Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Computational Statistics & Data Analysis 44 (2003) 313–338
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Transformations, background estimation, and process effects in the statistical analysis of microarrays☆

Karen Kafadar[a,*], Tzulip Phang[b]

[a] *Department of Mathematics, University of Colorado-Denver, P.O. Box 173364, CB170, Denver, CO 80217-3364, USA*

[b] *Department of Pharmacology, University of Colorado, Health Sciences Center, Denver, CO 80262, USA*

Received August 2002; accepted March 2003

Abstract

Microarray technology has made available large data sets that can provide information on gene expression when cells are subjected to various treatments. Before proceeding with a formal statistical analysis, many biological and procedural aspects should be considered. These aspects may guide the analysis and subsequent statistical inference. Several of these issues are discussed in connection with the analysis of oligonucleotide and cDNA microarray experiments. The particular focus in this article is on effects caused by the cDNA slide manufacturing process, appropriate transformations of the data, and on adjustments for background. A prescription for the analysis of microarray data is proposed and demonstrated using data from a cDNA experiment comparing the genetic expressions in two mouse cell lines; a candidate set of genes is identified for further study. The prescription may be modified for oligonucleotide microarray data.

Keywords: Background variation; Fluorescence; Lognormal distribution; Median polish; Process variation; Smoothing; Tukey's g-family of distributions

☆ This work is supported in part by a grant from the Agency for Healthcare Research and Quality R01-HS-10917-01, awarded to the University of Colorado-Denver (Kafadar), and by a grant from the National Institutes of Health AA13162-01, awarded to the University of Colorado Health Sciences Center (Phang).

* Corresponding author. Tel.: 3035562547; fax: 3035568550.
E-mail addresses: kk@math.cudenver.edu (K. Kafadar), tzu.phang@uchsc.edu (T. Phang).

0167-9473/03/$ - see front matter
doi:10.1016/S0167-9473(03)00069-0

1. Introduction

Increasing data storage and computational power have led to valuable collaborations between informatics and statistics, leading to enhanced knowledge in various scientific fields. Many statistical and computational issues arise in these studies, such as data collection, design, analysis, statistical and practical significance, and numerically efficient computational procedures. Biological information science, or bioinformatics, is one area where these issues arise, particularly in genomics. Microarray technology makes possible the quantitative analysis of gene expression and variation in thousands of genes from a single experiment. Some less obvious but potentially important statistical challenges relate to the biological and statistical significance of gene expression from a limited number of experimental conditions, transformations of non-normal data, weak signals amidst possibly high noise, estimation of and correction for background variation, instrumentation and measurement errors, variation in signal strengths both within and across experiments, and multiple dependent statistics used for testing significance of measured gene expression levels. Other important considerations precede formal hypothesis tests, such as the identification and quantification of errors in the cellular material, the manufacturing process that produces the slides, and the instrumentation that provides the measurements. These challenges arise in the analysis of data from both oligonucleotide microarrays and spotted cDNA microarrays. Both technologies are used in experiments to identify gene expression levels in the target sample(s), whose messenger RNA (mRNA) is labeled with a fluorescent dye. The actual measurements are fluorescence intensities as recorded by a laser scanner; higher intensities indicate higher mRNA concentrations, and thus more binding, or hybridization, and thus possibly more gene expression. Because genes contain the code for protein production, these measured intensities are believed to correlate somewhat with the proteins that the organism manufactures. [More direct measures of protein production are obtained from proteomics (Nicolls et al., 2003), but the analysis of data from such experiments is not discussed here.]

Knowledge of the process is essential for understanding the data and hence for an appropriate analysis from either type of experiment. This article describes some of these process issues, with particular emphasis on the biological and mechanical processes, on data transformations, and on valid estimation and adjustment for background intensities. Experimental design issues are not addressed here, as they have been well presented in other articles such as Lee et al. (2000), Kerr and Churchill (2001a,b), and Yang and Speed (2002b). General descriptions of both types of microarrays are available elsewhere; e.g., for cDNA slides, see articles by Yang et al. (2002a,c), Yang and Speed (2002b), Amaratunga and Cabrera (2001), Brazma and Vilo (2000), and the web site for the Speed group, `http://www.stat.berkeley.edu/terry/zarray/home/index.html`; for oligonucleotide arrays, see Affymetrix (2000, 2002), Efron et al. (2001), Tusher et al. (2001), and Irizarry et al. (2003). In a brief tutorial, Satagopan and Panageas (2003) explain the underlying science and describe the two types of arrays; the statistical analysis focuses on methods for clustering genes. Each technology offers advantages and disadvantages, and each yields measurements of mRNA concentration in a gene at the time that the mRNA was harvested, together with errors from various

sources that cause bias and uncertainty in these measurements. Brief descriptions of these two technologies follow in Section 2.

Section 3 describes several sources of error in microarray experiments. The implications of erroneous background adjustments in assessing statistical significance of a gene and transformations of the expression levels into a form that is more amenable to analysis are both discussed in Section 4. Section 5 describes the analysis of data from an experimental cDNA slide that contains 16,928 mouse genes. Comments and discussion for further work are presented in Section 6.

2. Brief description of cDNA and oligonucleotide arrays

A cDNA slide contains an array of "spots", arranged in "blocks" of rows and columns, where each spot consists of thousands of nominally identical probes. The process by which the probes, corresponding to specific genes, were obtained for these slides is quite complex and, once completed, yields a gene "library" for a given laboratory, which is then used for thousands of slides. Thus, errors in the gene library persist throughout all slides manufactured from it. A target sample contains a mixture of two types of cells, control and experimental, whose messenger RNA (mRNA), which degrades rapidly, is reverse-transcribed into the more stable cDNA (c is for complementary) and then labeled with two different fluorophores: the control cells are often labeled with Cyanine 3, or Cy3 (green dye), and the experimental cells (e.g., cells that have subjected to some sort of treatment, such as stress, heat, radiation, or chemicals) are often labeled with Cyanine 5, or Cy5 (red dye). When mRNA concentration is high in the genes of these cells, their cDNA will bind to their corresponding probes on the spotted cDNA slide; an optical detector in a laser scanner will measure the fluorescence at wavelengths corresponding to the green and red dyes (532 and 635 nm, respectively). Good experimental design will interchange the dyes in a separate experiment to account for imbalances in the signal intensities from the two types of fluorophores and the expected degradation in the cDNA samples between the first scan at 532 nm (green) and second scan at 635 nm (red). The ratio of the relative abundance of red and green dyes at these two wavelengths on a certain spot indicates relative mRNA concentration between the experimental and control genes. Thus, the gene expression levels in the target cells can be compared directly with those from the control cells.

Oligonucleotide arrays circumvent the possible inaccuracies that can arise in the preparation of a gene "library" and the control and experimental samples for spotted array slides, by using 16–20 predefined and pre-fabricated sequences of (usually 25) nucleotides for each gene. Rather than circular-shaped spots, these probes are deposited onto the chip in square-shaped cells. According to the Affymetrix© User's Manual 4.0 (Affymetrix, 2000), probe cells are 24×24 or 50×50 micrometers square and are divided in 8×8 pixels; the 28 border pixels are ignored, and the intensity is reported as the 27th largest of the remaining 36 pixel values (75th percentile). The oligonucleotide arrays avoid certain problems with the laboratory manufacturing process and with the use of probes of varying lengths, but these arrays also do not offer as

direct a comparison between experimental and control samples. The rough equivalent of a control for comparison is a "mismatch" probe, or "MM", which differs from the exact nucleotide sequences for a given gene, or "PM" for perfect match, at only the middle (13th) nucleotide. This "control" for the "PM" is only rough, since a target sample with elevated mRNA concentration for a certain gene may hybridize sufficiently to both the PM and MM probes.

Several authors have discussed analysis of gene expression data from various standpoints; e.g., experimental design of multiple cDNA slides (Kerr and Churchill, 2001a,b; Yang and Speed, 2002b), data transformations (Rocke and Durbin, 2001; Durbin et al., 2002), statistical methods (Dudoit et al., 2002), linear mixed-effects models (Wolfinger et al., 2001), mixture models (Pan, 2002), and multiple comparisons using false discovery rate (Reiner et al., 2003; Storey, 2001). Most of these papers assume that the data have been "pre-processed" with a sensible transformation to address partially the non-normality of the expression levels, and with normalization and background correction methods to adjust for different signal intensities across different microarray experiments and sources of variation arising from the chip manufacturing process and background intensity levels.

With either technology, the reported intensity level (spot, background, PM, or MM) is a summary of fluorescence measurements detected in a series of pixels. Because spot sizes are much more variable with cDNA slides, GenePix Pro© reports as "foreground" both the diameter of the spot (typically, about 130–150 μm) and the number of pixels where intensities are measured (usually 150–160), as well as the mean, median, and standard deviation of these pixel intensities. GenePix Pro© also reports the number of pixels used in the background calculation (typically, about 400–650, depending upon the location and size of spot for which the background measure is intended to correspond), along with the mean, median, and standard deviation of these background pixels. The definition of which pixels to include as "background" constitutes an important issue which strongly affects the analysis. Many algorithms to measure the background intensity for cDNA slides have been proposed. For example, GenePix Pro© assigns a pixel to the background of a particular target spot if it: (a) lies outside a two-pixel-wide boundary surrounding the circle of foreground pixels centered on the spot, (b) lies outside a two-pixel-wide boundary surrounding the circle centered on the foreground pixels of any one of the immediate neighboring spots (i.e., 8 neighbors for an interior spot, 5 neighbors for an edge spot, and 3 neighbors for a corner spot), and (c) lies inside a circle whose radius is three times that of the reported diameter of the circle centered on the target spot. [See p. 14 of the GenePix Pro© 4.0 User's Guide (Axon Instruments, 2001).] The software then reports the mean and median of these pixel values. Yang et al. (2002a) describe the background calculation (in their Table 5) from seven such algorithms provided by four gene chip software manufacturers (GenePix Pro©, Scanalyze©, Spot©, Quantarray©) and study the consequences of these algorithms on real data. They conclude that the `S.morph` algorithm from Spot© provides the least variable estimates of background. The image analysis method uses *morphological opening* (Soille, 2003), a nonlinear filtering operation that filters out the foreground pixels (cDNA spots, nominally circular, but in fact variously shaped), leaving only a background image, which is sampled near the target spot.

