Quality Control and Robust Estimation for cDNA Microarrays with Replicates

Raphael Gottardo, Adrian E. Raftery, Ka Yee Yeung and Roger E. Bumgarner*

Revised June 16, 2005

Abstract

We consider robust estimation of gene intensities from cDNA microarray data with replicates. Several statistical methods for estimating gene intensities from microarrays have been proposed, but there has been little work on robust estimation. This is particularly relevant for experiments with replicates, because even one outlying replicate can have a disastrous effect on the estimated intensity for the gene concerned. Because of the many steps involved in the experimental process from hybridization to image analysis, cDNA microarray data often contain outliers. For example, an outlying data value could occur because of scratches or dust on the surface, imperfections in the glass, or imperfections in the array production. We develop a Bayesian hierarchical model for robust estimation of cDNA microarray intensities. Outliers are modeled explicitly using a t-distribution, and our model also addresses classical issues such as design effects, normalization, transformation, and non-constant variance. Parameter estimation is carried out using Markov chain Monte Carlo. By identifying potential outliers, the method provides automatic quality control of replicate, array and gene measurements.

*Raphael Gottardo is Assistant Professor, Department of Statistics, University of British-Columbia, 333-6356 Agricultural Road, Vancouver, BC V6T 1Z2 (E-mail: `raph@stat.ubc.ca`; Web: `www.hajek.stat.ubc.ca/ raph`). Adrian E. Raftery is Professor of Statistics and Sociology, University of Washington, Department of Statistics Box 354322, Seattle, WA 98195-4322 (E-mail: `raftery@stat.washington.edu`; Web: `www.stat.washington.edu/raftery`). Ka Yee Yeung is Research Assistant Professor, Department of Microbiology, University of Washington, Box 358070, Seattle, WA 98195. (E-mail: `kayee@u.washington.edu`). Roger Bumgarner is Associate Professor, Department of Microbiology, University of Washington, Box 358070, Seattle, WA 98195. (E-mail: `rogerb@u.washington.edu`). The authors thank Julian Besag for helpful discussion and Angelique van't Wout for providing us with some of the data. This research was supported by NIH Grant 8 R01 EB002137-02, and Raftery's research was also partially supported by ONR Grant N00014-01-10745. Yeung and Bumgarner were supported by NIH-NIDDK grant 5U24DK058813-02.

The method is applied to three publicly available gene expression datasets. It is compared to three other methods: ANOVA normalized log ratios, the median log ratio, and estimation after the removal of outliers based on Dixon's test, and the between-replicate variability of the intensity estimates is lower for our method than for any of the others.

We also address the issue of whether the background should be subtracted when estimating intensities. It has been argued that one should not do so because it increases variability, while the arguments for doing so are that there is a physical basis for the image background, and that not doing so will bias the estimated log-ratios of differentially expressed genes downwards. We show that the arguments on both sides of this debate are correct for our data, but that by using our model one can have the best of both worlds: one can subtract the background without increasing variability by much.

KEY WORDS: Bayesian hierarchical model; Gene filtering; Heteroscedasticity; Markov chain Monte Carlo; Outlier; Quality control; t distribution.

1 INTRODUCTION

cDNA microarrays consist of thousands of individual DNA sequences printed in a high density array on a glass microscope slide using a robotic arrayer. A microarray works by exploiting the ability of a given labeled cDNA molecule to bind specifically to, or hybridize to, a complementary sequence on the array. By using an array containing many DNA samples, scientists can measure—in a single experiment—the expression levels of hundreds or thousands of genes within a cell by measuring the amount of labeled cDNA bound to each site on the array. In a typical two-color microarray experiment, two mRNA samples, from control and treatment situations, are compared for gene expression. Treatment is taken in a broad sense to mean any condition different from the control. Both mRNA samples, or targets, are reverse-transcribed into cDNA, labeled using different fluorescent dyes (red and green dyes), then mixed and hybridized with the arrayed DNA sequences. The hybridized arrays are then imaged to measure the red and green intensities for each spot on the glass slide. Image analysis is an important aspect of microarray experiments, whose purpose is to

provide estimates of the foreground and background intensities for both the red and green channels (Yang et al., 2002a). The estimates of the red and green intensities are the starting point of any statistical analysis.

In order to measure gene expression changes accurately, it is important to take into account the random and systematic variations that occur in every microarray experiment. One way to measure the variation is to use replicated experiments in which each gene is replicated several times. In recent years, there has been a considerable amount of work on the estimation of gene intensities and the detection of differentially expressed genes (Chen et al. 1997; Newton et al. 2001; Dudoit et al. 2002; Gottardo et al. 2003). Because of the large number of steps involved in the experimental process from hybridization to image analysis, cDNA microarray data often contain outliers, and there is a need for robust methods.

Some work has been done on quality measure and filtering. Such approaches consist of calculating individual quality measures (sometime referred to as quality indices), and low quality spots are usually removed. Such filtering is often done at the image analysis level (Brown et al. 2001; Dudoit et al. 2002). However, these methods do not remove all "bad" spots, and some remain. In addition, spots can fall anywhere in the range from "good" to "bad", and such uncertainty should be taken into account when computing the estimates, e.g. log ratio estimates. When two or more replicates are available, Ideker et al. (2000) remove replicate outliers using Dixon's test at the 10% level. Tseng et al. (2001) filter genes based on the coefficient of variation and Lönnstedt and Speed (2002) remove genes with low intensities, but they do not address the problem of replicate outliers. Li and Wong (2000) consider the problem of outliers for oligonucleotide arrays, which is quite different and not applicable to cDNA microarray technology.

In this paper we introduce a Bayesian hierarchical model to estimate the intensities in a robust way. The robustness is achieved using a hierarchical-t formulation (Besag and Higdon 1999), which is more robust than the usual Gaussian model. Our model also deals with classical issues such as normalization, data transformation and non-constant variance. We also propose a way to filter out gene outliers and flag array outliers, based on the parameter estimates from our model. This provides an automatic quality control method for replicate, gene and array measurements.

The paper is organized as follows. Section 2 introduces the data structure and the notation. In Section 3, we present the Bayesian hierarchical model used to estimate the intensities and the parameter estimation method. In Section 4, we apply our model to experimental data and compare our results to those from popular alternative estimators. We also discuss whether one should subtract the background intensities from the image analysis. Finally, in Section 5 we discuss our results and possible extensions.

2 DATA

We used three datasets that are fairly typical of data in this area. The first two were produced by the University of Washington Center for Expression Arrays. They have the advantage that in each case we know whether or not all or some of the genes were differentially expressed. The last one is a widely used dataset in the field of gene expression, first analyzed by Dudoit et al. (2002).

The Like-like data: This dataset consists of 8 experiments using the same RNA preparation on 8 different slides. The samples that were applied to the arrays were RNA isolations from a HeLa cell line. The expression levels of about 7680 genes were measured. The same RNA was used for both samples, and no genes should be differentially expressed.

The HIV1 data: This dataset consists of four experiments using the same RNA preparation on 4 different slides. The expression levels of 4,608 cellular RNA transcripts were assessed in CD4-T-cell lines at time $t = 1$ hour after infection with HIV virus type 1. Included in this number was a set of selected control genes spotted on each slide. These included HIV-1 genes used as positive controls, i.e. genes known in advance to be differentially expressed, and non-human genes used as negative controls, i.e. genes known not to be differentially expressed. Further details are given by van't Wout et al. (2003).

The HIV2 data: These data were collected in the same way and in the same laboratory as the HIV1 data, but using a different RNA preparation.

