

## Statistical Issues in cDNA Microarray Data Analysis

Gordon K. Smyth, Yee Hwa Yang, and Terry Speed

### 1. Introduction

Statistical considerations are frequently to the fore in the analysis of microarray data, as researchers sift through massive amounts of data and adjust for various sources of variability in order to identify the important genes among the many that are measured. This chapter summarizes some of the issues involved and provides a brief review of the analysis tools that are available to researchers to deal with these issues.

Any microarray experiment involves several distinct stages. First, there is the design of the experiment. The researchers must decide which genes are to be printed on the arrays, which sources of RNA are to be hybridized to the arrays, and on how many arrays the hybridizations will be replicated. Second, after hybridization, there follow a number of data-cleaning steps or “low-level analysis” of the microarray data. The microarray images must be processed to acquire red and green foreground and background intensities for each spot. The acquired red/green ratios must be normalized to adjust for dye bias as well as for any systematic variation other than that owing to the differences among the RNA samples being studied. Third, the normalized ratios are analyzed by various graphic and numerical means to select differentially expressed genes or to find groups of genes whose expression profiles can reliably classify the different RNA sources into meaningful groups. The sections of this chapter correspond roughly to the various analysis steps.

The following notation is used throughout. The foreground red and green intensities are written  $R_f$  and  $G_f$  for each spot. The background intensities are  $R_b$  and  $G_b$ . The background-corrected intensities are  $R$  and  $G$ , where usually  $R = R_f - R_b$  and  $G = G_f - G_b$ . The log differential expression ratio is

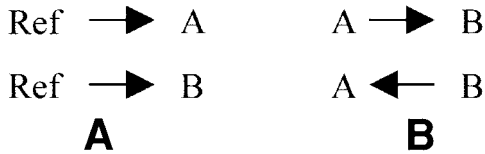


Fig. 1. Direct comparison (**B**) is more efficient than indirect comparison (**A**). Each arrow represents one microarray, the arrow by convention pointing toward the red labeled sample.

$M = \log_2 R/G$  for each spot. Finally, the log intensity of the spot is  $A = 1/2 \log_2 RG$ , a measure of the overall brightness of the spot. Note that the letter  $M$  is a mnemonic for minus, as  $M = \log R - \log G$ ; while  $A$  is a mnemonic for add, as  $A = (\log R + \log G)/2$ . It is convenient to use base 2 logarithms for  $M$  and  $A$  so that  $M$  is in units of twofold change and  $A$  is in units of twofold increase in brightness. On this scale,  $M = 0$  represents equal expression,  $M = 1$  represents a twofold change among the RNA samples,  $M = 2$  represents a fourfold change, and so on.

## 2. Experimental Design

Before carrying out a microarray experiment, one must decide how many microarray slides will be used and which mRNA samples will be hybridized to each slide. Certain decisions must be made in the preparation of the mRNA samples, such as whether the RNA from different animals will be pooled or kept separate and whether fluorescent labeling is to be done separately for each array or in one step for a batch of RNA. Careful attention to these issues will ensure that the best use is made of available resources, that obvious biases will be avoided, and that the primary questions of interest to the experimenter will be answerable. The literature on experimental design is still small. Kerr and Churchill (1) and Glonek and Solomon (2) apply ideas from optimal experimental designs to suggest efficient designs for some of the common microarray experiments. Pan et al. (3) consider sample size, and Speed and Yang (4) examine the efficiency of using a reference sample rather than direct comparison.

It is not possible to give universal recommendations appropriate for all situations, but the general principles of statistical experiment design apply to microarray experiments. In the simplest case in which the aim is to compare two mRNA samples, say  $A$  and  $B$ , it is virtually always more efficient to compare  $A$  and  $B$  directly by hybridizing them on the same arrays, rather than comparing them indirectly through a reference sample (**Fig. 1**) (4). In an experiment in which the intention is to compare several mutant types with

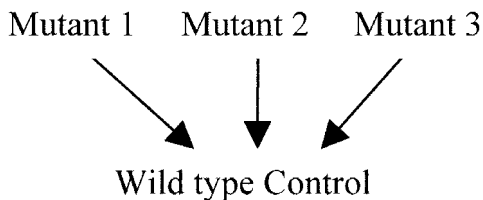


Fig. 2. WT RNA acts as a natural reference sample.

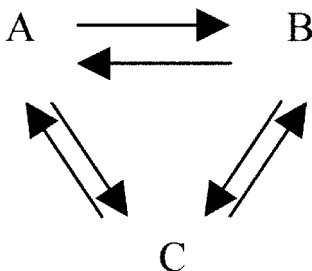


Fig. 3. Saturated design with dye-swap pairs.



Fig. 4. Possible design for a time-course experiment.

the wild type (WT), the obvious design treats the WT RNA effectively as a reference sample (Fig. 2). When more than two RNA samples are to be compared, and all comparisons are of interest, it may be appropriate to use a saturated design (Fig. 3). In time-course experiments, a loop design has been suggested (Fig. 4). For more complicated designs, with many samples to be compared, direct designs become more cumbersome, and it may be more appropriate to use a common reference sample. Factors to be considered in designing the experiment include the relative cost and availability of reference vs treatment RNA as well as the cost of the arrays themselves. In direct comparison experiments, it is generally advisable to use dye-swap pairs to minimize the effects of any gene-specific dye bias (Fig. 3).

The choice of experimental design depends not only on the number of different samples to be compared but on the aim of the experiment and on the comparisons that are of primary interest. For example, suppose the primary focus of an experiment involving a large series of tumor and normal tissues is on finding genes that are differentially expressed between the tumor and normal samples. Then direct tumor-normal comparisons on the same slide may

be the best approach. By contrast, if the focus of the analysis is to determine tumor subtypes as in **ref. 5**, then the use of a common reference RNA on each array may be better. Here the choice follows from the aim of the study, although considerations of statistical efficiency also play a role. In the first case, tumor-normal comparisons could be made indirectly, via a common reference RNA, but precision would be lost in doing so.

### 3. Image Analysis

The primary purpose of the image analysis step is to extract numerical foreground and background intensities for the red and green channels for each spot on the microarray. The background intensities are used to correct the foreground intensities for local variation on the array surface, resulting in corrected red and green intensities for each spot, which become the primary data for subsequent analysis. A secondary purpose of the image analysis step is to collect quality measures for each spot that might be used to flag unreliable spots or arrays or to assess the reproducibility of each spot value.

The first step is to image the array using an optical scanner. The array is physically scanned to produce a digital record of the red and green fluorescence emissions at each point on the array. This digital record typically takes the form of a pair of 16-bit tiff images, one for each channel, which records the intensities at each of a large number of pixels covering the array. Depending on the scanner, a number of settings can be varied to improve the sensitivity of the resulting image, one of the most common being the photomultiplier tube (PMT) voltage. The PMT voltage is usually adjusted so that the brightest pixels are just below the level of saturation ( $2^{16}$ ), thus increasing the sensitivity of the image analysis for the less bright pixels. Our own (unpublished) experiments with scanning a slide at varying PMT levels suggest that using different levels for the different channels has a negligible effect on the log ratios and ranks for the great majority of genes provided that an appropriate normalization method is used. In particular, any effect from varying the PMT levels is mitigated by using an intensity-based normalization method as described in **Subheading 4**.

