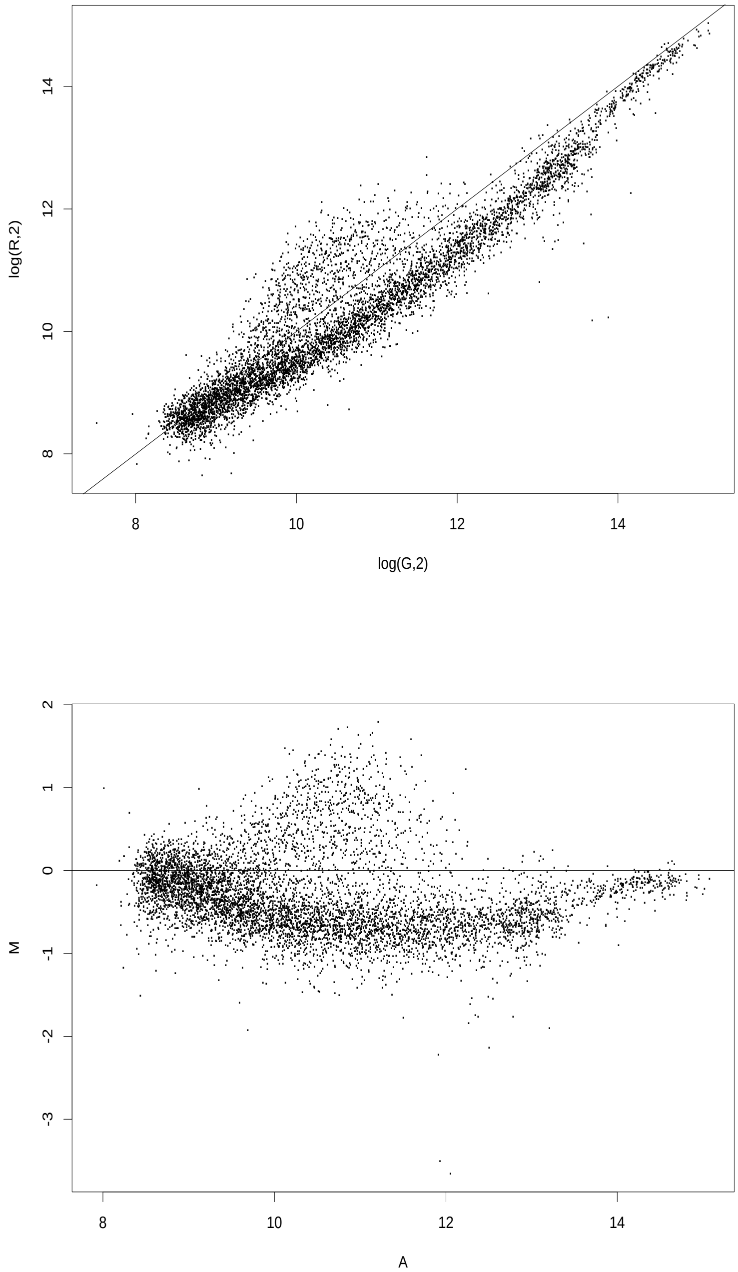

Figure 3: Apo AI. $\log_2 R$ *vs.* $\log_2 G$ plot and M *vs.* A plot for gene expression data from knock-out mouse 8. The $\log_2 R = \log_2 G$ and $M = 0$ lines are drawn for reference.