The above datasets are results of balanced dye-swap experiments. Half of the replicates were hybridized with the green dye (Cy3) for the control and the red dye (Cy5) for the treatment; then the dyes were reversed.

The Apo Data: The goal of the study is to identify genes with altered expression in the livers of two lines of mice with very low HDL cholesterol levels compared to inbred control mice

(Dudoit et al. 2002). This experiment resulted in two datasets: a control and a treatment. Here we use the control dataset, which consists of 8 "normal" C57B1/6 mice. The target cDNA is obtained from mRNA by reverse transcription and labeled using a red-fluorescent dye (Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations was prepared by pooling cDNA from the 8 control mice. There is no dye swap in this experiment.

The three datasets were preprocess using a global lowess normalization step (Yang et al., 2002b). The data take the form

$$y_{iscr},\ i = 1, \ldots, I;\ s = 1, 2;\ c = 1, 2;\ r = 1, \ldots, R,$$

where y_{iscr} are the preprocessed background-subtracted intensities of gene i in sample s with color c from replicate r. We use different indices for the color and the sample to allow for dye-swap experiments.

Data transformation is an important initial step in microarray data analysis. It is often assumed that transforming the raw data logarithmically makes the effects additive. This assumption is approximately correct for gene expression data (Li and Wong 2000; Kerr et al. 2000; Rocke and Durbin 2001; Dudoit et al. 2002). Throughout this paper, logarithms are to base 2, which is standard in the analysis of microarray data. Figure 1(a) shows that on the log scale the dye effect is approximately additive. The dye effect is the result of an imbalance between the red and green intensities, which is known to be nonlinear (Yang et al., 2002b). Figure 1(b) shows the log ratio intensity plotted against half the sum of the log intensities from the two channels; we refer to the latter quantity as the "overall intensity." Figure 1(b) is just a 45° counterclockwise rotation of Figure 1(a). The Locally Weighted Scatterplot Smoother, or lowess (Cleveland 1979), indicates that such a nonlinear trend is present in the Like-like data. We can see that the effects from the two different groups where the dyes have been swapped are almost identical but reversed. As a result, in a balanced dye-swap experiment, we expect the dye effect to be absent or at least greatly reduced when computing genewise averages. However, it is common to use lowess normalization (Yang et al., 2002b) to remove the nonlinear dye bias. This is particularly relevant for the APO data where there is no dye-swap.

The Like-like data presented here measure the relative expression of a group of genes using the same mRNA in the two samples. As a result, we expect to observe the technical

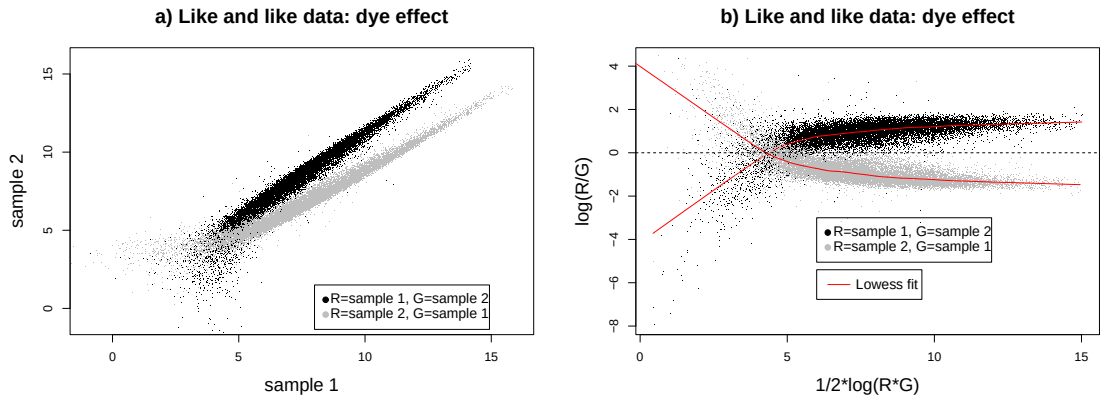

Figure 1: Effect of the Dye Swap on the Like-like Data. (a) Log(sample 1) versus log(sample 2). This shows that the dye effect is approximately additive on the log scale. (b) Overall intensity versus log ratio. This shows that the nonlinear effects approximately cancel one another out in a balanced dye-swap experiment. The overall intensity is half the sum of the log intensities in the two channels, and all logarithms are to base 2.

variation but not the biological variation. Figure 2 shows the normalized log intensities of ten different genes in two samples. Even though the data are normalized and there is no biological variation, we observe some replicate outliers in each sample. It is clear that these outliers can have a big effect on the intensity estimates.

3 ROBUST ESTIMATION AND QUALITY CONTROL

In this section, we introduce the Bayesian hierarchical model we use to estimate the intensities in each sample. We use a Bayesian linear model (Lindley and Smith 1972) with t-distributed sampling errors to allow for replicate outliers (Besag and Higdon 1999). We also explicitly model the non-constant variances by using an exchangeable prior for the gene precisions (Lewin et al. 2003). Our model includes design effects that deal with normalization issues (Kerr et al. 2000). We model the intensities on the log scale because the effects are close to additive on that scale, as shown in Section 2, and because log measurements have a simple interpretation.

3.1 The Model

We model $y^*_{iscr} = \log_2(y_{iscr} + \kappa)$ where κ is a positive additive constant. This shifted logarithmic transformation was proposed by Tukey (1957) and studied in detail by Box and Cox (1964); it is often used to analyze gene expression data (Kerr et al. 2000; Cui et al.

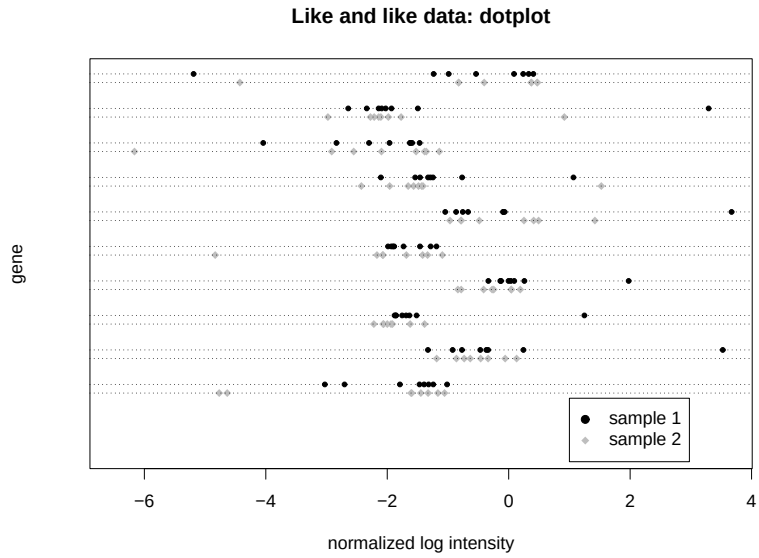

Figure 2: Dot Plots of 10 Genes from the Like-like Data. Even though the data were normalized, some replicate outliers are present in each sample.

2002). Rocke and Durbin (2003) showed that the shifted logarithm can be an approximate variance-stabilizing transformation for gene expression data. The purpose of introducing the shift κ is to avoid taking the logarithm of negative numbers and to reduce the variance at low intensities. The parameter κ is estimated beforehand and is treated as fixed in the estimation of the full model, as described in Section 3.3.

