5-16-2003

Linear Models for Microarray Data Analysis: Hidden Similarities and Differences

M. Kathleen Kerr
University of Washington, katiek@u.washington.edu

1 Introduction

In the past few years gene expression microarrays have become important tools in biology and genetics. Microarrays have had a comparably large impact on quantitative fields. A multitude of computational and statistical issues accompany microarrays, and quantitative scientists have been invigorated by the problems. Statisticians and other mathematical scientists have produced a variety of techniques specifically addressed to microarrays.

A biologist with microarray data who seeks guidance in the literature may be overwhelmed by the large number of different methods. As statisticians and other mathematical scientists promote methodologies that they helped develop, there is little guidance for a conscientious investigator who needs to decide what analyses to perform. In fact, many methodologies are substantially similar, but this is often not apparent in the literature. By understanding the similarities among methods, an investigator might then understand the differences and has a better chance of making a truly informed decision.

The goal of this paper is to conceptually organize some of the key methods for two-color microarray data analysis. It is intended to summarize some of the methodologies in a way that illuminates the hidden similarities and true differences among them. "True differences" refers to the fact that methods may be presented differently but yield the same effective analysis and have identical implications for experimental design. For example, it is not a coincidence that the comparisons of microarray designs using a regression model as in Yang and Speed (2002) reproduce the results in Kerr and Churchill (2002), where an

ANOVA model was employed. Although the data models in the two papers appear to be quite different, I will show such models actually differ only marginally. By understanding such connections, the differences between various methodologies should also become clearer.

A microarray experiment involves many decisions that can effect the conclusions, including

(1.) What RNA samples will be collected and which pairs will be hybridized together (experimental design)?

(2.) What method of image analysis will be used and will the data be adjusted for background (data extraction) ?

(3.) How will the raw data be used to estimate relative gene expression, and how will differential expression be decided (normalization, estimation, statistical inference)?

(4.) How can the data be explored to suggest high-order structure, or how can gene expression be used as predictors, classifiers, etc. (clustering, discrimination analysis, etc.)?

This paper concentrates on methodologies contained in (3.), but is not intended as a comprehensive review of all techniques described by (3.). Normalization methods are covered in detail elsewhere (Cui et al, 2002; Quackenbush, 2002; Yang et al, 2002; and many others). This paper is directed at a set of techniques whose aim is to combine information across arrays and estimate and infer relative expression. This covers a large set of analytical tools, but not everything. My purpose is to offer a framework in which to organize a substantial subset of the methodologies in use.

Brief reviews of two-color spotted microarray technology (Schena

et al, 1995) can be found in the introductions of many papers on microarrays. Nguyen et al (2002) give a thorough description for quantitative scientists.

2 Methods of Microarray Data Analysis

I present data methods organized into four groups. Methods 1 and 2 are explicitly intended to be applied one gene at a time. Method 3 is applied to all the data at once, across genes. Method 4 is a two-stage approach, where the first stage applies to all the data but the second stage is applied gene by gene. Method 1 is a model for log-ratios whereas Methods 2, 3, and 4 are applied to log red and green intensity values. However, as will be discussed, all four methods are related despite these apparent differences.

2.1 Notation

After image-processing, a microarray dataset is a set of Cy3 and Cy5 intensity values for the set of arrays that were hybridized and for the genes spotted on the arrays. These intensities are either background-adjusted or not depending on the decision at step (2.) above. For every gene g spotted on the arrays used in an experiment the data contain a Cy3 and Cy5 intensity measurement. I use the following notation throughout this paper. Let y_{ijkg} be the intensity for gene g on array i from dye j. The subscript k indicates which RNA sample the measurement represents. By the experimental design chosen by the investigator, i and j determine the variety k. In other words, the investigator has chosen which RNA sample to label with dye j for hybridization to array i. Thus the subscripts i, j, and g suffice to identify a data value in the data array.