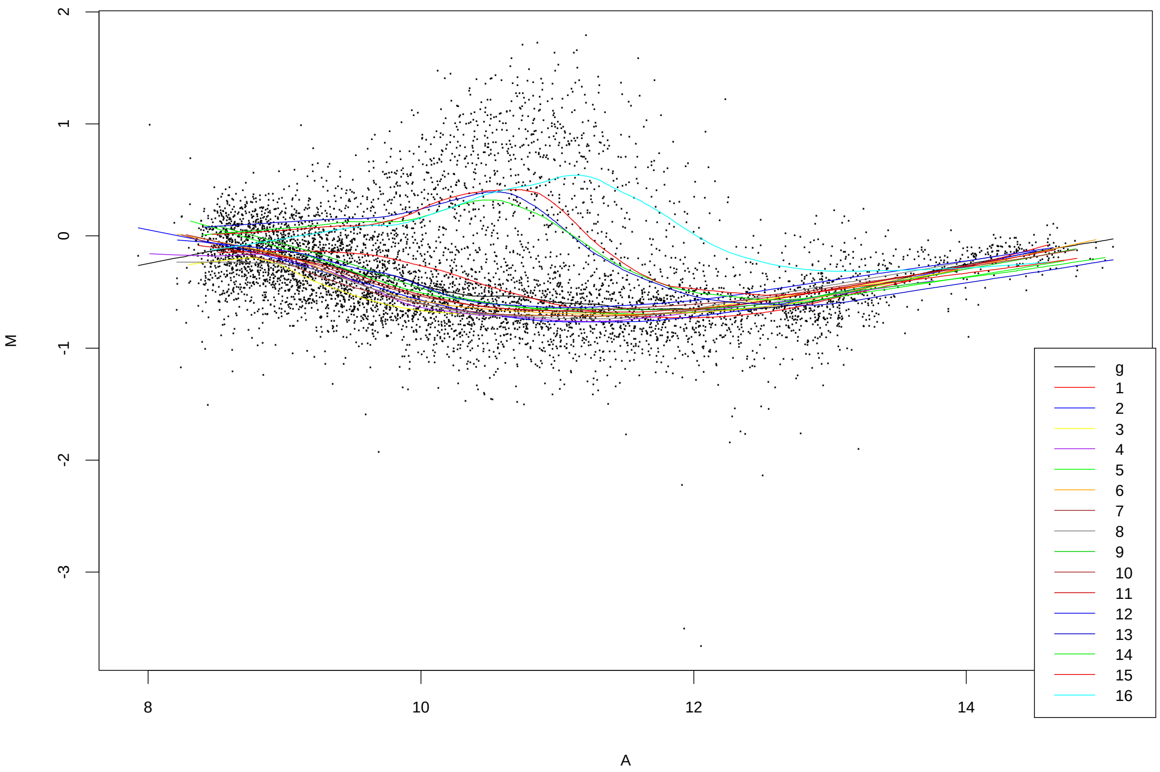

Figure 4: Apo AI. M *vs.* A plot for print-tip group normalization: `lowess` lines ($f = 20\%$) for each of the 16 print-tips (data from knock-out mouse 8). The curve labeled by "g" corresponds to the `lowess` fit to the entire dataset.