Conditionally on the parameters $(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{\gamma})$, it is assumed that the $(y^*_{i1cr}, y^*_{i2cr})'$ are independent and can be written as

$$\begin{aligned} y^*_{iscr} = g_\kappa(y_{iscr} - x_{iscr}) &= \mu + \alpha_s + \beta_c + \eta_r + \gamma_{is} + \delta_{sc} + \frac{\epsilon_{iscr}}{\sqrt{w_{icr}}}, && (1) \\ (\gamma_{is}|\lambda_{\gamma_s}) &\sim \mathrm{N}(0, \lambda_{\gamma_s}), && (2) \\ (\epsilon_{i1cr}, \epsilon_{i2cr})'|\mathbf{V}_i &\sim \mathrm{N}_2(\mathbf{0}, \mathbf{V}_i), \\ (w_{icr}|\nu_r) &\sim \mathcal{G}a(\nu_r/2, \nu_r/2), \end{aligned}$$

where w_{icr} and $(\epsilon_{i1cr}, \epsilon_{i2cr})'$ are independent. Since the w's are independent of the ϵ's, we have $\frac{\epsilon_{iscr}}{\sqrt{w_{icr}}} \sim \mathcal{T}_{(\nu_r, \mathbf{0}, \mathbf{V}_i)}$, i.e. the (bivariate) errors have a bivariate t distribution with ν_r degrees of freedom and covariance matrix $\mathbf{V}_i$. The advantage of writing the model this way is that, conditioning on the w_{icr}, the sampling errors are again normal but with different precisions, and estimation becomes a weighted least squares problem. The hierarchical structure of the model is summarized in the directed acyclic graph in Figure 3.

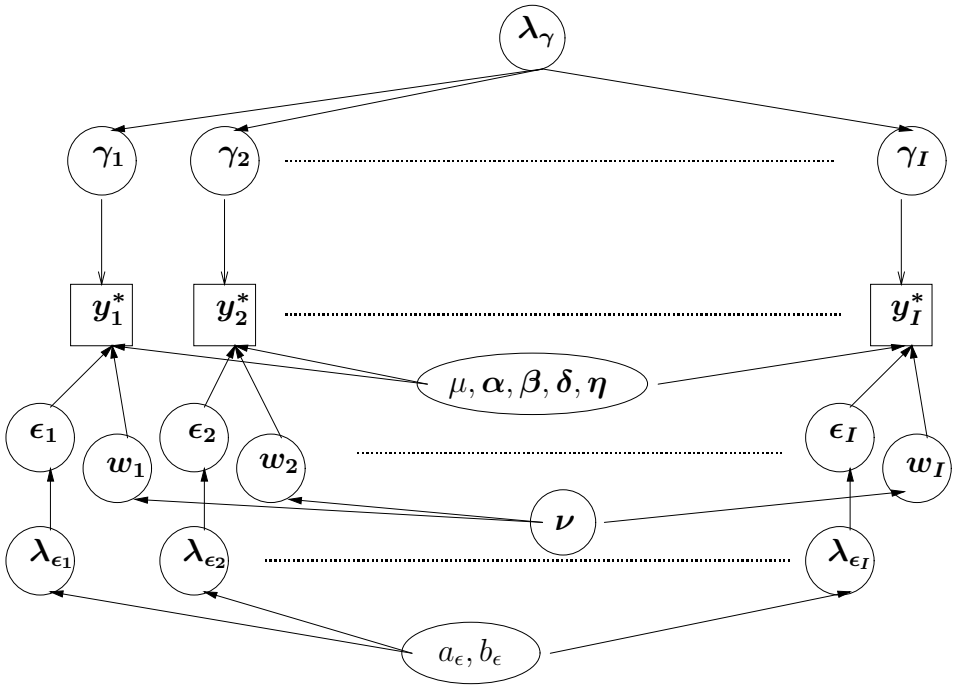

Figure 3: Directed Acyclic Graph of the General Model in Equation (1).

In (1), μ is the baseline intensity. The sample effect α_s is used to remove the bias between the two samples. If only a few of the genes are differentially expressed, the sample effect will measure only the sample bias and will not greatly affect the differentially expressed genes. The dye effect is represented by β_c, and accounts for the fact that the green dye tends to be brighter than the red dye (Yang et al., 2002b). The interaction of the sample s with the sample c is denoted by δ_{sc}, and is present because the different dyes tend to have different biases in different samples. The dye effects β_c and δ_{sc} are estimable only in a dye-swap experiment; otherwise they need to be removed from the model by setting $\beta_c = 0$ and $\delta_{sc} = 0$. The array effect of replicate r, η_r, is intended to normalize the overall intensity of each array across replicates. This parameter is needed because differences in overall intensity are frequent in microarray data. There are several reasons why this is so; for example, the amount of RNA solution used on each array might not be the same, leading to brighter arrays after the scanning process. Finally, γ_{is}, the effect of gene i in sample s, is the quantity of interest. We model it as a random effect with a Gaussian distribution as defined by (2).

For a given gene, the correlation matrix, $\mathbf{V}_i$, allows the measurements from the two samples to be correlated. Ideker et al. (2000) use a similar covariance structure in their

linear model. The log transformation usually stabilizes the variance for high intensity genes but low intensity genes can be highly variable. A model that allows gene-dependent variances seems more appropriate. The precision matrix (i.e. the inverse of the covariance matrix) is given by

$$\begin{aligned}(\mathbf{V}_i^{-1}|\rho,\lambda_{\epsilon_{1i}},\lambda_{\epsilon_{2i}}) &= \frac{1}{(1-\rho^2)}\begin{pmatrix}\lambda_{\epsilon_{1i}} & -\sqrt{\lambda_{\epsilon_{1i}}\lambda_{\epsilon_{2i}}}\rho\\ -\sqrt{\lambda_{\epsilon_{1i}}\lambda_{\epsilon_{2i}}}\rho & \lambda_{\epsilon_{2i}}\end{pmatrix},\\ (\lambda_{\epsilon_{si}}|a_\epsilon,b_\epsilon) &\sim \mathcal{G}a(a_\epsilon^2/b_\epsilon,a_\epsilon/b_\epsilon),\end{aligned}$$

where ρ is the correlation between samples, $\lambda_{\epsilon_{si}}$ is the precision of gene i in sample s, and $\mathcal{G}a(a_\epsilon^2/b_\epsilon,a_\epsilon/b_\epsilon)$ denotes a Gamma distribution with mean a_ϵ and variance b_ϵ. We use an exchangeable prior for the precisions, so that information is shared between the genes. This allows shrinkage of very small and very large variances.

3.2 Priors

We use a vague but proper prior for the precision of the random effects λ_{γ_s}, exponential with mean 200, so that $\lambda_{\gamma_s} \sim \mathcal{G}a(1, 0.005)$. Apart from the γ's, all the other effects are assumed to be random with a large variance, namely $\mathrm{N}(0, 25)$. They are fixed effects but are estimated in a Bayesian way, so that uncertainty about those parameters can be captured as part of the estimation process (Lindley and Smith 1972).

For identifiability, we impose the constraints $\alpha_1 = 0$ and $\beta_1 = 0$, $\delta_{11} = \delta_{12} = \delta_{21} = 0$, $\eta_1 = 0$, and $\eta_R = 0$. The constraint $\eta_R = 0$ is not needed if there is no dye-swap, because in that case $\beta_c = 0$. We also need two constraints on the γ_{is}, such as $\sum_i \gamma_{is} = 0$ for $s = 1, 2$. However, instead of including these constraints as part of the model definition, we let the $\boldsymbol{\gamma}$'s be "free" during the MCMC sampling process, and identify the parameters afterwards from the sampled values; see Section 3.3.

We also use vague but proper priors for the error precisions, namely $a_\epsilon \sim \mathcal{U}_{[0,10000]}$ and $b_\epsilon \sim \mathcal{U}_{[0,10000]}$. The prior for the correlation between the two samples is given by $\rho \sim \mathcal{U}_{[-1,1]}$.

