A method for normalizing microarrays using the genes that are not differentially expressed*

Cavan Reilly† Changchun Wang‡ Mark Rutherford§

January 27, 2003

Abstract

One of the more challenging, yet easily overlooked, aspects of the analysis of microarrays is how to normalize arrays so that comparisons can be made across arrays. Most studies that utilize microarrays to detect differential gene expression between samples find the data only enable one to conclude that a handful of genes are differentially expressed. The basic idea here is to use the genes that are not differentially expressed to conduct the normalization. Of course, since one can't determine which genes are differentially expressed until the normalization is conducted, this is a non-trivial problem. Here a general framework and computational method (using the Gibbs sampler) is devised to allow for such normalization. We apply the method to a gene expression experiment aimed at furthering our understanding of Porcine reproductive and respiratory syndrome virus (PRRSV), a major source of economic loss in the swine industry.

Keywords: Bayesian inference, gene expression, Gibbs sampler, M-A plots, microarray normalization, Porcine reproductive and respiratory syndrome virus.

1 Introduction

In recent years there has been an explosion in research in functional genomics. These studies aim to determine which genes are expressed in what cells and under what circumstances. This knowledge has the potential to reveal the functions of many genes simultaneously and to identify molecular markers of disease.

The key technology for conducting high-throughput experiments in functional genomics is the microarray: a technical means for assaying mRNA abundance for many genes simultaneously. We

*C. Reilly thanks the NIH for the Great Lakes Center for AIDS Research grant #1P30-CA79458-01.

†Division of Biostatistics, University of Minnesota, A448 Mayo Bldg., MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455-0378. e-mail: cavanr@biostat.umn.edu

‡Division of Biostatistics, Department of Veterinary PathoBiology, University of Minnesota

§Department of Veterinary PathoBiology, University of Minnesota

refer the reader to the rapidly expanding literature on the use of microarrays rather than trying to introduce those ideas here (see, for example, Efron et al. (2001) or Amaratunga and Cabrera (2001) for an introduction to the technology). Although there are 2 types in common use (i.e. oligonucleotide and spotted cDNA arrays), if one wants to compare intensities across arrays, the same problem often arises: array specific effects are confounded with treatment effects. One may argue that the solution is to have multiple arrays for each treatment. While such a strategy is certainly sensible, although not always feasible due to resource constraints, if we can estimate the array specific effects, we will obtain more accurate estimates than if we just treat the array specific effects as measurement variability. The process of estimating array specific effects, or, more properly, differences in array specific effects, is known as *normalization*. Although we here treat array specific effects as univariate random variables, we more generally think of these random effects as sets of random variables defined on the lattice of spots on the arrays. Normalization is only possible once the data have been obtained, and, as such, it is a data dependent procedure. Although normalization is a data dependent procedure (and hence contributes to the variability of estimates), the variability introduced by normalization is generally ignored in the statistical treatment of microarrays (even in papers which focus on normalization, e.g. Stadl et al. (2001)) and this can have implications for the reproducibility of findings.

1.1 An experiment to understand tropism in Porcine respiratory and reproductive syndrome virus (PRRSV)

Porcine reproductive and respiratory syndrome (PRRS) (European Commission, 1991) is an infectious disease in swine that was initially reported in the United States in 1987 (Keffaber (1989)) and in Europe in 1990. The disease is characterized by late-term abortions, stillbirths, weak piglets, and respiratory problems in pigs of all ages. PRRS causes substantial economic loss in the swine industry. The etiological agent, PRRS virus (PRRSV) was identified shortly after emergence of the disease (Terpstra et al., (1991), Wensvoort et al. (1991)). PRRSV predominantly infects and preferentially replicates in pulmonary alveolar macrophages (PAMs) in vivo (Rossow et al. (1996a), Duan et al. (1997a, b), Cheon et al. (1997a)).

There is considerable variation among PRRSV stains, and preliminary data from our laboratory suggested that PRRSV strains that show dissimilar pathologies differentially regulate swine viral response genes. Comparative sequence analysis of the single-stranded RNA viral genome confirmed the presence of two genotypic classes of PRRSV: a North American genotype and a European genotype (Wensvoort et al. (1991), Meulenberg et al. (1993)). There are striking differences between the PRRSV strains in terms of viral transmission, tropism, morphogenesis, clinical presentation,

antigenicity, virulence, serology, antibody neutralizing epitopes, the susceptibility to antibody dependent enhancement of infection, plaque size (Park et al. (1996)) and RNA sequence. There is also variation among isolates in the same genotype. For example, VR2332, which is the prototype of the Northern American isolate, and neurovirulent PRRSV (Neuro-PRRSV), a recently described Northern American strain with marked neurovirulence, exhibit documented differences in plaque size, clinical presentation and RNA genome, for example (Murtaugh, et al. unpublished data). Neuro-PRRSV can cause microscopic lesions characterized by severe meningoencephalitis, necrotic interstitial pneumonia and gastric muscular inflammation (Rossow et al. (1999)). It is known that genetic variation in PRRSV contributes substantially to virulence and pathogenesis. For example, there are only 41 nucleotide changes between VR2332 and its cell culture passaged descendent Resp-PRRS vaccine strain, yet the Resp-PRRS vaccine strain has greatly reduced virulence in swine (Yuan et al. (2001)).

In our initial effort to identify genes that impact macrophage function during viral infection, we used the differential display reverse transcription polymerase chain reaction (DDRT-PCR) to capture expressed sequence tags (ESTs, i.e. suspected genes with unknown functions) within PRRSV-infected macrophages over a 24 hour post infection period (Zhang et al. (1999), Wang et al. (2001)). From over 8000 DDRT-PCR amplicons examined, over 100 porcine-derived DDRT-PCR products induced or inhibited by PRRSV were identified and cloned. The ESTs isolated during DDRT-PCR analysis and certain other known swine genes represent good candidate genes for understanding PRRSV pathogenesis. Due to technical problems with DDRT-PCR, the results are usually confirmed with another method, such as northern blots or microarrays.

After identification and cloning of these candidate genes, we created spotted cDNA microarrays to explore gene expression in PRRSV infected PAMs at 4 and 24 hours post-infection. The arrays contained 139 ESTs (mostly identified through DDRT-PCR) spotted on each array in triplicate (417 spots per array). The swine PAMs were infected by one of three PRRSV strains of interest: VR2332, Neuro-PRRSV or the Resp-PRRS vaccine strain. In addition, a common reference sample was created by "mock infecting" a set of PAMs with CL2621 cell supernatent, the medium in which the virus is cultured. RNA from the virus infected samples was labeled during reverse transcription with the red-fluorescent dye Cy5 and mixed with a mock infected sample, labeled with the green-fluorescent dye Cy3, from the same time point. This design economizes on the use of viral infected samples compared to a design that considers all possible 2 way comparisons between the viral infected samples. After hybridization and washing, separate images were acquired for each dye, and fluorescence intensity was measured for all spots using the software QuantArray. Each hybridization was repeated twice for a total of 12 arrays in this experiment (2 time points and 3 viral strains

with 2 replicates for each condition). Finally, we note that the software which was used to report the intensity at each spot also reports values for background corrections for each spot, but for this experiment these corrections were always negligible and have no impact on the results. Nonetheless, we follow current practice and work with these background corrected values here.