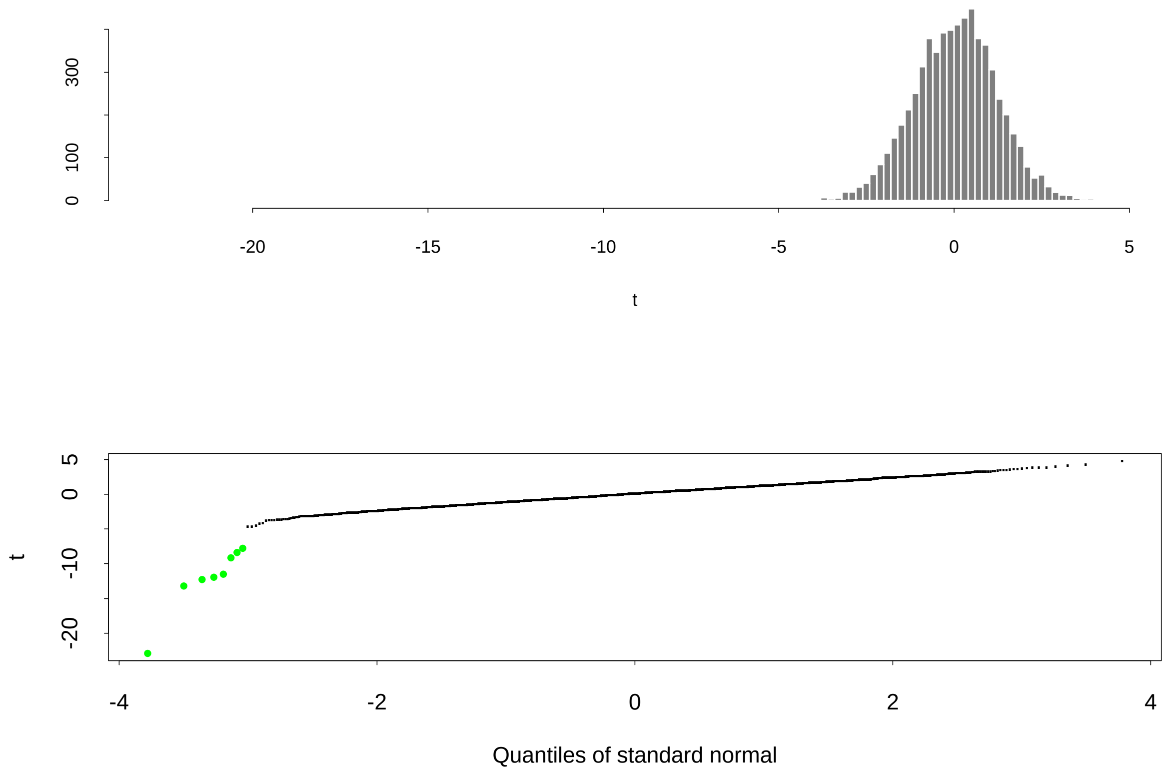

Figure 5: Apo AI. Histogram and Q-Q plot for two-sample t-statistics. The points corresponding to genes with adjusted p-value less than 0.01 are colored in green.

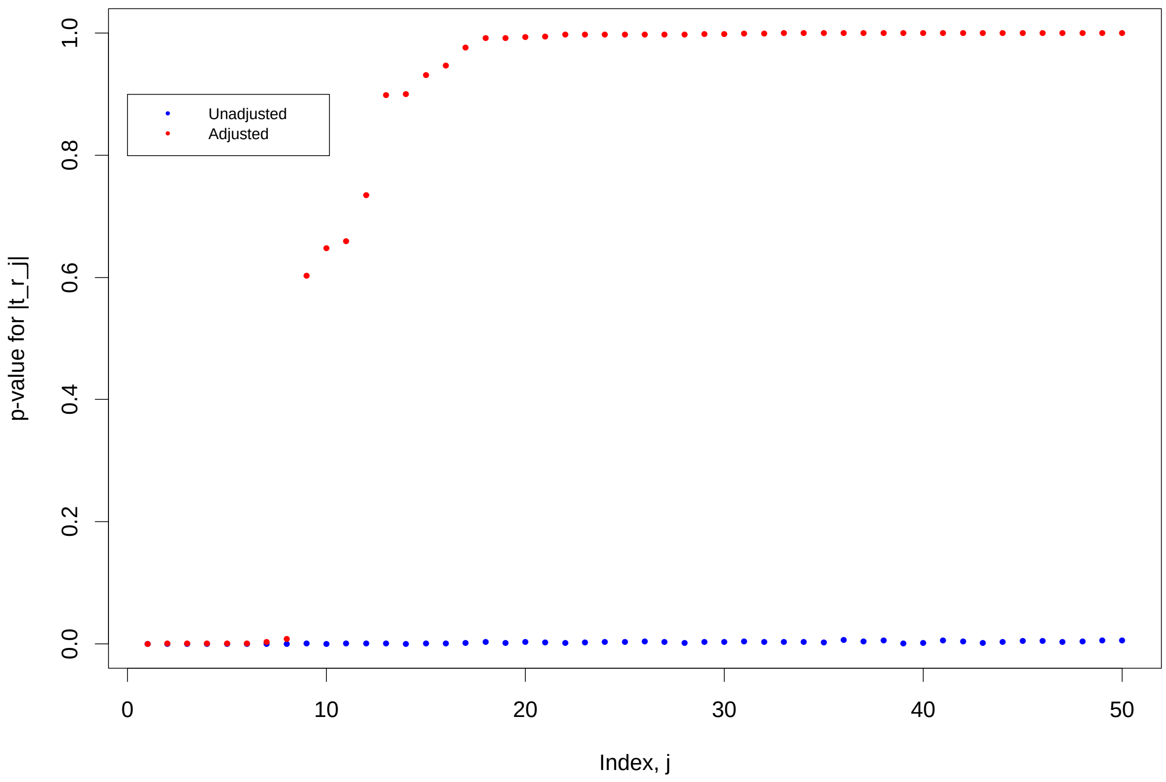

Figure 6: Apo AI. Adjusted and unadjusted p-values for the 50 genes with the largest absolute t-statistics.