The next step after scanning is to locate each spot on the slide. This is done mostly automatically by the image analysis software, using the known number and basic layout of spots on the slide, with some user intervention to increase reliability. Once a region containing a spot has been found, the image analysis software must segment the pixels into those in the spot itself (the foreground) and those in the background. There are a number of methods for doing this. The oldest method is the histogram method (**6**). A mask is chosen surrounding each spot and a histogram is formed from the intensities of the pixels within the mask. Pixels are classified as foreground if their value is greater than a threshold

and as background otherwise. Variations on this method are implemented in QuantArray software (7) for the GSI Lumonics scanner and in DeArray (8) by Scanalytics. The main advantage of this method is simplicity. However, the resulting foreground pixels are not necessarily connected, and the foreground and background intensities may be over and underestimated, respectively.

Other methods are designed to find spots as connected groups of foreground pixels. The simplest method is to fit a circle of constant diameter to all spots in the image. This is easy to implement and works nicely when all spots are circular and of the same size. In practice, this is not always the case. A generalization is to allow the circle's diameter to be estimated separately for each spot. GenePix (9) for the Axon scanner and Dapple (10) are two software programs that implement such algorithms. Dapple calculates the second differences (Laplacian) between the pixels in each small square and finds the brightest ring (circle) in the Laplacian images. Adaptive circle segmentation often works well, but spots are rarely perfectly circular, especially from noncommercial arrays.

Two methods for segmentation that do not assume circularity of the spot are the watershed method (11) and seeded region growing (12). Both methods require the specification of starting pixels or seeds. Adjoining pixels are then progressively added to the spot until adjacent spots appear to be distinctly less intense. Seeded region growing is implemented in the software Spot (13) and AlphaArray (14). Both the watershed and seeded region growing methods allow for spots of general shapes.

Once the foreground pixels have been identified, the foreground intensity for the spot is usually estimated as the average intensity of all foreground pixels, since this should be directly proportional to the number of RNA molecules hybridized to the spot's DNA. When estimating the background intensity, it is more common to use the median intensity, but first there is a decision to be made regarding which pixels to include in the local background.

One choice for the local background is to consider all pixels that are outside the spot mask but within the bounding box. Such a method is implemented by ScanAlyze (15). An alternative method used by QuantArray and ArrayVision (16) is to consider a disk between two concentric circles outside the spot mask. This method is in principle less sensitive to the performance of the segmentation procedure because the pixels immediately surrounding the spot are not used. Another method is to consider the valleys of the array, which are the background regions farthest from the nearest spot. The method is used by GenePix. It is also used by Spot as a quality control measure, although not for background correction. Since the valleys are farther from any spot than the other local background regions, the valley definition is less subject than the previous definitions to corruption by bright pixels affected by printed cDNA.

Any of the local background methods can result in background estimates that are higher than the foreground values either because of corruption by missegregated pixels or local artifacts or simply because of local variation.

The Spot software estimates the background using a nonlinear filter called morphological opening (17). The filter has the effect of smoothing the entire slide image so that all local peaks, including artifacts such as dust particles as well as the spots themselves, are removed, leaving only the background intensities. Technically, the filter consists of a local minimum filter (erosion) followed by a local maximum filter (dilation). This method of background estimation has several advantages over the use of local background regions. First, it is less variable because the background estimates are based on a large window of pixel values and yet are not corrupted by bright pixels belonging to the actual spots. Second, it yields background intensity estimates at the actual spot location rather than merely nearby. Another characteristic is that the morphological background estimates are usually lower than the local background estimates and very rarely yield background estimates that are greater than the foreground values. Yang et al. (18) compared various segmentation and background estimation methods. They found that the choice of background method has a larger impact on the log ratios of intensities than the choice of segmentation method and that morphological opening provides a more reliable estimate of background than the other methods.

After having estimated the background intensities, it is almost universal practice to correct the foreground intensities by subtracting the background,  $R = R_f - R_b$  and  $G = G_f - G_b$ . The adjusted intensities then form the primary data for all subsequent analyses. The motivation for background adjustment is the belief that a spot's measured intensity includes a contribution not specifically due to the hybridization of the target to the probe, such as nonspecific hybridization and fluorescence emitted from other chemicals on the glass. If such a contribution is present, one should measure and remove it to obtain a more accurate quantification of hybridization. An undesirable side effect of background correction is that negative intensities may be produced for some spots and hence missing values if log intensities are computed, resulting in loss of information associated with low channel intensities. Research has begun on more sophisticated methods of background adjustment that will produce positive adjusted intensities even when the background estimate happens to be larger than the foreground (19). Empirical experience suggests that local background estimates often overestimate the true background while the morphological method may underestimate, and these differences have a marked impact on the  $M$  values for less intense spots. There is a need for further research on adaptive background correction methodologies that can

produce intensities with consistent behavior regardless of the background estimator used.

#### 4. Graphic Presentation of Slide Data

It is a good idea to use routinely a variety of exploratory graphic displays to examine the results of any microarray experiment. Graphic displays can help assess the success of the experiment, guide the choice of analysis tools, and highlight specific problems.

The first and most obvious diagnostic graphic is the well-known image in which the scanned microarray output images of the Cy3 and Cy5 channels are false-colored green and red, respectively, with yellow representing an equal balance of the two. Coregistration and overlay of the two channels offer a quick visualization of the experiment, revealing information on color balance, uniformity of hybridization, spot uniformity, background, and artifacts such as dust or scratches. Overlay images also provide a rough impression of the number of genes that are differentially expressed between the two samples.

Other diagnostic plots involve plotting the numerical values of the red and green intensities. Since the raw intensities are strictly positive and vary by orders of magnitude, they should almost always be log-transformed before plotting or carrying out further analysis. There are a number of reasons for this. First, the intensities in a successful microarray experiment typically span the full 16-bit range from 0 to 65,535, with the vast majority in the lower range of values, less than 1000. If the data are not transformed, the data must, by necessity, be presented in very compressed form in the low range. Calculating log values spreads the values more evenly across the range and provides readier visualization of the data. Second, the random variation, as measured by the standard deviation (SD) of the intensities, typically increases roughly linearly with the average signal strength. Converting to logarithms tends to make the variability more constant. Third, logarithms convert the ratios  $R/G$  to differences  $M = \log R - \log G$ .

Any negative values of  $R$  or  $G$  will have to be excluded from any analysis on the logarithmic scale. Negative values can be made very rare by using an unbiased background estimator as described in **Subheading 2**. In any case, spots with negative values for either  $R$  or  $G$  are usually too faint to show evidence of differential expression and therefore tend to be of less interest in any subsequent analysis.