2 Normalization methods in current use

In the analysis of microarrays, 2 basic strategies that use the intensity at the spots are in popular use: global normalization and control gene normalization. In addition, there are methods that use the intensity of the array on some region surrounding the spots to correct the spot intensity (these methods are referred to a background correction methods), but the usefulness of these depends critically on the assumption that the hybridization which occurs around the spotted area is the same (or can be related to) the hybridization which occurs on the spot. The problem with this assumption is the "spot" contains biological material and the area around it does not, so it is not clear why we should use the background to correct for what is taking place at the spot. One may argue that the "spot" is not exact, so the region outside of the defined spot contains biological material, but this is really more of a criticism of segmentation algorithms. We see background correction so construed as a facet of segmentation, not normalization. Ultimately all these image processing tasks should be brought into the model used to detect differences across treatments so we can fully assess the uncertainty in our results. Here our intention is to incorporate normalization in the model used to assess differences across treatments.

To introduce the methods in current use and consider alternative approaches, we introduce a simple framework for thinking about normalization. For concreteness we will first consider the analysis of oligonucleotide arrays, but the treatment of cDNA arrays is very similar (we give a detailed example of the latter below). We use x_i to denote the log (base 2) transformed intensity for gene i from array 1 and y_i to denote the log transformed intensity from array 2. We assume

$$\mathrm{E}x_i = \phi_1 + \theta_i \quad \text{and} \quad \mathrm{E}y_i = \phi_2 + \eta_i.$$

Here the ϕ_j are array specific effects (common to all spots on the array). Generally researchers are interested in the fold change, $\delta_i = 2^{\theta_i - \eta_i}$, or its logarithm. Since

$$\begin{aligned} \theta_i - \eta_i &= \mathrm{E}x_i - \mathrm{E}y_i + \phi_2 - \phi_1 \\ &= \mathrm{E}x_i - \mathrm{E}y_i + \text{normalizing term}, \end{aligned}$$

the fold change is not identifiable. The role of normalization is to introduce assumptions that make these fold changes identifiable.

2.1 Global normalization

A method of normalization that is commonly used in the analysis of microarrays is based on expressing the intensity measured at each spot as a fraction of some estimate of the average intensity on the array. There are several estimates of the average in use: the mean, the median, and the geometric mean. We use the latter here to obtain unbiased estimates in the context of out simple model, but this is not crucial. Often some arbitrary scaling is used in addition to this normalization (for example, the default option in the Affymetrix software scales so that the trimmed average intensity is 1,000). Now, if N is the number of genes on the arrays, the estimate of the log fold change is

$$\hat{\log}\, \delta_i = x_i - y_i + \frac{1}{N}\sum_k (y_k - x_k).$$

This method has the potential to produce an inconsistent estimate. To see this, let $\overline{\theta}$ denote the average intensity for array 1 (over all spots), define $\overline{\eta}$ similarly, and consider the expected value of $\hat{\log}\, \delta_i$,

$$\begin{aligned} \mathrm{E}(\hat{\log}\, \delta_i) &= \phi_1 + \theta_i - \phi_2 - \eta_i + \phi_2 + \overline{\eta} - \phi_1 - \overline{\theta} \\ &= \theta_i - \eta_i - \overline{\theta} + \overline{\eta}. \end{aligned}$$

So the estimate is consistent when

$$\overline{\theta} = \overline{\eta}.$$

Unfortunately this method of normalization will generally make the estimates of all the log fold changes biased, and this bias is equal to the product of the proportion of genes that are differentially expressed and the mean difference in the log intensities of the differentially expressed genes. Hence this bias is not important when the large fold changes cancel the small fold changes or when the proportion of differentially expressed genes is small relative to the mean difference in log intensity for the genes that are differentially expressed. If N is large, the variance one obtains ignoring the normalization is only slightly larger than the actual variance of the estimate, hence ignoring the variability introduced by normalization is not so important when one uses this method.

While the assumption that $\overline{\theta} = \overline{\eta}$ is reasonable for many experiments, there are examples where we expect large changes in the overall quantity of every gene (e.g. measuring the halflives of mRNA after transcription is intentionally stopped, as in Lam et al. (2001) and Raghavan et al. (2002)). Another example where this method may be misleading is in the comparison of 2 samples where one sample has many genes up-regulated and no genes down-regulated relative to the other sample. The latter situation is quite likely to be common in practice (e.g. tissue undergoing repair compared to healthy tissue), especially with small arrays (as in many cDNA array experiments). In addition, if

the researcher first examines many transcripts using a tool like DDRT-PCR, as in our application, many of the genes on the array may be differentially expressed.

2.1.1 Resistant global normalization

There are several variations on global normalization that use other estimates than the sample mean to determine the total quantity of sample bound to the array. Such methods use some sort of trimmed mean (e.g. the median) in place of the sample mean and assume this trimmed mean is the same for both arrays. The optimal amount of trimming should balance the increase in variance due to using too few spots to estimate the normalization term against the bias introduced by global normalization with the mean. This optimal amount of trimming would depend on the proportion of genes that are not differentially expressed, and hence would vary from experiment to experiment. An optimal method that is robust with regard to the assumption of the proportion of genes that are differentially expressed would recognize that estimation of the extent of trimming depends on the proportion of genes that are differentially expressed. The method developed here is robust in this regard, and incorporates uncertainty regarding the extent of trimming in the final estimates.

2.2 Control gene normalization

In control gene normalization, we assume at least one gene on the array does not change under the experimental circumstances, and we use this gene (or genes) to normalize. The normalization consists of subtracting the log intensity of each gene on an array by the log intensity of the control gene on the same array. In the literature there is a distinction between housekeeping genes (genes whose expression level does not change across experimental treatments) and control genes (spiked controls) (see, for example, Bowtell and Sambrook (2003)). Here we don't distinguish between housekeeping genes and control genes because they play the same role in the analysis, hence we will call any gene whose expression level does not change between experimental conditions a control gene.

This method of normalization can greatly impact the variability of the estimated fold change. To see this, assume we know, prior to seeing the data, that the k^{th} gene doesn't change. A consistent estimate of the log fold change is then

$$\hat{\log \delta_i} = x_i - y_i + (y_k - x_k).$$

Suppose σ_i is the standard deviation of the i^{th} gene from array 1 and τ_i is the standard deviation for the same gene from the other array, then

$$\text{Var}(\hat{\log \delta_i}) = \sigma_i^2 + \tau_i^2 + \sigma_k^2 + \tau_k^2.$$

This exceeds the value that is typically used as the variance of the log fold change by $\sigma_k^2 + \tau_k^2$, hence, if we ignore the uncertainty introduced by the normalization, we will underestimate the variance of our estimates. Many treat normalization as an image analysis task, then test for differences using the normalized data, but clearly this will lead to incorrect false error rates when control gene normalization is used.

While the control gene approach provides a reasonable normalization method, there are several problems with this approach. First, it is often hard to say which genes will act as control genes, and sometimes alleged control genes don't seem to behave like control genes. In addition, using the outcome for a single gene to normalize thousands of other genes is potentially very dangerous if something "goes wrong" for that gene in an experiment.