In contrast, the oligonucleotide array does not have any real "background" on the chip, so a quantitative assessment of the intensity in the absence of hybridization must be derived in other ways. The Affymetrix© User's Manual 4.0 states that the software divides the chip into 16 sectors [approximately 16,384 small (24 μm size) or 4096 (50 μm size) probe cells], and reports the average intensity in the 2% smallest [approximately 328 small or 82 large] probe cells. The user may choose to summarize pixel values (which tend to be highly skewed) with the biweight instead of the mean. (The biweight was designed for data from symmetric, not skewed, populations.)

All algorithms to categorize pixels as foreground or background yield summaries of the pixel intensities. Background pixel summaries can be modeled, providing more stable estimates that account for spatial variability across the slide, and that are subtracted from the foreground counts (spot intensities) for further analysis. Thus, this prescription can be applied to the output of any algorithm for defining and summarizing background pixel intensities. The median is particularly robust to errors in both measurement and foreground/background assignment. In this article, R and G denote the median of the pixel values corresponding to "foreground" (spot), and r and g denote the median "background" pixel values, in the red and green channels, respectively.

3. Measurement and processing variation

Many sources of variation affect data from microarray experiments. Knowledge of these sources leads to better experiments designed to reduce their impact, to better characterization of the uncertainty in the reported result, and to possible modifications in the manufacturing process.

3.1. Oligonucleotide arrays

The production process for these arrays relies on photolithography, similar to that used in silicon chip manufacturing. Each probe, either perfect match (PM) or mismatch (MM) for a gene of interest, consists of 25 nucleotides. The chip is manufactured in layers with "screens"; four screens for each layer leave openings where one of the four nucleotides (single-ringed pyrimidines C = cytosine and T = thymine; double-ringed purines A = adenine and G = guanine) is to be deposited onto the chip. Variability in the chip production process may arise from screen registration, materials, and chemical impurities. In addition, chips can be scratched or otherwise marred. The purpose behind the one base change in the MM probe cell was so that it could serve as a control for the PM probe cell; i.e., if a piece of a gene sequence from a sample successfully binds to PM, then presumably it would not bind to the MM probe cell. In fact, some hybridization at the MM probe may well occur for various reasons; e.g., small gene pieces in the test sample may hybridize to one or both of the 12-base sequences at either end of the MM cell; the originally published gene sequence might be incorrect; the gene chip manufacturing process accidentally placed the correct, as opposed to changed, base at the 13th location; the MM sequence is a PM for an entirely different gene. To minimize the impact of any one of these potential errors on the determination of

the gene (absent, present, or marginal), Affymetrix© uses 16–20 such (PM,MM) probe pairs, and reports a binomial statistic (number of pairs for which PM exceeds MM).

In practice, the data values from the MM probe cells probably do contain some information, but do not provide as direct a control as the control sample does in the cDNA microarrays. For this reason, analyses have been proposed that use the MM cells as they are, ignore them altogether, or do something halfway in between, with strategies being justified on either biological or statistical grounds (Efron et al., 2001). Control gene probes help to address the problem, but currently the chips are manufactured with such controls in only one area (despite the potential for spatial variation across the chip) and with too few of them to adequately characterize their distribution. To date, no consensus has been reached on the most effective use of the MM values.

3.2. cDNA slides

The cDNA technology is more widely available since individual laboratories can manufacture their own slides more easily using their own constructed gene "libraries." Moreover, the sample mixture containing Cy3- and Cy5-labeled mRNA offers a direct control for each gene on each spot. However, these spotted arrays may be subject to additional sources of variability. Some of these sources affect only signal intensities, some affect only background intensities, and some affect both.

(1) *Gene libraries*: The spots on the cDNA array are pieces of specific genes that are manufactured in the laboratory using a complicated process with many steps. Sample cells are isolated, are placed in solutions that dissolve the nuclear wall and unravel the DNA, and then are subjected to restriction enzymes which splice the DNA in predictable locations. These spliced probes are then inserted into a vector (usually bacterial DNA), transformed back into the host cell, and placed on agar to grow and multiply the desired sequence. The restriction enzyme process is reversed to extract from the vectors the desired sequences, which are then freeze-dried for later use when the slides are prepared. The entire process involves many production and handling steps, and the final harvested probes will have different lengths and concentrations. In addition, the organism from which the gene library cells were harvested may not have been entirely "normal," or the genes may be altered after isolation from the genomic library due to various handling steps involved in making the slides. These same genes will be used repeatedly for further cDNA experiments in this laboratory, so any errors in preparing the library or during the subsequent isolation of the genes will reappear on all slides. The same set of errors can occur during the analogous process used to prepare the control (Cy3) and experimental (Cy5) samples which are placed on the slide with the spots from the gene library. Finally, gene expression may depend upon the choice of cell line for the control sample, arguing for the use of different cell line sources as controls to assess the apparent significance of gene expression levels.

(2) *Sample preparation*: Even in the ideal case of a perfect manufacturing process, the sample preparation is subject to errors. Ideally, the target sample to be analyzed contains equal amounts of control and experimental mRNA, labeled with equal

amounts of Cy3 and Cy5, each fluorescing to the same degree, and each being detected by the laser scanner with the same accuracy and precision. In reality, small departures from the 50–50 ratio in control and experimental mRNA will exist, some of the Cy3 and Cy5 dyes will wash away instead of bind to sample mRNA, the number of Cy3-labeled mRNA cells will not equal the number of Cy5-labeled mRNA cells, and the maximum intensity level in the 532-nm (green) channel may not equal that in the 635-nm (red) channel. This issue relates to "labeling efficiency" and is key to a successful experiment. If the red/green fluorophores do not bind well to their respective samples, then the measurements of "signal" will be substantially diminished, making it difficult to distinguish from background. Results from experiments to measure this labeling efficiency could provide information that may be used to adjust values from multiple experiments so they can be compared and combined with other experiments.

(3) *Slide coating*: Probes manufactured from the gene library are not placed directly on the raw glass slide; rather, the slide is treated first with a poly-L lysine coating to ensure gene probe adhesion. The coating thickness may vary across the slide and thus may affect both the signal intensity values as well as the reported measurement of background intensity.

(4) *Spot placement*: Spots are deposited on the glass slide using an inkjet-like technology. Suppose a production run involves the printing of 100 slides. A template of pins (e.g., 4×4 or 8×4) is dipped into wells containing (16 or 32) separate gene pools and then "spotted" onto the slide. Thus, genes appear on the slide in (16 or 32) separate "blocks"; the first spot is made in location [1,1], corresponding to the first row and first column, in each block. After all 100 slides have been spotted at the [1,1] locations in each block, the pins are washed and then are dipped into another set of (16 or 32) gene pools; these genes are deposited at location [1,2] in each block, again for all 100 slides. The process continues for a number of rows and number of columns (usually, 16–24 rows and columns, and often square). For the data in Section 5, the template was 8×4, with 23 rows and columns, leading to 32 blocks of 529 spots, or 16,928 spots per slide. Broken pin tips, incomplete washing, misaligned pins or slides, and time trends can all lead to additional variability. Yang et al. (2002c) describe a method that adjusts the data for print tip variability.

(5) *Probe concentrations and sample placement*: Wells containing the probes are not always uniform and homogeneous; thus probe concentrations in the spots vary across the array. Spots with high DNA concentrations can "smear" across the slide. Large air pockets render the slide unusable. For usable slides, the sample of Cy3- and Cy5-labeled mRNA is inserted between the slide and coverslip in one corner; capillary action distributes the sample across the slide. The possibly smeared spots, as well as the non-uniformity in the distribution of the sample across the slide, argue for replication and studies of process variability. For example, the samples may be inserted at different places for different replications, to assess the degree to which the insertion location affects the results.

(6) *Approximate hybridization*: Both cDNA and oligonucleotide arrays are subject to the phenomenon where some of the sample genes may settle on a spot whose

cDNA is a rough though imperfect match, creating the illusion of significant gene expression of unimportant genes and insignificant expression of important genes.

(7) *Instrumentation errors*: The optical device used to measure fluorescence typically scans and digitizes the slide or chip, and scales the result to utilize the entire range (e.g., for a 16-bit scanner, 65,536 gray levels). Thus, errors in scanning and digitization may occur and affect "detection efficiency." Some scanners focus automatically, based on perceived average depth of the slide; others expect the user to set it manually. Since target sample depth depends on the slide thickness, the poly-L lysine coating thickness, and the size of the spot, the scanner may not focus equally well on all parts of the slide. In addition, the biological samples on the slide degrade rapidly, so the second pass through the scanner (i.e., scanning at 532 nm followed by scanning at 635 nm) will lead to less precise measurements, and possible biases, in the second measurements (arguing again for dye-swapping).

(8) *Image processing*: The laser scanner records the intensity of reflected light at each pixel on the slide. An image processing algorithm designates pixels as either foreground, background, or neither (e.g., pixels that lie in a "buffer zone" surrounding the apparent spot). The final reported value is usually the mean or median pixel intensity from foreground, minus the mean or median background pixel intensity. While less efficient than other measures of location, the median is highly robust to both measurement errors and assignment errors (e.g., a pixel being called background when it should have been foreground, and vice versa). The data in Section 5 illustrate that background intensity can vary substantially across the slide, which can be modeled for purposes of data analysis, and ultimately suggest aspects of the process that might benefit from tighter control.