The prior for the degrees of freedom ν_r is uniform on the set $\{1, 2, \ldots, 10, 20, \ldots, 100\}$. A similar approach was taken by Besag and Higdon (1999). They used a uniform hyperprior on the set $\{1, 2, 4, 8, 16, 32, 64\}$ for the degrees of freedom. From a practical point of the view, the biggest difference between our approach and theirs is that we also include 3 in the set of possible values of ν_r. Our results suggest this to be useful, as there can be a noticeable

difference between results for low degrees of freedom, especially 2, 3 and 4, but much smaller differences for larger values of ν_r. By using a prior that allows degrees of freedom between 1 and 100, we allow a wide range of sampling errors from the heavy tailed Cauchy ($\nu = 1$) to nearly Gaussian ($\nu = 100$).

3.3 Parameter Estimation

Realizations were generated from the posterior distribution via Markov chain Monte Carlo (MCMC) algorithms (Gelfand and Smith 1990). We used Gibbs updates when the full conditionals had a simple form; otherwise we used slice sampling with the "stepping out" procedure (Neal 2003).

The model (1) does not allow the identification of all parameters because we do not impose any constraint on $\boldsymbol{\gamma}$. However, contrasts involving elements of $\boldsymbol{\gamma}$ are identified, and one could force all the parameters to be identified by imposing constraints such as $\sum_i \gamma_{is} = 0$ for $s = 1, 2$. For simplicity, we did not take such an approach. Instead we fitted the unconstrained model and postprocessed the MCMC output to identify all the parameters. By postprocessing, we mean that after running the MCMC algorithm, we changed the simulated values of γ_{is} so that $\sum_r \gamma_{sr} = 0$ for $s = 1, 2$, and then recomputed the corresponding other parameters at each iteration. A similar approach has been taken to solving the label-switching problem in Bayesian inference for finite mixture models using MCMC (Stephens 2000; Celeux et al. 2000). There exists other methods to overcome the lack of identifiability and we refer the reader to Vines, Gilks, and Wild (1996) for further details.

We started the Markov chain from the least squares estimates of the parameters. We used the method of Raftery and Lewis (1992, 1996) to determine the number of iterations, based on a short pilot run of the sampler. For each dataset presented here, this suggested that a sample of no more than about 50,000 iterations with 1,000 burn-in iterations was enough to estimate standard posterior quantities.

We estimated the shift κ in advance by fitting (1) with $w_{icr} \equiv 1$, $\rho = 0$ and $\lambda_{\epsilon_{is}} \equiv \lambda_{\epsilon_s}$ via MCMC, and treating κ as a parameter with a vague uniform prior $\kappa \sim \mathcal{U}_{[0,10000]}$. We then estimated κ by its posterior mean.

At first sight it would seem natural to estimate κ instead by including it as a parameter

in the MCMC estimation of the full model (1), but we did not do so, for the following reason. If we did so, the posterior distribution of the quantities of interest, the γ_{is}, would be averaged over different values of κ. However, when κ changes, so does the scale on which γ_{is} is measured and hence its interpretation, and so this would amount to averaging quantities denoted by the same symbol, but that actually have different interpretations. We therefore opted to estimate κ first and then estimate the other parameters conditionally on the resulting value of κ. A similar issue arises in making inference about regression parameters when a Box-Cox (1964) transformation has been used. Box and Cox (1982) pointed out that inflating the standard errors of regression parameters to take account of uncertainty about the transformation used amounts to averaging over inferences on different scales, and so is scientifically inappropriate. They recommended first estimating the transformation parameter, and then making inference about the regression parameters conditionally on the resulting estimate. In practice, in our datasets, the posterior distribution of κ was highly concentrated, and the results would have been similar had we treated κ as a parameter of the full model (1) in the MCMC estimation.

The full estimation, including estimation of the shift, took about 5 hours for the HIV data, about 8 hours for the Like-like data, and about the same for the APO data, on an Intel Xeon processor running at 3GHz. An R software package called `rama` implementing the method is be available from Bioconductor at `www.bioconductor.org`.

3.4 Quality control and gene filtering

Our model not only accounts for replicate outliers, but it can also be used to identify gene and array outliers. To identify such outliers we use the posterior mean of the w's, which we refer to as weights. For a single gene, if half or more of the replicates are "severely" downweighted, one might think that all the measurements from that same gene are unreliable. We propose to filter out a gene if $\lfloor (R+1)/2 \rfloor$ or more replicate measurements have associated weights smaller than a fixed threshold w_{min}, where $\lfloor \cdot \rfloor$ is the floor function. In the example explored here we use $w_{min} = 0.3$, which removes less than 1% of the data, while significantly improving the quality.

Similarly, one could flag array outliers, i.e. arrays containing too many downweighted measurements. The number of outliers on a single array directly affects the corresponding

array degrees of freedom, ν_r. One can therefore use the posterior mode of ν_r to determine if the rth array is a potential outlier. We do not necessarily recommend discarding the whole array in that case, but a small value of ν_r should certainly be used as a warning.

3.5 Combining biological and technical variation

The three datasets explored in this paper contain biological or technical replicates only. If an experiment contains both biological and technical replicates, it is possible to elaborate model (1) in order to account for each source of variation separately. We need to introduce a new subscript, b, indicating that replicate r comes from biological replicate b. We let y_{iscbr} be the log shift transformed intensity of replicate r, biological replicate b, of gene i in sample s and color c. Assuming that each replicate was spotted on a different slide, we modify model (1) as follows:

$$\begin{aligned} y^*_{iscbr} = g_\kappa(y_{iscbr}) &= \mu + \alpha_s + \beta_c + \eta_r + \gamma_{is} + \delta_{sc} + \psi_{iscb} + \frac{\epsilon_{iscbr}}{\sqrt{w_{icbr}}}, \\ \psi_{iscb} &\sim \mathrm{N}(0, \lambda^{-1}_{\psi_{is}}), \\ (\epsilon_{i1cbr}, \epsilon_{i2cbr})'|\mathbf{V}_i &\sim \mathrm{N}_2(\mathbf{0}, \mathbf{V}_i), \\ (w_{icbr}|\nu_r) &\sim \mathcal{G}a(\nu_r/2, \nu_r/2), \end{aligned} \tag{3}$$

where ψ_{iscb} is the biological error component, and all the other terms are as in model (1). For the model to be identifiable, we need additional constraints; we suggest $\sum_b \psi_{iscb} = 0$ for each i and s. We would use an exchangeable prior for the precisions of the biological error component, i.e. $\lambda_{\psi_{is}} \sim \mathcal{G}a(a^2_\psi/b_\psi, a_\psi/b_\psi)$, and we would again let a_ψ and b_ψ be uniform over broad intervals, which could be determined from previous experiments. If the number of replicates is small, the number of precision parameters could be reduced by assuming that one of the error components has a constant variance. Finally, the shift could be estimated by fitting (3) with $w_{icr} \equiv 1$, $\rho = 0$ and $\lambda_{\epsilon_{is}} \equiv \lambda_{\epsilon_s}$ via MCMC, and treating κ as a parameter with a vague uniform prior $\kappa \sim \mathcal{U}_{[0,10000]}$, as for model (1).

Model (3) is given only as a general guideline, and some designs might require modifications.

4 RESULTS

4.1 Methods to be compared

Here, we briefly review the different methods to be compared on the three datasets presented in Section 2.

