Analysis of Data From Viral DNA Microchips

Dhammika AMARATUNGA and Javier CABRERA

Viral DNA microchips, arrays of viral genes printed over a glass slide, are powerful tools for rapidly characterizing the expression pattern of these genes in an infection. The chips are exposed to a solution of fluorescently labeled cDNAs prepared from either mock or true infected human fibroblast cells and the expression levels of the various genes are recorded with the objective of detecting which viral genes are expressed to a significantly higher degree when exposed to the true infection as compared to the mock infection. The data were initially examined visually via image plots and scatterplots. These reveal that analysis of such data presents many challenges owing to, among other problems, high interchip and intrachip variability with low signal-to-noise ratio, differential intensity scales that have to be adjusted nonlinearly, nonGaussian data, data for a large number of genes with little replication, scratches and dark spots on the chips, dust, outliers, and an inability to quantitate intensities below a detection limit, or above a threshold. The first step of the analysis was to standardize the chips to a single intensity scale using a photograph analogy. Next, the average expression level of each gene was estimated using a highly resistant repeated median estimator to avoid being misled by aberrant values. Finally, a simulation-based approach was used to make a distribution-free assessment of significance.

KEY WORDS: Gene expression; Median; Mixed effects model; Monte Carlo simulation; Resistant location estimation; Standardization.

1. INTRODUCTION

DNA microchips, large arrays of DNA fragments attached to a glass surface in a highly dense format (Schena, Shalon, Davis, and Brown, 1995), are fast becoming an indispensable tool in the armamentarium of the research molecular biologist. A recent *Science* article states "One of the most important experimental approaches for discovering the function of genes promises to be gene chips and microarrays" (Somerville and Somerville, 1999). A recent development in the rapid evolution of this tool is the viral DNA microchip (Chambers, et al. 1999).

A viral DNA microchip is an array of genes from a viral genome printed over a dime-sized rectangular glass slide (a gene is a fragment of DNA that encodes a specific functional product, usually a protein). Such a chip can be used to quickly characterize the expression pattern of these viral genes during an infection.

To do this, the experimenter infects a sample of human foreskin fibroblast cells with the virus. RNA is harvested from this sample 72 hours later with the expectation that RNA corresponding to viral genes that are expressed owing to the infection will appear in large quantities in the sample. Fluorescently labeled cDNA is prepared from this RNA using reverse transcription. A probe is prepared from the resulting cDNA. The experimenter then exposes the chip to this probe for 4 hours, after which the chip is washed and dried. Finally the chip is scanned using a confocal laser microscope.

The fluorescently labeled cDNA of any viral gene that is expressed owing to the infection would have been present in the probe and would have hybridized to the corresponding DNA printed on the chip, resulting in fluorescence at that spot on the chip. Measuring the fluorescence at each spot thereby reveals which genes are expressed owing to the infection. The higher the level of expression, the higher the intensity of fluorescence.

In the experiment under consideration here, one set of chips was exposed to a probe as described above and a separate set of chips was exposed to a solution of fluorescently labeled cDNAs prepared from mock infected human fibroblast cells as a negative comparator. The objective of the analysis then, is to detect which genes were upregulated, that is, expressed to a significantly higher degree when exposed to the true infection as compared to the mock infection.

Analysis of data from microchips presents many challenges owing to, among other problems, high interchip and intrachip variability with low signal-to-noise ratio, nonGaussian data, differential intensity scales that have to be adjusted nonlinearly, data for a large number of genes with little replication, scratches and dark spots on the chips, dust, outliers, and an inability to quantitate intensities below a detection limit or above a threshold. Other work in this area, such as Chen, Dougherty, and Bittner (1997) and Slonim, Tamayo, Mesirov, Golub, and Lander (2000) take parameteric approaches that do not address many of these issues.

2. DATA AND GRAPHS

Viral DNA microchip technology is still in its infancy. Thus researchers in the field tend to utilize a variety of experimental designs for their experiments. The experiment of interest in this article involved six chips, C1–C6. On each chip was printed 233 viral genes, the entire known genome of the human cytomegalovirus (HCMV), the largest member of the herpes virus family and one of the largest known viral genomes. Also printed on the chip were 103 cellular genes to act as internal controls to standardize across chips (see, Section 3). There were three replicates of each of these 336 genes and 15 blank spots on each chip. Thus each chip has 1023 spots arranged in a 31×33 array. Chips C1, C2, and C3 were exposed to the solution of fluorescently labeled cDNAs prepared from the mock infected human fibroblast cells whereas chips C4, C5, and C6 were exposed to the probe

Dhammika Amaratunga is Research Fellow, Preclinical Biostatistics, The R.W. Johnson Pharmaceutical Research Institute, Raritan, NJ 08869 (E-mail: *damaratu@prius.jnj.com*). Javier Cabrera is Associate Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08855 (E-mail: *cabrera@stat.rutgers.edu*) and Visiting Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore. The microchip data analyzed in this article was provided by Peter Ghazal (Scripps Institute) and Mark Erlander (RWJPRI). The authors gratefully acknowledge the editor, associate editor, and referees for providing valuable feedback that improved the manuscript. The authors thank Jim Colaianne (RWJPRI) for his support during this project.

solution of fluorescently labeled cDNAs prepared from the true infected human fibroblast cells.

The measured quantity of interest is the intensity of fluorescence emitted at each gene in the array, the level of intensity indicating the level of expression of that gene as outlined above. Let X_{tgcr} denote the intensity of fluorescence measured at the rth replicate (1, 2, 3), cth chip (1, 2, 3), gth gene (1–336), and tth type of infection (mock(0) or true(1)).

Table 1 shows the data for two genes, a cellular gene CCND3 and a viral gene UL63. Intensities at spots that were impossible to distinguish from background were recorded as zeroes. The intensities that were above the upper measurement threshold were recorded approximately at the threshold; the threshold was about 64×10^6 for five of the six chips; the maximum intensity level of chip C4 was slightly higher, but all chips were adjusted to the same threshold of 64×10^6 (Table 1 shows data prior to this adjustment).

Table 1 illustrates some of the characteristics of the data that complicate their analysis. The three zeroes for UL63 in chip C2 is a case where, for that gene, the entire chip is an outlier as there are no other values that could not be quantitated for UL63. At the other extreme, the three values for the same gene in chip C5 are all measured near the threshold. There are also cases of sporadic individual outliers such as the third and perhaps the first values for gene CCND3 in chip C5. Despite these problems, however, it is reasonable to conclude from eyeballing the data that CCND3 is not expressed at a higher degree in chips C4–C6 compared to chips C1–C3, whereas UL63 is expressed; i.e., that cellular gene CCND3 is not affected by the infection but viral gene UL63 is affected.

