Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 20

Bayesian Hierarchical Model for Correcting Signal Saturation in Microarrays Using Pixel Intensities

Rashi Gupta* Petri Auvinen†

Andrew Thomas‡ Elja Arjas**

*Department of Mathematics and Statistics, P.O. Box 68 and Institute of Biotechnology, P.O. Box 56, University of Helsinki, FIN-00014, Helsinki, Finland, Rashi.Gupta@helsinki.fi

†Institute of Biotechnology, University of Helsinki, P.O. Box 56, FIN-00014, Helsinki, Finland, petri.auvinen@helsinki.fi

‡Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FIN-00014, Helsinki, Finland, andrew.thomas@rni.helsinki.fi

**Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FIN-00014, Helsinki, Finland, elja.arjas@helsinki.fi

Bayesian Hierarchical Model for Correcting Signal Saturation in Microarrays Using Pixel Intensities*

Rashi Gupta, Petri Auvinen, Andrew Thomas, and Elja Arjas

Abstract

Pixel saturation occurs when the pixel intensity exceeds the scanner upper threshold of detection and the recorded pixel intensity is then truncated at the threshold. Truncation of the pixel intensity causes the estimates of gene expression (i.e., intensity) to be biased. Microarray experiments are commonly affected by saturated pixels; as a result all higher level analyses are made on these biased gene expression estimates. In this paper, we propose a method for improving the quality of the signal for cDNA microarrays by making use of several scans at varying scanner sensitivities. For each spot, pixel level intensity readings are given as input to a Bayesian hierarchical model. The model uses the pixel intensities of the spot to provide a posterior distribution of the true expression level of the corresponding genes. The parameters of the hierarchical model are estimated jointly with these expression levels, thus performing an integrated analysis of the measurement data. The method improves in all ranges the accuracy with which intensities can be estimated and extends the dynamic range of measured gene expression at the high end. The method is generic and can be applied to data from any organism and for imaging with any scanner. Results from a real data set illustrate an improved precision in the estimation of the expression of genes compared to what can be achieved by applying standard methods and using only a single scan.

KEYWORDS: microarrays, gene expression, pixel censoring, Bayesian hierarchical model

---

*This work has been supported by the ComBi graduate school (RG), the Academy of Finland via the SYSBIO (207528) and POPGEN (211489) Research Programme (EA, AT), and the Institute of Biotechnology (PA). Contact: rashi.gupta@helsinki.fi

1 Introduction

Microarrays enable us to measure the expression pattern of a large number of genes (Schena *et al.,* 1995). These measurements provide insight into the functions of genes as well as their interactions. But before the data can be used to make inferences, they need to be carefully pre-processed. Pre-processing is done to account for the many sources of systematic and random variation involved in the microarray experiment. Even after pre-processing, the data are generally noisy and inaccurate, thus affecting the high-level analysis such as class prediction, class comparison and clustering. Though much research has been done on correcting the biases in the data (for instance, see Dudoit *et al.,* 2001, Yang *et al.,* 2001, Finkelstein *et al.,* 2002, , Yang *et al.,* 2002), little attention has been paid so far on improving the data quality by taking precautions at the time of data acquisition.

For dual-label cDNA microarrays, two RNA samples are labeled with two different fluorescent dyes (e.g. Cy3 and Cy5), after which they are mixed and allowed to hybridize on the array platform that typically contains thousands of gene probes. The degree of hybridization of each of the labeled samples with respect to gene probes is assessed by laser scanning of the hybridized array. This results in an image that is stored as pixel-level data, usually representing millions of fluorescence intensity measurements. The pixel intensities are determined by the image processing algorithm and then the mean or median of these pixel intensities corresponding to a spot is summarized as spot intensity. This is done for each channel. If one or more of the pixel measurements corresponding to a spot are saturated, the resulting spot summary will be biased.

The accuracy with which the spot intensities are summarized using the pixel intensities is affected by errors that may occur at the time of slide preparation, sample preparation or during scanning. However, our focus here is only on the errors that occur during data acquisition, and therefore assumes that the errors in wet-lab have been minimized. The range of gene expression that is measured on a single microarray is large and scanner parameters (photomultiplier tube, laser power) are manually adjusted to capture this wide range of gene expression on the array. Higher scanner sensitivities are desired to improve the signal-to-noise ratio of the low-intensity spots, but this generally leads to a saturation of the high-intensity spots.