The most common graphic display of data from a microarray slide is a scatter plot of the two channel intensities,  $\log_2 R$  vs  $\log_2 G$ . Although such a plot is straightforward, the very high correlation between the two channel intensities always dominates the plot, making the more interesting features of

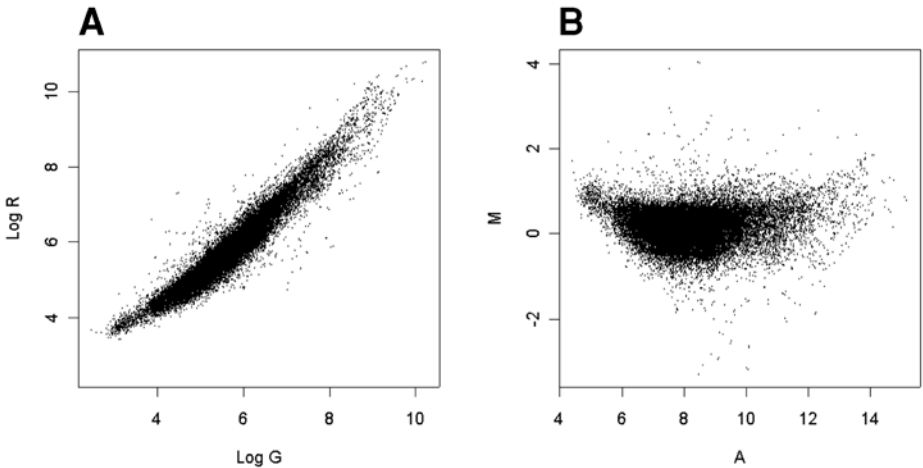


Fig. 5. (A) Scatter plot of  $\log R$  vs  $\log G$  and (B)  $MA$  plot. The central dip—an artifact—is more evident in (B) than in (A) and differentially expressed genes stand out more clearly. Data from the Nutt Lab, Walter and Eliza Hall Institute (WEHI).

the plot difficult to discern. Since the interest lies in deviations of the points from the diagonal line, it is beneficial to rotate the plot by  $45^\circ$  and to rescale the axes as in the  $MA$  plot of Dudoit et al. (20), which has the  $M$  values on the vertical axis and the intensity  $A$  values on the horizontal axis. The  $MA$  plot serves to increase the room available to represent the range of differential expression and makes it easier to see nonlinear relationships between the log intensities (Fig. 5). It also displays the important relationship between differential expression and intensity, which is used in later analysis steps.

Box plots can be useful for comparing  $M$  values among various groups. A box plot displays graphically the so-called five-number summary of a set of numbers: the three quartiles and the maximum and minimum. The central box of the plot extends from the first to the third quartile and therefore encompasses the middle 50% of the data. Figure 6 displays side-by-side box plots of the normalized  $M$  values for a series of six replicate arrays. The much longer box for array 5 shows that the interquartile range is much larger for this array. The different slides appear to be on varying scales, because of changes in PMT settings or other factors, and some rescaling seems to be called for to make the arrays more comparable.

A spatial plot of the background or  $M$  values can often reveal spatial trends or artifacts of various kinds. Figure 7 shows a spatial plot of red channel morphological background for one array. Each spot on the array corresponds to



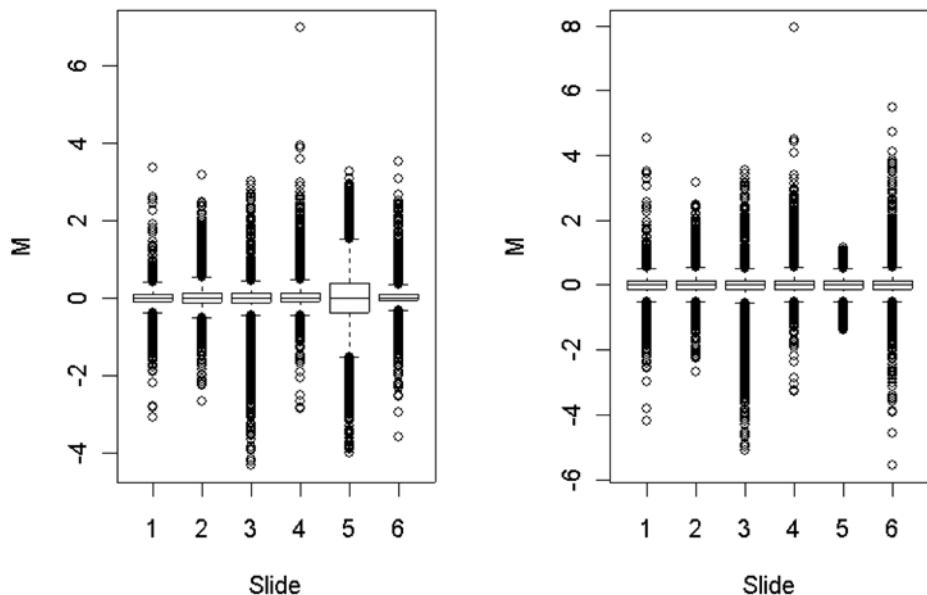


Fig. 6. (A) Side-by-side box plots of  $M$ -values from six arrays. The arrays are replicates except that three are dye-swap pairs of the others. Array 5 has a much larger spread than the others. (B) Box plots of the same arrays after scale normalization to equalize the median absolute deviation for each array. Data from the Corcoran Lab, WEHI.

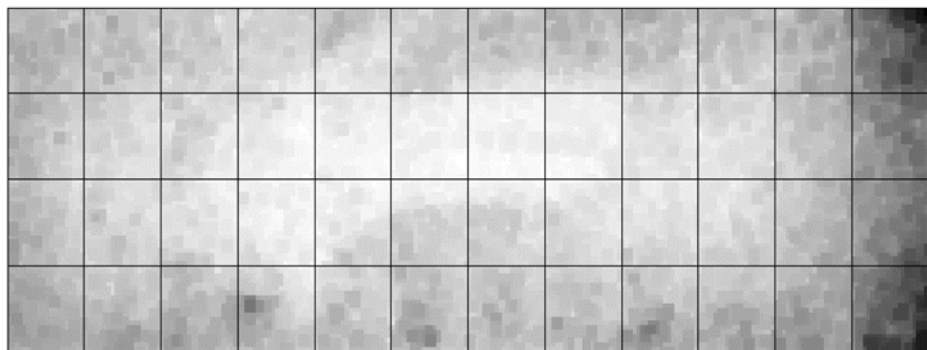


Fig. 7. Spatial plot of morphological red channel background for a microarray slide. The gray scale goes from white for low background to black for high. The background is much higher around the edges and near the right edge. The array contains 19,200 spots in a  $12 \times 4$  print-tip pattern. Data from the Scott Lab, WEHI.

one small square region on the plot. High background trends toward the edges of the plot stand out in the plot.

## 5. Normalization

The purpose of normalization is to adjust for any bias that arises from variation in the microarray technology rather than from biological differences among the RNA samples or the printed probes. Most common is red–green bias due to differences between the labeling efficiencies and scanning properties of the two fluorochromes complicated perhaps by the use of different scanner settings. Other biases may arise from variation among spatial positions on a slide or between slides. Positions on a slide may vary because of differences among the print tips on the array printer, variation over the course of the print run, or nonuniformity in the hybridization. Differences among arrays may arise from differences in print quality or from differences in ambient conditions when the plates were processed. It is necessary to normalize the intensities before any subsequent analysis is carried out.

The need for normalization can be seen most clearly in self–self experiments, in which two identical mRNA samples are labeled with different dyes and hybridized to the same slide. 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, location on the array, slide origin, and possibly other variables.