As we are dealing with a skewed distribution of intensities, the square roots of the intensities were observed to reveal the features of the data better than either the raw data or other power transformations we tried (see also Sec. 5). Let $X'_{tgcr} = \sqrt{X_{tgcr}}$. Henceforth all analyses will be done on the square root scale and, in the interest of brevity, we shall also refer to X'_{tgcr} as the intensity.

Because each microchip is an array of intensities, it is natural to display them on a color scale in an image plot (see Fig. 1). The upper 10 rows of the chip contain cellular genes and the lower 21 rows contain viral genes; the 15 blanks were placed in the rightmost column of the top 15 rows.

Chips C1–C3 (top row of Fig. 1) are mostly red and orange indicating low levels of expression throughout. In contrast, the lower portions of chips C4 and C5 have large swaths of green and blue indicating the viral genes that are being expressed by infection. At the first sight, chip C6, despite having been treated similar to these two chips, does not appear to show this.

3. STANDARDIZING THE MICROCHIPS

A troublesome difficulty one encounters with microchip data is that the overall intensity of a chip can vary quite substantially from chip to chip. The chip effect resembles the clarity of a photograph, some chips are darker whereas some are lighter. This is what we saw with chips C4–C6 in Figure 1. The image corresponding to chip C6 is darker than the other two and it is harder to see its features.

The chip effect can also be observed in Figure 2 which shows scatterplots for pairs of chips. It also can be seen that the intensity scale of chip C6 is quite different from that of the other two similarly treated chips, C4 and C5. It also appears that the intensity scale of chip C2 is different from that of the other two similarly treated chips, C1 and C3.

Prior to the analysis, it is therefore necessary to standardize the chips to have similar intensity scales. It is evident from Figure 2 that a linear transformation would not suffice. In keeping with the photograph analogy, we borrowed a methodology derived from techniques applied to images in computer vision (Cho, Meer, and Cabrera, 1997) to perform the standardization.

The intensity of a spot is a number in an interval $U = [0, M]$, where $M = 8,000 = \sqrt{(64 \times 10^6)}$ corresponds to the maximum measurable intensity. We expect that images from chips treated similarly would have similar distributions of intensities. In the field of computer vision, images are compared by graphing histograms of intensities (Cho, et al. 1997). If the histograms of intensities of a set of images show very different distributions, it suggests a need to standardize them. To standardize one or more images, we apply nondecreasing

Table 1. Fluorescence Measurements for Two Genes: CCND3 and UL63, Prior to Standardization (RAW) and After Standardization (STD)

				Mock infection			True infection		
Data	*Gene*	*Row*	*Col*	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*
RAW	CCND3	24	31	3411877	3079894	3024930	4184756	7329618	2251887
	CCND3	24	32	5796746	5729780	4357549	5458637	3552465	3508138
	CCND3	24	33	4989971	5904659	3657488	4525145	0	3518544
	UL 63	8	7	692385	0	2671360	58289754	65334938	44343919
	UL 63	8	8	868369	0	1494960	56203312	65029607	36928430
	UL 63	8	9	1623053	0	2651441	59093670	65175007	50646302
STD	CCND3	24	31	3840665	1279666	3890058	3832975	6154504	6489071
	CCND3	24	32	7426011	2772692	5883412	4821614	3070276	9430339
	CCND3	24	33	5883924	2899522	4561098	4274747	0	9435680
	UL 63	8	7	626499	0	3540575	37317340	64000000	44960908
	UL 63	8	8	688344	0	1750436	34127972	64000000	41003504
	UL 63	8	9	1625457	0	3511113	38851828	64000000	50679744

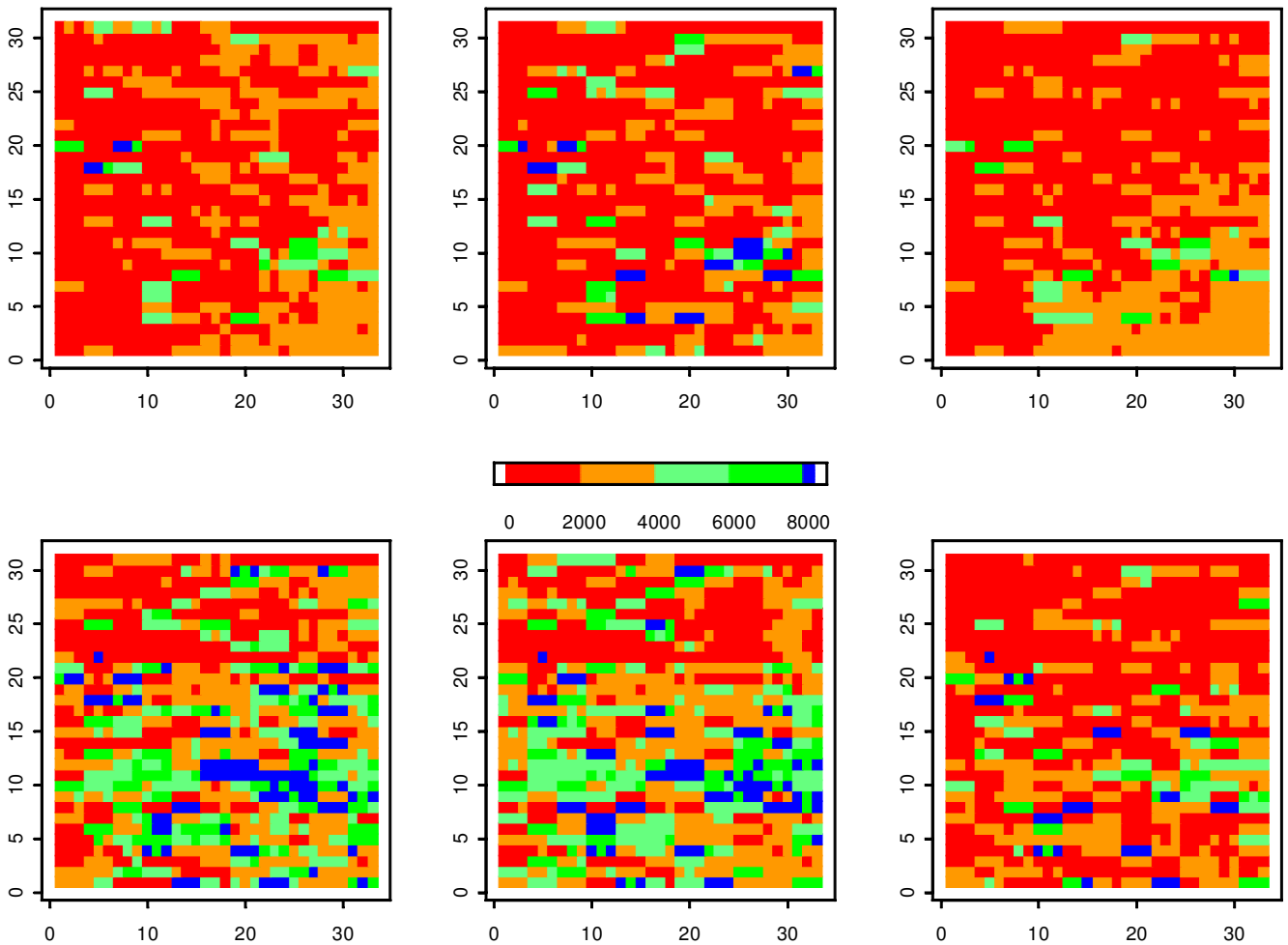

Figure 1. Image Plots of the Intensity Levels for the Six Chips Before Standardization. The top row shows the mock infected chips C1–C3. The bottom row shows the true infected chips C4–C6. The rows and columns of each image correspond to the rows and columns of the chip.