Previous works on correcting signal censoring include Dudley *et al.,* (2002) who proposed to hybridize the experimental and control samples against labeled oligos that would be complementary with respect to every microarray feature, rather than co-hybridizing the samples. However, their method can-

not be applied to experiments that follow the standard method of Schena *et al.,* (1995). Wit and McClure (2003) used spot summaries but made certain parametric assumptions that were difficult to verify in practice. Dodd *et al.,* (2004) demonstrated a method to tackle signal censoring but the method requires data from a dual channel and is therefore not suitable to single channel arrays like Affymetrix GeneChips. Nava *et al.,* (2004) and Khondoker *et al.,* (2005) proposed models for correcting signal censoring using the maximum likelihood approach for parameter estimation. Ekstrom *et al.,* (2004) on the other hand proposed models using spatial statistics of the spot to infer information about the saturated pixel intensities. In our earlier work, Gupta *et al.,* (2006) we also made an attempt to correct signal saturation and errors at the time of data acquisition by making multiple scans of the array and connecting spot summaries obtained from multiple scans via a link function to corresponding latent variables representing unsaturated signals. Bayesian inferential techniques were then applied to the estimation of these latent variables. Our modeling tackled saturation and was also able to improve the gene signal in all ranges. Better results using spot summary data encouraged us to seek for a deeper understanding of the problem by using pixel intensities.

In this paper, we focus on the statistical methods for correcting gene expression measurements obtained from cDNA microarrays in the presence of signal saturation using pixel intensities. As in Gupta *et al.,* (2006), the data from three scans are modeled within the Bayesian framework using a latent intensity model. Use of pixel intensities instead of their summary is intuitively appealing as it allows for a direct modeling of the saturation phenomenon, and has therefore the potential of further improving the accuracy of the results.

2 Material and Methods

2.1 Data

The DNA microarrays used for this study were produced by the Turku Centre for Biotechnology, University of Turku, Finland and contained 16000 human cDNAs spotted. HeLa cells were used for conducting these experiments which contained constructs expressing RhoG mutants (Rho-GTP and Rho-GDP). The aim of the study was to compare the effect of these RhoG mutants on the expression of genes in HeLa cells. Three time points were considered in the initial study, of which we considered only one time point for our study. The experiment was performed on two arrays (here named as A and B). Details about RNA extraction, probe labeling and microarray hybridization can be

found in Gupta *et al.,* (2006).

Array	**Scanner settings used for Cy3 (Cy5)**			
	PMT Gain	**Scan-1 (LP)**	**Scan-2 (LP)**	**Scan-3 (LP)**
A	98 (85)	100 (100)	90 (90)	80 (80)
B	85 (74)	100 (98)	90 (88)	80 (78)

Table 1: The combination of PMT and laser power (LP) used to obtain multiple scans for array-A and array-B are summarized for Cy3 and for Cy5 (in parenthesis).

2.2 Scanning procedure

The slides were scanned with ScanArray 5000 (GSI Lumonics) using an appropriate setting for samples labeled with Cy3 and Cy5. The line scan function of the scanner was used to equalize the intensities from Cy3 and Cy5 channels. A setting with the highest laser power was used to make the first scan (scan-1). Under this setting, due to high laser power, saturation occurred in spots with high RNA abundance, but spots with a low RNA abundance were captured accurately. This setting was chosen so that the background corrected intensity ($foreground - background$) from the spots with a low RNA abundance was at least 200 (200 is the mean background intensity and thus chosen as a threshold). Subsequent scans were performed with a lower laser power (reducing it by 10 units in each subsequent scan) for both channels but keeping PMT to be the same as in the first setting. With this lower laser power the number of spots with a saturated intensity decreases, whereas the number of spots with a low intensity (or below our threshold of 200 units) increases. Repeated scanning does not significantly damage the fluorescence (data not shown), nor does lowering the laser power affect the balance between the channels (see Table 1, Figure 1).

2.3 Quantification of spot intensities

The digital images were processed in Matlab to find spot boundaries and to calculate pixel intensities. Spot addressing was done using a grid and the same grid was used for all three scans of the array. Extraction of intensities was done on the merged image in order to associate the same pixels from the three different scans while analyzing the pixel intensities.