2.2.1 Finding the controls

Our approach to normalization is based on the idea that many genes are most likely not differentially expressed, hence these provide information on how to normalize. In a recent paper, Tseng et al. (2001) have attempted to normalize using a set of genes that is "most likely" not differentially expressed, although they do not demonstrate that the method they propose has any desirable statistical properties. Initially they attempted to use a non-parametric procedure, but apparently the procedure results in estimates that are "too variable". Although the method proposed there does not consider how the estimated fold changes are influenced by this data dependent normalization and uses a nonlinear normalization of dubious merit (see below), it is a similar idea. In that paper, the authors acknowledge that using too few genes results in highly variable fold changes, hence they introduce some arbitrary cut off points for the difference in the fold changes to define a set of genes that is to be used to conduct the normalization. While it is difficult to suggest a general method for determining these cut off points that is not totally *ad hoc*, we think the more serious problem is that there is not only uncertainty in the normalization introduced by using a data dependent normalization, there is also uncertainty due to the choice of genes deemed control genes. Within the context of the method proposed by Tseng et al. it is very difficult to incorporate this source of uncertainty in the final results.

2.3 Non-linear normalization

A normalization method that can be used with either of the above approaches (but usually associated with global normalization), first introduced by Dudoit et al. (2002) (in a widely distributed technical report first appearing in 2000) and subsequently used in several proposed approaches to normalization (see, for example, Schadt et al. (2001), Tseng et al. (2001) and Yang et al. (2002)), is

to allow the normalizing term to be a nonlinear function of the intensity level. That is, we assume $Ex_i = \phi_1(\theta_i) + \theta_i$ and similarity for y_i.

If we are to statistically assess the variability of the estimates produced using such a normalization we clearly need to make some assumptions regarding the relationship between ϕ and θ. One popular approach supposes

$$\phi_1(\theta_i) - \phi_2(\eta_i) = f\left(\frac{\theta_i + \eta_i}{2}\right) + \epsilon_i,$$

where ϵ_i is a sequence of i.i.d. errors. The method becomes more transparent when we assume $\theta_i = \eta_i$ for all i, because, if we use ϕ to represent the normalizing term, then

$$\phi(\theta_i) = f(\theta_i) + \epsilon_i.$$

Dudoit et al. did not really assume $\theta_i = \eta_i$ for all i: actually the working assumption there is that $\theta_i = \eta_i$ for "almost" all i. The idea of the "almost" all is brought into the analysis by using a robust, non-linear regression method to estimate f (e.g. lowess). One can estimate f using either all the genes (as in global normalization) or a set of control genes.

This method was motivated by plotting, for each gene, the difference in (log) intensity between the 2 arrays and the average (log) intensity across the 2 arrays, i.e. $x_i - y_i$ by $\frac{x_i+y_i}{2}$, and noting that sometimes there is a nonlinear relationship in the plot (these plots are referred to as M-A plots). In fact, such nonlinear relationships can arise when ϕ_k is independent of the underlying mean level. To see this, note that the expected curve falls along the points $\big((\phi_1 + \theta_i + \phi_2 + \eta_i)/2, \phi_1 - \phi_2 + \theta_i - \eta_i\big)$, so if no genes are differentially expressed, the curve should be $\big((\phi_1 + \phi_2)/2 + \theta_i, \phi_1 - \phi_2\big)$. That is, there should be a cloud of points situated about a horizontal line in these plots if no genes are differentially expressed. Now suppose there is a group of genes with a log fold change of, say 2. Then we would expect to see, superimposed over the scatter about the horizontal line mentioned above, a cloud of points centered at $\big(1 + \overline{\eta} + \frac{\phi_1+\phi_2}{2}, 2 + \phi_1 - \phi_2\big)$, where $\overline{\eta}$ represents the average η_i for this group of genes.

Depending on $\overline{\eta}$ (and, more generally the number of groups of genes with similar expression levels and the mean expression levels in these groups), we may think there is a positive, negative or non-linear relationship based upon the plot. That is, we expect to find nonlinearities in these sorts of plots even when the array specific effects are independent of the mean intensity if genes have intensities that occur in clusters. It is widely accepted that genes have expression levels that occur in clusters (e.g. Eisen et al. (1998)). Some think this is not a source of concern provided one uses a robust smoothing method because it is expected that only a small proportion of genes will be differentially expressed, but if the genes form clusters in intensity, then they can lead to systematic nonlinearities. Moreover, even if one doubts that many genes are differentially expressed and form

clusters in intensity levels, we could find "clusters" in intensity if there is a gradient or smooth spatial variation in intensity across the array even when no genes are differentially expressed. In Figure 1 we find an example. We generated data from a 5 component bivariate mixture model: for each of the 2 samples there are 50 genes very differentially expressed ($\mu_1 = (4, 1)$ and $\mu_2 = (0, -4)$) and 50 genes moderately differentially expressed ($\mu_3 = (2, 1)$ and $\mu_4 = (0, -2)$) in addition to 1000 genes with no difference in expression level ($\mu_5 = (0, 1)$). The very differentially expressed genes are indicated by an "R" or "G" depending on which sample is differentially expressed, lower case letters are used to designate the moderately differentially expressed genes and dots indicate genes with no difference in expression (the latter are not centered at zero horizontally due to the difference in array specific effects). In this simulation the "G" sample was more down-regulated than the "R" sample was up-regulated, hence the asymmetry in the plot.

There are also some troubling aspects regarding the implementation of this method. First, it is hard to give general recommendations for choosing the parameters that are necessary inputs for the algorithm. For example, if lowess is used one must specify the span, and while recommendations have been given (for example Yang et al. (2002) "typically use" 0.4, whereas Dudoit et al. (2002) suggest values between 0.2 and 0.4) clearly the choice should depend on the number of genes in some complicated, experiment specific manner. Another underappreciated aspect of this method is that x-outliers in the M-A plot will have substantial impact on the estimated curve. The robustness of the lowess smoother lies in not being mislead by a few outliers in the midst of many other points, but if the data is sparse the curve can overfit the data. This implies that if the average intensity across arrays for some gene is very large relative to the other genes, the lowess curve will go through this point (or nearly so) in the M-A plot, thereby producing an estimated log fold change of nearly 0 after normalization even if the difference in the gene intensities across arrays is large. Hence, the effect of using a nonlinear normalization method when one isn't necessary is to shrink the estimated fold changes towards one, as noted empirically in Schadt et al. (2001). Finally, no one takes account of the uncertainty introduced into the estimation of the fold changes that results from using a nonlinear smoothing algorithm to conduct the normalization, and this source of variability is potentially substantial for some values of the intensity due to sparseness in the plot at the extremes.

3 A general model for normalization

Here we propose a probability model embodying the idea that some genes are differentially expressed and that we should use the genes that aren't differentially expressed to normalize the others. The basic idea is to have a model for each array that has a gene specific mean and variance, and to add to this a parameter, ψ_i, which indicates if the gene is differentially expressed. By allowing the

variance of a gene to differ across arrays we can allow for different scales on the 2 arrays. The exact formulation depends on details of the application, so we present an example in the next section.