Knight (2001), Illouz (2002), and Finkelstein et al. (2002) discuss other cDNA slide effects.

4. Analyzing data: transformations and background

In addition to process variation, appropriate normalization of arrays for comparing results from different experiments, and multiple hypothesis testing, data transformations and proper adjustment for background intensities have important consequences when drawing inferences from a microarray experiment. This section discusses these two issues. Transformations are discussed first, to determine whether they should be applied to both the foreground and background intensities.

4.1. Transformations

Gene expression values exhibit an extremely wide range of gene expression levels. For example, in one experiment (Section 5), median foreground (spot) intensity ranged from 18 to 27,667 for Cy5-labeled mRNA and from 289 to 43,869 for Cy3-labeled mRNA. (A 16-bit scanner offers 65,536 possible values.) Moreover, the noise (measurement error, etc.) is usually related to the expression level. Thus,

measurements are often transformed via logarithms; e.g., $\log PM - \log MM$ for oligonucleotide arrays, or $\log R - \log G$ for cDNA arrays. There seems to be little use for raw differences (e.g., $PM - MM$ or $R - G$), despite their calculation in various software routines supplied by the manufacturer, except perhaps for purposes of understanding the measurement error.

Measurements of fluoresence share many of the same characteristics that Rocke and Lorenzato (1995) observed with measurements of chemical concentrations. In the latter article, the authors transformed large measurements via the logarithm, leaving small measurements unchanged. Durbin et al. (2002) note that the variance of gene expression values is a quadratic function of the mean; i.e., $Var(X) \equiv \sigma_x^2 = \alpha + \beta(\mu_x - \gamma)^2$, where X is R (red foreground), G (green foreground), PM (perfect match), or MM (mismatch), and μ_x denotes its mean value. By binning the expression levels, the parameters of this quadratic function can be estimated from the data. The data to be described in Section 5 also exhibit this phenomenon in both the red and green channels; Fig. 1 shows a subset of this data set in the red channel for one experiment only. Of 16,928 foreground values, 529 were sorted and divided into 23 bins whose bin means and variances are plotted in Fig. 1. Note that "value" is really a reported foreground median pixel intensity from the laser scanner, with higher numbers indicating higher intensities

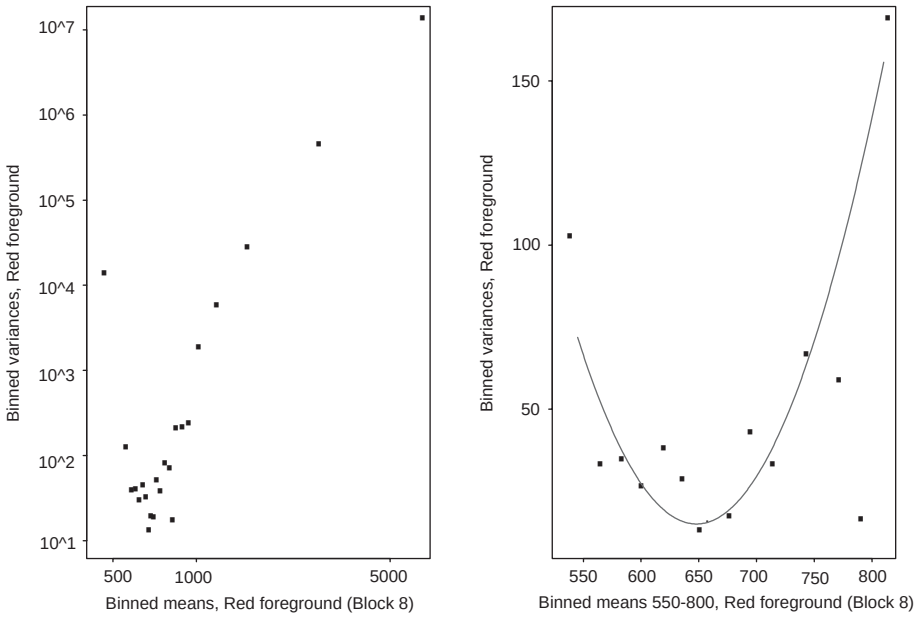

Fig. 1. Mean–variance plots for one set of 529 foreground counts in red channel of Block 8 (see Section 5 for description of the data). The 529 values were sorted and binned into 23 categories of 23 values each; the left panel plots the sample variance versus the sample mean for all 23 categories. The right panel shows the points from only those categories whose sample means lie between 550 and 800. The fitted quadratic is: $variance = 15 + 0.005358(mean - 648)^2$.

(and hence possibly higher mRNA concentrations when the mRNA was harvested). The left panel shows some very influential extreme mean-variance pairs among the 23 pairs, so the right panel focuses on only those mean–variance pairs whose means fall between 550 and 800. The fitted quadratic to these data is: $variance \approx 15 + 0.005358\,(mean - 648)^2$. A similar plot for the background counts in the red channel does not show this structure. Another plot shows little relationship between a spot's reported foreground median and its background median. Both plots, omitted for reasons of space, suggest that reported background medians need not be transformed.

Curtiss (1943) noted many years ago that a variance-stabilizing transformation can be obtained by using a first-order Taylor series expansion for the approximation

$$Var(f(X)) \approx [f'(\mu_x)]^2 \sigma_x^2, \tag{1}$$

(see also Ku, 1966) which, when $\sigma_x^2 = \alpha + \beta(\mu_x - \gamma)^2$, leads to

$$f(x) \equiv f(x; \alpha, \beta, \gamma) \equiv x^* = \log((x - \gamma) + \sqrt{\alpha/\beta + (x - \gamma)^2}). \tag{2}$$

(A referee noted that Curtiss did not derive this approximation but only formalized its use in transformations. The use of Eq. (1) probably appeared in the 19th century.) Durbin et al. (2002) follow this strategy and recommend this transformation.

In this formulation, $\alpha/\beta > 0$ (α represents the smallest fitted variance, and β, as half the second derivative of the convex variance function σ_x^2, will be positive), so the argument of the logarithm in (2) cannot be negative, whether x is R, G, PM, or MM, thus avoiding problems when (foreground–background) is negative. Expanding $f(x)$ in a Taylor series around $x = \gamma$ shows that $f(x) \approx \log c + (x - \gamma)/c$ where $c = \sqrt{\alpha/\beta}$, a linear function of x, and but when $|x - \gamma|$ is large, $f(x) \approx \log(x)$. (Tukey used to put two transformations together and call the result a "hybrid re-expression." The particular variance function for these data achieves this hybrid in a natural way.) This transformation to constant variance facilitates the comparison between the transformed values of R and G in a cDNA experiment, or between PM and MM in an oligonucleotide experiment. The constancy of the variance can be easily checked by calculating a robust estimate of the variance [e.g., $(1.5 \times MAD)^2$, where MAD = median absolute deviation from the median, or the Winsorized variance; cf. Horn and Kafadar, 2002] on the transformed values. Another benefit of this transformation is the reduced skewness in the distribution of the measurements.

A similar transformation having somewhat more interpretable parameters arises from Tukey's g-family of distributions (Hoaglin, 1985):

$$(X - a)/b = (e^{gZ} - 1)/g, \tag{3}$$

$$Z = z(X) \equiv z(X; g, a, b) \equiv g^{-1} \log[g(X - a)/b + 1], \tag{4}$$

where Z is a standard Gaussian random variable, g is the skewness parameter ($g = 0$ corresponds to the Gaussian, $g > 0$ yields a lognormal distribution, and increasing values of g lead to increasingly skewed distributions), and (a, b) represent location and

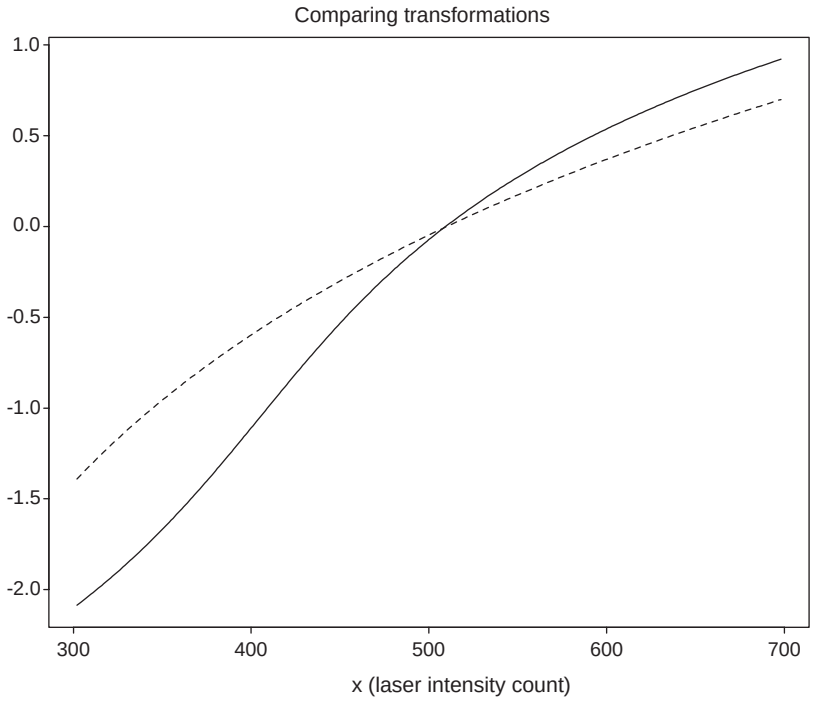

Fig. 2. Comparing two transformations for the data illustrated in Fig. 1. The solid line is the function $f(x)$ given by Eq. (2), where $\alpha = 15$, $\beta = 0.005358$, and $\gamma = 648$. The dotted line is the function $z(x)$ given by Eq. (4), where $g = 0.58$, $a = 510$, and $b = 218$. The two transformations are matched to coincide at $x = 500$.