transformations h: U $\rightarrow$ U. If h is convex, the transformation renders the image "darker," whereas if h is concave, the resulting image will be "lighter." The choice of h depends on the situation. Here we have two or more images we want to standardize to a target image that is generally the median (or mean) of the group, calculated pixel by pixel. We considered two groups of transformations:

(i) *Bilinear or spline transformation with one or more change points.* For the bilinear case, h is of the form $h(t) = at + b(t-c)_+$, where $a > 0, b > 0, 0 < c < M$, with the constraint that $h(M) = M$. The function $(t)_+$ is equal to t if $t > 0$ and 0 otherwise. This function is bilinear with a change point at c and is the simplest form of $h(t)$. This can be generalized to a multilinear function or to a cubic spline over a group of knots at a fixed set of ts. The function $h(t)$ can be estimated by least squares using the target image as the predictor and the original image as response and then inverting the transformation.

(ii) *Smoothed QQ-plot.* This transformation is obtained by smoothing the scatter plot of the quantiles of the original image versus the quantiles of the target image. Again we apply h^{-1} to the original image to obtain the standardized one.

Note that although we could have fitted the $h(t)$ function directly using the original image as predictor and the target image as the response, instead we chose to define $h(t)$ with the target image as predictor because, having been calculated as the average of a set of images, it exhibits less noise than the original image.

Here we have two sets of three images, [C1, C2, C3] and [C4, C5, C6]. Each set of three images corresponds to three replicates of the same experiment on three different chips. Based on the distribution patterns of intensities across the chips, it was enough to use the simplest h from group (i), that is, a bilinear function with a change point at c to standardize each set separately.

We calculated the median microchip for C1–C3 by taking the median of the three at each spot. This median microchip served as a "standard." The transformation h of the above form that transforms chips C1–C3 to this standard pattern was estimated using least squares. A similar exercise was performed with C4–C6. The change points, c, for C2–C6 (C1 did not need to be transformed) were, 7,426, 4,745, 6,849, 7,493, 3,856.

The next step was to standardize across all the six chips. The 103 cellular genes is essentially a selection of genes that were expected to be unaffected by viral infection, so they

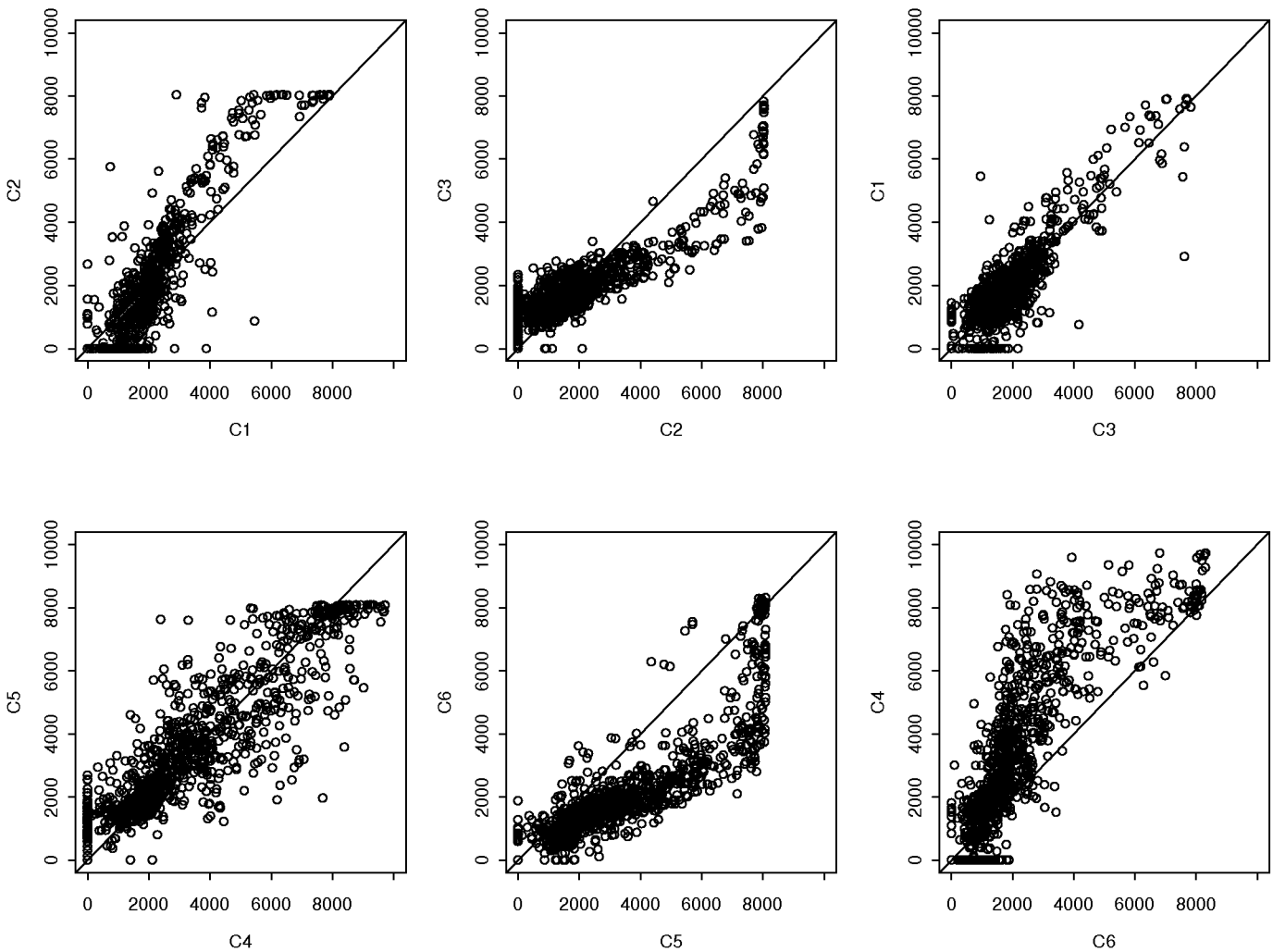

Figure 2. Pairwise Scatterplots of Chips C1–C3 (top row) and Chips C4–C6 (bottom row) Before Standardization.

should have similar expression levels across all the chips. Therefore the six chips were standardized to the median of the two median microchips described in the previous paragraph using the cellular genes only; each microchip was transformed using an h from group (ii), so that its percentiles (1st–99th) coincided with those of the median–median microchip.

