Statistical Design and the Analysis of Gene Expression Microarray Data[1]

M. Kathleen Kerr and Gary A. Churchill

The Jackson Laboratory

Bar Harbor, Maine 04609 USA

Short title: Statistical Design and Analysis of Microarrays

Proofs to be sent to: M. Kathleen Kerr, The Jackson Laboratory, 600 Main Street, Box 303, Bar Harbor, ME 04609 USA; phone (207)-288-6000 x1281; fax (207)-288-6077; e-mail mkk@jax.org

Summary: Gene expression microarrays are an innovative technology with enormous promise to help geneticists explore and understand the genome. Although the potential of this technology has been clearly demonstrated, many important and interesting statistical questions persist. We relate certain features of microarrays to other kinds of experimental data and argue that classical statistical techniques are appropriate and useful. We advocate greater attention to experimental design issues and a more prominent role for the ideas of statistical inference in microarray studies.

[1]Direct reprint requests to M. Kathleen Kerr.

1. INTRODUCTION

Gene expression microarrays are an exciting new tool in molecular biology (Brown & Botstein, 1999). Geneticists are intrigued by the prospect of collecting and mining expression data for thousands of genes. Statisticians have taken a correspondingly enthusiastic interest in the many quantitative issues that arise with this technology. These issues begin with analyzing scanned array images and extracting signal (Yang *et al.*, 2000a). After one has estimates of relative expression in hand, there are problems in data visualization, dimension reduction (Hilsenbeck *et al.*, 1999), and pattern recognition (Brown *et al.*, 2000). In the world of gene expression, a lot of attention has been focused here, particularly on clustering tools. In contrast, our focus is on the analysis that takes place after image analysis and before clustering. Namely, how does one get from fluorescence readings off an array to valid estimates of relative expression, and how does one put error bars on those estimates?

With the rush to embrace microarray technology and its potential, we believe a number of fundamental experimental principles have been neglected. In addition, we believe there are some common misconceptions about which quantitative issues with microarrays are truly novel. Accordingly, we have somewhat different ideas about the areas of statistical research that are key to improving microarray data analysis. This article discusses spotted cDNA arrays because this is the data we have the most experience with. There are similar issues with oligonucleotide arrays, but these are outside the scope of this article.

2. "IT'S ALL RELATIVE"

In the context of microarray technology, the word "spotted" refers to the process by which sequences of DNA are attached to a glass slide or other surface. By various mechanisms, a robotic arm with blocks of pins places DNA strands as spots on an array. Any given spot contains one particular DNA sequence, although the same sequence may be spotted multiple times per array. The mRNA samples under study are reverse-transcribed into cDNA and a dye label is incorporated. One sample is labeled with a "green" dye and the other with "red." The samples are then mixed and washed over the array, where the dye-labeled cDNA strands can hybridize to their complementary sequences on the array. Unhybridized cDNA is washed off,

and the green and red fluorescence are measured from each spot on the array. There is little information in a single fluorescence measurement from a spot because there is poor control over the amount of DNA target in each spot. What is interesting is the *relative* fluorescence of red and green from a spot, because the sample that contained more transcript should produce the higher signal. This makes the two-dye system integral to the process. An alternative proposal is to use radioactive labeling in place of dye labeling (Friemert *et al.*, 1989). However, with single reads from each spot each measurement becomes confounded with spot-to-spot variation. This adds to error and can only be overcome with extensive replication and careful randomization. As we will discuss, spot effects can be accounted for with the two-dye system. This results in greater precision and more power to test for differences in expression.

Although there is no question that relative comparisons are the meaningful quantities, we disagree with two common conclusions drawn from this fact. The first is that all the relevant information is captured in the ratio of the two signals from a spot. The second is that the "relativeness" of microarray data is a novel feature of this technology and thus traditional statistical analyses are inadequate.