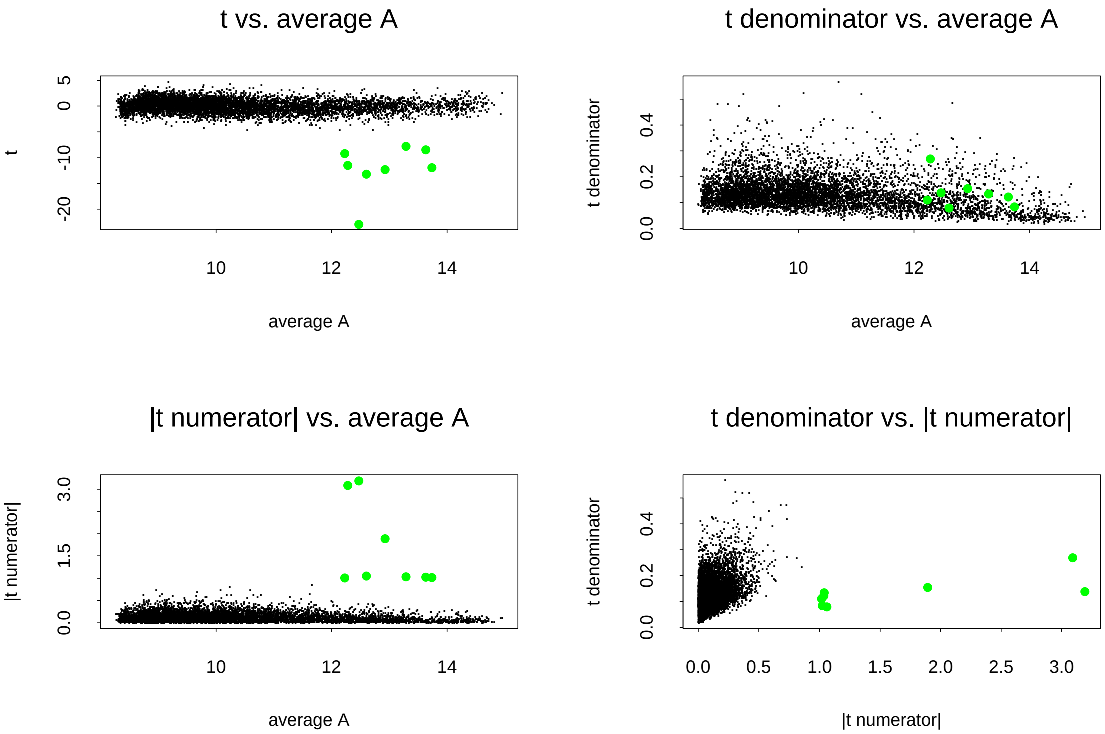

Figure 7: Apo AI. Plots of t-statistics, numerator, and denominator, against overall intensity $\bar{A}$. The points corresponding to genes with adjusted p-value less than 0.01 are colored in green.