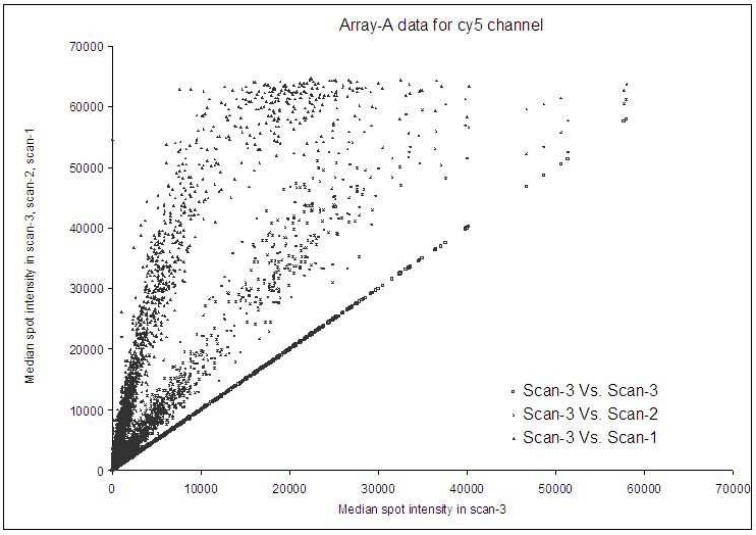

Figure 1: Plot shows multiple scans of array-A for data from Cy3 channel. The median spot intensities from scan-3, scan-2, and scan-1 are plotted against scan-3. Censoring at the upper end of the intensities can be clearly seen. Very similar behavior was seen for the data from Cy5 dye.

2.4 Latent variable model with multiple scans

Suppose that there are g genes on the array. Let Y_{ijs} denote the pixel intensity for the j^{th} pixel of the i^{th} spot observed under scanner setting s, where $i = 1, \cdots, g$ (genes), $j = 1, \cdots, p$ (pixels), and $s = 1, 2, 3$ (settings). Our model aims at estimating the *true underlying expression* of genes, using the three sets of measurements. Therefore we assume, for each gene i, the existence of a corresponding latent random variable T_i; $i = 1, \cdots, g$. Also, corresponding to the three scanner settings, we assume for each gene i and under each setting s, the existence of a corresponding latent random variable Q_{is}, $i = 1, \cdots, g$, $s = 1, 2, 3$. In order to combine the latent intensities under different settings (Q_{is}) and the true latent intensities (T_i), we use a simple calibration and treat the latent intensity under scan-1 as the true latent variable to be estimated, that is, set $T_i = Q_{i1}$ for all i. We assume that the latent intensities under settings $s = 2$ and $s = 3$ are linked to T_i by simple functional relationships, say $Q_{i2} = f_2(T_i)$ and $Q_{i3} = f_3(T_i)$. To use this notation also for $s = 1$, we let $f_1 =$ identity function. For computational convenience, we divide the full range of intensity into l intervals $I_1, I_2, \cdots, I_l$ and within each interval we

assume a simple multiplicative form for f_2 and f_3. In other words, we set

$$\begin{array}{rcll} Q_{i2} & = & b_k T_i, & \\ Q_{i3} & = & b_k d_k T_i, & \text{for } T \in I_k; k = 1, 2, \cdots, l; i = 1, 2, \cdots, N. \end{array} \quad (1)$$

where the parameters $(b_k, d_k), k = 1, 2, \cdots, l$ are restricted to have values in the interval (0,1). As a result, the latent intensities corresponding to the three scans are ordered as $T_i = Q_{i1} \geq Q_{i2} \geq Q_{i3}$. In the hypothetical case in which there were no measurement errors, we could write the observed intensity, for given i, j and s, as $Y_{ijs} = f_s(T_i)$, whenever this intensity is below the value 65535 (where 65535 value is the scanner's upper threshold of detection). Realistically however, extraction of intensities for genes from a scanned array always involves some measurement errors. Having explored a few possibilities, we arrived at postulating the relationship between the true and the observed pixel intensities in terms of a hierarchical model involving both an additive and a multiplicative component:

$$\begin{array}{rcll} Y_{ijs} & = & f_s(T_i)\ e^{\epsilon_{is}} + \eta_{ijs} & \text{if } f_s(T_i)\ e^{\epsilon_{is}} + \eta_{ijs} < 65535, \\ Y_{ijs} & = & 65535 & \text{if } f_s(T_i)\ e^{\epsilon_{is}} + \eta_{ijs} \geq 65535, \end{array} \quad (2)$$