scale parameters, respectively. In practice, these three parameters are fit robustly, so that the *majority* of the transformed values look *roughly* Gaussian with mean 0 and variance 1. A connection between these two transformations is shown in Fig. 2, which plots the function $f(\cdot)$ in (2) and $z(\cdot)$ in (4). Also, since

$$[z'(\mu_x)]^2 = [b(g(\mu_x - a)/b + 1)]^{-2} = [g(\mu_x - a) + b]^{-2} \tag{5}$$

it follows, again from (1), that

$$1 = Var(Z) = Var(z(X)) \approx [z'(\mu_x)]^2 \sigma_x^2 \Rightarrow \sigma_x^2 \approx [g(\mu_x - a) + b]^2 \tag{6}$$

which also is a quadratic. [This approximation can be derived directly without resorting to (1), since the density of $Y = gX + (b - ga)$ is lognormal with parameters $\ln(b)$ and g^2: $\mu_y \equiv E(Y) = be^{g^2/2} = g\mu_x + (b - ga)$ and $Var(Y) = b^2 e^{g^2}(e^{g^2} - 1) \approx b^2 e^{g^2} g^2 = \mu_y^2 g^2$ since fitted values of g are small, so $Var(X) = g^{-2}\, Var(Y) = (b/g)^2 e^{g^2}(e^{g^2} - 1) \approx b^2 e^{g^2} = \mu_y^2 = [g\mu_x + (b - ga)]^2$, as in (6).] For values of (α, β, γ) in (2) and (g, a, b) in (4) estimated from the same set of data, these two transformations are often similar; see Fig. 2. The difference between these two approaches is part philosophical and

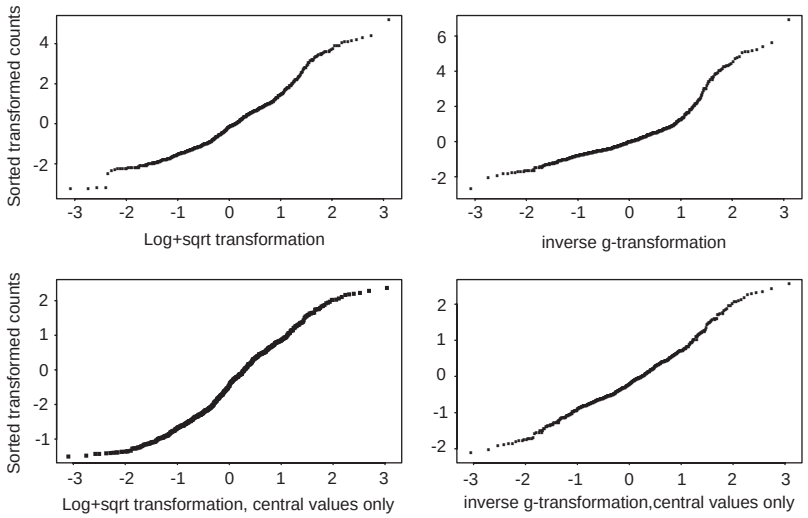

Fig. 3. Quantile–quantile plots (Wilk and Gnanadesikan, 1968) of transformed data values in Fig. 1 via $f(x)$ [Eq. (2)], where $\alpha = 15$, $\beta = 0.005358$, and $\gamma = 648$, denoted as "Log+sqrt transformation", and $g_y(x)$ [Eq. (4)], where $g = 0.58$, $a = 510$, $b = 218$, denoted as "inverse-g transformation". The top row shows the result applied to all 529 values; the bottom row shows the result in only those 486 values whose g-transformed value lies between -2.5 and 2.5. The inverse g transformation results in a distribution that appears roughly Gaussian with a long right tail, which may indicate genes with significantly high expression (see Section 5).

part interpretive. Transformation (2) follows by assuming that the mean–variance relationship really is quadratic, and then one finds the transformation so that the variance approximation (1) holds exactly. The g-transformation (4) assumes that the data can be represented precisely as a location-and-scale transformation of the standard Gaussian Z, whose variance is approximately a quadratic function of the mean. Neither assumption ("the variance is exactly quadratic" versus "location-and-scaled X is exactly a transformed Gaussian") is likely to hold precisely. However, the $z(\cdot)$ transformation offers convenient interpretations for the three parameters g, a, b (skewness, location, scale). A further benefit is shown in Fig. 3, where 486 of the 529 resulting transformed values do indeed appear more Gaussian than those from the transformation in (1). This approach is related to the "soft thresholding" method in Finkelstein et al. (2002), who select a constant which, when added to the foreground values, results in a distribution that appears as lognormal as possible; (4) involves three parameters.

Many possibilities exist for the treatment of negative values that arise when the background intensity exceeds the spot intensity in cDNA arrays, or when the value of MM exceeds PM in oligonucleotide arrays. When they are few in number, they can be replaced by a small positive value, trusting the robustness of the overall analysis to give them little weight. Alternatively, one could fill in a value using an EM algorithm based on a model for the transformed data. (Under the assumption that the transformed data are Gaussian, this value is likely to be 1, so that its logarithm is 0.) Negative

values provide opportunities for identifying potential sources of errors in the biology, manufacturing, or measurement process. This situation actually occurred for two of the 529 values shown in Fig. 1; they are described further in Section 6.

4.2. Background estimation

Yang et al. (2002a) investigate in detail various methods for estimating background, which arises when fluoresence is detected at locations on the array where no spot or probe occurred. The importance of a correct adjustment for background intensity can be seen as follows.

Transformations (2) and (4) both resemble the logarithm for large values of x, so, for this purpose, consider the quantity $\log(red) - \log(green)$ [or $\log(PM) - \log(MM)$ for oligonucleotide arrays]. The *red* and *green* counts comprise three signal intensities: spot intensities R and G, background intensities r and g, and errors in these background adjustments, x and y (assumed optimistically to be independent of each other). The effect of these errors can be seen easily using a Taylor series expansion around $(x, y) = (0, 0)$ (Ku, 1966, p. 269; Vardeman, 1994, p. 257):

$$\begin{aligned}\log[(R + r + x)/(G + g + y)] \approx \log[(R + r)/(G + g)] \\ +[x/(R + r)][1 - 0.5x/(R + r)] \\ -[y/(G + g)][1 - 0.5y/(G + g)].\end{aligned}$$

When the errors are equal in the two channels ($x = y$) and the correct values r and g are subtracted, the comparison between *red* and *green* is unbiased. For any other case, however, and particularly when x and y are large in magnitude, the comparison will be biased. An error in $(R + r)$ and $(G + g)$ of $e\%$ renders an error of slightly over $2e\%$ in the comparison of two equally expressed genes.

Reported background intensities may exhibit considerable structure, due to processing and material variation. Both the glass on which the array is printed and the poly-L lysine coating to which the spots adhere may have non-uniform thickness across the slide. The laser scanner that measures intensity levels operates using filters tuned to the appropriate wavelengths (532 nm and 635 nm) and by moving the slide so that eventually all pixels fall directly under the laser beam. Consequently, accuracy and precision may vary across the slide. Finally, both the fluorophores and the cDNA degrade rapidly (Cy5 faster than Cy3), and this degradation will affect the measurements made on the second scan (632 nm) more than the first (535 nm). To estimate the effects within a block of this process variation, we propose in the next section to fit the background count using median polish (a robust form of two-way analysis of variance that relies on medians instead of on means; see Tukey, 1977), and subtract the fitted value from the foreground count which is then transformed.

5. Illustration

A cDNA experiment comparing the mast cells in two mouse cell lines, UCOZ-22 (immature) and MC9 (mature), was conducted at the University of Colorado Health

Table 1

	Stripe number			
	1	2	3	4
Layer 1	1	2	3	4
Layer 2	5	6	7	8
⋮				
Layer 7	25	26	27	28
Layer 8	29	30	31	32

Sciences Center. Mast cells are derived from bone marrow stem cells which are present in a variety of tissues, especially in skin and the gastrointestinal tract. They play an important role in the immune system, specifically in allergic inflammation (they contain large amounts of histamine and heparin, which trigger the inflammation response, thus leading to the allergic reaction). While it is known that mast cells are regulated by the Stem Cell Factor (SCF), the differentiation from immature (precursor) mast cell to its mature stage is poorly understood. In vitro experimentation between these two stages is difficult under the establishment of these two stable cell lines, UCOZ-22 and MC9. UCOZ-22 is a mast cell precursor which requires stem cell factor (SCF) plus interleukin-3 (IL-3) for growth and proliferation, whereas MC9 responds to either SCF or IL-3 alone. The purpose of the cDNA microarray experiment is to determine the gene expression levels between these two stages. Such knowledge will help elucidate mast cell maturation mechanisms.

As indicated in Section 3, slides are printed in blocks, corresponding to the template of pins used by the particular laboratory. The template used at the University of Colorado Health Sciences Center has 32 pins, arranged in 8 layers of 4 pins in each layer. Genes are printed onto the slide by placing the pin template at a particular location, which prints 32 spots at one time, for all slides. Then the template is washed, the pins are dipped into the next set of 32 gene wells, and 32 more spots are printed successively on the slide alongside the first set of spots. Denoting the spots printed by one single pin as a "block", the slide contains 32 blocks, with spots within each block printed row by row. Thus, if the time between successive prints on adjacent slides is δ, and 100 slides are printed, then the time between the printing of spots in [row, column] locations $[1,1]$ and $[1,2]$ of each block is 100δ time units. The blocks are formatted on the slide in 8 "layers" of 4 "stripes" as shown in Table 1.