Let Y_{tgcr} denote the standardized value of X'_{tgcr}. The results of the complete standardization are shown in Figures 3 and 4 (Table 1 gives the particular values for CCND3 and UL 63). We observe that the region on the top third of the microchips that corresponds to the cellular genes, now appears homogeneous across the six chips. Also, the dark chip C6 is now in line with its sister chips C4 and C5. The standardization process has transformed the six chips to a single scale on which the data can be modeled and inferences can be derived. Interestingly, it is possible to perceive a slight color trend from left to right in a way reminiscent to what can often be seen in photographs.

The standardization process was performed using S-PLUS. We are also working on a graphical user interface that implements this part of the method interactively. This would enable a practitioner to eyeball the results of various attempts at standardization. We find that the use of an interactive graphical tool greatly facilitates communication with the scientists. The interface will include many other types of transformations and algorithms to estimate them.

4. ESTIMATING THE AVERAGE EXPRESSION LEVEL FOR A GENE

The average expression level of each gene for each type of infection is estimated resistantly. High resistance is crucial for a number of reasons. The standardization was a global adjustment that did not quite eliminate local effects and aberrations. Thus the standardized data contain numerous outliers caused by deficiencies of the chip, such as scratches or grey patches, nonquantifiable observations, and inexplicable individual large or small measurements. Use of a resistant approach reduces the impact of such problems.

Thus, the average expression level of the gth gene with the tth treatment is estimated as $Y_{tg\cdot\cdot} = \text{median}_c \text{median}_r [Y_{tgcr}]$. Although the breakdown point of this estimator is only 4/9 for the 3×3 layout for each gene in the experiment, this estimator is actually more resistant than this against certain patterns of outliers, including one entire set of bad values from one chip and one more from each of the other two chips.

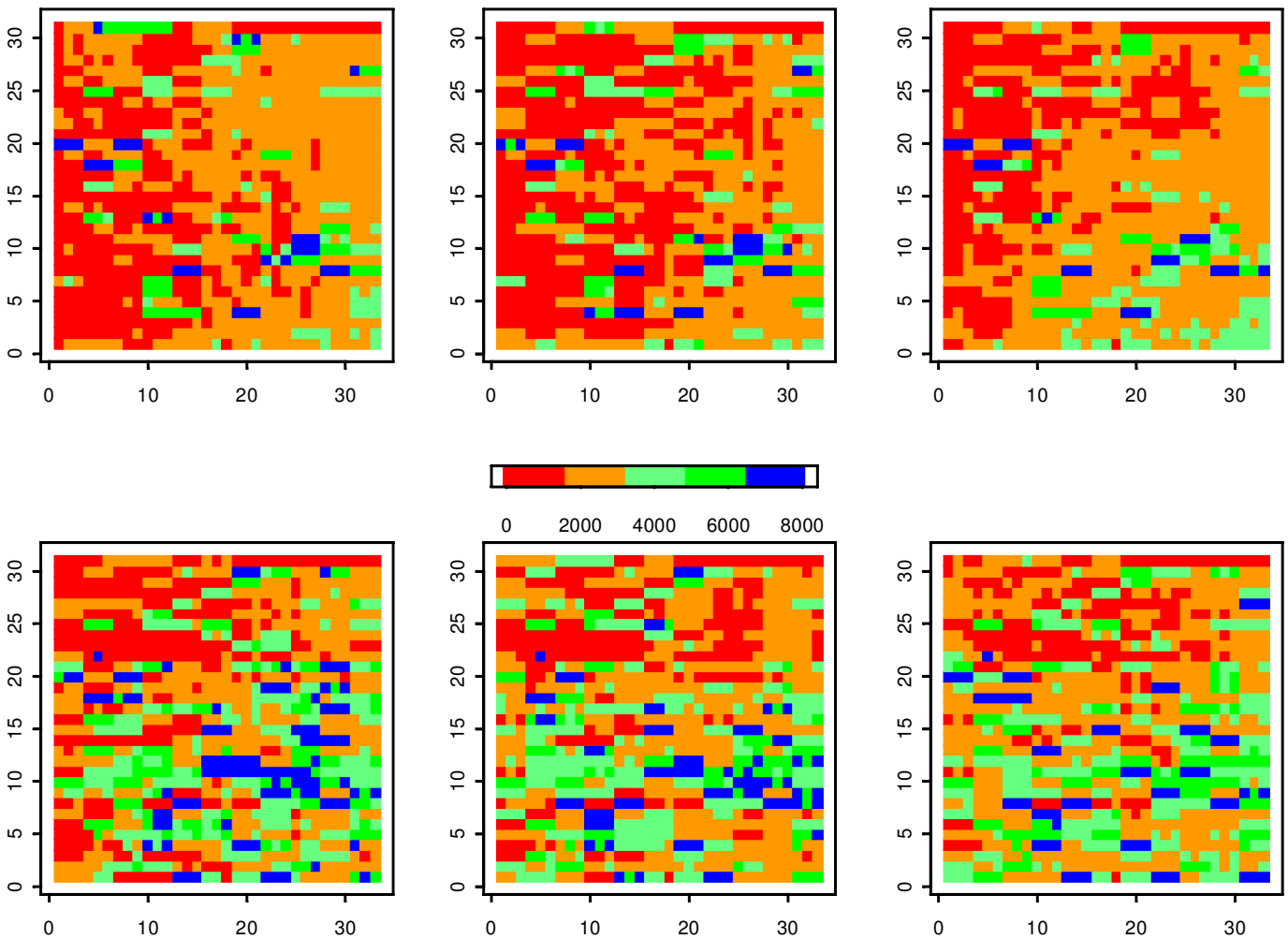

Figure 3. Image Plot of the Intensity Levels for the Six Chips After Standardization. The top row shows chips C1–C3; the bottom row shows chips C4–C6.