Normalization can be carried out within each array or between arrays. The simplest and most widely used within-array normalization method assumes that the red–green bias is constant on the log scale across the array. The log ratios are corrected by subtracting a constant  $c$  to get normalized values  $M = M - c$ . The global constant  $c$  is usually estimated from the mean or median  $M$  value over a subset of the genes assumed to be not differentially expressed, but many other estimation methods have been proposed. Chen et al. (6) proposed iterative estimation of  $c$  as part of one of the first proposed normalization procedures. Kerr et al. (21) and Wolfinger et al. (22) have proposed the use of analysis of variance models for normalization. These methods are equivalent to subtracting a global constant as already discussed. Global normalization is still the most widely used in spite of evidence of spatial and intensity-dependent biases in numerous experiments. We favor more flexible normalization methods based on modern regression that take into account the effects of predictor variables such as spot intensity and location.

The next level of complication, which we have always found necessary, is to allow the correction  $c$  to vary between spots in an intensity-dependent

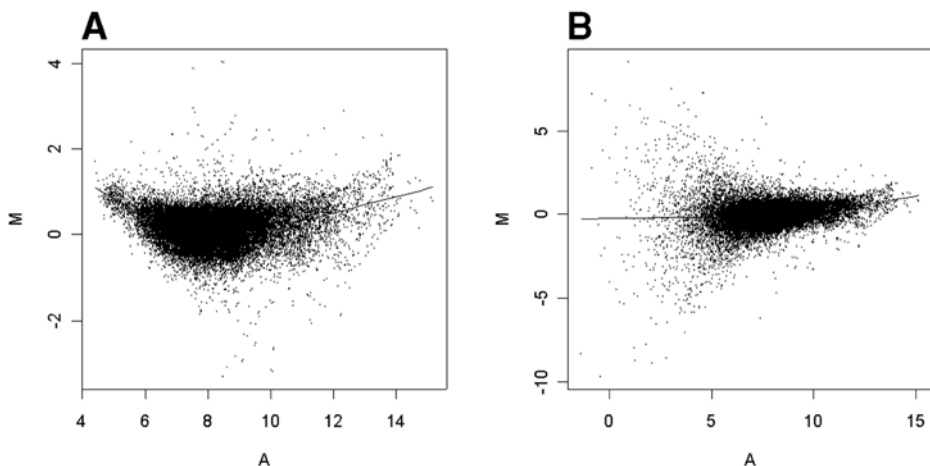


Fig. 8. Two MA plots of same microarray, (A) with morphological background and (B) with local median valley background. Data from the Nutt Lab, WEHI.

manner. In Fig. 8, a constant value for  $c$  would imply no trend between  $M$  and  $A$ . Instead, it can be seen that the majority of points lie on a curve, showing that the red–green bias depends on the intensity of the spot. Write  $c(A)$  for the height of the curve at each value of  $A$ . We normalize the  $M$  values by subtracting this curve,  $M = M - c(A)$ . The curve is estimated using a suitable robust scatter plot smoother, such as local weighted regression (loess) (23,24). A few other intensity-dependent methods have been proposed. Finkelstein et al. (25) proposed an iterative linear regression method that is essentially equivalent to what is known as robust linear regression in the statistical literature. This is similar to the aforementioned intensity-dependent normalization except that the curve  $c(A)$  is constrained to be linear. Kepler et al. (26) proposed an intensity-dependent normalization that is similar to the loess method above but uses a different local regression method.

A further generalization is to use a different curve for different regions of the array,  $M = M - c_i(A)$ , in which  $i$  indexes the region of the array. We have found subarray normalization based on the print-tip groups to be particularly useful. Not only does this allow for physical differences among the actual tips of the printer head but the print-tip groups act as a surrogate for any spatial variation across the slide (Fig. 9).

There are often substantial scale differences among microarrays, because of changes in the PMT settings or other reasons. In these circumstances, we have also found it useful to scale-normalize among arrays, a simple scaling of the  $M$  values from a series of arrays so that each array has the same median absolute deviation (Fig. 6).

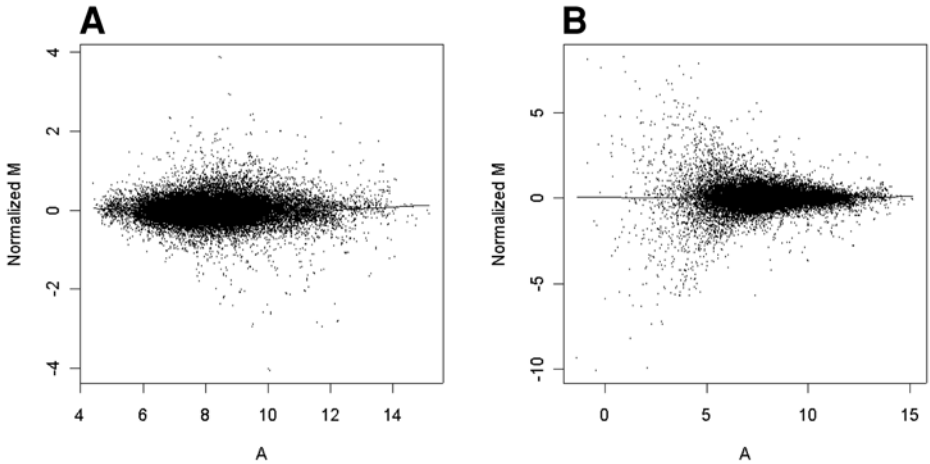


Fig. 9. Same two MA plots as in Fig. 8 after print-tip loess normalization.

In all of the discussed normalization methods, it is usual to use all or most of the genes on the array. It can be useful to modify the normalization method if a suitable set of control spots is available. A traditional method is to use housekeeping genes for normalization. However, housekeeping genes often do show sample-specific bias. Housekeeping genes are also typically highly expressed, so they will not allow the estimation of dye biases for less expressed genes when the dye bias is intensity dependent. Housekeeping genes may also not be well represented on all parts of the plate, so spatial effects may not be well estimated. The most satisfactory set of controls is a specially designed microarray sample pool (MSP) titration series. MSP is analogous to genomic DNA as control with the exception that noncoding regions are removed. Typically, a concentration titration is done to span as wide an intensity range as possible. Theoretically all labeled cDNA sequences could hybridize to this mixed probe sample, so it should be minimally subject to any sample-specific biases. On the other hand, the use of all genes for normalization offers the most stability in terms of estimating spatial and intensity-dependent trends in the data. In some cases, it may be beneficial to use a compromise between the subarray loess curves and the global titration series curve (24).

An alternative method is to select an invariant set of genes as described for oligonucleotide arrays by Schadt et al. (27) and Tseng et al. (28). A set of genes is said to be invariant if their ranks are the same for both red and green intensities. In practice, the set of invariant or approximately invariant genes is too small for comprehensive normalization. When there are sufficient invariant

genes, the use of invariant genes is similar to global intensity-dependent normalization as described above.

Subarray loess normalization is able to correct for a variety of spatial and intensity-dependent biases. It is advisable, however, using the exploratory plots mentioned in **Subheading 4.** to check whether other systematic effects exist in the data of which account should be made before the primary data analysis is carried out.

## 6. Quality Measures

### 6.1. Array Quality

It is important to assess the quality of the data obtained from each microarray experiment on a global array basis as well as on an individual spot basis. The quality of the results from each microarray will vary with cDNA purity, variations in the printing process, RNA quality, success in carrying out the hybridization protocols, and scanning effectiveness. A simple global assessment of quality is found in the distribution of log-intensity values in each of the two channels across the spots on the slide. Pixel intensities are usually scaled to be between 0 and 16 on the log base 2 scale. If the observed intensities fail to use the greater part of this scale, this is a strong indication that something is wrong; possibly the hybridization has failed. More precisely, we expect the intensity  $A$  values to span the majority of the response range. Control spots should be represented in this spread: null control spots such as blanks and printing buffers should have low intensities while housekeeping genes and titration series spots should show a range of higher intensities. At the same time, the intensity values should not be too dense around the largest value, suggesting that the scanner has been set too high and pixels have been saturated. This will lose discrimination and linearity of response on the log scale.