The analysis here will be illustrated mostly on measurements in Block 8, which occurs in the second layer of blocks near the right-most edge of the glass slide.

The effects of these time trends, as well as variations in the glass and slide coating, appear in the signal intensities, as can be seen from a simple two-way additive fit to the background intensities in the red and green channels. [All analyses here were conducted using S-Plus, Version 6.0.1 Release 1 for Linux 2.2.12 from Insightful (2001).] The red and green background intensity of row i, column j, and block k, namely r_{ijk} and g_{ijk} (the subscripts i and j depend upon block number k), can be fit by a two-way

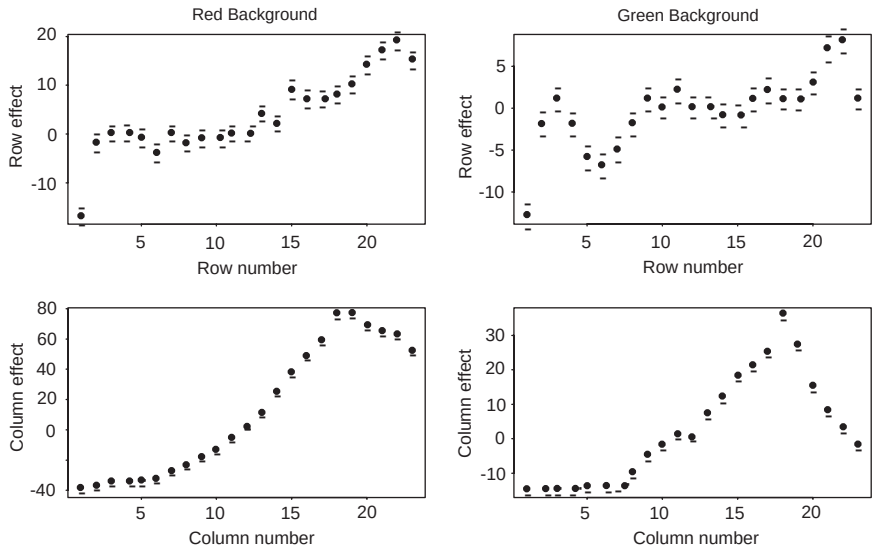

Fig. 4. Results of applying median polish to the background counts in the red and green channels. The top (respectively, bottom) plots show the effect of the row (respectively, column) number (red on left and green on right); i.e., the estimated number of laser scanning units of light intensity above or below the overall term $m_{8r} = 223$ (respectively, $m_{8g} = 337$). Limits of 1 standard error based on 200 bootstrap replications of the residuals are shown.

additive model via median polish (Tukey, 1977):

$$r_{ijk} = m_{kr} + row_{ikr} + col_{jkr} + res_{ijkr},$$

$$g_{ijk} = m_{kg} + row_{ikg} + col_{jkg} + res_{ijkg},$$

where all effects (m, row, col, res) include an additional subscript, r or g, to denote red or green channel. For block 8, $m_{8r} = 223$ ($se = 1.9$) and $m_{8g} = 337$ ($se = 1.8$); standard errors (se) have been estimated using 200 bootstrap replications of the residuals. (A non-parametric bootstrap was used: the 529 residuals were sampled with replacement, added to the overall, row, and column effects, re-fit by median polish, and repeated 200 times. Standard errors were obtained by calculating standard deviations in the usual way from the 200 sets of estimates.) The row and column effects for this block, with limits of one bootstrap-estimated standard error, are shown in Fig. 4. Clear patterns exist: variation is greater across the block (columns) than down (rows), reflecting perhaps the effects of the coverslip on background accuracy of the laser scanner or varying thickness of the glass slide or poly-L Lysine coating on the slide. In the vertical direction, background counts in row 1 tend to be considerably lower than those in successive rows in both channels. An estimate of the residual standard deviation ($1.5 \times MAD$) is about 5 ($se = 1$) in both channels; the standard errors of the column effects (bottom panel plots in Fig. 4) look smaller simply because the range of the column effects is about twice that for the row effects.

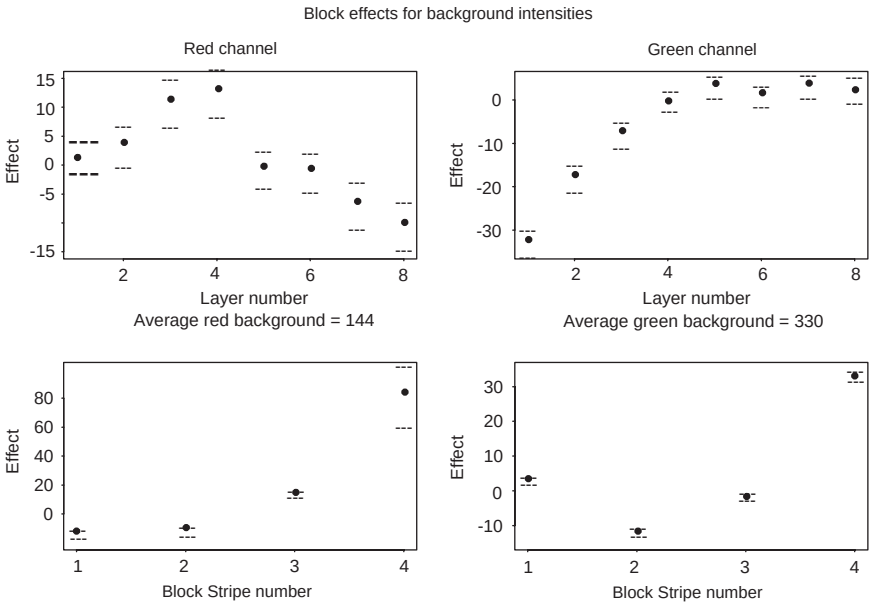

Fig. 5. Layer and stripe effects resulting from a median polish of the overall terms, m_{kr} and m_{kg}, where $k = 1, 2, 3, 4$ refers to the blocks in the first layer, and $k = 29, 30, 31, 32$ refers to the blocks in the eighth layer, and $k = 4, 8, 12, \ldots, 32$ refers to the fourth (rightmost) stripe on the slide. Limits of 1 standard error based on 200 bootstrap replications of the residuals are shown.

In addition, the 32 values of m_{kr} and m_{kg} from the 32 median polish fits in each channel can be arranged as a matrix of 8 layers and 4 stripes, corresponding to the block locations on the slide, and then fit with layer and stripe effects:

$$m_{kr} = M_r + layer_{lr} + stripe_{sr} + res_{lsr},$$

$$m_{kg} = M_g + layer_{lg} + stripe_{sg} + res_{lsg},$$

where $l = \lfloor (k-1)/4 + 1 \rfloor$ and $s = \mathrm{mod}\,(k, 4)$. These effects are plotted in Fig. 5, as a function of the block row (layer) number and the block column (stripe) number. The fitted common terms M_r and M_g are 144 and 330, respectively. (Standard errors are again estimated via 200 bootstrap replications of the residuals.) In both the red and the green channels, layer and stripe effects are evident; for example, values in the last stripe of blocks numbered $4, 8, \ldots, 32$ (bottom panels) tend to be higher than the others. Estimates of the residual standard deviation in these two fits (again as $1.5 \times$ MAD) tend to be around 4. These sorts of trends appear in other blocks on this slide also. The print tip adjustment via loess in Dudoit et al. (2002) may be viewed as a block adjustment, since different blocks are spotted with different pins.

Finally, consider the residuals from the simple additive fit exhibits structure. Fig. 6 shows a coded plot of the magnitude of the red residuals in the 23 rows and 23 columns of block 8. The radii of the circles are proportional to the magnitudes of

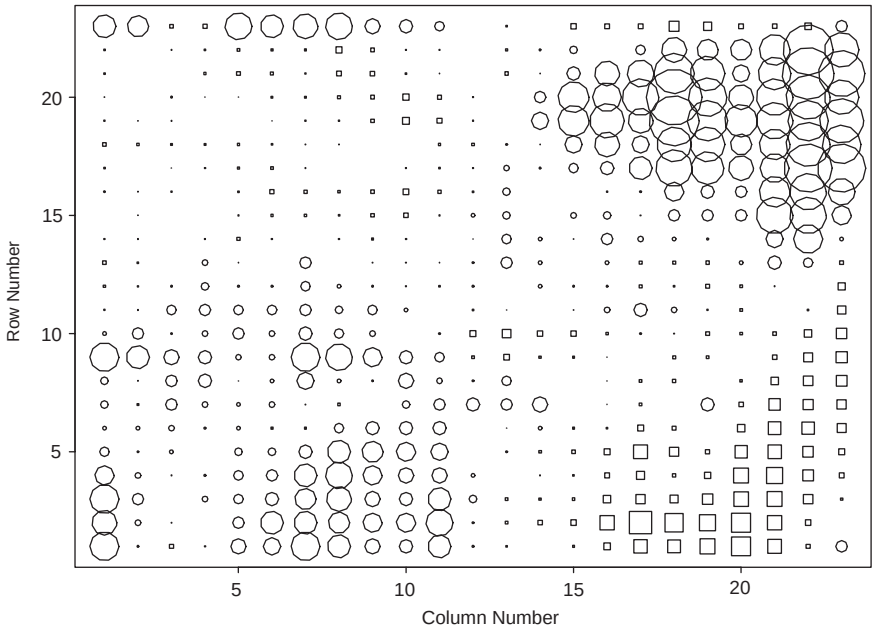

Fig. 6. Coded residuals from the median polish fit to the 529 red background values in Block 8. Radii of circles are proportional to magnitudes of negative residuals; sides of squares are proportional to magnitudes of positive residuals. Structure in residuals indicates need for an additional term in the additive fit.

the negative residuals, and the sides of the squares are proportional to the magnitudes of the positive residuals. The plot shows clear structure in these residuals and the inadequacy of the simple additive fit described above. (For some reason, many of the eight blocks in this region (blocks $4, 8, 12, \ldots, 32$) showed spatial patterns, while the residuals in most of the other 24 blocks showed little structure.) A better fit for this block is

$$r_{ijk} = m_{kr} + row_{ikr} + col_{jkr} + T\, row_{ikr}\, col_{jkr} + res_{ijkr}, \tag{7}$$

also known as Tukey's "plus-one fit", the "one" referring to one degree of freedom (T) for non-additivity (Tukey, 1949; Tukey, 1977; Emerson and Hoaglin, 1983), here estimated as $T = 0.0224$. The final coded residual plot (Fig. 7) indicates much less structure.