To provide a unified treatment of normalization for oligonucleotide and spotted cDNA arrays we here introduce the idea of a "scan specific" effect. This is a random effect that is measured during the scanning of the array. For oligonucleotide arrays, the scan specific effect is what we have referred to as the array specific effect, while for cDNA arrays, there is a scan specific effect for each of the 2 dyes, since one scans the array once for each wavelength (so the scan effect is the combination of the array effect and the dye effect). Let

$$x_{ijk} = \text{ log intensity for gene } i, \text{ scan } j, \text{ replicate } k$$

then our model postulates that

$$x_{ijk} \sim \mathrm{N}\left(\phi_j + \theta_{ij}, \sigma_{ij}^2\right)$$

where ϕ_j is the scan specific effect, and θ_{ij} is the log intensity of the i^{th} gene under experimental conditions of the j^{th} scan. By supposing

$$\theta_{ij} = \theta_{ij} + \psi_{ijj'}(\theta_{ij'} - \theta_{ij}),$$

where

$$\psi_{ijj'} = \begin{cases} 0 & \text{if} \quad \text{gene } i \text{ differentially expressed in scans } j \text{ and } j' \\ 1 & \text{if} \quad \text{else} \end{cases}$$

then treating the $\psi_{ijj'}$ as parameters, we build the possibility that each gene is a control gene for certain scan comparisons into the model. The presence of such genes allows for control normalization.

We note that the details of the application often impose relations between the θ_{ij}s or different parameterizations and relations between the parameters in these alternative parameterizations. These are easily accommodated in the context of the current model. For example:

1. Time series effects, such as $\theta_{ij} = f_i(t_j)$, where t_j is the time of the j^{th} scan and f_i is some function of time;

2. Demographic characteristics, such as $\theta_{ij} = \beta_{0i} + \beta_{1i} z_j$, where z_j is some characteristic of the samples such as age of the patient;

3. Reference sample structures, such as in the application presented below.

Our parametric modeling of the intensities is similar in spirit to the parametric modeling of Newton et al. (2001), Li and Wong (2001) and Ibrahim, Chen and Gray (2002), although these authors

did not consider normalization extensively. Newton et al. use a Gamma distribution to model the intensities, whereas here we use the log normal distribution (in practice it is difficult to distinguish between observations from these 2 distributions). While much of the literature surrounding the analysis of microarrays has focused on non-parametric methods (e.g. Efron et al. (2001)), we think that parametric models are useful because one can focus on setting up models that embody relevant biological facts.

3.1 Incorporating information about known control genes

Depending on the context, we may have information regarding the proportion of genes that are control genes. We can incorporate such information through a prior, for example

$$\frac{1}{N}\sum_i \psi_{ijj'} \sim \text{Beta}(\alpha, \beta),$$

where α and β are selected to reflect information regarding the proportion of control genes. Given the exploratory nature of most microarray experiments, researchers may feel more comfortable using a non-informative prior for the proportion of genes differentially expressed, and this leads to proper posteriors for the fold changes in many situations.

Other priors are possible. For example, we may want to suppose the probability is one that some gene is a control gene. If we further suppose all other genes are not control genes with probability one, then we find the control normalization approach is a special case of this model. The proportional normalization approach can also be cast as a special case of the general model in which we use the prior that no genes are controls, but

$$\overline{\theta}_j = \overline{\theta}_{j'},$$

for all j, j'. We think that if a gene is typically considered to be a control gene, the best approach is to suppose that gene is a control gene with high probability and use a hierarchical prior for the other genes. In this way, we utilize information regarding likely control genes while allowing for additional controls, but by not supposing the control gene is a control with probability 1, we allow the data to indicate if one of these control genes really is not a control.

4 Application to cDNA arrays with a common reference sample

To illustrate the general model, we will consider the application to a common design encountered in the use of spotted cDNA arrays (a type of experiment also considered in Amaratunga and Cabrera (2001)). It is widely recognized that the dyes used to identify samples in the scans impact the binding of the samples to the arrays (see for example, Tseng et al. (2001)), hence researchers often

make the desired comparison by using a reference sample, labeling the reference sample with one dye, then mixing some of this reference sample with each of the samples of interest (the samples of interest are labeled with the other color dye). In these experiments, the dye effect is confounded with the reference sample effect, but by having the same reference sample on the 2 arrays one can hope to sort out the array effect from the treatment effect, at least under some set of assumptions.

We can use the general normalization model to account for differences in dye labeling efficiency. For now, consider just comparing 2 arrays. If we let x_{1ij} and x_{2ij} denote the log of the intensities for gene i on the j^{th} replicate for dye 1 on arrays 1 and 2, and we let y_{1ij} and y_{2ij} denote the corresponding quantities for dye 2, then we propose the following model

$$x_{1ij} \sim \mathrm{N}(\phi_1 + \theta_{1i} + \psi_{1i}(\eta_i - \theta_{1i}), \sigma^2_{1i})$$

$$y_{1ij} \sim \mathrm{N}(\phi_2 + \eta_i, \tau^2_{1i})$$

$$x_{2ij} \sim \mathrm{N}(\phi_3 + \theta_{2i} + \psi_{2i}(\eta_i - \theta_{2i}), \sigma^2_{2i})$$

$$y_{2ij} \sim \mathrm{N}(\phi_4 + \eta_i, \tau^2_{2i}).$$

where

$$\psi_{ki} = \begin{cases} 0 & \text{if else} \\ 1 & \text{if } i \text{ is a control gene on array } k \end{cases}$$

for $k = 1, 2$. We use standard non-informative priors for all parameters in the model and use the techniques of Bayesian estimation to conduct exact inference (up to Monte Carlo error).

Often researcher advocate dye swap experiments (see for example, Kerr and Churchill (2002) and Wolfinger et al. (2001)) as a method for estimating dye specific effects, but we see that this model allows one to run fewer arrays and still estimate such effects. Additionally, we can estimate array by dye interactions, and this is not possible within the context of the analysis proposed by these authors.

4.1 Reference sample normalization

The usual application of global and control gene normalization is slightly more complicated than the methods outlined in Section 2 when there are 2 scans per array. In terms of the model outlined in the previous section, to estimate the fold changes we need an estimate of $\phi_3 - \phi_1$. Suppose the only difference between 2 scans of the same array is a multiplicative dye effect (there is no array by dye interaction). If we let α represent the dye effect, so that $\phi_1 = \phi_2 + \alpha$ and $\phi_3 = \phi_4 + \alpha$, then we can consistently estimate $\phi_3 - \phi_1$ with $\overline{y}_2 - \overline{y}_1$. We refer to this method as the reference sample normalization method. This estimate of the normalizing term will be very precise, and

uncertainty in its estimation can most likely be safely ignored when N is large. As mentioned above, interactions between dye and array effects seem to be quite common in practice, hence this method of normalization is likely to be misleading.