In most experiments, the great majority of genes should not be differentially expressed, so the range of  $M$  values for the bulk of genes should be much less than the range of  $A$  values. On an  $MA$  plot, the bulk of points should follow an elongated shape if the  $M$  and  $A$  axes are on a similar scale. If morphological background estimation has been used, the  $MA$  plot will typically follow an elongated comet shape with a long tail on the right. If a local background estimate has been used, the  $MA$  plot will typically follow a fan shape with, again, a long tail on the right (**Figs. 8 and 9**).

The ability of normalized intensities to follow a full range of values partly depends on the background level. A good-quality array will typically have relatively low background intensities and in particular a low average ratio of background to foreground intensity across the spots on the array.

The exploratory plots described in **Subheadings 4.** and **5.** will give an impression of array quality. The false-colored and spatial plots are particularly useful for judging spatial variation. Marked variation in the red–green dye bias across different parts of the array is an indication of quality problems. Although the subarray normalization will partly correct for spatial variation, strong variation will persist even after normalization and is an indication of problems with the experimental protocol.

## **6.2. Spot Quality**

If the overall quality of an array is satisfactory, then it becomes relevant to assess the quality of individual spots. There are two broad approaches to this. The first is to assess the quality of a spot according to its physical characteristics. The second is to assess the quality of a spot according to whether the observed intensities for that spot are in general agreement with those from other spots printed with the same gene and hybridized with the same RNA. The first approach is an attempt to predict the repeatability of each spot's  $M$  value. Spots with low-quality scores are supposed to be less repeatable and are typically removed from subsequent analysis. The second is a data-based approach that observes repeatability empirically given a suitable series of replicate arrays or duplicate spots on the same array. A fully integrated approach to quality will include both approaches.

The first approach constructs quality measures for each spot from information collected by the image analysis program. Most image analysis programs routinely record a variety of spot details. These might include heterogeneity measures, such as SDs or interquartile ranges across pixels in the foreground and local background, as well as more basic details, such as spot area, perimeter, and location. Further quantities, such as circularity (area/perimeter<sup>2</sup>) or interpixel coefficient of variation (CV) (SD/mean), can obviously be derived from the basic measures. In general, spots can be expected to be unreliable if they are very small or very large relative to the bulk of spots on the array, if they are markedly noncircular, if the background intensities are high, if the signal-to-noise ratio (SNR) is low, or if the foreground or background regions are very heterogeneous. Examples of such work can be found in **refs. 10, 14, 29, and 30.** Buhler et al. (**10**) reject or accept spots based on brightness and position of the spot center. Brown et al. (**29**) consider pixel-level variability for each spot. Yang et al. (**30**) omit points with low intensities. Wang et al. (**14**) measure spot quality using a composite index involving spot size, SNR, level and heterogeneity of background, and saturation of pixels.

Examples of the more empirical quality approach can be found in **refs. 28 and 31.** Nadon et al. (**31**) reject spots that are judged to be outliers relative to a normal distribution for a series of  $M$  values from replicate slides. Tseng

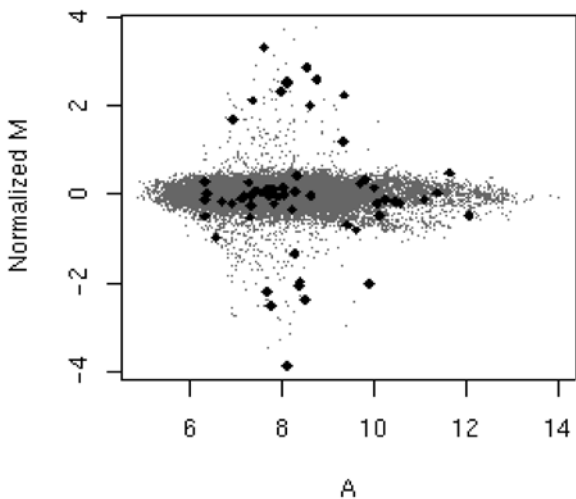


Fig. 10. Normalized  $MA$  plot for one microarray showing that very small spots are more variable than larger spots. Spots with areas  $<75$  pixels are highlighted. Data from the Corcoran Lab, WEHI.

et al. (28) filter out genes according to the variability of duplicate spots on the same slide.

In all of these works, spots that are flagged as low quality are omitted from the primary analysis. Naturally this improves the look of the data, as indicated by a range of visual diagnostics. However, spots do not go from “good” to “bad” in a sharp way, and the cutoffs that are used to judge low quality are inevitably somewhat arbitrary. A more satisfactory approach would be to give less weight to lower-quality spots in a graduated way, with excellent spots getting full weight, down to excluding really bad spots entirely.

In the empirical quality approach, a more systematic approach to handling outlier spots can be achieved by using the robust estimation procedures discussed in **Subheading 7**. Robust estimators of location and scale will automatically downweight any  $M$  value that is discordant with other comparable values. Robust methods downweight outlying  $M$  values in a graduated way and avoid the need to choose an arbitrary cutoff.

In the physically based approach to quality, a graduated approach is more difficult. Ideally, quality measures should be found that predict the between-slide variance of the  $M$  values. Spots can then be weighted inversely according to the predicted variances. An obvious treatment of spot area, e.g., would be to weight small spots directly proportional to their area, such as  $w = a/a_F$ , in which  $w$  is the weight,  $a$  is the area of the spot in pixels, and  $a_F$  is the area of a full-sized spot. **Figure 10** demonstrates empirically that spots with small areas

can be substantially more variable than larger spots. The correct treatment of other measures such as SNR is less obvious, although Brown et al. (29) and Wang et al. (14) have promising results. Wang et al. (14) demonstrated graphically an increasing trend relationship between spot variance and the spots' composite quality measure. However, we have not observed the same variance trends using data from our own institutions, and the results may be sensitive to the particular image analysis and background correction method that was used.

## 7. Selecting Differentially Expressed Genes

### 7.1. Ranking Genes

One of the core goals of microarray data analysis is to identify which of the genes show good evidence of being differentially expressed. This goal has two parts. The first is to select a statistic that will rank the genes in order of evidence for differential expression, from strongest to weakest evidence. The second is to choose a critical value for the ranking statistic above which any value is considered to be significant. The first goal is the more important of the two and, as it turns out, also the easier. The primary importance of ranking arises from the fact that only a limited number of genes can be followed up in a typical biological study. In many microarray studies, the aim is to identify a number of candidate genes for confirmation and further study. It will usually be practical to follow up only a limited number of genes, say 100, so it is most important to identify the 100 most likely candidates. The complete list of all genes that can be considered statistically significant may be of less interest if this list is too large to be followed up.