The median polish fits to blocks can be useful in two ways. First, the time trends across rows, columns, and blocks (pin tips) can be communicated to the production facility, and those aspects of the process that influence these patterns can be identified and modified. Second, the present data can be adjusted by a *fitted* value of the background intensity, rather than by the median background intensity value that is reported by the software. For example, rather than subtracting the unadjusted background counts in the red and green channels, r_{ijk} and g_{ijk}, from the reported foreground counts

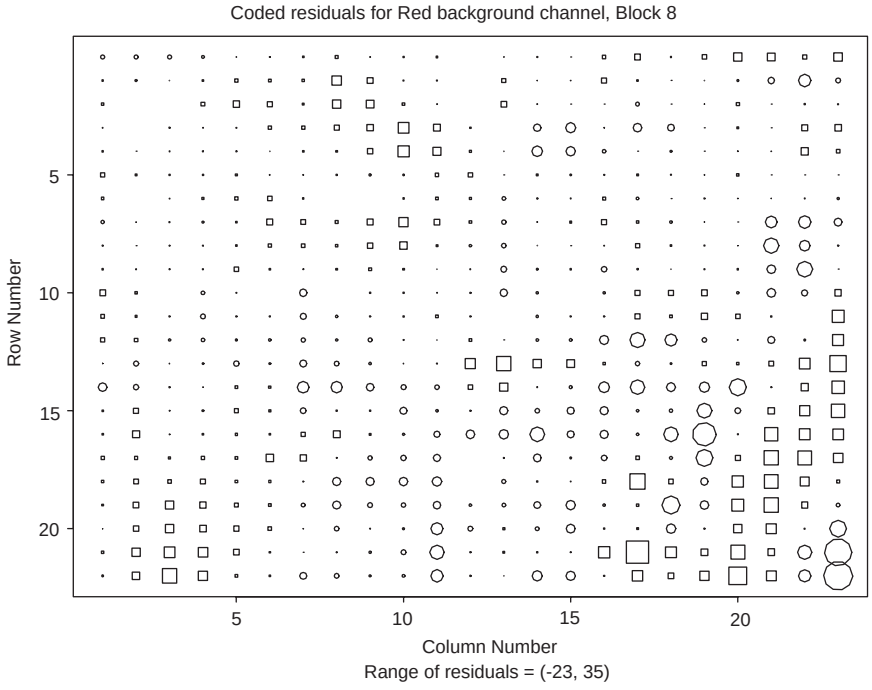

Fig. 7. Coded residuals from the median polish fit plus one degree of freedom for non-additivity [Eq. (7)] to the 529 red background values in Block 8. Radii of circles are proportional to magnitudes of negative residuals; sides of squares are proportional to magnitudes of positive residuals. Residuals range in size from –23 to +35; lack of structure in the plot indicates success of the "plus-one fit."

(spot intensities) R_{ijk} and G_{ijk}, one subtracts instead

$$\hat{r}_{ijk} = m_{kr} + row_{ikr} + col_{jkr},$$

$$\hat{g}_{ijk} = m_{kg} + row_{ikg} + col_{jkg}$$

or, if the additional nonlinear term is needed to achieve patternless residuals,

$$\hat{r}_{ijk} = m_{kr} + row_{ikr} + col_{jkr} + T_{kr} row_{ikr} col_{jkr},$$

$$\hat{g}_{ijk} = m_{kg} + row_{ikg} + col_{jkg} + T_{kg} row_{ikg} col_{jkg}.$$

For more stable estimates, one might further smooth (e.g., via loess or running medians) the row and column effects as a function of row (i) and column (j) before using them in the fit (but this was not done here). The value of these fits is shown by comparing two measures of the residuals: for the unfitted background counts in the red channel, 1.5 $MAD(residuals)$ and $[mean(residuals^2)]^{1/2}$ are 54 and 243.8; for the residuals from the additive fit, 9 and 12.9; for the residuals from the additive-plus-one fit, 6.7 and 9.2. Thus the fit accounts for a substantial portion of the variation in background counts associated with their locations on the slide, and these fitted values can be used to more reliably estimate the background intensities affecting the spot signal intensities.

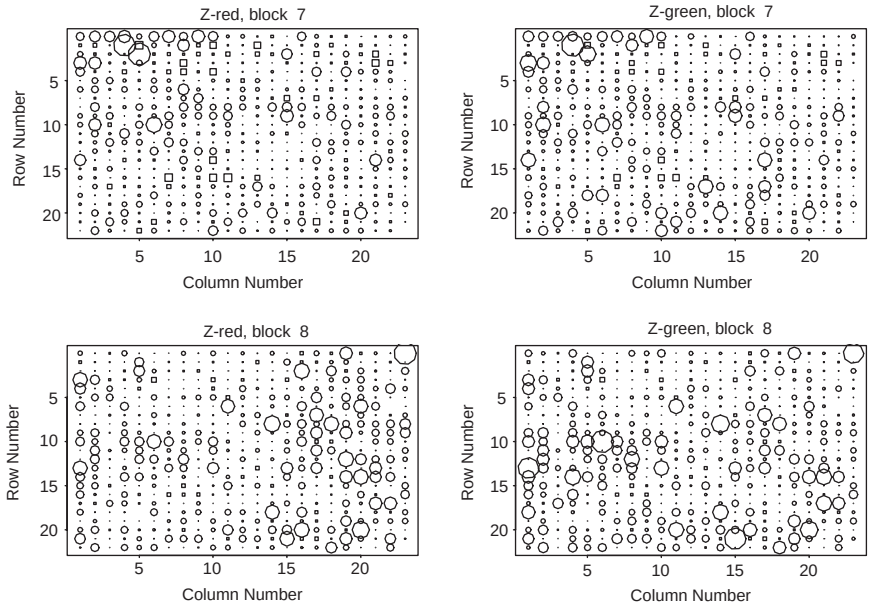

Fig. 8. Coded plot of Z_R and Z_G in blocks 7 and 8, obtained via transforming the foreground values for fitted background in the red and green channels. Radii of circles are proportional to magnitudes of negative Z-values; length of side squares are proportional to magnitudes of positive Z-values. Lack of structure indicates success of the model for background in removing spatial structure.

Statistical analysis can now proceed on the transformed and adjusted signal intensities, $Z_R \equiv z_R(R^*_{ijk})$ and $Z_G \equiv z_G(G^*_{ijk})$, where $R^*_{ijk} = R_{ijk} - \hat{r}_{ijk}$ and $G^*_{ijk} = G_{ijk} - \hat{g}_{ijk}$, and z_R and z_G denote the two transformations defined in (4) using values of (g, a, b) specific to channel and block. The procedure for estimating these parameters is described in detail by Hoaglin (1985, pp. 468–471). For block 8 in the red channel, the parameters are $g = 0.58$, $a = 510$, $b = 218$; in the green channel, the parameters are $g = 0.60$, $a = 763$, $b = 276$. (Maximum likelihood estimates of g, a, b could be derived but would be very sensitive to extremely high values, some of which may represent significant gene expressions.) The success in the removal of the spatial effects in the background values is evident from the plot in Fig. 8, which shows no patterns in the magnitudes of Z_R and Z_G in blocks 7 and 8 by row and column.

One motivation for these transformations is to obtain values whose distribution is roughly Gaussian, so that data in all slides can be "normalized" and therefore combined, and standard cut-off values can be used to assess significance. That is, this procedure can be conducted on all duplicates of this experiment, and the average and standard deviation after differences $Z_R - Z_G$ corresponding to specific gene can be calculated. Note that Z_R and Z_G are very highly correlated; in the left panel of Fig. 9, a plot of Z_G versus Z_R indicates very high correlation (Pearson correlation coefficient is 0.94; a robust correlation coefficient (see Section 6) is 0.94; both least squares and robust regression lines from which these correlations were estimated are shown on the plot). The right panel

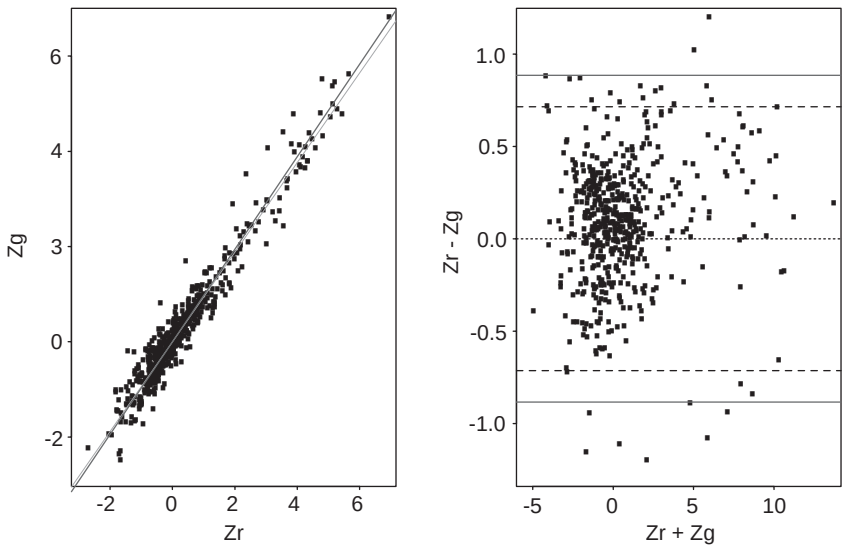

Fig. 9. Left panel: Plot of Z_G versus Z_R (labeled on the plot as `Zg` and `Zr`), the transformed and background-adjusted spot intensities in block 8. Estimated correlation coefficient is 0.97 (Pearson) and 0.94 (robust); lines from which they are estimated are shown (but barely distinguishable) on the plot. Right panel: Plot of $Z_R - Z_G$ versus $Z_R + Z_G$, to remove visual effect of extreme correlation. Limits of 3 estimated standard deviations based on the two correlation estimates are shown (solid=Pearson; dashed=robust).