The y_{ijkg} are assumed to be on log or similar scale and any pre-processing is assumed to be complete. Informally, the y_{ijkg} are "normalized log intensities," where quotation marks acknowledge that this designation is imprecise because a multitude of different data transformations are currently in use. Regardless of these transformations, I refer to within-spot differences in the y_{ijkg} as "log-ratios" in line with convention.

2.2 Methods

Before describing Methods 1–4, I first describe a simple microarray analysis based on comparisons of log-ratios. For gene g on the i^{th} array, the log-ratio is

$$\text{log-ratio}_{ig} = y_{i2-g} - y_{i1-g}. \tag{1}$$

The RNA-identifying subscripts k are omitted in (1) since they are determined by the array i and the dye j.

For simple experimental designs some very straightforward microarray analyses can be performed using log-ratios. Suppose an experimental design uses a "reference" RNA in one channel of every array (say channel 1), as depicted in Figure 1. For example, suppose a pool of tonsil RNA has been used as the reference RNA. Then gene expression in the RNAs of interest are measured in "tonsil" units. With all measurements in comparable units, simple statistical tests, such as t-tests, can be performed on the log-ratios if there are suitable replicates or repeated measures. This kind of procedure has been used by many researchers (Callow et al 2000; Geiss et al 2000; and many others). Simplicity is the major advantage of this approach. The

greatest disadvantage is that one is severely limited in the experimental designs that can be employed. Therefore, for designed experiments one needs to use a more general method, such as those described next.

Method 1. Linear Combinations of Log-Ratios. The straightforward procedure of making simple comparisons of log-ratios can be re-formulated as a linear model. This re-formulation has the advantage that it allows an investigator to consider other experimental designs that may hold advantages over the reference design. I first describe Method 1 for a reference design, then give the generalization to other experimental designs.

For reference designs (Figures 1 and 2), the parameters of the Method 1 model are the differences in log gene expression between the RNAs of interest and the reference RNA. In the y_{ijkg} notation, let $k = 0$ represent the reference RNA and $k = 1, 2, 3, \ldots$ represent the RNAs of interest. The notation k_i refers to the RNA in the non-reference channel of array i. Let d_{kg} be the difference in gene expression between RNA k and the reference RNA for gene g. We have the model

$$\text{log-ratio}_{ig} = d_{k_i g} + \epsilon_{ig}. \tag{2}$$

The left-hand side of (2) is the log ratio for gene g from array i. The parameters of the model are the differences in gene expression d_{kg}.

Model (2) simply re-states the basic analysis described in the beginning of this sub-section as one-way analysis of variance (ANOVA) model, which is a well-known correspondence. If one applies least-squares estimation for the parameters of this model to the data from a simple "reference design" (Figure 1), the d_{kg} parameters are estimated with the log ratios $y_{i2k_ig} - y_{i10g}$. If there are no biological or

technical replicates in this design, no kind of inference can be made because one cannot estimate error to assess statistical significance. Applying model (2) to a reference design with replicate arrays (Figure 2) leads to simple averages of replicate log-ratios. For example, if arrays 1–3 are replicate hybridizations of RNA 1 with the reference RNA, then the estimated parameter d_{1g} is just the average of the three log-ratios from arrays 1–3.

An advantage using this model framework, rather than just comparing log-ratios, is that one then is able to consider designed experiments. For example, consider data from a 3-loop microarray design (Figure 3). The goal in analyzing data from such a design is to appropriately combine all the data relevant to a particular comparison. In the 3-loop, RNAs A and B are compared directly on array 1 but also indirectly on arrays 2 and 3. The array 1 comparison is direct, and thus more precise, and should be given more weight.

For data from a designed experiment like the 3-loop, one could derive the optimal way to combine all the information for a given comparison. This means finding optimal linear combinations of different estimates that give higher weight to more precise estimates. An appropriate linear model does this automatically, drawing on the theory of least-squares estimation.