If we suppose there is an interaction between dye and array, then both common methods of normalization will not use the reference samples, y_{1ij} and y_{2ij}, unless we model the correlation across channels. Considering global normalization, we ultimately assume $\overline{\theta}_1 = \overline{\eta} = \overline{\theta}_2$, hence the reference samples play no role in the normalization. For the control gene method, suppose the k^{th} gene is a control (so $\theta_{1k} = \eta_k = \theta_{2k}$), then we can estimate $\phi_3 - \phi_1$ with $\mathrm{E}\overline{x}_{2k} - \mathrm{E}\overline{x}_{1k}$, and once again the reference samples are of no use. We could use these facts to write the likelihood for the parameters of interest ignoring the reference samples, but we will conduct inference for all parameters here so as to investigate some of the assumptions commonly made in the analysis of these experiments. If we build a correlation between x_{1ij} and y_{1ij} into the model (a sensible extension) then the reference samples would be informative about the fold change between the 2 samples of interest.

5 A Gibbs sampling methodology for drawing simulations from the posterior distribution of the fold changes

Given the probability model and the prior distributions, we can find the posterior distribution of the parameters. This distribution is proportional to

$$\prod_{i=1}^{N}\prod_{j=1}^{n} \mathrm{N}(x_{1ij}|\phi_1 + \theta_{1i} + \psi_{1i}(\eta_i - \theta_{1i}), \sigma_{1i}^2)\mathrm{N}(y_{1ij}|\phi_2 + \eta_i, \tau_{1i}^2)$$

$$\times \mathrm{N}(x_{2ij}|\phi_3 + \theta_{2i} + \psi_{2i}(\eta_i - \theta_{2i}), \sigma_{2i}^2)\mathrm{N}(y_{2ij}|\phi_4 + \eta_i, \tau_{2i}^2)/(\sigma_{1i}^2\sigma_{2i}^2\tau_{1i}^2\tau_{2i}^2).$$

We then obtain credible intervals for the fold changes by drawing simulations from the posterior distribution of all the parameters. The easiest way to draw samples from the probability distribution given above is with the Gibbs sampler. The Gibbs sampler is convenient for this model because all of the conditional distributions we need to sample are either normal ($\phi_k, \theta_{1i}, \theta_{2i}$ and η_i), inverse-gamma (σ_{li}^2 and τ_{li}^2) or Bernoulli (ψ_{li}). We use multiple chains and the Gelman-Rubin $\sqrt{\hat{R}}$ statistic to diagnose convergence of the log fold changes (see, for example, Gelman and Rubin (1992)), and the model behaves well converging very quickly (within a couple of hundred of iterations). Computation time is very short (within a minute for the comparisons analyzed here) and increases linearly with the number of genes. Moreover, computation time could be halved by ignoring the reference samples.

6 Application to the PRRSV experiment

As an illustration of the method, we will apply this normalization technique to the experiment described in Section 1.1. As we will see, the model is easily extended to accommodate the replicate arrays, and the presence of multiple arrays allows for more sophisticated hierarchical models. First, we show that the choice of normalization method can have a dramatic impact on our conclusions. Then we apply our normalization method to compare just 2 arrays in order to demonstrate the method. Finally we apply the method, with some extensions, to analyze all of the arrays from the experiment.

6.1 Comparing normalization methods for the PRRSV data set

Here we use several normalization methods to demonstrate that the choice of normalization method can have a tremendous impact on the results. First we used global normalization, ignored the uncertainty in the estimation of the normalization factor (as is commonly done) and calculated independent t-statistics for each gene. A common choice for a housekeeping gene is the gene GAPDH, hence we used this as a control gene and conducted control normalization. The control gene normalization calculations were performed using the obvious extension of the formulas from Section 2.2 in addition to the Gibbs sampler (as a check for coding errors and as a basis for assessing Monte Carlo error). Figure 2 displays the relationship between the t-statistics produced using the 2 normalization methods for 6 array comparisons. There is substantial disagreement between the 2 methods for some array comparisons. Figure 3 shows the relationship between t-statistics produced by the method that uses the control gene but ignores uncertainty in the normalization term and the t-statistics produced when uncertainty is included in calculating the normalization term. As we would expect, there can be large differences between the 2 estimates. For the comparison in the middle of the top row, the normalizing term increases the variance of all log fold change estimates by 0.07. For some genes, the variability in the estimated log fold change is quite small, hence the relative impact of ignoring the normalizing term can be substantial. For example, one gene has an estimated log fold change with variance 6.96×10^{-4} if we ignore the normalizing term, thus the ratio of the variances is 98.7 (the p-value goes from 7.2×10^{-4} to 0.064 if we use the t_2 distribution for reference). The median variance ratio for this array comparison is 2.8 with a mean of 7.2, thus the extent of undercoverage of interval estimates that ignore estimation of the normalizing term can be substantial in practice.

6.2 Comparing 2 arrays

To illustrate the new model, we now just consider comparing a single Neuro-PRRSV array to a single Resp-PRRS vaccine array (the comparison in the middle of the top row of Figure 2 and Figure 3).

While there were replicate arrays in this experiment, replicate arrays are not always available, hence normalization methods are necessary for this situation. If we use control gene normalization we find many genes differentially expressed (61 up-regulated and 1 down-regulated with at least 90% probability). If we use a non-informative prior for the proportion of genes differentially expressed, we come to very different conclusions. In contrast to the control gene method, we now find 42 genes are up-regulated and 17 down-regulated with probability at least 90%. The reason for the discrepancy is clear: a central 95% credible interval for the fold change of GAPDH is (0.4, 0.9, 1.7), thus, although this gene is most likely not differentially expressed (as we suspect based on substantive grounds), the best point estimate finds that it is down-regulated, hence if we normalize using this gene all fold changes will be too high.

Figure 4 displays credible intervals for the log fold changes. In this figure the lower x-axis shows the log fold changes produced by using the control gene method (so the credible interval for GAPDH is at zero on this axis) while the upper x-axis displays the log fold changes produced using global normalization (the 2 methods only differ in regard to the placement of zero). The dotted lines indicate cutoffs for a 2-fold change (a fold change that is often taken as the cutoff for differential expression in less statistical treatments). We see that there are genes other than GAPDH that have much narrower confidence intervals centered at zero, indicating that there are genes other than GAPDH that act as better controls. These genes are better controls because they are less variable across arrays and do not differ across arrays.

By allowing the data to select the controls, the zero of the scale is placed between the zeros from the control gene method and the global normalization method (despite the fact that neither of the sets of assumptions that justify these methods were employed in the analysis). This indicates that the global normalization is under-estimating all the fold changes. Figure 5 provides insight into the discrepancy between the data selected control method and the global normalization method. In the latter figure, we find histograms of the posterior distribution of the difference in the average intensity (convergence of the chain was also monitored with respect to these quantities). Recall from Section 2.1 that the validity of the global normalization method rests on these histograms being centered at zero, but in fact this is not what is taking place (the probability that the intensity is higher on array 1 exceeds 0.99). Next, if there is no interaction between the dye effect and array, as discussed in Section 4, we would expect $(\phi_2 - \phi_1) - (\phi_4 - \phi_3)$ to be centered at zero. In fact, a central 95% credible interval for this difference is (0.67, 1.05), hence there does appear to be an interaction between dye and array, and so the precise estimate of the normalizing term mentioned in that section is not warranted either. In short, while GAPDH is most likely a control, it is not the best control, and the assumptions of the global normalization method in addition to the reference

sample normalization method are almost surely not satisfied for this comparison.