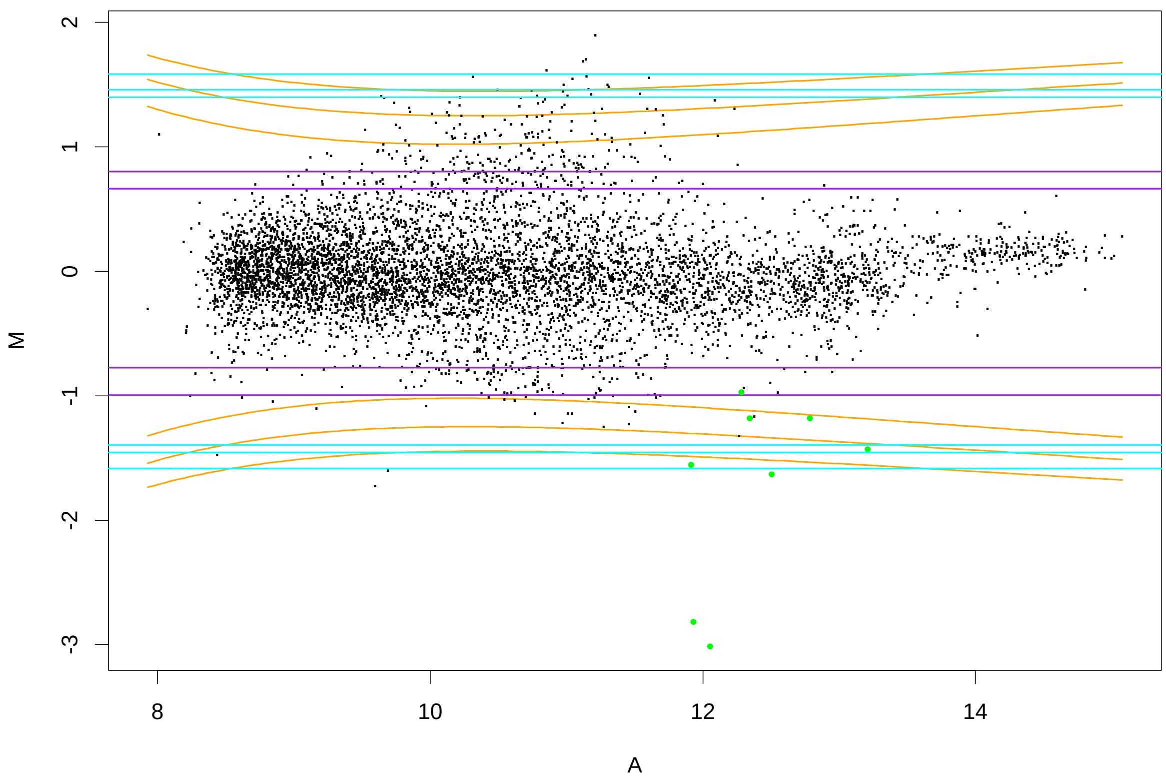

Figure 8: Apo AI. Single-slide methods: For the M *vs.* A representation, contours for the methods of Newton *et al.* (orange, odds of change of 1:1, 10:1, and 100:1), Chen *et al.* (purple, 95% and 99% “confidence”), and Sapir and Churchill (cyan, 90%, 95%, and 99% posterior probability of differential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

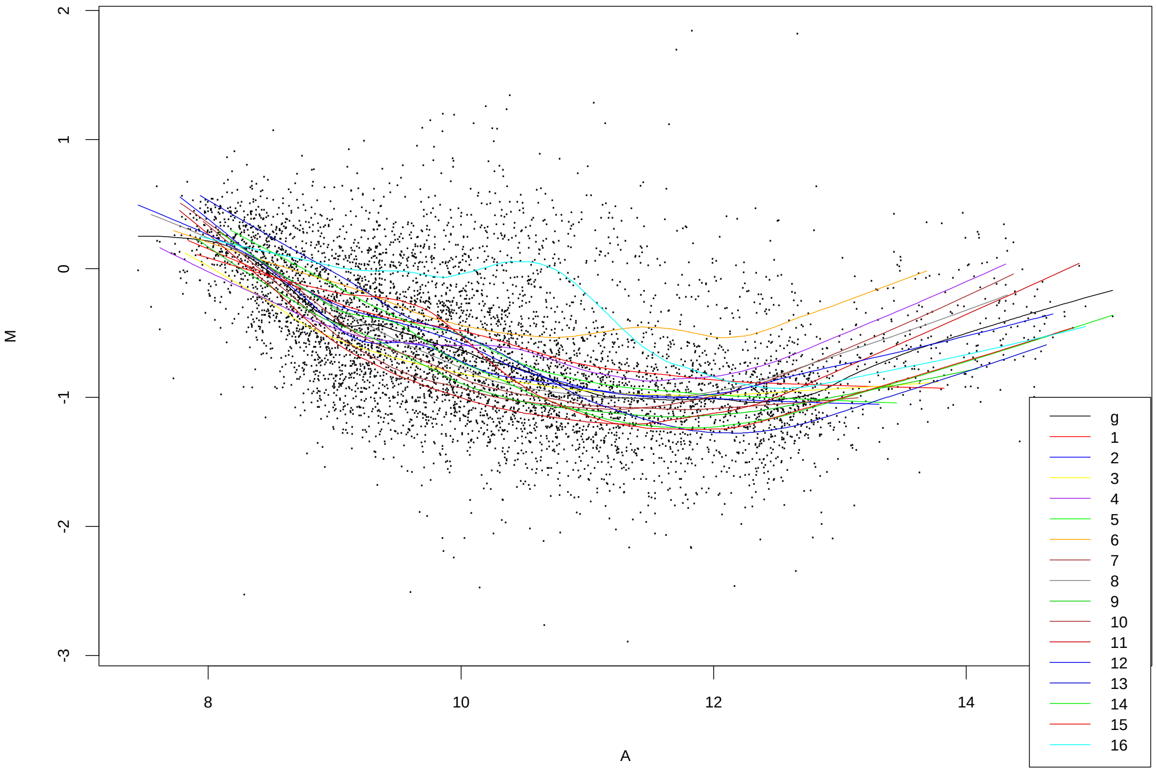

Figure 9: SR-BI. M *vs.* A plot for print-tip group normalization: `lowess` lines ($f = 20\%$) for each of the 16 print-tips (data from transgenic mouse 8). The curve labeled by "g" corresponds to the `lowess` fit to the entire dataset.