To demonstrate the model with the 3-loop, notice there are three comparisons between the three pairs of RNAs (A-B, B-C, and A-C). Momentarily suppressing the subscript g, let d_{AB} be the difference between A and B, d_{BC} be the difference between B and C, and d_{AC} be the difference between A and C. The log-ratio (dye 2 minus dye 1) from Array 1 estimates expression in B relative to A, $-d_{AB}$. The log-ratio

from Array 2 estimates the gene expression in C relative to B, $-d_{BC}$. The log-ratio from Array 3 estimates the gene expression in A relative to C, d_{AC}. Since $d_{AC} = d_{AB} + d_{BC}$, the model is over-parameterized if all three terms are included, but any two suffice. Arbitrarily choose d_{AB} and d_{BC} as the model parameters. In summary:

$$\begin{aligned}
\text{log-ratio}_1 &= -\,d_{AB} \qquad\quad + \epsilon_1, \\
\text{log-ratio}_2 &= \qquad -\;d_{BC} + \epsilon_2, \\
\text{log-ratio}_3 &= \;d_{AB} + d_{BC} + \epsilon_3.
\end{aligned}$$

The parameters of the linear model are the quantities of interest, differences in log gene expression. Thus the linear model extends the logic of the basic analysis of reference designs to designed experiments. Yang and Speed (2002) use these kind of linear models on log-ratios (and give a similar example for a 3-loop).

Method 2. Single gene ANOVA Performing an analysis as described in Method 1 for designed experiments involves choosing and applying a model parameterization. This was quite simple in the 3-loop analysis, but can become more cumbersome for larger designs. A more traditional formulation of this model is as an analysis of variance (ANOVA) model. An ANOVA model for microarray data includes A_i as a parameter for array i and V_k as a parameter for RNA k. (The 'V' stands for "variety" — a generic term for the different RNAs in the study.) Unlike Method 1, where a parameterization needs to be worked out for every design, the ANOVA model is easy to state in general:

$$y_{ijkg} = \mu_g + A_{ig} + V_{kg} + \epsilon_{ijkg} \tag{3}$$

The model is applied separately for each gene. Note the subscript g appears in every term in (3) and could be suppressed. In contrast to Method 1, the data for this analysis are the individual Cy3 and Cy5 log intensities rather log-ratios. This might seem like a drastic change, but it is not. Estimates of expression differences among samples are linear combinations of log-ratios, just like with the linear model on log-ratios. For example, starting from (3) and taking within-spot differences gives

$$\begin{aligned} y_{i1k_{i1}g} - y_{i2k_{i2}g} =& [\mu_g + A_{ig} + V_{k_{i1}g} + \epsilon_{i1k_{i1}g}] - \\ & [\mu_g + A_{ig} + V_{k_{i2}g} + \epsilon_{i2k_{i2}g}] \\ =& V_{k_{i1}g} - V_{k_{i2}g} + [\epsilon_{i1k_{i1}g} - \epsilon_{i2k_{i2}g}]. \end{aligned} \tag{4}$$

Notice the left-hand side of (4) is a log-ratio and has expectation $V_{k_{i1}g} - V_{k_{i2}g}$. Differences in the estimated values of the V_{kg} parameters estimate differences in expression and are derived from log-ratios. Such examination shows that models (2) and (3) produce identical estimates of gene expression differences.

ANOVA models such as (3) are well-known in classical statistics as models for "block" designs. Besides convenience and tradition, ANOVA models have other potential advantages over models on log-ratios (Method 1). First, model (3) has some important generalizations. For example, one can include a "dye-effect" D_{jg} in (3) to account for genes that exhibit a dye-bias (Kerr et al, 2002b):

$$y_{ijkg} = \mu_g + A_{ig} + D_{jg} + V_{kg} + \epsilon_{ijkg}. \tag{5}$$

Another generalization (and potential advantage) of the ANOVA formulation is that it allows one to consider treating the "spot effects,"

A_{ig}, as random effects rather than fixed effects. This approach acknowledges that spot-to-spot variation is not pre-defined but arises from a series of random processes. There is precedent for random-effects modeling in classical block design, where it is sometimes referred to as "recovering interblock information" (Cochran and Cox, 1992). Wolfinger et al (2001) and Jin et al (2001) treat spot effects in microarrays as random (although in a two-step procedure — see Method 4 below). For Jin et al (2001), random effects modeling enabled gene expression comparisons between RNAs that were not "connected" in the design (see Figure 4).