6.3 Results based on all arrays

We can easily extend the previous model to incorporate the 2 replicate arrays for each condition and conduct inference for all 3 viral strains at both time points. We make this extension by specifying distributions for each viral strain as follows. Consider just the 2 arrays for one of the viral strains. We suppose

$$x_{1ij} \sim \mathrm{N}(\phi_1 + \theta_{1i} + \psi_{1i}(\eta_i - \theta_{1i}), \sigma_i^2)$$

$$y_{1ij} \sim \mathrm{N}(\phi_2 + \eta_i, \tau_i^2)$$

$$x_{2ij} \sim \mathrm{N}(\phi_3 + \theta_{1i} + \psi_{1i}(\eta_i - \theta_{1i}), \sigma_i^2)$$

$$y_{2ij} \sim \mathrm{N}(\phi_4 + \eta_i, \tau_i^2).$$

where

$$\psi_{ki} = \begin{cases} 0 & \text{if else} \\ 1 & \text{if } i \text{ is a control gene for strain } k \end{cases}$$

We have a set of these distributions for each of the 3 viral strains at the 2 time points (assuming independence across strains and time). Here we suppose the variances do not depend on array or viral strain but do depend on gene (or EST) and channel. For this larger model, it also makes sense to introduce a hierarchical structure for some of the parameters. Hence we also supposed that the scan specific effects ϕ_i are from a common distribution

$$\phi_i \sim \mathrm{N}(\phi, \zeta^2),$$

and we conduct inference for the hyperparameters ϕ and ζ using standard non-informative priors for this case. We also supposed

$$\psi_{ki} \sim \mathrm{Ber}(\pi_k),$$

where $\pi_k \sim \mathrm{Beta}(1,1)$ for $k = 1, 2, 3$. By having a hierarchical distribution for the control gene indicators we needn't concern ourselves with specifying a prior for the proportion of control genes, and we obtain more efficient estimates. In addition, we suppose 4 genes commonly used as housekeeping genes (GAPDH, cyclophilin, β-actin and HPRT) are control genes with prior probability 0.95. An examination of the posterior predictive distribution of the residuals indicated no serious inadequacies of the model. Moreover, fitting of lowess curves to posterior predictive draws of the M-A plots reveals consistent non-linearities even though our model does not have a non-linear normalizing term.

Comparative microarray analysis confirms our hypothesis that PRRSV strains that show dissimilar pathologies differentially regulate viral response genes. As an example, consider the swine gene Mx-1 (accession number AF102506). For this gene the 95% credible interval for the fold change in expression for VR2332 relative to Neuro-PRRSV is (1.4, 4.82). This gene is known to be induced by interferon-α (IFN-α) and some viruses. PAMs do not express IFN-α, so viral stimulation is the likely source of expression of this gene. Moreover, the protein produced by this gene confers resistance to influenza in mice. Hence, the fact that this gene is expressed at lower levels in PAMs infected with Neuro-PRRSV may partially account for the ability of this strain to infect tissues VR2332 can not infect. Another example is provided by the gene GRP78 (accession number AW231967), with credible intervals of (1.24, 4.26) for VR2332 relative to Neuro-PRRSV and (1.93, 6.26) for VR2332 relative to the Resp-PRRS vaccine strain. It is known that GRP78 plays a role in the conformational maturation of some viruses. We have previously shown that this gene is expressed in both unstimulated and mock infected PAMs, and during VR2332 infection, the GRP78 gene is up-regulated compared to mock infection (Wang et al. (2001)). GRP78 is an intraluminal chaperone protein that can sense stress in the endoplasmic reticulum (ER). This suggests that GRP78 may be involved in the process of PRRSV morphogenesis in the lumen of the ER. Differences between strains in the expression of this gene suggest that different strains have different dynamics of replication. Dozens of other differences were found between the various samples across strains and time points: some of these seem quite reasonable given our current understanding of PRRSV while others suggest interesting, new findings.

In order to confirm our array results, we conducted several northern blots. Figure 6 shows the results of these experiments. Northern blotting showed that GRP78 and 24G29N1 have higher expression in VR2332 infection than Neuro-PRRSV infection at 24 hours after infection, as we found using our microarray data. In contrast, the microarray results did not find a significant difference in expression for the ESTs RIHV and 24A28N3, while the northern blot indicates these ESTs are differentially expressed (but the microarray point estimates are in the right direction). Both methods were in agreement for the other ESTs represented in the figure (i.e. no difference between the samples).

7 Discussion

These very basic methods for normalization provide a means to estimate differences in array or scan specific effects. With replicate arrays within groups, the method will provide more accurate estimates of the fold changes across groups (how much is gained is still an open question). Although the basic idea has been presented in the literature previously, we have developed an implementation with a

rigorous basis, indicated why non-linear normalization may be problematic and introduced a method that not only finds the genes to conduct the normalization, but determines how much variability such estimation will add to the final estimates. In addition, the simple framework for thinking about normalization has provided some insights into the process, but this formulation is most useful when one starts considering extensions of the basic methods presented here. For example, we may want to allow for nonlinear normalizations. Indeed, although we have argued nonlinear relationships can arise in the M-A plot when the array specific effect doesn't interact with the intensity at the spots, such effects are easy to build into the current framework (hence we can assess the usefulness of such normalizations). Of more interest is the potential to build spatial effects into these models by allowing ϕ_k to depend on spatial location on the array (and bivariate spatial models when there are 2 dyes). This extension would allow us to achieve what is desired from background correction in a statistically rigorous fashion and combine the process with normalization. Another useful extension of the simple model presented here is to utilize hierarchical models for the intensities that allow for clusters of genes. In fact, this method could be easily combined with many of the existing parametric procedures for the analysis of microarrays.

The method presented here is most useful for comparing arrays in which a large proportion of the genes are differentially expressed and there is no compelling reason to think that there are as many genes up-regulated as there are genes down-regulated. This method is superior to the simple control gene approach because it incorporates the uncertainty in the estimation of the control intensity and allows for the maximum number of controls. Placing many control genes on the array also has the disadvantage of crowding out genes of interest (or replicates of such genes). This method can also be extended to analyze the results from proteomics experiments, whether the experiments utilize 2 dimensional electrophoresis, differential in gel electrophoresis or antibody arrays, since normalization is a necessary component of any of these technologies and the number of proteins that can be analyzed with these technologies is typically relatively small (currently less than 1,000 and often only a few hundred). Moreover, there are potentially other applications where the model developed here may assist in the estimation of treatment effects when such effects are confounded with block effects.

Finally, this method is able to detect important differences in gene expression in a biologically verified manner. While this exploratory work is still ongoing, the results we have found thus far are consistent with our understanding of certain aspects of disease mechanism and suggest other genes that may be important for understanding the tropism of PRRSV. By examining differences in tropism between strains we can start to understand the molecular mechanisms of infection. Such understanding is an important first step towards developing treatments for this costly disease.

References

Albina, E., Carrat, C., Charley, B. (1998), "Interferon-alpha response to swine arterivirus (PoAV), the porcine reproductive and respiratory syndrome virus", *Journal of Interferon and Cytokine Research*, 18, 485–490.