where ϵ_{is} is the error associated with the i^{th} spot under setting s and η_{ijs} is the error associated with the j^{th} pixel of the i^{th} spot under setting s. We assume that η_s and log-transformed ϵ_s (i.e. $log\epsilon_s$) are a priori $i.i.d$ and Normally distributed with mean 0 and respective variances $\sigma^2_{\eta_{sk}}$ and $\sigma^2_{\epsilon_{sk}}$, $s = 1, 2, 3$, $k = 1, \cdots, l$. Denoting $\mu_{is} = exp\big(log[f_s(T_i)] + \epsilon_{is}\big)$ and using the definition of ϵ_s and η_s , (2) can be rewritten as

$$[Y_{ijs}|\mu_{is}, \sigma^2_{\eta_{sk}}] = N(\mu_{is}, \sigma^2_{\eta_{sk}}) \quad (3)$$

where μ_{is} is the mean of spot i under scan s and

$$[log(\mu_{is})|Q_{is}, \sigma^2_{\epsilon_{sk}}] = N(log(Q_{is}), \sigma^2_{\epsilon_{sk}}) \quad (4)$$

2.5 Joint probability distribution

Our data set comprises of pixel intensities corresponding to spots from scans attained at varying scanner sensitivities. A proportion of pixels associated with spots with a high signal can be saturated in any of the three scans. Let p_{is} represent the total number of pixels forming spot i under scan s (= 69 in our case) and let c_{is} denote the number of saturated pixels within spot i under

scan s. The likelihood contribution from the saturated and unsaturated pixels of spot i under scan s is:

$$\phi\left(\frac{(Y_{ijs})-\mu_{is}}{\sigma_{\eta_{sk}}}\right)^{p_{is}-c_{is}}\left[1-\Phi\left(\frac{(threshold)-\mu_{is}}{\sigma_{\eta_{sk}}}\right)\right]^{c_{is}} \tag{5}$$

where ϕ is the standard Normal density, Φ denotes the corresponding cumulative distribution function and $threshold = 65535$. Our primary interest is in estimating the true underlying latent intensities T_i jointly with the intensities Q_{i2}, Q_{i3}, $i = 1, 2, \cdots, N$ and the model parameters b_k, d_k, $\sigma^2_{\epsilon_{sk}}$, $\sigma^2_{\eta_{sk}}$, $s = 1, 2, 3$, $k = 1, 2, \cdots, l$. Let Θ denote all the variables that need to be estimated. Then the joint distribution of $\Theta = (T_i, Q_{i2}, Q_{i3}, b_k, d_k, \sigma^2_{\epsilon_{sk}}, \sigma^2_{\eta_{sk}}$, $i = 1, 2, \cdots g$ (spots), $k = 1, 2, \cdots, l$ (intervals), $s = 1, 2, 3$ (scans)) given the data $Y = (Y_{ijs}, i = 1, 2, \cdots g, j = 1, 2, \cdots, p$ (pixels), $s = 1, 2, 3)$ is obtained via Bayes' rule

$$p(\Theta \mid Y) \propto p(\Theta)p(Y \mid \Theta). \tag{6}$$

To complete the specifications of the model, we assign here to T_i the Uniform prior distribution over the interval (200, 1000000). The standard-deviation of the log-transformed ϵ_{sk} errors under three settings and over l intervals ($\sigma_{\epsilon_{sk}}$; $s = 1, 2, 3$; $k = 1, 2, \cdots, l$) were all assigned the Uniform prior over (0,2). The precision (inverse of variance) of η_{sk} errors under the three settings and over l intervals ($\sigma_{\eta_{sk}}$; $s = 1, 2, 3$; $k = 1, 2, \cdots, l$) were assigned vague Gamma priors with parameters (0.001, 0.001).

2.6 Sampling algorithm

The numerical computations were done using the Gibbs sampling Markov chain Monte Carlo (MCMC) algorithm. If a pixel is saturated, the integration implicit in the cumulative Gaussian function in (5) can be easily carried out via a data augmentation technique by imputing a value for Y_{ijs} from the tail of the Normal distribution with mean μ_{is} and variance $\sigma^2_{\eta_{sk}}$. Further, by restricting the maximum value attained by T_i (via the prior), we restrict the maximum imputed value for the saturated pixels. Each variable is sampled in turn from its conditional distribution, holding all other variables fixed. These conditional distributions are derived from the joint distribution by extracting those factors that depend on the variable to be sampled and multiplying them together. Our algorithm can be summarized in to the following steps:
Step 1: Specify initial values of μ_{is}, $\sigma_{\epsilon_{sk}}$, $\sigma_{\eta_{sk}}$, b_k, d_k, T_i;
Step 2: Sample values for the censored pixels Y_{ijs};
Step 3: Sample values for μ_{is};