For simplicity, we assume in this section that we have data from the simplest possible experiment. We assume that we have a series of  $n$  replicate arrays on which samples A and B have been hybridized, and we wish to identify which genes are differentially expressed. Many data analysis programs sort the genes according to the absolute level of  $\bar{M}$ , in which  $\bar{M}$  is the mean of the  $M$  values for any particular gene across the replicate arrays. This is known to be a poor choice because it does not take into account the variability of the expression levels for each gene (32,33). The shortcoming of the method is that the variability of the  $M$  values over replicates is not constant across genes and genes with larger variances have a good chance of giving a large  $\bar{M}$  statistic even if they are not differentially expressed. A better choice is to rank genes according to the absolute value of the  $t$  statistic:

$$t = \frac{\bar{M}}{s/\sqrt{n}}$$



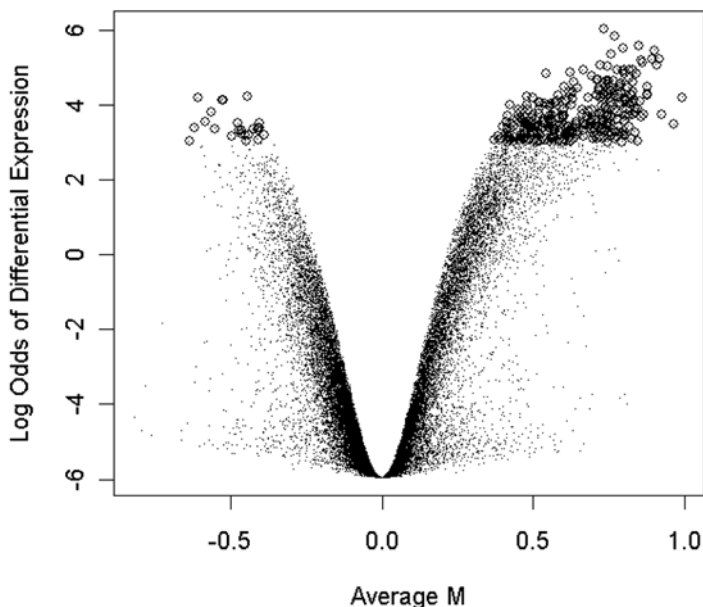


Fig. 11. Volcano style plot of the empirical Bayes  $B$  statistic for a series of six replicate arrays. Genes with log odds of differential expression greater than three have been highlighted for follow-up and confirmation. Data from the Corcoran Lab, WEHI.

in which  $s$  is the SD of the  $M$  values across the replicates for the gene in question, since this incorporates a different variability estimate for each gene. An added advantage of the  $t$  statistic is that it introduces some conservative protection against outlier  $M$  values and poor-quality spots. Any  $M$  value that is an outlier will give rise to a large SD  $s$ , which will usually prevent the gene in question from being spuriously identified as differentially expressed.

The ordinary  $t$  statistic is still not ideal because a large  $t$  statistic can be driven by an unrealistically small value for  $s$ . The shortcoming of the  $t$  statistic is the opposite of that of  $\bar{M}$ . Genes with small sample variances have a good chance of giving a large  $t$  statistic even if they are not differentially expressed. A suitable compromise between the  $\bar{M}$  and  $t$  statistics is therefore desirable.

Lönnstedt and Speed (33) adopt a parametric empirical Bayes approach to the problem of identifying differentially expressed genes. They produce a  $B$  statistic that is an estimate of the posterior log odds that each gene is differentially expressed (Fig. 11). Subject to the parametric assumptions being valid for the data, values for the  $B$  statistic greater than zero correspond to a greater than 50–50 chance that the gene in question is differentially expressed.

The  $B$  statistic is equivalent for the purpose of ranking genes to the penalized  $t$  statistic:

$$t = \frac{\bar{M}}{\sqrt{(a + s^2)/n}}$$

in which the penalty  $a$  is estimated from the mean and SD of the sample variances  $s^2$ . Tusher et al. (32) and Efron et al. (34) have used penalized  $t$  statistics of the form

$$t = \frac{\bar{M}}{(a + s)/\sqrt{n}}$$

when assessing differentially expressed genes for oligonucleotide microarrays. This differs slightly from the previous statistic in that the penalty is applied to the sample SD  $s$  rather than to the sample variance,  $s^2$ . Tusher et al. (32) choose  $a$  to minimize the coefficient of variation of the absolute  $t$ -values while Efron et al. [34] choose  $a$  to be the 90th percentile of the  $s$  values. These choices are driven by empirical rather than theoretical considerations. Efron et al. (34) use the above  $t$  value as the basis for a nonparametric empirical Bayes method leading to an estimated log odds that each gene is differentially expressed. Lönnstedt and Speed (33) show in a simulation that both forms of penalized  $t$  statistic are far superior to the mean  $\bar{M}$  or to the ordinary  $t$  statistic for ranking differentially expressed genes.

The penalized  $t$  statistics can be extended in several natural ways to apply to more general experimental situations. If there are missing values for some arrays, perhaps because low-quality spots have been flagged for removal, then the value  $n$  in the denominator will reflect the actual number of observations for each gene rather than the total number of arrays.

The  $t$  statistic also extends naturally to more complicated experiment designs. For example, we might use a penalized two-sample  $t$  statistic if we are comparing samples A and B through a reference rather than directly on the same arrays. In that case, there will be  $n_A$  replicate arrays comparing sample A with reference RNA and  $n_B$  replicate arrays comparing B with the same reference and a two-sample  $t$  statistic:

$$t = \frac{\bar{M}_A - \bar{M}_B}{s_* \sqrt{1/n_A + 1/n_B}}$$

in which  $s_* = \sqrt{a + s^2}$  is the penalized pooled sample SD might be used. Here,  $\bar{M}_A$  and  $\bar{M}_B$  are the average of the  $M$  values for the two groups of arrays. For more complicated experimental designs, a multiple regression model will in general be estimated for each gene as, e.g., in ref. 35. In the general

case, differential expression can be judged using a penalized  $t$  statistic of the form

$$t = \frac{b}{s_* \text{se}}$$

in which  $b$  is a regression coefficient estimated by the multiple regression that discriminates between the RNA samples of interest;  $\text{se}$  is the unscaled standard error for  $b$  returned by the multiple regression; and  $s_* = \sqrt{a + s^2}$ , in which  $s$  is the residual SD returned by the multiple regression. See Lönnstedt et al. (36), who indicate the extension of the empirical Bayes  $B$  statistic to general experimental designs.

Another direction in which the  $t$  statistic can be generalized is to replace the sample mean  $\bar{M}$  and sample SD  $s$  with location and scale estimators that are robust against outliers. This extension is very useful for microarray data because it is impossible to guarantee or adjust for the data quality of every individual spot. The general idea of robust estimation is to replace  $\bar{M}$  and  $s$  with values that behave very much like  $\bar{M}$  and  $s$  when the data actually are normally distributed but are insensitive to a small proportion of aberrant observations (37) (38). For general microarray experiments, a robust multiple regression can be computed for each gene and a penalized  $t$  statistic formed from the robust versions of  $b$ ,  $s$ , and  $\text{se}$ .

## 7.2. Assigning Significance

Having ranked the genes on the basis of a suitable statistic, the next step is to choose a cutoff value above which genes will be flagged as significant. The crux here is the need to control for the massive level of multiple testing inherent in the need to conduct a test for each gene.