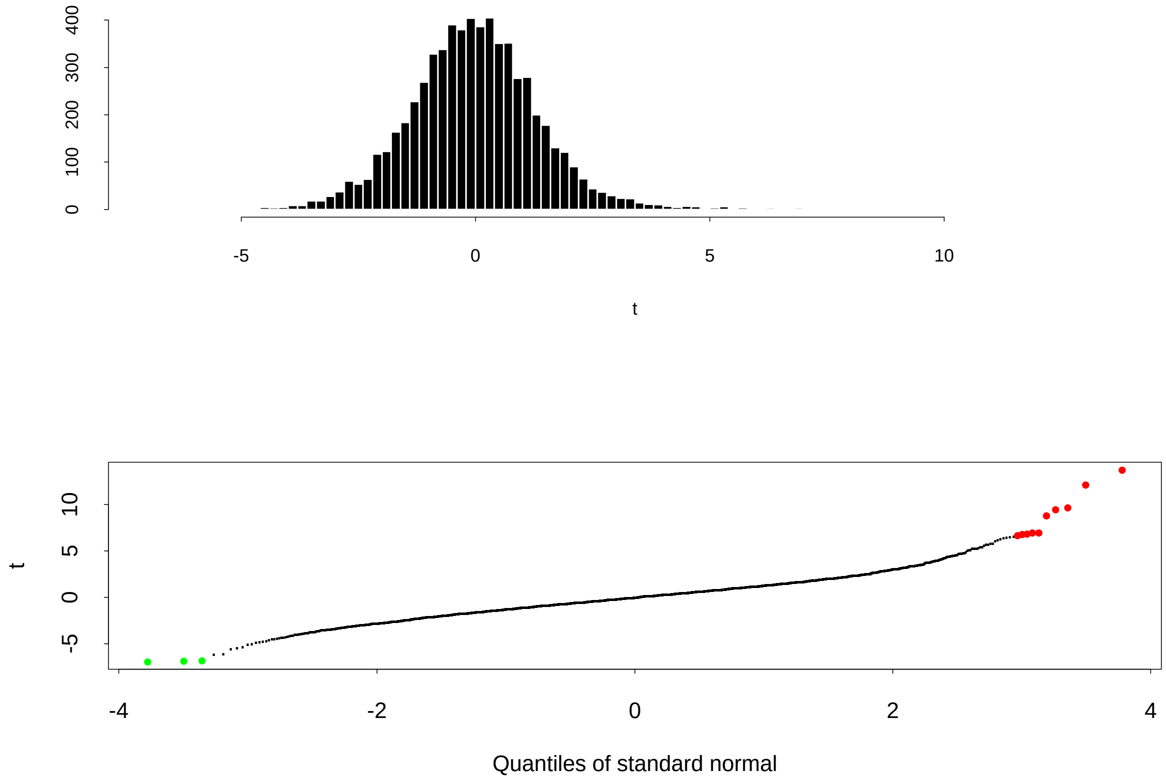

Figure 10: SR-BI. Histogram and Q-Q plot for two-sample t-statistics. The points corresponding to genes with adjusted p-value less than 0.05 are colored in green (negative t-statistic) and red (positive t-statistic).