ANOVA log ratios without shift: A popular estimate is the ANOVA normalized log ratio, $(\hat{\gamma}_{i1} - \hat{\gamma}_{i2})$, where $\hat{\gamma}_{is}$ is the least squares estimate of the effect of gene i in sample s, which corresponds to fitting Model (1) with fixed effects, Gaussian errors and constant variance, to the non-shifted log measurements ($\kappa = 0$). The name "ANOVA normalization" was introduced by Kerr et al. (2000), and we use the same terminology even though our model is slightly different.

ANOVA log ratios: This is the same as above but on the log shifted measurements, where the shift κ is estimated from our model.

Median log ratios: A more robust alternative to the mean is the median. For each single gene, we compute the median of the log ratios (with shift) across replicates.

Robust Analysis of MicroArrays (RAMA): From model (1), we estimate the effect of gene i in sample s by the posterior mean of γ_{is}, denoted by $\bar{\gamma}_{is}$. The log ratio of gene i is estimated by $\bar{\gamma}_{i1} - \bar{\gamma}_{i2}$. In addition we filter out gene outliers by removing genes that have half or more of their weights smaller than 0.3, as described in Section 3.4.

ANOVA log ratios with Dixon's test: These are the same as the ANOVA log ratios, except that Dixon's test (Dixon 1950) at the 10% level is now applied to the shifted log ratios to filter out replicate outliers. Dixon's test was used by Ideker et al. (2000) to remove replicate outliers from cDNA microarray data.

4.2 Application to Experimental Data

In this section, we use specific genes from the HIV data described in Section 2 to demonstrate the potential of our model. Table 1 summarizes the estimated coefficients when our model is applied to the HIV1 data. The posterior modes of the degrees of freedom of the t-distribution, ν_r, ranged from 3 to 6, indicating that the sampling errors are heavier-tailed than the Gaussian distribution. There is substantial between-sample correlation, estimated as 0.73, even after removing design effects and gene effects. Our model also captures the

nonconstant variance with posterior means 7.31 and 1.91 for a_ϵ and b_ϵ respectively. The posterior mean of 1.91 for the variance of the gene precisions, b_ϵ, allows the gene precisions to be quite different and should capture the larger variance at low intensity. The estimated shift is 11.47, which is relatively small on the raw scale which runs from 0 to 65535.

Table 1: Estimation of the Coefficients from our Bayesian Model on the HIV1 Data. The posterior modes of the degrees of freedom of the t-distribution, ν_r, ranged from 3 to 6, indicating that the sampling errors are heavier-tailed than the Gaussian distribution.

Parameter	Effect	Bayesian estimate	Posterior sd	$q_{0.025}$	$q_{0.975}$
μ	baseline intensity	7.71	0.006	7.70	7.73
α_2	sample effect	-0.02	0.006	-0.03	-0.01
β_2	dye effect	0.19	0.009	0.18	0.21
δ_{22}	dye×sample interaction	-0.009	0.005	-0.015	0.0022
η_2	array 2 effect	-0.50	0.008	-0.52	-0.48
η_3	array 3 effect	-0.09	0.009	-0.11	-0.08
λ_{γ_1}	gene precision sample 1	0.29	0.006	0.28	0.31
λ_{γ_2}	gene precision sample 2	0.30	0.007	0.29	0.31
ρ	correlation between samples	0.73	0.004	0.72	0.74
a_ϵ	mean of error precisions	7.31	0.10	7.11	7.51
b_ϵ	variance of error precisions	1.91	0.49	1.08	3.01
ν_1	df for array 1	6	0.23	6	7
ν_2	df for array 2	5	0.05	5	5
ν_3	df for array 3	4	0	4	4
ν_4	df for array 4	3	0	3	3

Note: The Bayesian estimate is the posterior mean, except for ν_1, ν_2, ν_3 and ν_4, for which it is the posterior mode.
$q_{0.025}$: 0.025 quantile. $q_{0.975}$: 0.975 quantile

The methods compared in Tables 2–5 are all applied to the log shifted measurements where the shift is estimated from our model. Tables 2–4 show the effect of the t-distribution when replicate outliers are present. The weights correspond to the posterior mean of the w_{icr} for each pair of observations. Conditioning on the w_{icr}'s, the posterior mean can be seen as a weighted mean with the w_{icr} as weights. Table 2 shows that Dixon's test at the 10% level fails to remove a clear replicate outlier. Because of the outlier, the difference between the estimates of the effect of the same gene from the HIV1 data and the HIV2 data is quite large, a difference of 0.46. Our model clearly downweights the outlier, and as a result the difference between the two estimates is much smaller with our method, at 0.29. The median

also performs well in this case.

Table 2: Log Ratios of One Gene of the HIV Datasets. Dixon's test fail to remove a clear replicate outlier whereas it is downweighted by our model.

	Replicates						posterior	Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	-0.01	0.11	0.12	-0.10	0.03	0.06	0.03	0.03
weights	1.16	1.26	1.26	0.55				
HIV 2 (log ratio)	-1.31	-0.56	0.06	0.10	-0.43	-0.25	-0.26	-0.43
weights	0.25	0.83	0.87	0.82				
difference					0.46	0.31	0.29	0.46

Note: ⋆gene removed by Dixon's test on the normalized measurement at the 10% level. The Dixon mean is the sample mean after removing the replicate measurements identified as outliers by Dixon's test (if any).

In Table 3, there are clear replicate outliers in the HIV1 and HIV2 data that are removed by Dixon's test and are also downweighted by our model. In this case our method gives estimates that differ between experiments by 0.06, which is better than the three other methods.

Table 3: Log Ratios of One Gene of the HIV Datasets. Dixon's test remove two clear replicate outliers. The same two outliers are also downweighted by our model.

	Replicates						posterior	Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	0.26	-0.85⋆	0.68	0.59	0.17	0.43	0.34	0.52
weights	1.07	0.31	0.76	1.21				
HIV 2 (log ratio)	-4.09⋆	0.29	0.57	0.21	-0.75	0.25	0.28	0.36
weights	0.02	1.23	0.87	1.17				
difference					0.92	0.18	0.06	0.16

Note: ⋆gene removed by Dixon's test on the normalized measurement at the 10% level. The Dixon mean is the sample mean after removing the replicate measurements identified as outliers by Dixon's test (if any).

Even though the median and the Dixon mean are more robust than the usual mean, they can be quite inefficient when the number of replicates is small. In particular, Dixon's test often falsely identifies replicate outliers. For example, in an experiment with four replicates, it is not unusual for some of the genes to have three replicate measurements close together, by chance only, making the last measurement look like an outlier (Table 4).

Table 4: Log Ratios of One Gene of the HIV Datasets. A non-outlying replicate is incorrectly removed by Dixon's test.

	Replicates						posterior	Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	-0.17*	0.18	0.19	0.19	0.10	0.18	0.08	0.19
weights	1.08	1.24	1.29	1.26				
HIV 2 (log ratio)	0.09	-0.06	-0.13	-0.05	-0.03	0.06	-0.05	-0.03
weights	1.03	1.22	1.05	0.91				
difference					0.13	0.24	0.13	0.22

Note: *gene removed by Dixon's test on the normalized measurement at the 10% level. The Dixon mean is the sample mean after removing the replicate measurements identified as outliers by Dixon's test (if any).

Finally, Table 5 shows a gene outlier removed by our filtering method. For this particular gene, the posterior mean performs worst due to the fact that half of the replicates are greatly downweighted. However, the gene is easily identified as a gene outlier, and removed from further analysis.

Table 5: Log Ratios of an Outlying Gene of the HIV Datasets. This gene is removed by our filtering method since it contains two replicates with weights smaller than 0.3. The posterior mean is not performing well, but the gene is easily identified as a gene outlier.