A simple graphic method for assigning significance that is applicable even for single microarray experiments is to display the sorted genewise test statistics in a normal or  $t$  distribution probability plot. The bulk of the genes should follow an approximate straight line on the plot. Genes whose points deviate markedly from the line are identified by the method as significantly differentially expressed. Unfortunately, this remains an informal method because the implicit assumptions of normality for the  $M$  values and independence among genes are unlikely to be satisfied. The method tends in practice to overestimate the number of differentially expressed genes somewhat because the null distribution of the  $M$  values tends to have heavier tails than does the normal distribution. Tusher et al. (32) do use a variant of this method in conjunction with other multiple testing methods.

Shaffer (39) has reviewed the issues involved in multiple testing. The most stringent approach to multiple testing is to control for familywise error rate,

which is the probability of at least one false positive among the genes selected as differentially expressed regardless of what configuration of the genes truly are differentially expressed. Dudoit et al. (20) consider a design for which two-sample  $t$  statistics are appropriate, comparing two RNA samples indirectly through a reference sample. They give a rigorous method for controlling the familywise error rate using a resampling method (40) that computes a step-down adjusted  $p$  value for each gene. Unfortunately, this method requires a moderate to large number of microarrays to give useful results. If, e.g., there are 15,000 distinct genes to be tested, then the method requires at least 16 microarrays to be able to detect differentially expressed genes because of the granularity of  $p$  values computed by resampling.

It can be argued that controlling the familywise error rate is unnecessarily stringent in the microarray testing context, because falsely selecting a handful of genes as differentially expressed will not be a serious problem if the majority of significant genes are correctly chosen. A less stringent and therefore more powerful method is to control the false discovery rate, defined to be the expected proportion of errors among the genes selected as significantly differentially expressed (41). Tusher et al. (32), Efron et al. (34) and Storey and Tibshirani (42) take an alternative approach to the false discovery rate. Rather than trying to control the false discovery rate, they treat it as an exploratory tool. After choosing the subset of differentially expressed genes by other means, they estimate the false discovery rate among this subset using a resampling method. Estimation of the false discovery rate, which is described in detail in ref. 42, relies formally on some assumptions about dependence among the genes that are difficult to verify in practice. However, this is a very promising approach.

The empirical Bayes methods of Efron et al. (34) and Lönnstedt and Speed (33) do not allow absolute cutoff values because the overall proportion of differentially expressed genes is an indeterminate parameter in the models. In using these methods, one has to specify a value in advance, say 1%, for the overall proportion of differentially expressed genes, including those that are detected and those that are not. Moving this value up or down will move all the posterior odds of differential expression up or down by a similar amount but will not change the order in which the genes are ranked. Efron et al. (34) suggest that the posterior odds can be calibrated post hoc by estimating the false discovery rate as discussed.

There exist other methods that can in principle give absolute cutoff values for differential expression (6,22,43,44), in some cases even for as few as a single microarray in the experiment (6,44). The price that is paid to achieve

such results is that strong global distributional assumptions must be made about the red and green intensities. These assumptions are inevitably more simple than reality and seem to us too strong for routine data analysis.

## 8. Classification

Two very important uses for microarray data are to generate gene expression profiles that can (1) discriminate among different known cell types or conditions, e.g., between tumor and normal tissue or among tumors of different types, and (2) to identify different and previously unknown cell types or conditions, e.g., new subclasses of an existing class of tumors. The same problems arise when it is genes that are being classified: one might wish to assign an unknown cDNA sequence to one of a set of known gene classes, or one might wish to partition a set of genes into new functional classes on the basis of their expression patterns across a number of samples.

These dual tasks have been described as class prediction and class discovery in an influential article by Golub et al. (45). In the machine-learning literature they are known as supervised and unsupervised learning, the learning in question being of the combinations of measurements—here gene expression values—that assign units to classes. In the statistical literature, they are known as discrimination and clustering. The distinction is important. Clustering or unsupervised methods are likely to be appropriate if classes do not exist in advance. If the classes are preexisting, then discriminant analysis or supervised learning methods are more appropriate and more efficient than clustering methods.

There are many powerful techniques for class prediction in the statistical and machine-learning literatures (46–48). Such techniques invariably begin with data for which the existing class assignments are known, the so-called training set of units. These techniques can be effective even if some of the existing class assignments of the units are wrong or if there are unknown subclasses that would refine the existing classes. Indeed, there are well-established methods of evaluating the quality of prediction methods that, at the same time, check the assignment of individual units in the training set.

Cluster methods tend to be overused in microarray data analysis relative to discrimination methods. A common practice, e.g., is to suppress existing class assignments, use an unsupervised learning technique to define new classes and assign the units to these classes, and then see how well the existing class assignments are reflected in the new classes. A more direct and efficient approach would be to use a supervised method to discriminate the classes in conjunction with a method such as cross validation to evaluate the repeatability of the

results on new data. The efficiency of direct discrimination over clustering becomes increasingly important as the prediction problem becomes more challenging.

Discrimination methods include linear discriminant analysis in various forms (46), nearest-neighbor classifiers (48), classification trees (49), aggregating classifiers (50–51), neural networks (48), and support vector machines (52–53). The first three methods are simple to apply once the genes have been filtered. The other methods are more sophisticated and require considerable skill in their application.

Dudoit et al. (54) compared the performance of different discrimination methods for the classification of tumors using gene expression data from three recent studies. Their main conclusion is that simple classifiers such as linear discrimination and nearest neighbors performed remarkably well compared to more sophisticated prediction methods such as aggregated classification trees.

There are factors other than accuracy that contribute to the merits of a given classifier. These include simplicity and insight gained into the predictive structure of the data. Linear discriminant methods are easy to implement and had low error rates in the study by Dudoit et al. (54) but ignore interactions among genes. Nearest-neighbor classifiers are simple, intuitive, and had low error rates compared with more sophisticated classifiers. While they are able to incorporate interactions among genes, they do so in a “black-box” way and give very little insight into the structure of the data. By contrast, classification trees are capable of exploiting and revealing interactions among genes. Trees are easy to interpret and yield information on the relationship between predictor variables and responses by performing stepwise variable selection. However, classification trees tend to be unstable and lacking in accuracy. Their accuracy can be greatly improved by aggregation (bagging or boosting). As more data become available, one can expect to observe an improvement in the performance of aggregated classifiers relative to simpler classifiers, because trees should be able to correctly identify interactions.

The use of clustering methods to identify group coregulated genes is an area of very active research, stimulated by influential articles such as those by Eisen et al. (55) and Alizadeh et al. (5). The most popular clustering methods are nicely reviewed by Quackenbush (53). Recent work includes that of Hastie et al. (56), who formed clusters around the largest principal components of the data; Lazzeroni and Owen (57), who proposed models in which each gene can belong to more than one cluster to none at all as different characteristics are considered; Parmigiani et al. (58), who considered more general probabilistic models; and Lin et al. (35), who clustered genes on the basis of regression coefficients estimated by a linear model.

## 9. Conclusion

Attention to statistical issues at each stage of microarray data analysis can ensure that the best use is made of available resources, that biases of various sorts are avoided, and that reliable conclusions are made. Software to carry out the analyses discussed in this chapter is described by Dudoit et al. (59) and Dudoit and Yang (60).

## Acknowledgments

Thanks to Matthew Ritchie and Natalie Thorne for discussions and to Matthew Ritchie for **Fig. 10**. We are also grateful to Drs. Lynn Corcoran, Stephen Nutt, and Hamish Scott for permission to use data from their laboratories at WEHI.