Step 4: Sample the T_i, from their conditional distributions;
Step 5: Sample $\sigma_{\epsilon_{sk}}$ from its conditional distribution;
Step 6: Sample $\sigma_{\eta_{sk}}$ from its conditional distribution;
Step 7: Sample b_k, d_k from their conditional distribution;
Step 8: Repeat step 2 to step 8 till convergence is achieved.
The model was formulated in the BUGS language (see Appendix) and parameter estimation was performed using WinBUGS (Spiegelhalter *et al.*,1999). The current model runs in OpenBugs on an Intel Pentium processor 2.80 GHz with 1 GB RAM and takes approximately 10 hours to do 400000 iterations using two chains in parallel. Two chains of 100000 iterations each were generated to check convergence of the parameters under consideration. Thereafter a sample of size 100000 was generated and every 10^{th} iteration was used to make inference.

3 Results

Our approach was tested on several real data sets of varying dimensions. For a demonstration purpose we considered 4814 out of 16000 spots from the data set described in section 2.1, where the RNA from Rho-GTP was labeled with Cy5 and the control was labeled with Cy3. Each spot contained 69 pixels and was measured thrice at varying scanner sensitivities, thus leading to a total of 996498 pixel intensities to be modeled. (Each spot was composed of approximately 100 pixels; however, we considered only the central 69 pixels per spot and excluded the pixels near the spot boundaries.)

The hierarchical model presented in (2) involves error variances at both spot and pixel levels as well as parameters b and d, none of which can be assumed to have a constant value across the entire intensity range. Therefore, the intensity range here was divided into four intervals: I_1 = [0, 2000), I_2 = [2000, 5000), I_3 = [5000, 11000), I_4 = [11000, -). These breakpoints were selected using visual inspection (see Figure 1). Another possibility would be to estimate these breakpoints jointly with the model parameters from a Bayesian model, but this was not done here because of the additional computational burden that would result. The very simple linear form for the functions f_2 and f_3 on intervals was chosen for the same reason.

The Uniform prior distribution for T over [200, 1000000] allows one to sample even very large values of T. Pixels belonging to spots with extremely high signals could be saturated even in scans made at the lowest sensitivity level (scan-3). Such saturated pixel intensities are taken care of by sampling values corresponding to them from the tail of a Normal distribution. Imputed

values for the censored pixels along with the measurements taken from the uncensored pixels from all three scans help in estimating the latent variables T_i. The posterior distributions of the estimated T_i for two genes with very different expression levels are shown in Figure 2. Figure 2a illustrates a situation in which there is no saturation in scan-1 measurements, whereas for the gene considered in Figure 2b, the signal from scan-1 is completely saturated, and also a fraction of the pixel measured from scan-2.

Also the benefit of using data from three scans, compared to using data from only 2 scans (i.e. scan-1 and scan-2), is illustrated in Figure 2. It can be seen that the spread of the posterior distribution using data from all three scans is less than that obtained when using data from two scans (here scan-1 and scan-2). This tendency persisted for a large number of genes which we considered, and naturally corresponds to our expectations as long as the three measurements do not provide conflicting information.