is rotated by 45° to clarify the structure, $Z_R - Z_G$ versus $Z_R + Z_G$ (Tukey sum-difference plot; cf. Cleveland, 1985, pp. 118–123). Using the formula $Var(Z_R - Z_G) = Var(Z_R) + Var(Z_G) - 2\,Cov(Z_R, Z_G) \approx 1 + 1 - 2(0.97)$, an estimated standard deviation of the differences between the background-adjusted and transformed intensities in the two channels is 0.24. A conservative limit of three standard deviations is indicated on the right panel in Fig. 9, based on the Pearson correlation (dashed line) and on the robust correlation (solid line). (However, the uncertainties in the parameter estimates have not been taken into account when estimating the variance of the difference, so these limits may not be overly conservative.) Alternatively, one can compute $\Phi^{-1}((Z_R - Z_G)/0.24)$ as a sort of "p-value" for each gene, rank them, and select those that appear significant on the basis of the false discovery rate (FDR) criterion (Benjamini and Hochberg, 1995).

This analysis was repeated on all 32 blocks:

(1) Apply median polish separately to background medians in each block and each channel (possibly with smoothing of the fitted row and column effects, and possibly with the extra term for non-additivity), yielding 32 sets of fitted background counts $\hat{r}_{ijk}$ and $\hat{g}_{ijk}$ $(i = 1, \ldots, 23,\ j = 1, \ldots, 23,\ k = 1, \ldots, 8)$ for both red and green channels.

(2) Adjust foreground medians R_{ijk}, G_{ijk} by subtracting fitted background counts in Step 1 from foreground medians, yielding $R^*_{ijk} = R_{ijk} - \hat{r}_{ijk}$, $G^*_{ijk} = G_{ijk} - \hat{g}_{ijk}$.

(3) Estimate for each block the parameters in the g transformation (4) to the adjusted foreground counts obtained in Step 2, yielding (g_{kr}, a_{kr}, b_{kr}) and (g_{kg}, a_{kg}, b_{kg}).
(4) Transform the adjusted foreground counts via Eq. (4) to obtain approximate Gaussian distributed quantities

$$Z_R \equiv Z_{R,ijk} = g_{kr}^{-1} \log[g_{kr}(R^*_{ijk} - a_{kr})/b_{kr} + 1],$$

$$Z_G \equiv Z_{G,ijk} = g_{kg}^{-1} \log[g_{kg}(G^*_{ijk} - a_{kg})/b_{kg} + 1].$$

(5) Estimate the correlation $\hat{\rho}_k$ between Z_R and Z_G in each block.
(6) Calculate an approximate standard error for the difference $Z_R - Z_G$ as $[2(1-\hat{\rho}_k)]^{1/2}$. Combine values of $Z_R - Z_G$ for the same gene on different slides (experiment) via (weighted) averages, and
(7) Denote as "significant" those differences that either exceed a set number of standard deviations, or achieve significance via FDR.

For Block 8, the false discovery rate criterion on 529 genes using an FDR of 0.0015 (= 0.05/32) identifies 10 significant genes, with p-values ranging from 8.0×10^{-8} to 0.0001. (An eleventh gene just barely missed "significance", with a p-value of 0.00149, just over the maximum allowed by the FDR procedure, 0.0015(1–519/529) = 0.00147.) Coincidentally, this turns out to be the same number of genes identified by using a

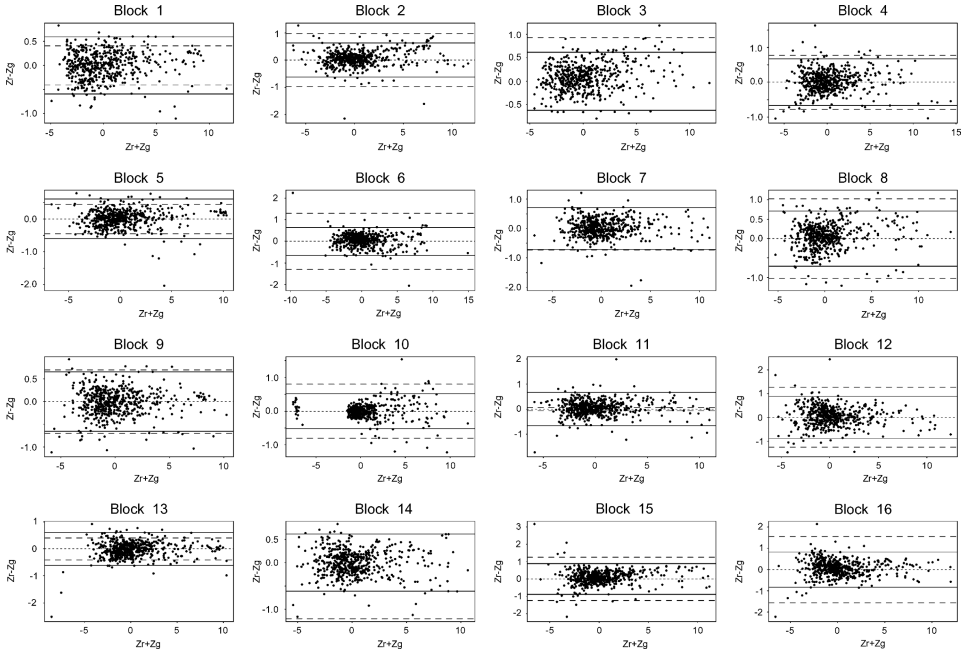

Fig. 10. Plots of $Z_R - Z_G$ versus $Z_R + Z_G$ in Blocks 1–16. Limits of 3 estimated standard deviations based on the two correlation estimates are shown (dashed=Pearson; solid=robust).

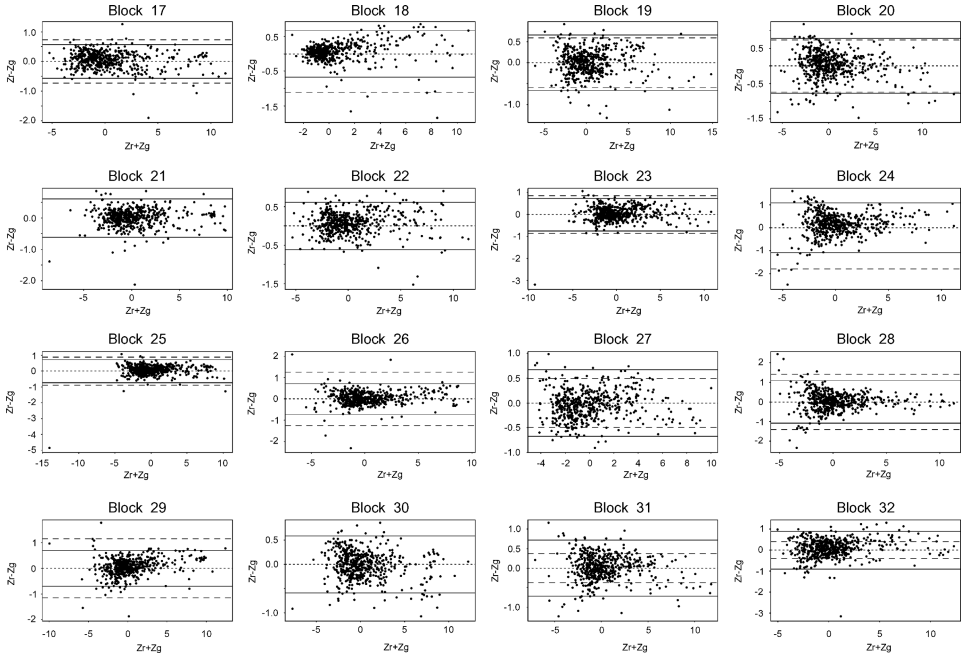

Fig. 11. Plots of $Z_R - Z_G$ versus $Z_R + Z_G$ in Blocks 17–32. Limits of 3 estimated standard deviations based on the two correlation estimates are shown (solid = Pearson; dashed = robust).

3-standard deviation limit on $Z_R - Z_G$. Plots of $Z_R - Z_G$ versus $Z_R + Z_G$, with limits of 3 standard deviations for reference, are shown for all 32 blocks in Figs. 10 and 11.

In total, 168 genes out of the $32 \times 529 = 16,928$ genes were identified as "significant", and were communicated to the biologists who conducted the experiments. These results are very tentative, and need to be confirmed with other experiments.

6. Conclusions and further issues

This article suggests a strategy for analyzing microarray data, with particular attention to estimation of background variation and appropriate transformations to approximate Gaussian variables so that familiar inference procedures can be applied. Other issues can affect the analysis as well.