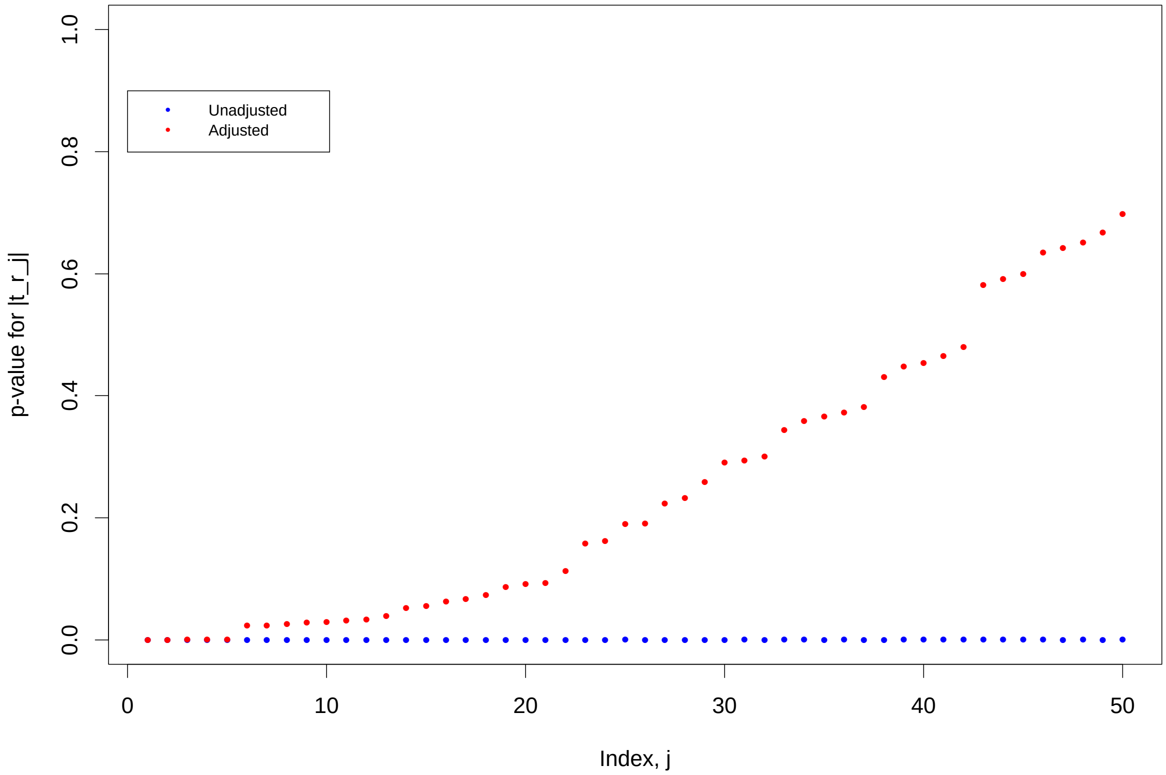

Figure 11: SR-BI. Adjusted and unadjusted p-values for the 50 genes with the largest absolute t-statistics.