5. MODELING THE EXPRESSION LEVEL

The intensities may be modeled as $Y_{tgcr} = \mu_{tg} + \alpha_{tgc} + \varepsilon_{tgcr}$, with $\alpha_{tgc} \sim \text{dist}(0, \gamma^2)$ and $\varepsilon_{tgcr} \sim \text{dist}(0, \sigma^2)$, where the notation "dist(a, b)" refers to an unspecified symmetric unimodal distribution with mean a and variance b. This is essentially a mixed effects model, in which the average expression level for the gth gene and the tth type of infection, μ_{tg}, is a fixed effect and the between chip effect α_{tgc} and the within chip effect ε_{tgcr} are random effects with variance components γ^2 and σ^2, respectively. The between chip effect refers to any "local" differences between chips that still remain after the standardization, which is a nonlinear "global" adjustment for overall chip differences. Observe that this model assumes that both the interchip differences and the intrachip differences are homoscedastic. The square root reexpression of the intensities was successful in bringing the data into conformance with both this assumption and the assumption of symmetry.

To see this, we examine the "error residuals", $R_{tgcr} = Y_{tgcr} - \text{median}_r[Y_{tgcr}]$, and the "chip residuals" $D_{tgc} = \text{median}_r[Y_{tgcr}] - Y_{tg..}$. The plots of R_{tgcr} versus $Y_{tg..}$ and D_{tgc} vs $Y_{tg..}$ (top row of Fig. 5) show the validity of the homoscedasticity assumption. A small boundary effect caused by the nonquantifiable observations can be seen in the second of these plots; our analysis will take this into account. The plots of the upper quantiles of R_{tgcr} versus the absolute values of the lower quantiles of R_{tgcr} and the same for D_{tgc} (bottom row of Fig. 5) show the validity of the symmetry assumption.

6. ESTIMATING THE INFECTION EFFECT FOR A GENE

The quantity $Y_{tg..}$ is a median-unbiased consistent estimator of μ_{tg}. If we assume an additive treatment effect, $\mu_{tg} = \nu_g + \tau_{tg}$, where ν_g refers to the expression level of the gth gene under mock infection, a natural estimate of the increase τ_{tg} in its expression level owing to infection is $T_g = Y_{1g..} - Y_{0g..}$. A value of T_g substantially exceeding 0 indicates that the gth gene is upregulated owing to infection. Of the 233 viral genes, 205 (88% versus 58% for cellular genes) had T_g values exceeding 0.

7. TESTING FOR UPREGULATION DUE TO INFECTION

The next step in analysing the data is to assess, for each gene, the significance of the difference in level of gene expression between the true and mock infection. We shall use T_g as the test statistic for testing the null hypotheses H_g: $\mu_{1g} = \mu_{0g}$

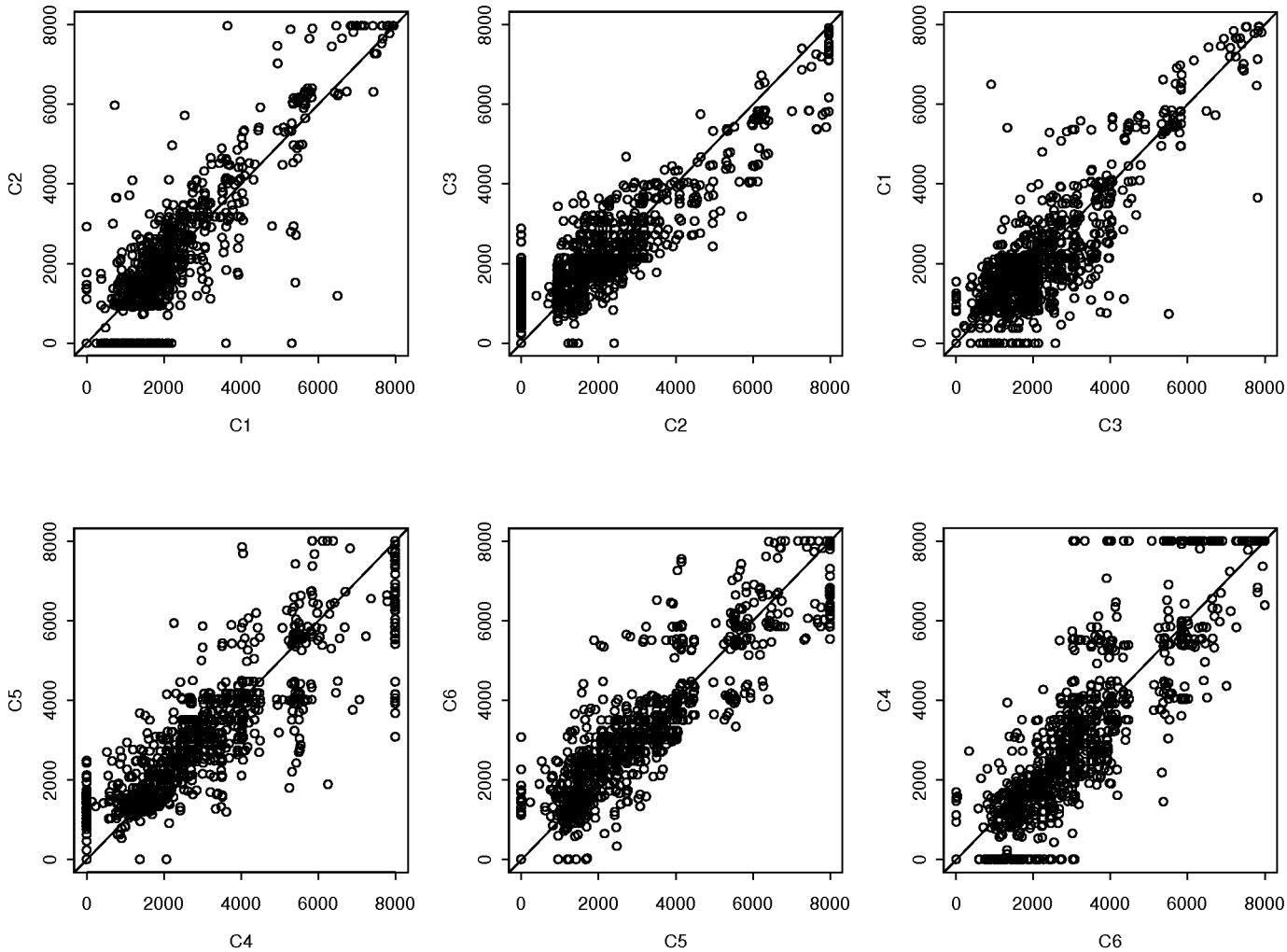

Figure 4. Pairwise Scatterplots of Chips C1–C3 (top row) and Chips C4–C6 (bottom row) After Standardization.

versus K_g: $\mu_{1g} > \mu_{0g}$, that is, for testing whether the gth gene is more highly expressed in the true infection than in the mock infection, larger values of T_g indicating less fidelity with H_g.

If we assume that the error distributions are Gaussian and use a conventional mixed effects modeling procedure, then the outliers and other aberrations in the data will inflate the variance estimates, resulting in tests with very low power. Many upregulated genes will be missed if we took this approach. In addition, the presence of nonquantifiable and truncated values will complicate the analysis. We therefore decided to take a nonparametric Monte Carlo simulation-based approach that overcomes these problems.