	Replicates						posterior	Dixon
	1	2	3	4	mean	median	mean	mean
HIV 1 (log ratio)	-1.37	-1.44	1.51	-0.16	-0.36	-0.76	-1.21	-0.36
weights	0.98	0.99	0.06	0.17				
HIV 2 (log ratio)	-0.31	0.55	0.35	-0.97	-0.10	0.02	0.03	-0.10
weights	0.90	0.84	0.99	0.35				
difference					-0.26	-0.78	-1.24	-0.26

Note: *gene removed by Dixon's test on the normalized measurement at the 10% level. The Dixon mean is the sample mean after removing the replicate measurements identified as outliers by Dixon's test (if any).

4.3 Between-Replicate Variability of Estimates

In this section, we compare the different log ratio estimates introduced in Section 3.3, by dividing each dataset into two groups of four replicates. We first compare our estimates with the ANOVA normalized log-ratios without shift obtained from the first four replicates of the Like-like data, as shown in Figure 4. These are a natural first point of comparison, because

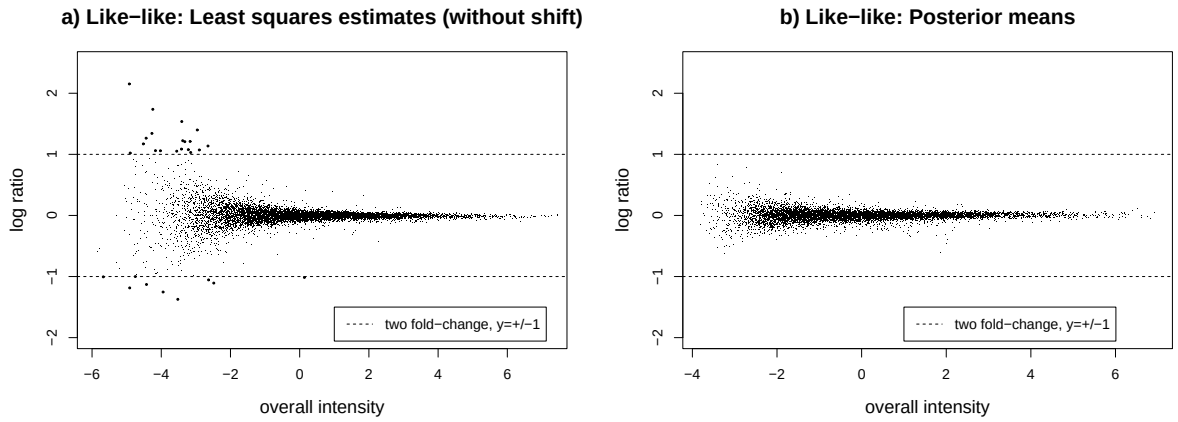

Figure 4: Log Ratio Estimates as a Function of the Overall Intensity, i.e. the average of the gene effects in each sample, on the Like-like Data (first 4 replicates). Using our method (b), 26 gene outliers were filtered out. The dashed lines show a two-fold change. The number of false positives for the normalized log ratios is 21, as against only 0 for the posterior means, an 100% reduction.

they are essentially the averaged log ratios, since the data are lowess normalized. In theory, the Like-like data should not show any differentially expressed genes. The dashed lines in Figure 4 show a two-fold change. The ratio of two is sometimes used as a rule of thumb for selecting differentially expressed genes (Schena et al. 1995; Yang et al. 1999). Because of the high variability at low intensity, some of the genes show a greater than two-fold change in expression. Using the two-fold change rule, the number of false positives for the ANOVA normalized log ratios is 21, as against zero for the posterior means, an 100% reduction.

We highlighted two groups of genes in the two HIV datasets. The first group consists of HIV genes (positive controls) that are known to be differentially expressed, and the second one consists of non-human genes (negative controls), which are known not to be differentially expressed. Figure 5 shows that our model enhances the identification of the differentially expressed genes. It shrinks the low intensity (highly variable) genes, but does not modify the differentially expressed genes too much. Note that one of the 13 HIV positive control genes, which we expected to be highly expressed, has a small log ratio estimate using all of the methods, suggesting that the corresponding probe did not properly hybridize (Figure 5).

We now compare all the methods described in Section 3.3 by computing an estimate of the log ratio for each gene from each of the two groups, and computing the mean squared differences (MSD) between the two estimates, averaged over genes. The eight replicates of

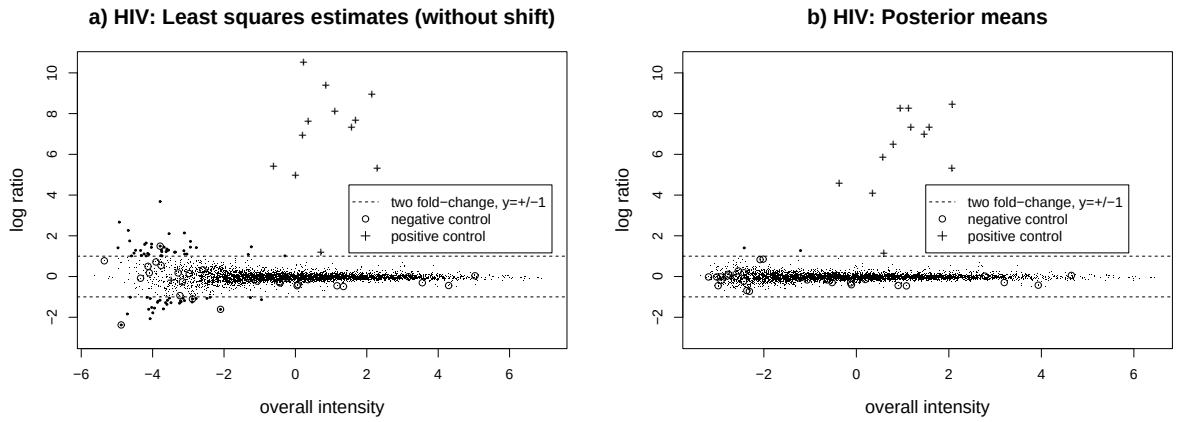

Figure 5: Log Ratio Estimates as a Function of the Overall Intensity, i.e. the sum of the gene effects in each sample, on the HIV2 Data. The dashed lines show a two-fold change. Using our method (b), eight gene outliers were filtered out. The number of false positives at low intensity is greatly reduced. The log ratio estimates of the true differentially expressed genes stay about the same.

the HIV data were separated into two groups of four, consisting of the HIV1 and HIV2 replicates respectively. Even though those two groups are biological rather than technical replicates, we expect the log ratios from each group to be similar. For the Like-like and Apo data, we used five randomly chosen partitions of the eight replicates into two groups of four. As part of our method, we remove the genes that were identified as gene outliers in either dataset. This removed on average 24 genes from the HIV data, 35 genes from the Apo data and 55 from the Like-like data, which represents less than 1% in each case.

The ANOVA log ratios without shift perform poorly, with high variability between estimates. The ANOVA log ratios with shift perform much better, as shown in Table 6. Removing replicate outliers using Dixon's test does not improve things; it actually does worse even though it removed more than 500 measurements in each dataset. The median performs slightly better for the Like-like data but not the other two.

The between-replicate variability is substantially decreased when using our model. Our method reduced the mean squared difference between estimates by 6% for the Like-like data, 3% for the HIV data, and 6% for the APO data compared to the ANOVA estimates with shift. In Table 6, we use the ANOVA estimates with shift estimated by our model as the baseline, in order for the data to be on the same scale. However, our method provides an overall framework in which estimation of the shift is included, and this turned out to be an

important feature (Figures 4 and 5).