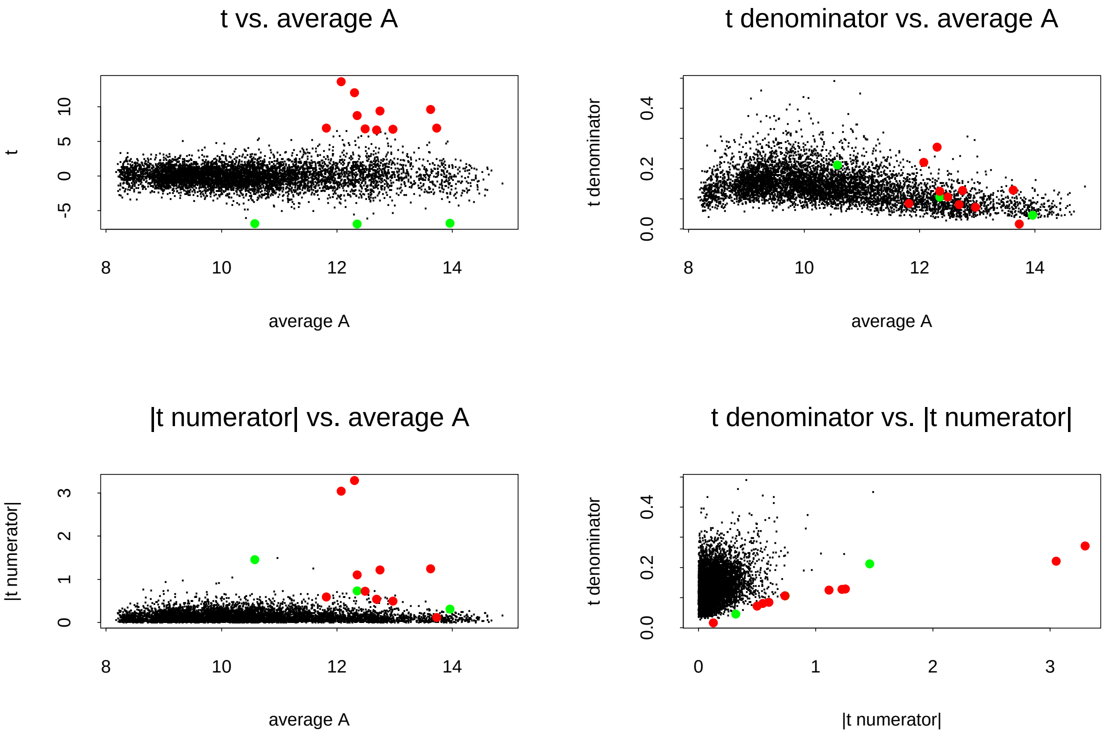

Figure 12: SR-BI. Plots of t-statistics, numerator, and denominator, against overall intensity $\bar{A}$. The points corresponding to genes with adjusted p-value less than 0.05 are colored in green (negative t-statistic) and red (positive t-statistic).

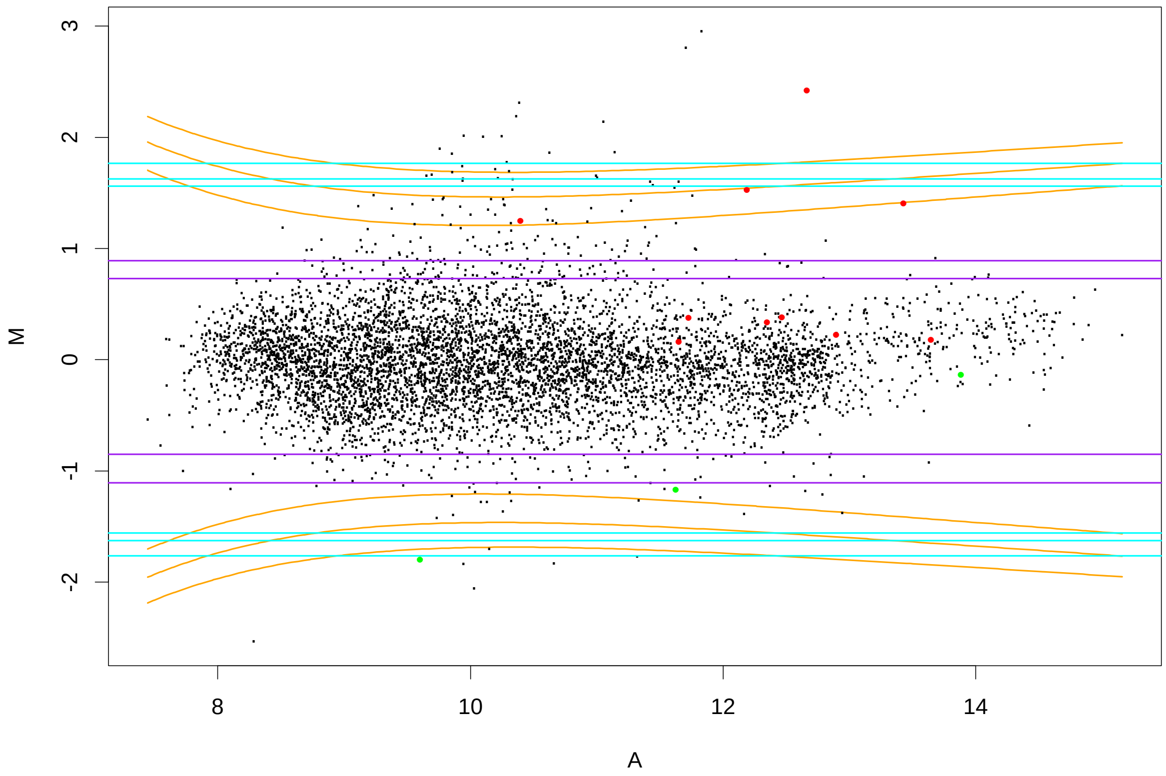

Figure 13: SR-BI. Single-slide methods: For the M *vs.* A representation, contours for the methods of Newton *et al.* (orange, odds of change of 1:1, 10:1, and 100:1), Chen *et al.* (purple, 95% and 99% "confidence"), and Sapir and Churchill (cyan, 90%, 95%, and 99% posterior probability of differential expression). The points corresponding to genes with adjusted p-value less than 0.05 (based on data from 16 slides) are colored in green (negative t-statistic) and red (positive t-statistic). The data are from transgenic mouse 8.