Amaratunga, D., Cabrera, J. (2001), "Analysis of data from viral DNA microchips", *Journal of the American Statistical Association*, 96, 1161–1170.

Bowtell, D. and Sambrook, J., ed. (2003), *DNA Microarrays*, CSHL Press, Cold Spring Harbor, New York.

Cheon, D., Chae, C., Lee, Y. (1997), "Detection of nucleic acids of porcine reproductive and respiratory syndrome virus in the lungs of naturally infected piglets as determined by in-situ hybridization", *Journal of Comparative Pathology*, 117, 157–163.

Duan, X., Nauwynck, H., Pensaert, M. (1997a), "Effects of origin and state of differentiation and activation of monocytes/macrophages on their susceptibility to porcine reproductive and respiratory syndrome virus (PRRSV)", *Archives of Virology*, 142, 2483–2497.

Duan, X., Nauwynck, H., Pensaert, M. (1997b), "Virus quantification and identification of cellular targets in the lungs and lymphoid tissues of pigs at different time intervals after inoculation with porcine reproductive and respiratory syndrome virus (PRRSV)", *Veterinary Microbiology*, 56, 9–19.

Dudoit, S., Yang, Y., Callow, M. and Speed, T. (2002), "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", *Statistica Sinica*, 12, 111–139.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment", *Journal of the American Statistical Association*, 96, 1151–1160.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998), "Cluster analysis and display of genome-wide expression patterns", *Proceedings of the National Academy of the Sciences*, 95, 14863–14868.

Gelman, A., and D. Rubin (1995), "Inference from iterative simulation with multiple sequences (with discussion)", *Statistical Science*, **7**, 457–511.

Ibrahim, J., Chen, M., Gray, R. (2002), "Bayesian models for gene expression with DNA microarray data", *Journal of the American Statistical Association*, 97, 88–99.

Keffaber, K. (1989), "Reproductive failure of unknown etiology", *Am Assoc Swine Pract Newsletter*, 1–9.

Kerr, M., Churchill, G. (2002), "Experimental design for gene expression microarrays", *Biostatistics*, in press.

Lam, L. Pickeral, O., Peng, A., Rosenwald, A., et al. (2001), "Genomic measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol", *Genome Biology*, **2(10)**, 1–11.

Li, C., Wong, W.H. (2001) "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection", *Proceedings of the National Academy of the Sciences*, **98**, 31–36.

Meulenberg, J., Hulst, M., de Meijer, E., Moonen, P., den Besten, A., de Kluyver, E., Wensvoort, G., Moormann, R. (1993), "Lelystad virus, the causative agent of porcine epidemic abortion and respiratory syndrome (PEARS), is related to LDV and EAV", *Virology*, 192, 62–72.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F. and Tsui, K. (2001), "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data", *Journal of Computational Biology*, v. 8, n. 1, 37–52.

Park, B., Yoon, I., Joo, H. (1996), "Pathogenesis of plaque variants of porcine reproductive and respiratory syndrome virus in pregnant sows", *American Journal of Veterinary Research*, 57, 320–323.

Raghavan, A., Ogilvie, R., Reilly, C., Abelson, M., Raghavan, S., Vasdewani, J., Krathwohl, M. and Bohjanen, P. (2002), "Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes", *Nucleic Acids Research*, 30, 24, 5529–5538.

Rossow, K., Shivers, J., Yeske, P., Polson, D., Rowland, R., Lawson, S., Murtaugh, M., Nelson, E., and Collins, J. (1999), "Porcine reproductive and respiratory syndrome virus infection in neonatal pigs characterized by marked neurovirulence", *Veterinary Research*, 144, 444–448.

Rossow, K., Benfield, D., Goyal, S., Nelson, E., Christopher-Hennings, J., and Collins J. (1996a), "Chronological immunohistochemical detection and localization of porcine reproductive and respiratory syndrome virus in gnotobiotic pigs", *Veterinary Pathology*, 33, 551–556.

Schadt, E., Li, C., Ellis, B., and Wong, W. (2001), "Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data", *Journal of Cellular Biochemistry Supplement*, **37**, 120–125,

Terpstra, C., Wensvoort, G., and Pol J. (1991), "Experimental reproduction of porcine epidemic abortion and respiratory syndrome (mystery swine disease) by infection with Lelystad virus: Koch's postulates fulfilled", *Veterinary Quarterly*, 13, 131–136.

Tseng, G., Oh, M., Rohlin, L., Liao, J. and Wong, W. (2001), "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects", *Nucleic Acids Research*, **29**, 2549–2557.

Wang, C., Hawken, R., Larson, E., Zhang, X., Alexander, L., and Rutherford M. (2001), "Generation and mapping of expressed sequence tags from virus-infected swine macrophages", *Animal Biotechnology*, 12, 51–67.

Wensvoort, G., Terpstra, C., Pol, J., ter Laak, E., Bloemraad, M., de Kluyver, E., Kragten, C., van Buiten, L., den Besten, A., and Wagenaar, F. (1991), "Mystery swine disease in The Netherlands: the isolation of Lelystad virus", *Veterinary Quarterly*, 13, 121-30.

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L. et al. (2001), "Assessing gene significance from cDNA microarray expression data via mixed models", *Journal of Computational Biology*, **8**, 625–637.

Yang, Y. Dudoit, S., Luu, P., Speed, T. (2001), "Normalization for cDNA microarray data", In Bittner, M., Chen, Y., Dorsel, A. and Dougherty, E. (ed.s), *Microarrays: Optical Technologies and Informatics*, V. 4266 proceedings of SPIE.

Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J. and Speed, T. (2002), "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation", *Nucleic Acids Research*, 30, 4, 1–10.

Yuan, S., Mickelson, D., Murtaugh, M., and Faaberg, K. (2001), "Complete genome comparison of porcine reproductive and respiratory syndrome virus parental and attenuated strains", *Virus Research* , 9, 189–200.

Zhang, X., Shin, J., Molitor, T., Schook, L., and Rutherford, M. (1999), "Molecular responses of macrophages to porcine reproductive and respiratory syndrome virus infection", *Virology*, 262, 152–162.

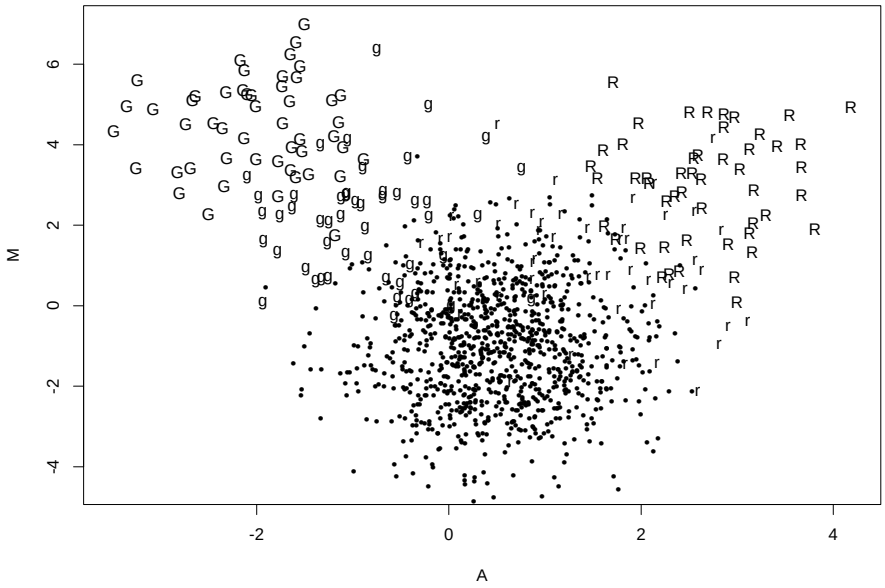

Figure 1: Clustering of genes leads to nonlinear M-A plots. An "R" indicates sample 1 is very up-regulated, an "r" indicates sample 1 is moderately up-regulated, "G" and "g" are analogous but sample 2 is down-regulated and a dot indicates no difference.