To the contrary, relative data is about as old as statistics itself. The "grandfather" of statistics, R.A. Fisher, worked with agricultural field trials. In controlled experiments with clear objectives, scientists sought to determine the productivity of different varieties of a crop, for example different strains. They recognized that there is no such thing in absolute terms as *the* yield of a variety because productivity depends on soil fertility, sunlight, rainfall, and myriad other factors. They understood that the only meaningful *direct* comparisons are for strains grown on the same block of land. Consider a hypothetical experiment to study three varieties. Suppose there are three blocks of land available, but each block only has room for two varieties. Table 1 gives possible experimental plans. It is easily accepted that the yield data contain information about the varieties grown in the same block. However, there is a corresponding fact relying on the same logic that can be overlooked. Namely, there is also information about the blocks of land because they have varieties in common. Fisher recognized this duality and realized one could simultaneously estimate the relative yield of varieties and the relative effects of the blocks

of land. The quantitative tool for doing this is a simple linear model,

$$y_{ij} = \mu + B_i + V_j + \epsilon_{ij}, \tag{1}$$

where y_{ij} is the measured yield for variety j grown on block i, μ is the overall mean, the block effect B_i is the effect of block i, and V_j is the effect of variety j. The term ϵ represents random error. In a large experiment with many varieties and blocks, unbiased yield comparisons can be made, even for varieties not grown on the same block of land.

Returning to microarrays, consider the spots for a particular gene on different arrays (or reproduced within arrays). The spots vary in size, shape, and concentration, analogous to the variation in fertility of blocks of land. Using the same principles as in the agricultural experiment, we can simultaneously measure the relative transcription level of the corresponding gene and the "fertility" of the spots. However, this is only possible if we use all the information in the data and do not reduce to ratios.

Of course, microarray experiments are considerably more complicated than the simple agricultural experiment just described. With the agricultural experiment we can speak of the plots of land as the experimental units, but with microarrays there are different sizes of experimental units. Spots may be organized into blocks or pin groups, which in turn are nested within the slides themselves, which may also belong to batches or print runs. The hypotheses that motivate a particular experiment may also necessitate a complex design structure to the mRNA samples. For example, mRNA samples may be obtained from tissues from different strains of mice and different treatment conditions. We note, however, that the comparisons of interest are across samples but within genes. That is, one makes inferences about the relative levels of expression for a gene in the different samples but not about the level of expression of one gene with respect to another. This is because hybridization properties differ from sequence to sequence so that the correspondence between fluorescent signal and any kind of absolute measure of transcript level is unknown. Because the meaningful comparisons are within genes, the fundamental experimental units are the spots.

3. EXPERIMENTAL DESIGN FOR MICROARRAYS

Once we recognize that microarray data contains information about both expression and spot characteristics, new possibilities arise for designing these experiments. In particular, it is

not necessary to use a reference sample on every array. The logic behind using reference samples is that if the expression of a gene is twice the level in sample 1 compared to the reference, and six times higher in sample 2 compared to the reference, then the expression is three times as large in sample 2 compared to sample 1. However, the same logic that allows these indirect comparisons also allows one to consider experimental designs that only include the samples of interest. Introducing a reference as an intermediate step is unnecessary and generally inefficient because it means fully half of the data are dedicated to an extraneous sample. These precious resources could be better allocated to gather more information about the samples of interest so that they may be compared with the best possible precision.

We return to the hypothetical agricultural experiment to illustrate this idea. Consider the two experimental plans in Table 1. To compare three varieties of interest in three blocks of size two, the "reference" design uses a fourth, arbitrary reference variety in each block. This is analogous to a common experimental strategy used with microarrays. An alternative is the "balanced design," which does away with the reference variety. The balanced design doubles the amount of data on the varieties of interest without requiring any additional resources. Using the model in (1) and assuming independent error with constant variance, the least-squares estimate of the yield difference for two varieties of interest will have variance one-third as large for the balanced design compared to the reference design. In other words, the balanced design allows one to compare the varieties of interest with much greater precision.

<<< Table 1 approximately here. >>>

Although statistics is commonly viewed as primarily dealing with post-experimental data analysis, statistical experimental design is one of the oldest sub-disciplines. Fisher (1951; p.3) noted that "statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation." With the two-dye system, microarrays are effectively blocks of size two (Kerr & Churchill, 2000b). When there are more than two samples under study, a microarray design is necessarily an *incomplete block design* (Cochran & Cox, 1957). While it is not a trivial task to find a good incomplete block design, the topic has been under study for decades and there is a body of research to help find efficient experimental plans.

The designs that are most suitable for any particular experiment depend on the questions of interest and the hypotheses to be investigated.

A different aspect of good design that is conceptually simple to incorporate is replication. As Fisher noted, replication serves two purposes. The first is to increase the precision of estimation. The second purpose, "which there is no alternative method of achieving, is to supply an estimate of error by which the significance of these comparisons is to be judged" (Fisher, 1951; p.60). Fisher lamented that "it is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available" (Fisher, 1951; p. 34). Without the ability to estimate error there is no basis for statistical inference. One has the experience of seeing some data and can look for patterns, but there is no way to decide whether patterns are real or spurious. As noted by Lee *et al.* (2000), many microarray experiments are currently completed without replication. Replication can be incorporated at several levels: genes can be spotted multiple times per array, mRNA samples can be used on multiple arrays, and mRNA samples can be taken from multiple specimens to account for inherent biological variability. The last example represents true replication while the others are probably more accurately described as repeated measures.