Table 6: Mean Squared Differences (MSD) Between the Estimates when Dividing each dataset into Two Groups of Four Replicates. For the Like-like data and the Apo data the MSD numbers are averages from five random splits. The posterior mean reduces the MSD by 6%, 3% and 6% in each dataset, respectively, relative to the MSD from the ANOVA log ratios with shift estimated from our model (MSD_b). The lowest MSD for each dataset is indicated in bold.

	Like-like		HIV		APO	
Estimates	MSD	MSD/MSD_b	MSD	MSD/MSD_b	MSD	MSD/MSD_b
ANOVA (No Shift)	0.062	2.77	0.181	2.39	0.038	1.15
ANOVA (Baseline)	0.022	1.00	0.074	1.00	0.033	1.00
ANOVA (w/ Dixon)	0.023	1.01	0.077	1.04	0.036	1.09
Median	0.022	0.99	0.077	1.04	0.034	1.03
Posterior mean	**0.021**	**0.94**	**0.072**	**0.97**	**0.031**	**0.94**

Note: MSD_b: MSD from the baseline, i.e. ANOVA normalized log ratios.
ANOVA (No Shift): ANOVA log ratios without shift
ANOVA (Baseline): ANOVA log ratios (with shift estimated from our model)
ANOVA (w/ Dixon): Same as ANOVA and Dixon's test at the 10% level was used to remove replicate outliers.

Table 7 gives the MSD values computed from variants of our model. We calculated the MSD when fitting our model with Gaussian errors, which can be seen as a limiting case as ν goes to infinity, and with t errors as in (1) with and without gene filtering. Note that it is not clear how to do gene filtering with the Gaussian model since in that case the weights are all equal to one. In each case, we give two numbers for each dataset, one obtained from the log transformed data and one obtained from the log shift transformed data where the shift is estimated from our model. It can be seen that the shift brings a large improvement for the Like-like and HIV data, but it is not as large for the APO data. The t distribution alone improves the MSD for the APO data over the Gaussian model, but the gene filtering is necessary for the other two datasets to improve the MSD over the Gaussian model. This is consistent with the results shown in Table 5. Gene outliers can lead to large squared differences with the t model due to the downweighting of too many replicates; this does not occur with the Gaussian model. However, using our model with t errors, gene outliers are easily identified and removed from the dataset. Overall, our model provides a framework for shift estimation, replicate outlier accommodation, and gene outlier filtering. All these

Table 7: Mean Squared Differences (MSD) Between Estimates (posterior means) obtained from variants of our model when Dividing each Dataset into Two Groups of Four Replicates. For the Like-like data and the Apo data the MSD is an average over five random divisions. The first line corresponds to the MSD computed from our model replacing the t errors with Gaussian errors. The second line is the MSD obtained from the t model as described in Section 3.1. The last line corresponds to the MSD from the t model obtained after removing gene outliers as described in Section 3.4. For each dataset, we provide two numbers, one where the shift was set to zero (No Shift) and one where the shift was estimated from our model (Shift). For each dataset, the lowest MSD is shown in bold.

	Like-like		HIV		APO	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
Gaussian	0.046	0.022	0.128	0.074	0.040	0.035
t	0.049	0.023	0.161	0.076	0.036	0.032
t with gene filtering	0.038	**0.021**	0.086	**0.072**	0.036	**0.031**

features combined give the best results in terms of MSD on all three datasets.

Finally, we compared our approach to the quality filtering of Tseng et al. (2001). They filter so-called low quality genes, based on genewise coefficient of variations. A gene whose coefficient of variation (on the raw scale) is too large is removed from the dataset. We used their software, which can be downloaded at `http://biosun1.harvard.edu/~tseng/download.html`. In their method, the user has to decide on a threshold used to filter the genes, which corresponds roughly to the proportion of genes filtered out. Tseng et al. (2001) recommended using a 90% threshold, i.e. removing about 10% of the data. We used this threshold and it seemed to be too aggressive, for example filtering out all the HIV genes. We then used a 1% threshold, which is closer to the proportion of genes removed by our gene filtering method, but it still removed three HIV genes in the HIV 1 data and four in the HIV 2 data, while our filtering method removed only one gene from the HIV 1 data and none from the HIV 2 data. It thus seems that for this dataset, our gene filtering methods performed better. In their paper, Tseng et al. (2001) also recommend looking for replicate outliers when a gene fails to pass the quality filtering. However, they do not mention how to decide if a replicate is an outlier. This is a crucial point especially when the number of replicates is small.

4.4 Should the Background be Subtracted?

The image from which individual gene expression levels are estimated takes the form of roughly circular spots superimposed on a background. Expression levels are estimated by measuring the average intensity in the spot. The measured intensity in the background is often greater than zero for various physical reasons, including fluorescence of the glass substrate, amplifier offset, dark current, and so on. Thus the estimate of the intensity in a spot is often modified by subtracting the estimated background intensity. In general, background intensities vary spatially within a slide, and so it is common to estimate the background for each spot separately (e.g. Yang et al., 2002a).

However, some authors have pointed out that subtracting the background can have the negative effect of increasing the variability, especially at low intensities (Rocke and Durbin 2001; Cui et al. 2002; Glasbey and Ghazal 2003). A counterargument to this is that by not subtracting the background one increases all the intensity measurements, and so one tends to reduce the estimates of ratios that are large, thus biasing the ratio estimates of differentially expressed genes downwards. This debate about whether or not to subtract the background remains unresolved.

A comparison of standard unshifted ANOVA log ratios with and without background subtraction (Figures 5 and 6) shows that the arguments on both sides of the debate are correct for our data. Figure 5(a) shows the estimates with background subtraction, and the high variance at low intensities is clear; using the two-fold change rule of thumb, this seems likely to lead to a considerable number of false positive assessments of differential expression. Figure 6 shows the same plot, but without background subtraction. The variance is indeed considerably reduced. But this comes at a high price in terms of bias. For the 13 genes known to be differentially expressed, the median log ratio is 6.3 with background subtraction, and 3.3 without. Such a level of bias could lead us to miss genes that are moderately differentially expressed, and indeed one of our differentially expressed genes now falls below the two-fold threshold when the background is not subtracted. On the other hand, four of the 29 genes known not to be differentially expressed exceed the threshold when the background is subtracted, but none do so when the background is not subtracted.

Inspection of the posterior means in Figure 5 suggests that our method allows one to have the best of both worlds: one can subtract the background without paying such a high price in

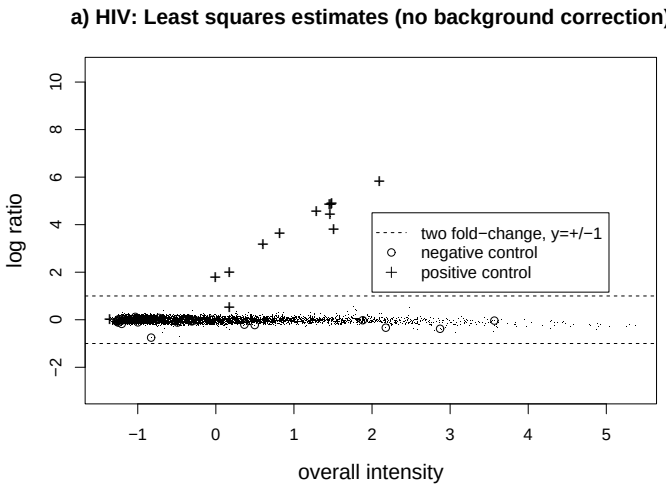

Figure 6: ANOVA Normalized Log Ratio Estimates Without Background Subtraction for the HIV Data. Not subtracting the background shrinks the estimates towards the x-axis. One of the genes known to be differentially expressed shows a less than two-fold change whereas it does not with background-subtracted data.

terms of variance. After background subtraction, the variance at low intensities is much less than with the standard non-robust method, and the median log ratio for the 13 differentially expressed genes is 5.6, much closer to the non-robust estimates with background subtraction than without. With our method, the same known differentially expressed genes as with the ANOVA log ratios without shift exceed the two-fold threshold, while none of the 29 genes known not to be differentially expressed do so.