## References

1. Kerr, M. K. and Churchill, G. A. (2001) Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.
2. Glonek, G. F. V. and Solomon, P. J. (2002) Factorial designs for microarray experiments. Technical Report, Department of Applied Mathematics, University of Adelaide, Australia.
3. Pan, W., Lin, J., and Le, C. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.* **3**(5), research0022.1–0022.10.
4. Speed, T. P. and Yang, Y. H. (2002) Direct versus indirect designs for cDNA microarray experiments. Technical Report 616, Department of Statistics, University of California, Berkeley.
5. Alizadeh, A. A, Eisen, M. B., Davis, R. E., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**(6769), 503–511.
6. Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997) Ratio based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **2**, 364–374.
7. QuantArray Analysis Software. <http://lifesciences.perkinelmer.com>.
8. Scanalytics MicroArray Suite. [www.scanalytics.com](http://www.scanalytics.com).
9. GenePix Pro microarray and array analysis software, Axon Instruments [www.axon.com](http://www.axon.com).
10. Buhler, J., Ideker, T., and Haynor, D. (2000) Dapple: improved techniques for finding spots on DNA microarrays. CSE Technical Report UWTR 2000-08-05, University of Washington.
11. Beucher, S. and Meyer, F. (1993) The morphological approach to segmentation: the watershed transformation: mathematical morphology in image processing. *Opt. Eng.* **34**, 433–481.
12. Adams, R. and Bischof, L. (1994) Seeded region growing. *IEEE Trans. Pattern Anal. Machine Intelligence* **16**, 641–647.

13. Buckley, M. J. (2000) *Spot User's Guide*, CSIRO Mathematical and Information Sciences, Sydney, Australia. [www.cmis.csiro.au/iap/Spot/spotmanual.htm](http://www.cmis.csiro.au/iap/Spot/spotmanual.htm).
14. Wang, X., Ghosh, S., and Guo, S.-W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.* **29**(15), E75–5.
15. Eisen, M. B. (1999) *ScanAlyze User Manual*, Stanford University, Palo Alto. <http://rana.lbl.gov>.
16. ArrayVision, Imaging Research. <http://imaging.brocku.ca>.
17. Soille, P. (1999) *Morphological Image Analysis: Principles and Applications*, Springer, New York.
18. Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002) Comparison of methods for image analysis on cDNA microarray data. *J. Computat. Graph. Stat.* **11**, 108–136.
19. Kooperberg, C., Fazio, T. G., Delrow, J. J., and Tsukiyama, T. (2002) Improved background correction for spotted cDNA microarrays. *J. Computat. Biol.* **9**, 55–66.
20. Dudoit, S., Yang, Y. H., Speed, T. P., and Callow, M. J. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–140.
21. Kerr, M. K., Martin, M., and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. *J. Computat. Biol.* **7**, 819–837.
22. Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Computat. Biol.* **8**, 625–637.
23. Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001) Normalization for cDNA microarray data, in *Microarrays: Optical Technologies and Informatics* (Bittner, M. L. Chen, Y. Dorsel, A. N. and Dougherty, E. R., eds.), Proceedings of SPIE, vol. 4266.
24. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**(4), E15.
25. Finkelstein, D. B., Gollub, J., Ewing, R., Sterky, F., Somerville, S., and Cherry, J. M. (2001) Iterative linear regression by sector, in *Methods of Microarray Data Analysis. Papers from CAMDA 2000*. (Lin S. M. and Johnson, K. F., eds.) Kluwer Academic, pp. 57–68.
26. Kepler, T. B., Crosby, L., and Morgan, K. T. (2000) Normalization and analysis of DNA microarray data by self-consistency and local regression, Santa Fe Institute Working Paper, Santa Fe, NM.
27. Schadt, E. E., Li, C., Ellis, B., and Wong, W. H. (2002) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* **84**(Suppl. 37), 120–125.
28. Tseng, G. C., Oh, M.-K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557.



29. Brown, C. S., Goodwin, P. C., and Sorger, P. K. (2000) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl. Acad. Sci. USA* **98**, 8944–8949.
30. Yang, M. C., Ruan, Q.-G., Yang, J. J., Eckenrode, S., Wu, S., McIndoe, R. A., and She, J.-X. (2001) A statistical procedure for flagging weak spots greatly improves normalization and ratio estimates in microarray experiments. *Physiol. Genomics* **7**, 45–53.
31. Nadon, R., Shi, P., Skandalis, A., Woody, E., Hubschle, H., Susko, E., Rghei, N., and Ramm, P. (2001) Statistical methods for gene expression arrays, in *Microarrays: Optical Technologies and Informatics Proceedings of SPIE*, vol. 4266, (Bittner, M. L., Chen, Y., Dorsel, A. N., and Dougherty, E. R. eds.), pp. 46–55.
32. Tusher, V., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5124.
33. Lönnstedt, I. and Speed, T. P. (2002) Replicated microarray data. *Statistica Sinica* **12**, 31–46.
34. Efron B., Tibshirani, R., Storey J. D., and Tusher V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160.
35. Lin, D. M., Yang, Y. H., Scolnick, J. A., Brunet, L. J., Peng, V., Speed, T. P., and Ngai, J. (2002) A spatial map of gene expression in the olfactory bulb, Department of Molecular and Cell Biology, University of California, Berkeley.
36. 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. Technical Report, Department of Mathematics, Uppsala University, Sweden.
37. Huber, P. J. (1981) *Robust Statistics*, Wiley, New York.
38. Marazzi, A. (1993) *Algorithms, Routines and S Functions for Robust Statistics*, Wadsworth & Brooks/Cole, CA.
39. Shaffer, J. P. (1995) Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–576.
40. Westfall, P. H. and Young, S. S. (1993) *Re-Sampling Based Multiple Testing*, Wiley, New York.
41. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser.* **57**, 289–300.
42. Storey, J. D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence with applications to DNA microarrays, Technical Report, Department of Statistics, Stanford University.
43. Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Computat. Biol.* **7(6)**, 805–817.
44. Newton, M. A., Kenzioriski, 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. Biol.* **8**, 37–52.

45. 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.
46. Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979) *Multivariate Analysis*, Academic, London.
47. McLachlan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.
48. Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.
49. Breiman, L., Friedman, J. H., Olsen, R. A., and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth, Monterey, CA.
50. Breiman, L. (1996) Bagging predictors. *Machine Learning* **24**, 123–140.
51. Breiman, L. (1998) Arcing classifiers. *Ann. Stat.* **26**, 801–824.
52. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97**, 262–267.
53. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427.
54. Dudoit, S., Fridlyand, J., and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77–87.
55. Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14,863–14,868.
56. Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000) “Gene shaving” as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **1**(2), 0003.1–0003.21.
57. Lazzeroni, L. and Owen, A. B. (2002) Plaid models for gene expression data. *Statistica Sinica* **12**, 61–86.
58. Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002) A statistical framework for expression-based molecular classification in cancer, Technical Report, Department of Biostatistics, Johns Hopkins University.
59. Dudoit, S., Yang, Y. H., and Bolstad, B. (2002) Using R for the analysis of DNA microarray data. *R News* **2**(1), 24–32.
60. Dudoit, S. and Yang, Y. H. (2003) Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data, in *The Analysis of Gene Expression Data: Methods and Software* (Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L., eds.), Springer, New York, in press.