In addition to finding a good pairing of samples on arrays and incorporating replication, a third aspect of design we recommend is balance with respect to dyes. That is, we recommend each sample be labeled with both the red and green dyes and both aliquots be incorporated into the experimental plan. In multiple datasets from different labs we have seen genes that exhibit higher expression when labeled with one dye or the other, regardless of the sample. This difference is beyond any overall dye effect. If samples are only labeled with one dye, not only can this phenomenon not be corrected, it cannot be detected. Then, if these effects are present, they will lead to biased estimates of relative expression and misleading results. Although we do not yet have a satisfactory explanation for these gene-specific dye effects, at this stage we advocate incorporating balance into designs until the issue is resolved.

4. MICROARRAY DATA ANALYSIS AND STATISTICAL INFERENCE

The analysis of variance (ANOVA) is a natural tool for studying data from experiments with multiple categorical factors (Kerr *et al.*, 2000). Borrowing terminology from agriculture,

the term *variety* refers to the mRNA samples under study. Varieties can be different strains, tissue types, timepoints in a biological process, etc. Let y_{ijkgr} be the signal on the appropriate scale from the r^{th} spot for gene g on array i for dye j and variety k. A typical ANOVA model for a microarray experiment is

$$y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \epsilon_{ijkgr}. \quad (2)$$

Here, μ represents the average signal across all the factors in the experiment. The "global" effects A_i, D_j, and $(AD)_{ij}$ account for overall variation in arrays and dyes. These terms saturate the design space of arrays and dyes, indirectly accounting for overall effects of varieties. These effects are not of interest, but accounting for them amounts to data normalization. We advocate, as much as possible, normalizing data as an integral part of the data analysis so that it is done systematically and the degrees of freedom are explicitly acknowledged. In addition to these "global" normalization terms, there are sources of variation to consider at the level of individual genes. The gene effects G_g account for the average signal for gene g across arrays, dyes, and varieties. The $(AG)_{igr}$ terms in the model are the "spot" effects. These are analogous to the block effects B in the agricultural experiment. The $(DG)_{jg}$ terms are gene-specific dye effects. As mentioned in the previous section, we did not anticipate such effects, but they have appeared repeatedly and are a potential source of bias if ignored. The terms $(VG)_{kg}$ account for the expression of gene g specifically attributable to variety k. Contrasts in the $(VG)_{kg}$ for fixed g are the quantities of interest.

A fundamental assumption of ANOVA is that there exists a scale on which the various effects are additive. We share with others a bias for the logarithmic scale. Biological phenomena tend to be thought of in terms of multiplicative effects, such as fold-changes in expression. The log transforms these into additive effects, providing interpretability that is a clear advantage of the log. However, the log transform is problematic with image analysis programs that produce non-positive data values. Ad-hoc "flooring" gets around the problem but produces undesirable artifacts in the data. We have seen cases where ANOVA modeling of the log data produced residual plots with systematic trends. In some cases a simple shift in the data before taking logs solves the problem (manuscript in preparation). Figure 1 demonstrates this simple adjustment, but a more complicated correction may be required in some cases (Yang *et al.*, 2000b). Another

problem that occurs is truncation at the high end of the data, either because every probe molecule is hybridized or because the dynamic range of the scanner has been reached. We routinely use the diagnostic plots shown in Figure 1 and advocated by Yang *et al.* (2000b) to monitor for this phenomenon.

An additional issue is whether the error ϵ in the model can reasonably be treated as homoscedastic. If not, one must consider the nature of the heteroscedasticity and its consequences for estimation and statistical inference. In some data we have analyzed there was no clear evidence against homoscedasticity (Kerr *et al.*, 2000; Kerr & Churchill, 2000a). In these cases we used a straightforward application of bootstrapping (Efron & Tibshirani, 1994) to produce non-parametric confidence intervals for differences in gene expression across samples (Figure 2). We have observed heteroscedasticity in other data (manuscript in preparation), although the nature of the heteroscedasticity has not been obvious. For example, we are uncertain whether the error variance depends on intensity or is perhaps more particular to genes. In either case, we prefer randomization techniques such as bootstrapping over classical t-tests and F-tests because we have consistently observed non-normality in residual distributions.