5 DISCUSSION

We have developed a Bayesian hierarchical model for estimating cDNA microarray intensities in a way that is robust to outlying measurements caused by things such as scratches, dust, imperfections in the glass and imperfections in the array production. The robustness is achieved by using a hierarchical t-distribution and allowing the data to choose the number of degrees of freedom. Our model borrows strength from all the genes when deciding if a replicate measurement is an outlier. This is essential: it is hard to detect outliers based only on the four measurements for a single gene. Classical robust estimators, such as M-estimators, would be inefficient with a small number of replicates. For example a trimmed mean with four replicates would remove at least two observations and the estimate would

then be based on two replicates. Our model works well with four replicates, thanks to the borrowing of strength. Our model also deals with the classical issues of design effects, normalization, transformation and non-constant variance. Our framework requires more computing than some other methods because it involves MCMC, and users would need to decide whether the improved results are worth the additional computing time. However, MCMC methods allow us to sample from the full posterior distribution of all the parameters and to focus on any quantity of interest. For example, in addition to point estimates, one could compute measures of uncertainty such as log ratio posterior standard deviations, or even the joint posterior distribution of the ranks of the log ratios, etc.

We have specified our model on the scale of log transformed intensities. Durbin et al. (2002) and Huber et al. (2002) independently proposed a transformation that stabilizes the variance very well. However, this transformation is somewhat complex, and Rocke and Durbin (2003) have shown that the shifted log transformation provides a good approximation while keeping the ease of interpretation of log ratios.

We estimate the dye bias assuming that most of the genes are not differentially expressed. This assumption is usually at least approximately correct, and without it one cannot distinguish a poor RNA preparation from differential expression. If one does not accept this assumption, the term α, the dye effect, should be removed from the model. If we still want to estimate the sample bias, a technique similar to that of Tseng et al. (2001), where a group of genes that are believed not to be differentially expressed is selected, could be used.

In our application we have considered only simple designs. However, our model could easily be modified to take account of other designs such as those proposed by Kerr and Churchill (2001) and Dobbin et al. (2003). For example, our model could be extended to the loop design introduced by Kerr and Churchill (2001). Note that a loop design with only two samples (or varieties) is just a dye swap experiment, and so a more general loop design is a natural extension of the present work.

In our comparison with the ANOVA method, we assumed that the variance was constant, which is commonly done in practice (Kerr et al. 2000; Churchill 2002; Chen et al. 2003). However, as pointed out by Kerr and Churchill (2001), if there were evidence for heteroscedasticity, it would be possible to use weighted least squares to fit gene specific variance models. In our comparison, the datasets used have four replicates, and gene-specific

variance estimates would typically be noisy. We have tried fitting ANOVA models with gene specific variances and it did not improve things. One alternative would be to use the shrunken variance estimates of Cui et al. (2003), which would be closer to our exchangeable prior for the gene precisions.

In our model we assumed that the gene effects arise from a common Gaussian distribution. It seems to work well with the dataset explored in this paper, but this could be modified to allow more flexibility. For example one could easily use a t distribution if the gene effects are heavier tailed. If parametric assumptions are still to restrictive one could use a non-parametric prior as in Newton et al. (2004), or even model the effects as fixed.

References

Besag, J. E. and D. M. Higdon (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society, Series B 61*, 691–746.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B 26*, 211–252.

Box, G. E. P. and D. R. Cox (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association 77*, 209–210.

Brown, C. S., P. C. Goodwin, and P. K. Sorger (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Science 98*, 8944–8949.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*, 957–970.

Chen, J. J., R. R. Delongchamp, C. Tsai, H. Hsueh, F. Sistare, K. L. Thompson, V. G. Desai, and J. C. Fuscoe (2003). Analysis of variance components in gene expression data. *bioinformatics 20*(9), 1436–1446.

Chen, Y., E. R. Dougherty, and M. L. Bittner (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics 2*, 364–374.

Churchill, G. A. (2002). Fundamentals of experimental design for cdna microarrays. *Nature Genetics 32*, 490–495.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Assiciation 74*, 829–836.

Cui, X., J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill (2003). Improved statistical tests for differential gene expression by skrinking variance components. *Statistical Applications in Genetics and Molecular Biology 2*(1).

Cui, X., M. K. Kerr, and G. A. Churchill (2002). Data transformations for cDNA microarray data. Technical report, The Jackson Laboratory.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics 21*, 488–506.

Dobbin, K., J. H. Shih, and R. Simon (2003). Questions and answers for indentifying differentially expressed genes. *Journal of the National Cancer Institute 95*, 1362–1369.

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

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

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Glasbey, C. A. and P. Ghazal (2003). Combinatorial image analysis of DNA microarray features. *Bioinformatics 19*, 194–203.

Gottardo, R., J. A. Pannucci, C. R. Kuske, and T. Brettin (2003). Statistical analysis of microarray data: A Bayesian approach. *Biostatistics 4*, 597–620.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*, S96–S104.

Ideker, T., V. Thorsson, A. F. Siegel, and L. E. Hood (2000). Testing for differentially

expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology 7*, 805–817.

Kerr, M. K. and G. A. Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics 2*, 183–201.

Kerr, M. K., M. Martin, and G. Churchill (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology 7*, 819–837.

Lewin, A., S. Richardson, C. Marshall, A. Glazier, and T. Aitman (2003). Bayesian modelling of differential gene expression. Technical report, Imperial College, London.

Li, C. and W. H. Wong (2000). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science 98*, 31–36.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B 34*, 1–41.

Lönnstedt, I. and T. P. Speed (2002). Replicated microarray data. *Statistica Sinica 12*, 31–46.

Neal, R. (2003). Slice sampling. *The Annals of Statistics 31*, 705–767.

Newton, M. A., C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology 8*, 37–52.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*, 155–176.

Raftery, A. E. and S. M. Lewis (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (edited by J. M. Bernardo et al.), pp. 763–773. Oxford: Oxford University Press.

Raftery, A. E. and S. M. Lewis (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice.* (edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), Chapter 7, pp. 115–130. London: Chapman and Hall.

Rocke, D. M. and B. Durbin (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology 8*, 557–569.

Rocke, D. M. and B. Durbin (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics 19*, 966–972.

Schena, M., D. Shalon, R. W. Davis, and P. Brown (1995, Oct). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science 270*, 467–470.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 795–809.

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

Tukey, J. W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics 28*, 602–632.

van't Wout, A. B., G. K. Lehrma, S. A. Mikheeva, G. C. O'Keeffe, M. G. Katze, R. E. Bumgarner, G. K. Geiss, and J. I. Mullins (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of $\mathrm{CD4}^+ - \mathrm{T} - \mathrm{Cell}$ lines. *Journal of Virology 77*, 1392–1402.

Vines, S. K., W. R. Gilks, and P. Wild (1996). Fitting bayesian multiple random effects models. *Statistics and Computing 6*, 337–346.

Yang, G. P., D. T. Ross, W. W. Kuang, P. O. Brown, and R. J. Weigel (1999). Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acid Research 27*, 1517–1523.

Yang, Y. H., M. J. Buckley, S. Dudoit, and T. P. Speed (2002a). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics 11*(1), 108–136.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002b). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30*, e15.