A final potential advantage of an ANOVA formulation is that the error is modeled on the raw intensity measurement. Arguably, it makes more sense to think of error as added to the measurements that are actually made rather than to differences in measurements. However, this distinction may be largely academic. More importantly, the ANOVA formulation allows one to consider appropriate error structures for experiments that include multiple sources of error. For example, an experiment may include biological replicates, replicated hybridizations, and repeated spots (Churchill, 2002). Methods for such error structures are well-developed in an ANOVA framework. In a linear model for log-ratios such as Method 2, such structures cannot be accommodated appropriately without restricting the design options.

Method 3. Global ANOVA Methods 1 and 2 are approaches to microarray data analysis that are applied one gene at a time. Another option is an ANOVA model that is "global" in the sense that it applies to the data for all the genes at once. Kerr et al (2000) introduced these

models for microarray data. Such a model is:

$$y_{ijkg} = \mu + A_i + D_j + (AD)_{ij} + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg}. \quad (6)$$

The parameter μ in the model is the overall mean across all factors – arrays, dyes, and genes. A_i is the overall effect of array i, D_j is the overall effect of dye $j = 1, 2$, and $(AD)_{ij}$ is a "channel" effect for dye j on array i. Notice none of these terms has a g subscript — they are "global" effects, describing variation across genes. These terms produce the sorts of linear normalizations that are often done informally. G_g is the overall effect of gene g across the other factors and corresponds to the term μ_g in Method 2. The $(AG)_{ig}$ terms capture spot effects, and correspond to the A_{ig} terms in Method 2. The $(VG)_{kg}$ effects represent levels of signal intensity for genes that can specifically be attributed to the RNA varieties under study. These correspond to the V_{kg} terms in Method 2 and are the effects of interest. Differences in these terms estimate gene expression differences between varieties of RNA, i.e. for RNAs k and k', relative gene expression is estimated as $(VG)_{kg} - (VG)_{k'g}$.

What are the differences in estimates of expression differences between Method 1 and 2 (which are equivalent) and Method 3? Method 3 is actually single-gene ANOVA on the data that has been "centered," meaning that the average intensity from every channel on every array is set to 0. (Technically, this equivalence is mathematically exact only if the same set of genes is spotted on every array in the experiment with no missing data.) In other words, global ANOVA correspond to single-gene ANOVA on the data $x_{ijkg} = y_{ijkg} - y_{ijk\cdot}$ (a $\cdot$ indicates averaging over a subscript). In practice, the adjustment $y_{ijkg} \rightarrow x_{ijkg}$ is

very small because $y_{ijk\cdot}$ is usually small due to normalization processes that are typically done prior to these analyses.

Method 4. Two-stage ANOVA As already suggested, Method 3 can be re-written as a two-stage model. First, fit a "centralization" model

$$y_{ijkg} = \mu + A_i + D_j + (AD)_{ij} + x_{ijkg}. \tag{7}$$

The parameters A_i, D_j, and AD_{ij} in (7) are interpreted as in (6). The residuals of this model, x_{ijkg}, become the data in the second stage, which is applied one gene at a time:

$$x_{ijkg} = \mu_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg}. \tag{8}$$

Such a two-stage analysis is the approach of Lee et al (2002). Under typical conditions, Methods 3 and 4 produce mathematically identical estimates of gene expression differences. Specifically, if the same set of genes is spotted on every array with no missing data, then Methods 3 and 4 produce identical results if all effects are treated as fixed effects. This is because all of the gene-specific effects in (6) are orthogonal to all the global effects under these conditions. If missing data cause an imbalance in which genes are effectively represented on different arrays, Methods 3 and 4 are no longer mathematically equivalent but their difference is typically miniscule. As described in the discussion of Method 3, estimates from Methods 3/4 tend to vary little from the estimates produced by Methods 1/2.