A crucial issue with ANOVA for microarrays is deciding whether effects should be treated as fixed or random. Generally, fixed effects are those thought of as unknown constants. In contrast, random effects are thought to arise from some random process, such as sampling from an effectively infinite population. We have used fixed effect models as a starting point, but agree with those who argue that random effects are more appropriate in some cases (Wolfinger *et al.*, 2000). Our hesitation has been that the empirical distributions of parameter estimates and residuals we have seen have been decidedly non-normal, but standard methods for random effects assume underlying normality. We believe this area is ripe for further research.

Wolfinger *et al.* (2000) present a microarray data analysis using ANOVA methods and random effects. In line with classical techniques, they use random effects for the "blocking effects" such as the spot effects. We believe there is a case to be made to go a step further and treat the gene effects and gene interactions, including the effects of interest, VG, as random. This may be appropriate because many microarray experiments are exploratory. The genes spotted on the arrays are not specifically suspected to relate to the differences in the varieties under study. Rather, they are a set of clones that happen to be available. The result would be some

"shrinkage" in the estimates of differential expression, reducing the bias in the most extreme estimates. The prospect of treating the effects of interest as realizations of a random process is not without controversy. Robinson (1991) provides an excellent discussion of the topic.

5. DISCUSSION

The first microarray experiments demonstrated the promise of the technology by estimating patterns of gene expression and showing these patterns were in good agreement with prior knowledge of the systems under study (Chu *et al.*, 1998; DeRisi *et al.*, 1997). This was a remarkable and noteworthy achievement. As we move forward, the purpose of experiments is not to confirm known properties of well-studied genes. Rather, it is to acquire information about unknown genes and unknown gene functions. To this end, one cannot consider just the genes with large changes in expression that are obvious to detect without the aid of statistics. This practice overlooks important genes that may have small, but reproducible, changes in expression. In order to detect such genes, scientists need statistically designed experiments and data analysis that not only produce estimates of relative expression but, in addition, error bars for those estimates. Error bars provides a basis to decide which features in the data likely represent interesting biology and which are likely to have arisen by chance. Measures of confidence should be incorporated whether one is simply looking to identify differentially expressed genes or evaluating the results of higher-order analyses such as clustering (Kerr & Churchill, 2000a).

Ignoring potential sources of experimental bias, such as assuming the two dyes behave the same, can yield misleading results. An advantage of model-based data analysis such as ANOVA is that a model helps the analyst explore the data. If one finds a model inadequate, discovering *why* it is inadequate can help the analyst identify sources of variation and bias. On the other hand, sources of variation cannot be measured if they are confounded by the experimental design. Any finite body of data contains a limited amount of information, which cannot be increased by any amount of ingenuity expended by statisticians (Fisher, 1951; p.39).

Statistics has a long-standing relationship with biology, probably more than with any other natural science. Biological data is inherently variable, and statistical inference is required in order to draw conclusions from data and add to the body of knowledge. Collecting data and

acquiring knowledge are not the same thing. Good design and sound statistical inference will be a crucial factor in determining whether microarrays fulfill their potential.

References

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Area, M. Jr., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences, USA* **97**, 262-267.

Brown, P.O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**, 33-37.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., & Brown P.O. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705.

Cochran, W.G. & Cox, G.M. (1957). Experimental designs. Wiley, New York.

DeRisi, J.L., Iyer, V.R., & Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686.

Efron, B. & Tibshirani, R.J. (1994). An introduction to the bootstrap. Chapman and Hall, London.

Fisher, R.A. (1951). The design of experiments, 6th edn. Edinburgh: Oliver and Boyd Ltd.

Friemert, C., Erfle, V., & Strauss, G. (1989). Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. *Methods in Molecular Cell Biology* **1**, 143-153.

Hilsenbeck, S.G., Friedrichs, W.E., Schiff, R., O'Connell, P., Hansen, R.K., Osborne, C.K., & Fuqua, S.A.W. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *Journal of the National Cancer Institute* **91**, 453-459.

Kerr, M.K. & Churchill, G.A. (2000a). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Submitted.

Kerr, M.K. & Churchill, G.A. (2000b). Experimental design for gene expression microarrays. *Biostatistics*, to appear.

Kerr, M.K., Martin, M., & Churchill G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, to appear.

Lee, M.-L.T., Kuo, F.C., Whitmore, G.A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the USA* **97**, 9834-9839.

Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15-51.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., & Paules, R.S (2000). Assessing gene significance from cDNA microarray expression data via mixed models. Submitted.

Yang, Y.H., Buckley, M.J., Dudoit, S., & Speed, T.P. (2000a). Comparison of methods for image analysis on cDNA microarray data. Technical report 584, Department of Statistics, University of California, Berkeley.
http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html

Yang, Y.H., Dudoit, S., Luu, P., & Speed, T.P. (2000b). Normalization for cDNA microarray data. Technical report 589, Department of Statistics, University of California, Berkeley.
http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html

Reference Design

Block 1	Block 2	Block 3
A	B	C
R	R	R

Balanced Design

Block 1	Block 2	Block 3
A	B	C
B	C	A

Table 1: Experimental designs to study three varieties of interest in three blocks of size two. Varieties A, B, and C are of interest, while R represents a fourth variety introduced to serve as a reference. The second design is balanced in the sense that every pair of varieties in the design appears together once.

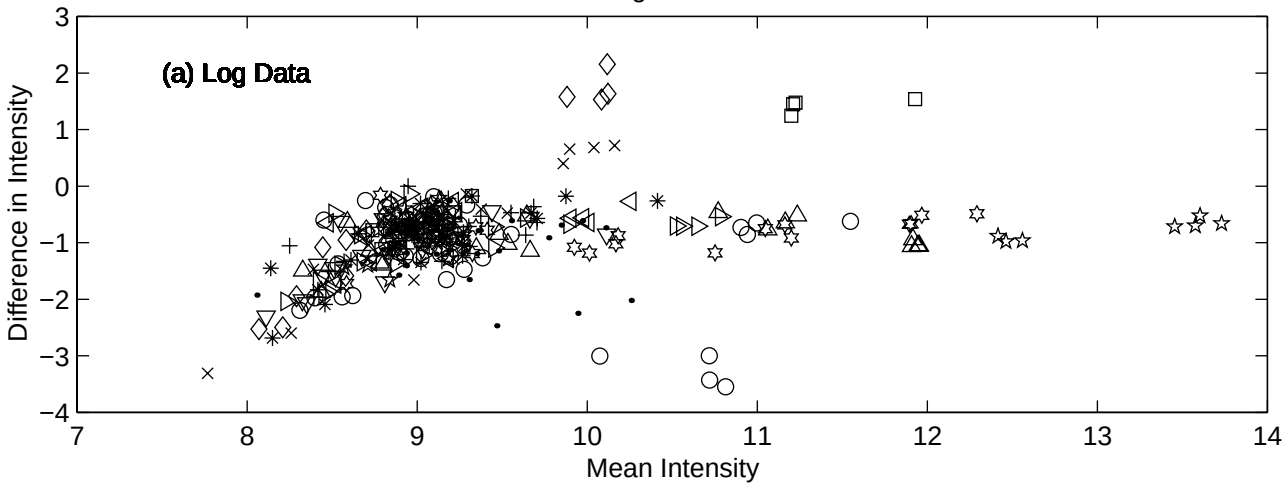
Figure 1: Difference vs. Mean Plots

A cDNA array used to compare a drug-treated with a control mouse liver sample. For each spot on the array, the difference in the two signal intensities is plotted against the mean. Assuming there is a uniform difference in the dyes and most genes are not differentially expressed, most points should fall along a horizontal line. However, notable curvature at the low end is seen in plot (a) for the data on the log scale. This curvature is removed by using a uniform shift in each channel before taking log, as seen in plot (b). Each gene was spotted with 4x replication on the array; like symbols were used for the four spots for the same gene (symbols had to be re-used). Symbols tend to cluster in groups of four, but also show the inherent variability in the signal.

Figure 2: Temporal Patterns of Expression with Bootstrap 99% Confidence Bounds

We analyzed the gene expression time series data of Chu *et al.* regarding the yeast sporulation cycle using ANOVA. Confidence intervals on the estimates of relative expression serve as a gauge for which changes in expression should be regarded as "real" and which changes are attributable to noise. For example, there is strong evidence there is increased expression of YSW1 between 2 and 5 hours but little evidence for changes in expression between 5 and 12 hours.