To do this, we need to generate a null distribution of T_g under the null hypothesis H_g. To be able to borrow strength across genes and residuals, rather than the actual observations, would be the values that are resampled. Because there are two sources of variability, there are two sets of "residuals," $\{Y_{tgcr}\}$ and $\{D_{tgc}\}$, that must be resampled.

In a standard regression problem, the resampling may be done by regarding the empirical distribution of the residuals as an estimator of the error distribution. In our case this would be incorrect because the empirical distributions of the residuals (both $\{R_{tgcr}\}$ and $\{D_{tgc}\}$), which are comprised of 1/3 zeroes, 1/3 positive values, and 1/3 negative values, do not adequately approximate the error distributions. Thus a modification is necessary.

8. RESAMPLING ZERO HEAVY RESIDUALS

It turns out that there is a surprisingly simple, yet effective, solution to this problem. Keep in mind that the observed positive residuals are the third-order statistic (maximum) minus the second-order statistic (median) for three observations. It is well-known that, if F is the error distribution, then the distribution of the third-order statistic of three observations is $F_{(3)}(t) = F^3(t)$. The distribution of the third-order statistic minus the second-order statistic for three independent observations has a more complicated functional form, but this basic result suggests that one way to estimate $F(t)$ is by some transformation of the empirical distribution functions of the observed positive and negative residuals. In our case, as the error distributions are symmetric, we define $R^+(t)$ as the empirical distribution of absolute values of the residuals excluding the zeros. We had great success with transformations of the form

$$\widehat{F}(t) = 0.5 + R^+(\alpha t^\beta)/2 \quad t > 0,$$
$$\widehat{F}(t) = 0.5 - R^+(\alpha t^\beta)/2 \quad t < 0.$$

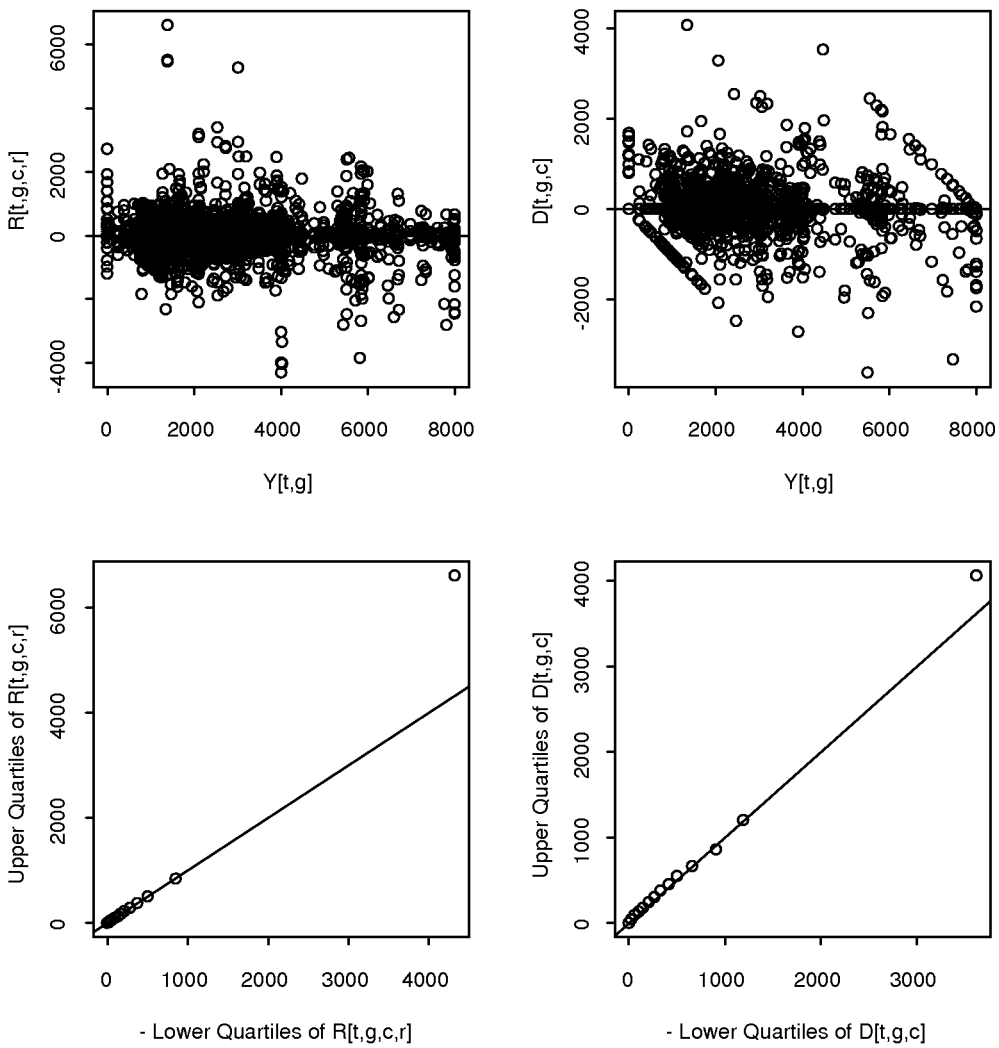

Figure 5. Residual Plots. The plot on the top left shows the error residuals R_{tgcr} versus the average expression level of the gth gene under the tth type of infection, $Y_{tg..}$. The plot on the top right shows the chip residuals D_{tgc} vs. the average expression level of the gth gene under the tth type of infection, $Y_{tg..}$. The plot on the bottom left shows the upper 2.5, 5.0, 7.5,. . . ,97.5 quantiles of $\{R_{tgcr}\}$ vs. the absolute values of the corresponding lower quantiles of $\{R_{tgcr}\}$. The plot on the bottom right shows the upper 2.5, 5.0, 7.5,. . . ,97.5 quantiles of $\{D_{tgc}\}$ vs. the absolute values of the corresponding lower quantiles of $\{D_{tgc}\}$.

The justification for this type of transformation can be found in the ideas of the exploratory data analysis. Tukey (1977) proposed this family of transformations as it produces a rich class of shapes derived from an initial distribution. This class provides a wide array of distribution tail shapes, ranging from light to heavy. This transformation can be used to adjust the tails and scale of the residual distribution to behave like those of the original error distribution.