3 Data Example

Data come from a study of gene expression in three RNAs (Unpublished data, Sam Bruschi, University of Washington Department of Medicinal Chemistry). The experimental design is a 3-loop (Figure 3). The three RNAs were derived from the same mouse hepatocyte cell lines, TAMH (Transforming growth factor - alpha overexpressing mouse hepatocyte cell line). Each RNA sample was prepared and treated on a different day. The treatment was a 4 hour exposure to 200 uM Tetrafluoroethylcysteine, a toxic metabolite of the industrial gas Tetrafluoroethylene. Since the only difference between the RNAs is the day of preparation and treatment, few gene expression differences were expected.

Each of 7680 clones was double-spotted on each array in two separate grids so that each array contained two sub-arrays. Substantial intensity-dependent and spatial variation was observed in the log ratios, but the pattern of spatial variation differed on the two sub-arrays contained each array. Therefore normalization was done separately for each sub-array. Normalization was done with via a "loess" procedure to adjust for intensity-dependent and spatial artifacts (Cui et al, 2002). In the following analyses, I also treated sub-arrays as independent arrays, implying a repeated 3-loop design as in Figure 3(b). A more appropriate analysis might account for the dependence between repeated spots on the same microarray, but treating the sub-arrays as independent serves the purposes of this example.

For obtaining point estimates of relative gene expression, Methods 1 and 2 are mathematically equivalent, as are 3 and 4. Because nor-

malization necessarily results in data with log-ratios roughly centered around 0, the difference between Methods 1/2 and 3/4 is very small. Figure 5 presents the estimates of gene expression differences between samples A and B using Methods 1/2 and Methods 3/4. The figure is uninteresting because the only difference in the estimates is a constant 0.0007 shift. For the B-C and A-C comparisons (data not shown), the constant differences are -0.0004 and 0.0004 respectively.

As mentioned, ANOVA models such as Methods 2, 3, and 4 can be varied by treating effects as random instead of fixed. For example, Wolfinger et al (2001) and Jin et al (2001) use a variant of Method 4 where spot effects are treated as random effects. This decision – random or fixed effects – has a much more noticeable effect on estimates of relative expression than the choice among methods. To illustrate this, I compared estimates from Method 2 with spot effects treated as fixed and random. (The mixed effects model was estimated in the statistical package R (Ihaka and Gentleman, 1996) using the 'lme' function and the default REML methodology for estimating variance components.) As seen in Figure 6, this change makes little difference in estimation for comparing RNAs B and C. However, for the other pairwise comparisons, A vs. B and A vs. C, this change results in some large differences for a handful of genes.

When spot effects are treated as fixed effects, estimates of relative expression are linear functions of "within-spot" differences, i.e. log-ratios. For example, in a simple 3-loop (Figure 3(a)), the comparison of RNAs A and B for a given gene is given by

$$\hat{V}_A^F - \hat{V}_B^F = \frac{2}{3}(y_{1A} - y_{1B}) + \frac{1}{3}(y_{3A} - y_{3C} + y_{2C} - y_{2B}). \qquad (9)$$

The numerical subscript on the y's refers to the array (as denoted in Figure 3(a)) and the letter in the subscript refers to the RNA. The superscript F denotes the model with fixed spot effects. For the design in the data example (Figure 3(b)), the estimate can be expressed in the same form by averaging measurements from replicate hybridizations. Notice that each expression inside parentheses in (9) has expectation $V_A - V_B$, but the first is given double weight since its variance is half as much.

When spot effects are treated as random effects, let α^2 denote their variance. The estimate of relative gene expression is more complicated than in the fixed-effects case, as it now depends on α^2 as well as the error variance σ^2:

$$(3\alpha^2 + 2\sigma^2)(\hat{V}_A^R - \hat{V}_B^R) = 2\alpha^2(y_{1A} - y_{1B}) + \alpha^2(y_{3A} - y_{3C} + y_{2C} - y_{2B}) + \sigma^2(y_{1A} + y_{3A} - y_{1B} - y_{2B}). \quad (10)$$