Figure 1


(a) Log Data
Difference in Intensity
Mean Intensity
3
2
1
0
−1
−2
−3
−4
7
8
9
10
11
12
13
14

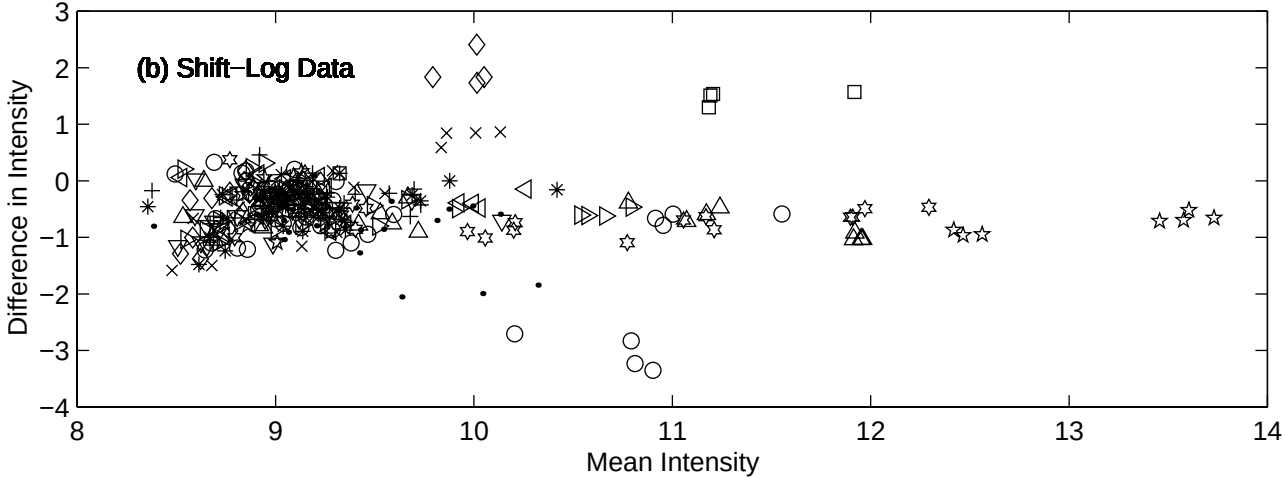
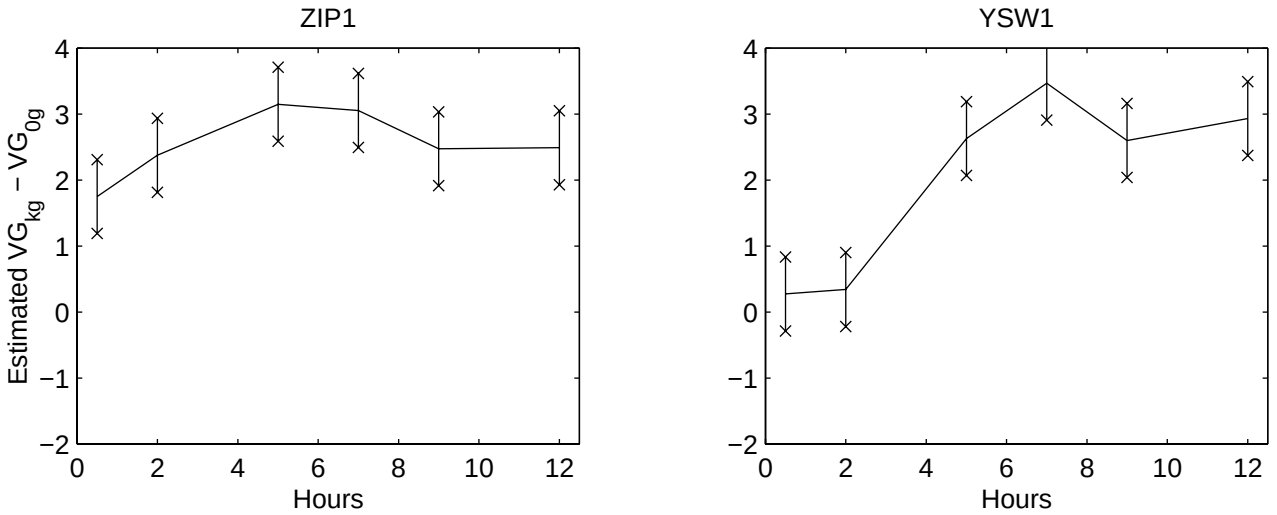

(b) Shift−Log Data
Difference in Intensity
Mean Intensity
3
2
1
0
−1
−2
−3
−4
8
9
10
11
12
13
14

Figure 2


ZIP1
YSW1
Estimated $VG_{kg} - VG_{0g}$
Hours
Hours
4
3
2
1
0
−1
−2
0
2
4
6
8
10
12