To estimate α and β, we resample the errors by replacing F with the observed R^+. That is, we construct a distribution R, the empirical distribution of the combination of the absolute values of the residuals and the same values with negative signs. R is a symmetric version of R^+. The functional form for R is

$$R(t) = 0.5 - R^+(t)/2 \quad \text{if } t < 0$$

$$R(t) = 0.5 + R^+(t)/2 \quad \text{if } t > 0.$$

The algorithm for estimating F is as follows:

(i) Simulate $3m$ errors from the distribution R with replacement.
(ii) Arrange them into m samples of three and calculate the third-order statistic minus the second-order statistic (i.e., the maximum minus the median) and the second minus the first-order statistics; call this vector V; V serves as the resampling analog of R^+.
(iii) Calculate the percentiles from 1% to 99% from V and the same percentiles from R^+; call these two vectors q_v and q_r.
(iv) Fit a robust regression line to the model $\log(q_r) = \log(\alpha) + \beta \log(q_v)$ and replace α and β in the above equation with these estimates.

The idea of the algorithm is that the same α and β that transforms V into R^+ will transform R^+ into the positive and negative sides of F.

Note that in (iii), the use of percentiles 1–99 is just a convenient choice that, combined with the robust regression fit in (iv), have to give a reasonable estimate of the transformation.

We tried this procedure with simulated examples from several symmetric distributions with different degrees of tail weight: a standard Gaussian distribution, a t_5 distribution, a Cauchy distribution, and a centered Beta (2, 2) distribution, for sample sizes of the chip data. Figure 6 shows the QQ-plot of the error distribution estimated by this algorithm compared to the true error distribution for the simulated Gaussian example. All examples produced nearly perfect matches like Figure 6. Thus we are satisfied that this procedure will generate a reasonable simulation error distribution for the chip data.

We note in passing that this procedure is very general and can be applied to simulation residuals in other problems that rely on medians of a few observations, two other examples of which are median polish and smoothing by medians. A more general version of the method, including the nonsymmetric case, will be published elsewhere.

9. TEST RESULTS

Now we proceed as follows. We draw a random sample $\{R^*_{tgcr}\}$ with replacement from $\{R_{tgcr}\}$ using the procedure outlined above and a random sample $\{D^*_{tgc}\}$ with replacement from $\{D_{tgc}\}$ again using the procedure outlined above. Keeping in mind that we must generate a simulated version $\{Y^*_{tgcr}\}$ of $\{Y_{tgcr}\}$ under H_g, we take $Y^*_{tgcr} = D^*_{tgc} + R^*_{tgcr}$. The value of Y^*_{tgcr} is restricted to be between 0 and $M = 8{,}000$, in analogy with the observations. From $\{Y^*_{tgcr}\}$, we calculate a simulated version T^*_g of T_g.

By repeating this process a large number of times, we generate a simulated null distribution of $\{T^*_g\}$, from which we determine the critical value c_α for an α level test as the value c_α such that $100(1-\alpha)\%$ of $\{T^*_g\}$ values exceed c_α.

We could make a Bonferroni adjustment for multiple testing by dividing α by the number of tests performed prior to determining the critical value. This protects the experimentwise error rate, but given the large number of tests being performed, results in the individual tests having low power against moderate upregulation. Therefore we analyze the data both with and without adjustment. Genes found significantly upregulated with the adjustment are the ones that we are most confident are truly upregulated. Genes not significant with the adjusted analysis but significant with the unadjusted analysis are also useful for the researcher who did not want to miss moderately upregulated genes.

Table 2 summarizes the findings from this analysis. Of the 233 viral genes, 141 (61% versus 14% for the cellular genes) were found to be significantly upregulated at the 5% unadjusted level and 50 (21% vs. 1% for the cellular genes) at the 5% adjusted level. Figure 7 is an image plot that shows the significance levels of the genes on a color scale, with blue and green representing genes significant at the 1% and 5% levels, respectively and red representing genes that were not significant at the 5% level. The top third of the chip, which contains the cellular genes, appears mostly red. The lower two-thirds is mostly blue and red indicating that the viral genes are either highly expressed or hardly expressed at all.

Some additional comments pertaining to Table 2 follows:

1. More cellular genes than would be expected owing to chance alone appear to be significantly upregulated. In fact, it was anticipated that a few would be, with at most a handful expressed to as high a degree as an upregulated viral gene.

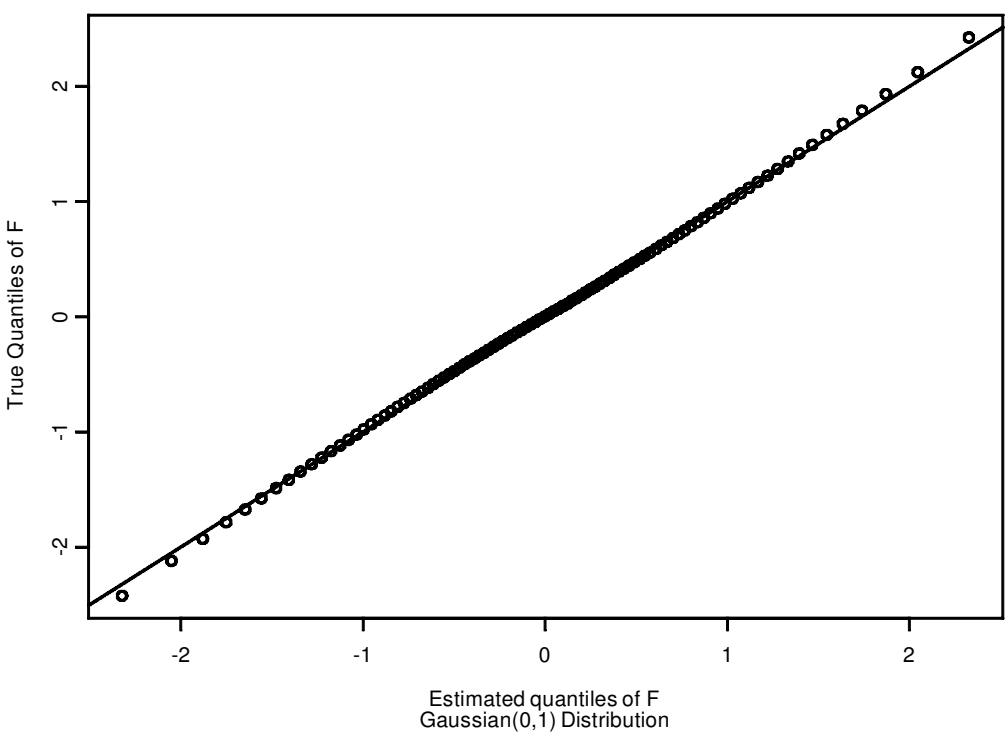

Figure 6. QQ-plot for Comparing the Estimated Error Distribution With the True Error Distribution for Simulated Data From the Standard Gaussian Distribution.