The superscript R denotes the estimate with random spot effects. An instructive form in which to express the estimate is:

$$\hat{V}_A^R - \hat{V}_B^R = \frac{3\alpha^2(\hat{V}_A^F - \hat{V}_B^F) + 2\sigma^2((y_{1A} + y_{3A} - y_{1B} - y_{2B})/2)}{3\alpha^2 + 2\sigma^2} \quad (11)$$

Thus $\hat{V}_A^R - \hat{V}_B^R$ is a weighted average of the fixed effects estimate $\hat{V}_A^F - \hat{V}_B^F$ from (9) and the quantity $\frac{1}{2}(y_{1A} + y_{3A} - y_{1B} - y_{2B})$. Note that this latter quantity is a simple contrast of the observations from variety A and the observations on variety B. It is an unbiased estimate of $V_A - V_B$ since the spot effects are random variables with expectation 0. Notice that as $\frac{\sigma^2}{\alpha^2} \to 0$, then $\hat{V}_A^R - \hat{V}_B^R$ converges to $\hat{V}_A^F - \hat{V}_B^F$. In other words, if spot variation is much larger than measurement error, then the estimate of $V_A - V_B$ using random spot effects will be close to

the estimates using fixed spot effects. This is the case in the analyzed data. The estimated ratio $\frac{2\sigma^2}{3\alpha^2+2\sigma^2}$ is very small for most genes, with a median of 0.0075 and 75th quantile 0.0182. Out of 7680 genes on the array, only 82 have the estimated ratio $\frac{2\sigma^2}{3\alpha^2+2\sigma^2} > 0.5$. Thus $\hat{V}_A^R - \hat{V}_B^R$ is heavily weighted towards $\hat{V}_A^F - \hat{V}_B^F$, and the two have little opportunity to differ. Even when $\frac{2\sigma^2}{3\alpha^2+2\sigma^2}$ is large, $(y_{1A} + y_{3A} - y_{1B} - y_{2B})/2$ must also differ from $\hat{V}_A^F - \hat{V}_B^F$ in order for $\hat{V}_A^R - \hat{V}_B^R$ to deviate substantially.

4 Discussion

An advantage of single-gene models (Methods 1, 2, and 4) is computational practicality. General statistical software cannot handle models such as (6) because of the large number of parameters. On the other hand, model-fitting algorithms specialized for microarrays get around this problem rather easily (Wu et al, 2003) by capitalizing on the relationships described in this paper. Method 4 can be viewed as a computationally tractable re-formulation of Method 3. Parameter estimates from the two-stage model in Method 4 can be pieced together to construct the global model. This can be done in a statistical programming language (Wolfinger et al, 2001), or through software specialized for microarray data analysis (Wu et al, 2002).

Section 2.2 discusses the mathematical equivalence of several different methods for microarray data analysis for estimating differences in gene expression. All of the models discussed in this paper are linear. As such, they reflect the assumption that relative fluorescence (properly normalized) is proportional to the relative amount of transcript in the dye-labeled cDNA pool. The validity of combining data across arrays to estimate relative gene expression relies on this assumption.

As shown in Section 2.2, Methods 1 and 2 are equivalent, as are 3 and 4. Further, the practical difference between Methods 1/2 and 3/4 is very small, as illustrated in Section 3. One might conclude from these results that it does not matter which method is used. However, although the methods produce the same (or nearly the same) *point estimates* of gene expression differences, they can differ substantially in *statistical inference*. That is, they can lead to very different

conclusions about what genes are differentially expressed. Momentarily setting aside the issue of fixed and random effects, quite different conclusions can be reached solely due to how measurement error is modeled. This is the error denoted by ϵ throughout Section 2.2.

Methods 1, 2, and 4 are explicitly one-gene-at-a-time analyses. This means the inference of whether a particular gene is differentially expressed is based only on the data for that gene. If the only replicates are technical replicates, then this assumption is that the technical or measurement error is different for every gene. While the generality of this assumption is appealing, it is problematic. There is usually not enough data for individual genes to get an accurate estimate of the error variance. Across thousands of genes, many will have small error variances by chance. Empirical evidence suggests that a one-gene-at-a-time approach leads to many "false positives" when differential expression is assessed (Efron et al, 2001; Tusher et al, 2001). Various ways to get around this problem have been used but only a little theoretical work has been done on the problem (Lönnstedt and Speed, 2002).