(1) *Negative values*: While transformation (2) successfully handles negative values (because $(\alpha/\beta + y^2)^{1/2} > y$ when $\alpha/\beta > 0$, even when $y < 0$), transformation (4) may be undefined if the adjusted value is less than $a - b/g$. Two such values arose in blocks: one in $[\text{row}, \text{column}] = [7, 22]$ and one in $[23, 23]$. The foreground and

background values around these two locations are shown below:

Red foreground and background values around row 7, column 22

	Foreground				Background			
	20	21	22	23	20	21	22	23
5	677	684	600	676	273	267	263	250
6	658	1302	1149	869	276	267	258	250
7	557	630	**255**	786	278	272	**260**	249
8	660	774	1840	676	283	269	264	258
9	653	661	626	575	273	267	263	263

Red foreground and background values around row 23, column 23

	Foreground				Background			
	20	21	22	23	20	21	22	23
20	694	705	696	669	345	344	328	301
21	547	828	844	848	347	344	327	297
22	695	784	609	767	361	344	319	293
23	872	858	654	**277**	357	339	311	**283**

In both instances, the foreground counts (255, 277) are substantially lower than the surrounding values, and in fact are lower than some of the surrounding background values. One suspects that spots at these locations failed to be deposited at all. For this analysis, both values were replaced by a, which transforms to a value of 0. The robustness of the procedures used in the analysis renders the results insensitive to these two replacements.

(2) *Fitting g transformations simultaneously*: An argument could be made for using the same skewness parameter g and scale parameter b in both the red and green adjusted foreground counts. (A common location parameter a would not be wise, for reasons mentioned earlier, concerning the laser scanner electronics and the possible degradation of cDNA between times of the two scans.) In fact, one might consider fitting a common g for all blocks. In this analysis, separate g values were fitted for each channel in each block (i.e., 64 separate g parameter estimates). A plot of the g-values fitted to the green channel data versus those fitted to the red channel data indicated a near linear relationship between them, with the green g being just slightly smaller than the red g (the regression slope coefficient is 0.95 with a standard error of 0.01). Alternatively, a bivariate version of (4) (i.e., where Z is bivariate Gaussian) might be possible to derive.

(3) *Correlation estimates*: Because Z_R and Z_G are so highly correlated, robust estimates of the correlation coefficient were also calculated, in addition to the usual Pearson correlation coefficient. Mosteller and Tukey (1977, p. 211) propose a robust estimate of the correlation $cob(x, y) = sgn(slope)\,[1 + s_{bi}^2(y - slope\,x)/(slope^2\, s_{bi}^2(x))]^{-1/2}$ where *slope* is a robust estimate of the slope in the regression of y on x, and s_{bi}^2 is a robust estimate of the variance (of either the data x or

the residuals $y - slope\, x$) based on the ψ-function for the biweight. The efficiency of such an estimate has not been studied, particularly for such highly correlated variables, so Pearson's correlation coefficient was used except those where one or two outliers clearly affected the estimates. Usually the difference between the two estimates was less than 3%, with the Pearson estimate typically being larger than the robust estimate.

(4) *Automation*: The 7-step prescription for this analysis was developed only after intensive scrutiny of the data in one of the 32 blocks. Subsequent analysis on the remaining 31 blocks was quick and required little intervention. However, the coded residual plot from the median polish fits did require individual inspection (although presumably one could apply tests for spatial randomness to flag the existence of patterns in the residuals), as did the choice between the robust and the Pearson correlation coefficient.

(5) *Oligonucleotide arrays*: Many of these same principles could be applied, with modification, to the data from oligonucleotide arrays, with "foreground" and "background" being replaced by "PM" and "MM". Finkelstein et al. (2002) note the inadequacy of a single parameter fit for oligonucleotide probe pairs, so the three-parameter fit in (4) might work well. A procedure for such arrays is currently being investigated.

Many other issues surrounding effects of process and material variation, spatial and time trends, and common fitting of data transformations will likely lead to a more sensitive analysis. As with any exploratory analysis, the "statistical significance" of these results must be confirmed with replication, before claiming "biologically significance." These experiments have been replicated; a complete analysis of all of the data that attempts to account for the issues raised above will be explored in further work.

Acknowledgements

We are grateful to Dr. Brian Soriano and Jeannette Gaydos for their help with many of the details involved in cDNA slide manufacturing, and especially for the use of the data analyzed in Section 5. We also thank Cary Miller (UCHSC), Kim Kafadar (Stanford), and Layne Watson, Naren Ramakrishnan, Jonathan Watkinson, and Roger Ehrich (Virginia Tech) for useful discussions. Finally, we acknowledge with great appreciation the very insightful and constructive comments from the referees and from Dr. David C. Hoaglin, which led to a much improved version of this paper. Part of this work was conducted while Kafadar was a visiting Professor at Virginia Tech, whose support is also acknowledged.

References

Affymetrix©, 2000. User's Manual 4.0.
Affymetrix©, 2002. Statistical Algorithms Description Document.

Amaratunga, D., Cabrera, J., 2001. Analysis of data from viral DNA microchips. J. Amer. Statist. Assoc. 96 (456), 1161–1170.

Axon Instruments, 2001. GenePix Pro© 4.0, Array Acquisition and Analysis Software for the GenePix 4000B User's Guide and Tutorial, Part Number 2500-137 Rev E. Union City, California.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B 57 (1), 289–300.

Brazma, A., Vilo, J., 2000. Gene expression data analysis. Federation European Biochem. Soc. Lett. 480, 17–24.

Cleveland, W.S., 1985. The Elements of Graphing Data. Wadsworth, Pacific Grove, CA.

Curtiss, J.H., 1943. On transformations used in the analysis of variance. Ann. Math. Statist. 14, 107–132.

Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P., 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statist. Sinica 12, 111–139.

Durbin, B., Hardin, J.S., Hawkins, D.M., Rocke, D.M., 2002. A variance-stabilizing transformation for gene expression microarray data. Bioinformatics 18, 105S–110S.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96, 1151–1160.

Emerson, J.D., Hoaglin, D.C., 1983. Analysis of two-way tables by medians. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), Exploring Data Tables, Trends, and Shapes. Wiley, New York, pp. 212–242 (Chapter 6).

Finkelstein, D.B., Gollub, J., Cherry, J.M., 2002. Normalization and systematic measurement error in cDNA microarray data. Unpublished manuscript.

Hoaglin, D.C., 1985. Summarizing shape numerically: the *g*- and *h*-distributions. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), Exploring Data Tables, Trends, and Shapes. Wiley, New York, pp. 461–513 (Chapter 11).

Horn, P.S., Kafadar, K., 2002. Trimming and Winsorization. In: El-Shaarawi, A.H., Piegorsch, W.W. (Eds.), Encyclopedia of Environmetrics, Vol. 4. Wiley, New York, pp. 2264–2267.

Illouz, K., 2002. Microarray technology: challenges and possibilities. Unpublished manuscript.

Insightful, Inc., 2001. S-Plus Version 6.0.1 Release 1, for Linux 2.2.12. Seattle, Washington.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.C., Antonellis, K.J., Scherf, U., Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4 (2), 249–264.

Kerr, M.K., Churchill, G., 2001a. Statistical design and the analysis of gene expression microarray data. Genetics Res. 77, 123–128.

Kerr, M.K., Churchill, G., 2001b. Experimental design for gene expression arrays. Biostatistics 2, 183–201.

Knight, J., 2001. When the chips are down. Nature 410, 860–861.

Ku, H.H., 1966. Notes on the use of propagation of error formulas. J. Res. Natl. Bureau Standards-C, Eng. Instrum. 70C (4), 263–273 (Reprinted In: Ku, H.H. (Eds.), Precision Measurement and Calibration: Selected NBS Papers on Statistical Concepts and Procedures. NBS Special Publication 300, Vol. 1, 1969, pp. 331–341).

Lee, ei-Ling Ting, uo, F.C., Whitmore, G.A., Sklar, J., 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. 97, 9834–9839.

Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA.

Nicolls, M.R., D'Antonio, J.M., Hutton, J.C., Gill, R.G., Czuornog, J.L., Duncan, M.W., 2003. Proteomics as a tool for discovery: proteins implicated in Alzheimer's disease are highly expressed in normal pancreatic islets. J. Proteome Res., 2, 199–205.

Pan, W., 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18, 546–554.

Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19, 368–375.

Rocke, D.M., Durbin, B., 2001. A model for measurement error for gene expression arrays. J. Comput. Biol. 8, 557–569.

Rocke, D.M., Lorenzato, S., 1995. A two-component model for measurement error in analytical chemistry. Technometrics 37, 176–184.

Satagopan, J.M., Panageas, K.S., 2003. A statistical perspective on gene expression data analysis. Statist. Med. 22, 481–499.

Soille, P., 2003. Morphological Image Analysis: Principles and Applications. 2nd Ed. Springer, New York.

Storey, J.D., 2001. The positive false discovery rate: a Bayesian interpretation and the q-value. Submitted for publication.

Tukey, J.W., 1949. One degree of freedom for non-additivity. Biometrics 5, 232–242 (Reprinted in: Cox, D.R. (Eds.), The Collected Works of John W. Tukey, Vol. VII, Factorial & Anova, 1949–1962. Wadsworth, Monterey, CA, 1992, pp. 1–13.).

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Tusher, V., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. Proc. Natl. Acad. Sci. 98, 5116–5121.

Vardeman, S.B., 1994. Statistics for Engineering Problem Solving. PWS Publishing, Boston, MA.

Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika 55, 1–17.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R.S., 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. 8, 625–638.

Yang, Y.H., Buckley, M.J., Dudoit, S., Speed, T.P., 2002a. Comparison of methods for image analysis on cDNA microarray data. J. Comput. Graph. Statist. 11, 108–136.

Yang, Y.H., Speed, T.P., 2002b. Design issues for cDNA microarray experiments. Nature Rev. 3, 579–588.

Yang, Y.H., Dudoit, S., Luu, P., Speed, T.P., 2002c. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl. Acids Res. 30, E15.