Table 2. Summary Results of Tests

			Simulation method significance level			Wilcoxon test significance level		
Data	*Adj*	*Type of gene*	*NS*	*5%*	*1%*	*NS*	*5%*	*1%*
RAW	No	Cellular	87	4	12	82	12	9
		Viral	92	22	119	73	28	132
STD	No	Cellular	89	7	7	79	9	15
		Viral	92	34	107	78	17	138
STD	Bonf	Cellular	102	0	1	98	0	5
		Viral	183	17	33	159	12	62

NOTE: RAW refers to the unstandardized data, STD to the standardized data. Adj indicates whether or not a Bonferroni multiplicity adjustment was used.

Using the cellular genes for the second step of the standardization thus results in a slightly conservative procedure; borderline genes may appear to be not upregulated when they actually are. The use of a resistant analysis partially offsets this effect. It is comforting that few cellular genes come out significant, suggesting that any biasing effect is minor. Unfortunately, there is really no choice other than to use the cellular genes for standardizing across chips treated differently. The selection of appropriate standards for DNA microchip experiments remains an active research area.

2. Because the data are homoscedastic and symmetric, a simple alternative to the computer intensive simulation method that is yet resistant to outliers is the Wilcoxon test. Table 2 shows the summary results of applying it to test, for each gene g, whether the nine observations $\{Y_{1gcr}\}$ are generally larger than the nine observations $\{Y_{0gcr}\}$; 44 findings are different. Several of these are because the Wilcoxon test produced significance whereas the simulation method did not due to chip outliers. A chip outlier occurs when all the values on a chip for a gene differ from the values for that gene on the two other chips treated similarly. Such cases were not uncommon owing to local effects such as scratches and grey areas on the chip and are not adjusted by standardization (which adjusts for the global difference in intensity scales between the chips). If we omit the data for that chip, we get nonsignificance but if we retain it, Wilcoxon shows significance. As the simulation method is based on an estimator that is a median (across chips) of medians, it is resistant against chip outliers and did not show significance. In a few cases, the Wilcoxon test did not show significance when there were several values tied at the limits, 0 or M; the simulation method avoided this problem and gave significance.

3. Table 2 also shows the summary results of applying the simulation and Wilcoxon tests to the raw unstandardized data. For the simulation method, 42 findings are different between the raw and standardized data. This is a sizeable number (that it is not larger attests to the resistance of the simulation method) and shows the value of standardizing the chips to a single intensity scale prior to an analysis, even a resistant one.

All in all, we found the simulation method applied to the standardized data to be the most satisfying approach for this data. Several of the genes identified by this method as being upregulated were new findings, a few of which were subsequently confirmed by the more definitive but harder to perform Northern blot analysis.

10. A FINAL GENERAL COMMENT

Statisticians face fresh challenges as new technologies emerge. A conventional statistical approach may encounter difficulties in such instances because of characteristics and

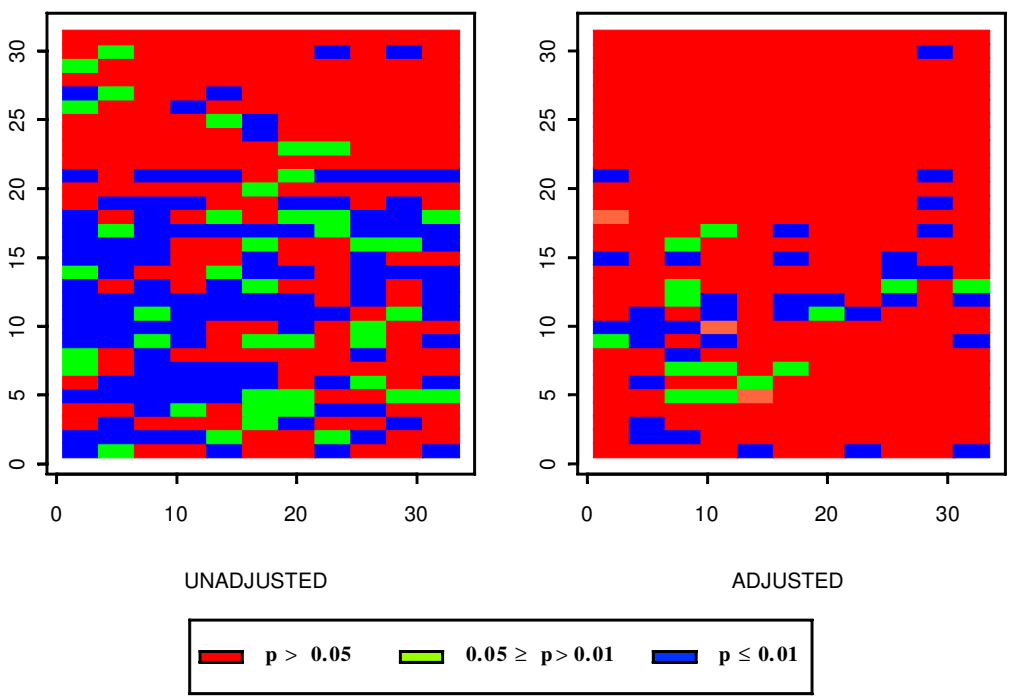

Figure 7. Image Plot of the Significance Levels of the Genes (a) Without and (b) With the Bonferroni Adjustment. Blue and green represent genes that were significant at the 1% and 5% levels, respectively. Red represents genes that were not significant at the 10% level.

complexities of the new data. In this article, we have described how techniques based on exploratory data analysis ideas, particularly graphics, together with simulation, were effectively used to analyze a messy dataset from viral DNA microchips. Once microchip technology has evolved into a more stable stage and the data quality has improved, a more formal approach will be appropriate.

[Received August 1999. Revised January 2001.]

REFERENCES

Chambers, J., Angulo, A., Amaratunga, D., Gao, H., Khaleghi, M., Wittig, J., Bittner, A., Frueh, K., Jackson, M., Petersen, P., Erlander, M., and Ghazal, P. (1999), "DNA Microarrays of the Complex Human Cytomegalovirus Genome, Profiling Kinetic Class by Drug Sensitivity of Viral Gene Expression," *Journal of Virology*, 73, 5757–5766.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997), "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images," *Journal of Biomedical Optics*, 2, 364–374.

Cho, K., Meer P., and Cabrera, J. (1997), "Performance Assessment Through Bootstrap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 1185–1198.

Manly, B. F. J. (1997), *Randomization, Bootstrap and Monte Carlo Methods in Biology* (2nd ed.), London: Chapman and Hall.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995), "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science,* 270, 467–470.

Slonim, D., Tamayo, P., Mesirov, J. P., Golub, T. R., and Lander, E. S. (2000), "Class Prediction and Discovery Using Gene Expression Data," in *Proceedings of the Fourth Annual Conference on Computational Molecular Biology*, 263–272.

Somerville, C., and Somerville, S. (1999), "Plant Functional Genomics," *Science*, 380–383.

Tukey, J. W. (1977), *Exploratory Data Analysis*, New York: Addison-Wesley.