Global models of microarray data analysis enable one to combine data across genes to estimate error distributions. For example, Kerr et al (2002a) observed larger error for low-intensity genes. They combined information for genes at similar intensity levels for estimating error variances and making inferences. This produced more robust inference than modeling error separately for every gene. Global models are conducive to combining data across genes for realistic and robust models of error.

Combining information across genes may also be useful if random

effects are modeled. For example, Kerr et al (2002a) observed spot-to-spot variation that was normally distributed across genes. As seen in Section 3, the decision of whether to treat some effects, such as spot effects in an ANOVA model, as fixed or random can make a substantial difference in the results. Ideally, we would like to remove (through normalization) or model all the systematic effects so that variation in spot intensity could be treated as random. However, in practice it is difficult to evaluate whether this has been accomplished. For the data examined in Section 3, there was a noticeable trend for less intense spots in the lower half of one array. Modeling or attempting to correct such trends through normalization increases the risk of overfitting the data. Therefore, although random spot effects may be philosophically preferable, the assumption may still not be reasonable. This is an outstanding issue in microarray data analysis. A general conclusion applicable to every microarray dataset might not be possible.

A general conclusion is that microarray data should be analyzed by a conscientious statistician or other scientist who understands the implications of the choices that are made. Modeling assumptions should be evaluated in any analysis. This means, for example, looking for systematic trends in the model residuals that might cast doubt on the results. Residual analysis can also reveal whether the data have been analyzed on the proper scale (Kerr et al, 2002a). A potential advantage for global modeling over single-gene methods is that model evaluation and residual analysis is feasible. In contrast, such evaluation is not practical for gene-by-gene analyses because it is not feasible for a data analyst to examine residual plots for every gene.

Acknowledgments. I thank Ruth Etzioni, David Rocke and other colleagues as well as an anonymous reviewer for comments and suggestions that improved this article. I also thank Sam Bruschi for the data used in Section 3.

References

[1] Callow, M.J., Dudoit, A., Gong, E.L., Speed, T.P. and Rubin, E.M. 2000. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.* 10, 2022-2029.

[2] Churchill, G.A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32, 490-495.

[3] Cochran, W.G., and Cox, G.M. 1992. Experimental Design, 2nd edition, John Wiley & Sons, New York.

[4] Cui, X., Kerr, M.K., Churchill, G.A. 2002. Data transformations for cDNA microarray data, submitted.

[5] Efron, B., Tibshirani, R., Storey, J.D., Tusher, V. 2001. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* 96, 1151-1160.

[6] Geiss, G.K., Bumgarner, R.E., An, M.C., Agy, M.B., van't Wout, A.B., Hammersmark, E., Carter, V.S., Upchurch, D., Mullins, J.I., Katze, M.G.2000. Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology* 268, 8-16.

[7] Ihaka, R., Gentleman R. 1996. R: A Language for Data Analysis and Graphics. *J. Graphical and Computational Statistics* 5, 299-314.

[8] Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., Gibson, G. 2001. The contributions of sex, genotype

and age to transcriptional variance in Drosophila melanogaster. *Nat. Genet.* 29, 389-395.

[9] Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J. Walker, N.J., and Churchill, G.A. 2002a. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* 12, 203-217.

[10] Kerr, M.K. and Churchill, G.A. 2001. Experimental design for gene expression microarrays. *Biostatistics* 2, 183-201.

[11] Kerr, M.K. Leiter E.H., Picard L., Churchill G.A. 2002b. Sources of variation in microarray experiments. In Zhang, W., and Smulevich, I., eds., *Computational and Statistical Approaches to Genetics*, Kluwer, Boston.

[12] Kerr, M.K., Martin, M., and Churchill G.A. 2000. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7, 819-837.

[13] Lönnstedt, I., Speed, T. 2002. Replicated microarray data. *Statistica Sinica* 12, 31-46.