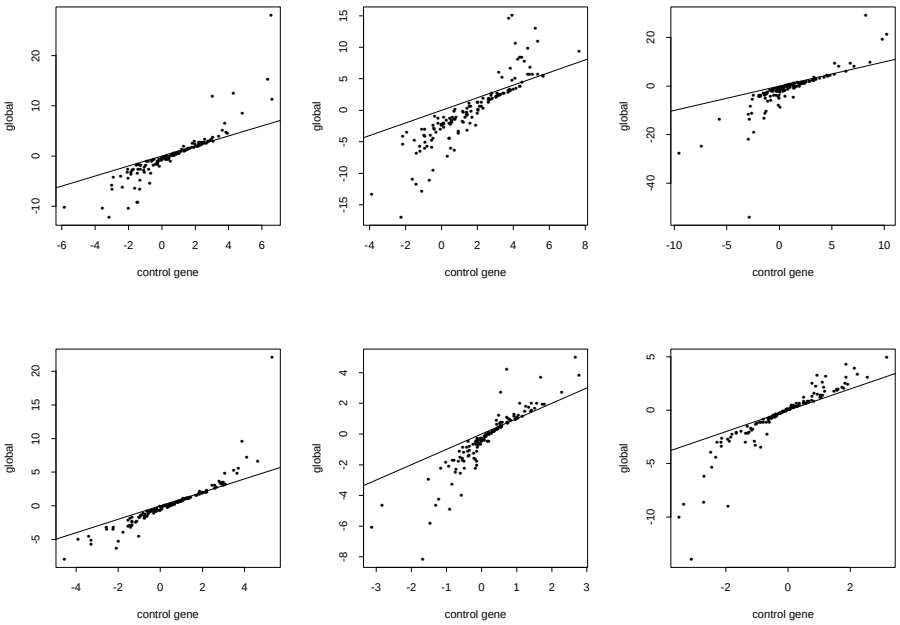

Figure 2: Comparison of global and control gene normalization (incorporating uncertainty in the estimation of the normalizing term) via t-statistics for 6 array comparisons. The lines are $y = x$ lines.

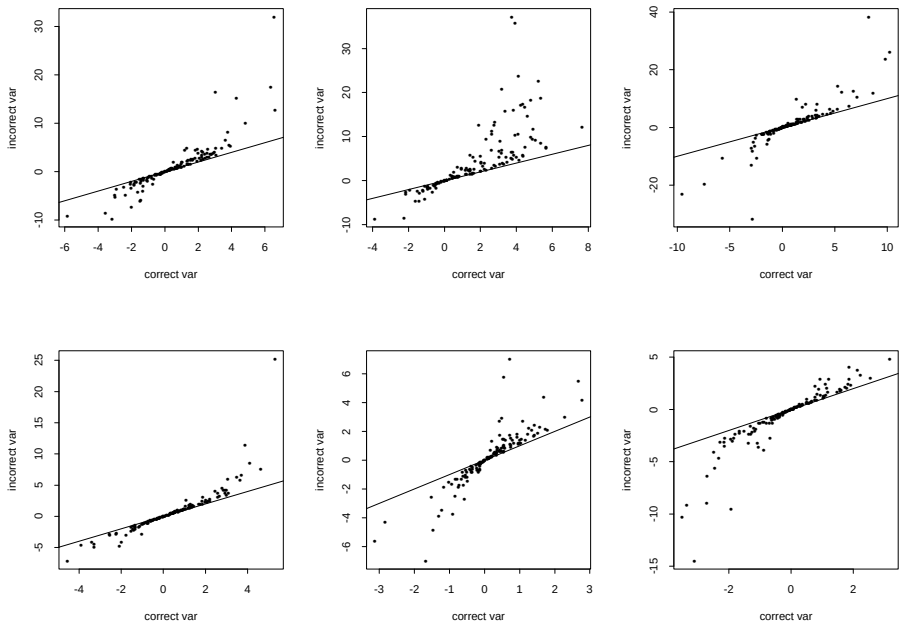

Figure 3: Comparison of control gene normalization ignoring uncertainty to control gene normalization incorporating uncertainty via t-statistics for 6 array comparisons. The lines are $y = x$ lines.

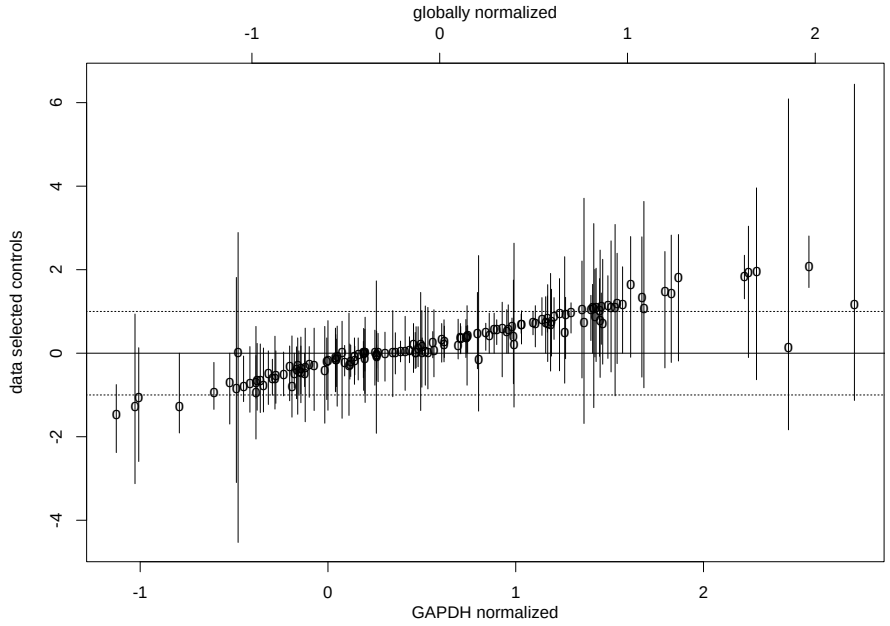

Figure 4: Comparison of global and control gene normalization incorporating uncertainty to the model based control gene normalization that estimates the set of control genes with 95% credible intervals of the log fold change for each gene. Dotted lines indicate a fold change of 2.