[14] Lee, M.-L.T., Lu, W., Whitmore, G.A., and Beier, D. 2002. Models for microarray gene expression data. *J. Biopharm. Stat.* 12, 1-19.

[15] Nguyen, D.V., Arpat, A.B., Wang, N., and Carroll, R.J. 2002. DNA microarray experiments: biological and technological aspects. *Biometrics* 58, 701-17.

[16] Quackenbush, J. 2002. Microarray data normalization and transformation. *Nat. Genet.* 32, 496-501.

[17] Schena, M., Shalon, D., Davis, R., Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470.

[18] Tusher, V.G., Tibshirani, T., Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98, 5116-5121.

[19] 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, 625-637.

[20] Wu, H., Kerr, M.K., Cui, X., and Churchill, G.A. 2003. MAANOVA: A software package for the analysis of spotted cDNA microarray experiments. In Parmigiani, G., Garrett, E.S., Irizarry, R.A, and Zeger, S.L., eds., *The analysis of gene expression data: methods and software*, Springer, New York.

[21] Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30.

[22] Yang, Y.H., Speed, T. 2002. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3, 579-588.

5 Figure Captions

Figure 1 A reference design. The nodes represent RNA samples and arrows represent microarrays, with the head and tail of an arrow representing dye-labeling with Cy3 and C5. See Kerr and Churchill (2001) or Yang and Speed (2002) for an explanation of this graphical depiction of microarray designs.

Figure 2 Reference-type designs. A reference design with (a) repeated hybridizations and (b) dye-swap arrays.

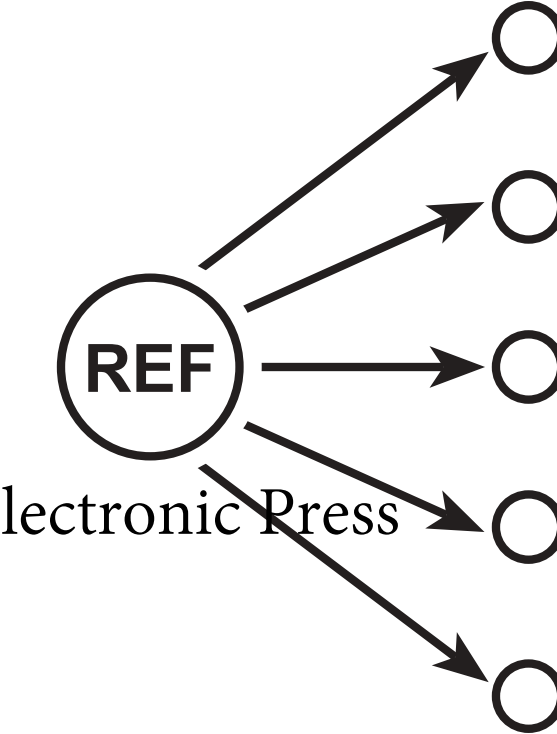
Figure 3 Loop designs. (a) A loop design for three RNAs of interest. (b) The design for the data analyzed in Section 3, where subarrays of the physical arrays were analyzed as independent arrays.

Figure 4 The design used by Jin et al (2001). There were three binary factors in this study of fruitflies, sex (male and female), strain (O and S), and age (1 week vs. 6 weeks old). The numbers on the arrays indicate the number of replicate arrays of each type. All microarrays were comparisons across age within a sex and strain. Random-effects modeling of spot-to-spot variation enabled comparisons across strain and sex.

Figure 5 A comparison of estimates of relative expression using different linear models. For the data analyzed in Section 3, the figure compares estimates of $\log_2$ differences in gene expression between samples A and B using Methods 2/3 and Methods 3/4. The difference is a constant 0.0007 for every gene, so the points in the graph

fall almost along the line of identity.

Figure 6 A comparison of estimates of relative expression treating spot effects as fixed and random. For the data analyzed in Section 3, the figures compare estimates of the $\log_2$ difference in gene expression for Method 2 when spot effects are treated as fixed or random.



lectronic Press