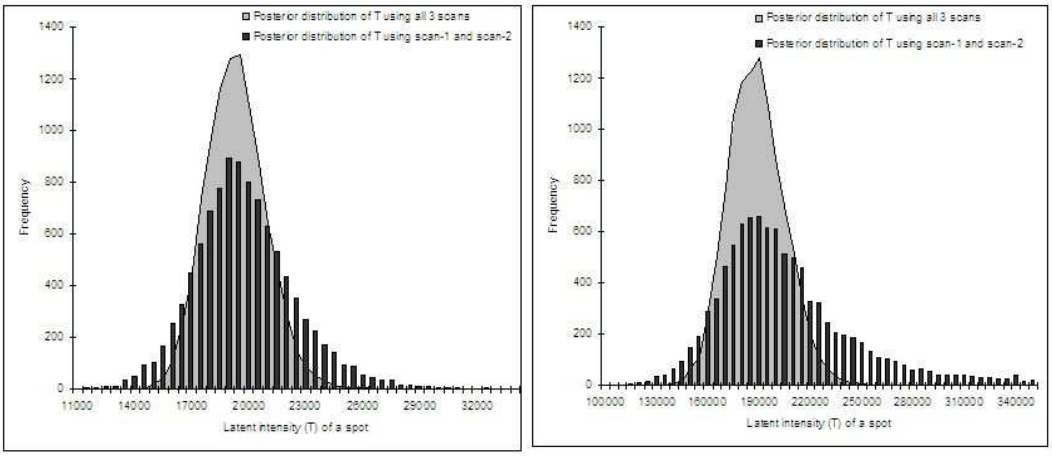

Figure 2: Posterior distribution of the latent variable T for two different genes, representing different intensity ranges in scan-1 in array-A, obtained using all three scans (shown by grey area) and using only two scans (scan-1 and scan-2, shown in black bars). The observations (median) from (scan-1, scan-2, scan-3) for the two genes are: (a) (19917, 7778, 2657), (b) (65000, 51723, 25372). Different representations have been used to enhance the visibility.

The robustness of our model could be easily tested by comparing the estimated latent intensity for replicated spots, both within an array and across different arrays. The posterior distribution of the latent intensity, when estimated for some genes that were replicated within an array, demonstrated good consistency between the measured signals (Figure 3a). No normalization was

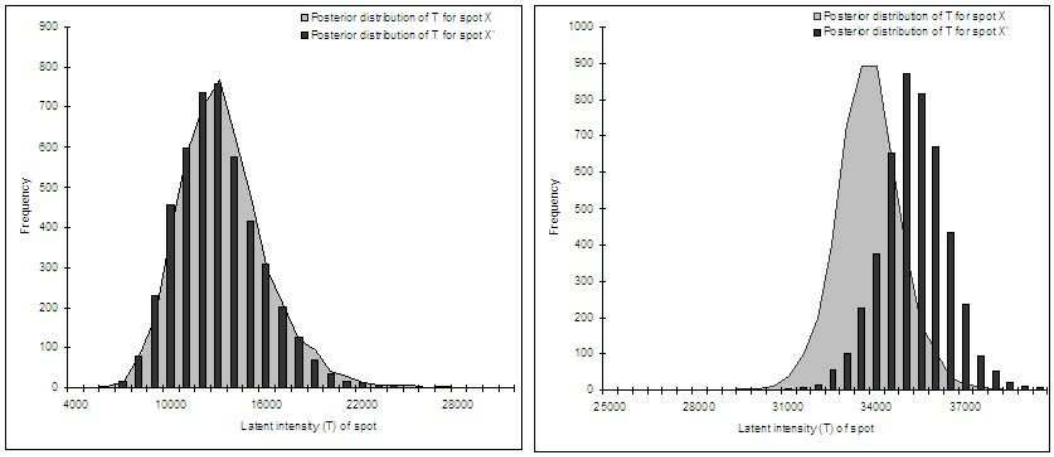

Figure 3: Posterior distribution of true latent intensity for replicated genes (a) on the same array-A (b) on different arrays (array-A and array B). Different representations have been used to plot the posterior distribution of T for the two genes to enhance visibility.

applied for producing these plots. However, for spots replicated on different arrays, there was sometimes only little overlap between the latent intensity distributions, thereby implying a need of some form of normalization (Figure 3b). Again this was in line with what we had expected, being a consequence of an "array effect". To check the correction to the signal made using our model, we plotted for 250 randomly chosen genes the estimated posterior median of the latent variables using signals from all three scans, and their corresponding measured scan-1 signal (Figure 4). The percentage of saturated pixels from scan-1 is plotted on the secondary Y-axis and the spots are arranged in an ascending order relative to the saturation percentage. It can be seen that for low intensity spots which contain less than 25% saturated pixels, the estimated posterior medians of $T's$ are close to the corresponding measured scan-1 signals. However, there is a clear increase in the estimated signal for spots which have more than 25% saturated pixels.

Figure 4 also compares the performance of the current model utilizing pixel intensities to our earlier work (Gupta *et al.,* (2006)) utilizing spot summaries (i.e. mean or median values obtained using GenePix software (Axon Instruments, Inc.)). In addition to displaying the estimated posterior median of the latent variables of spots using pixel intensities (current model), the measured scan-1 signal and the percentage of saturation in scan-1, the figure also shows the estimated posterior medians of the latent variables of spots using

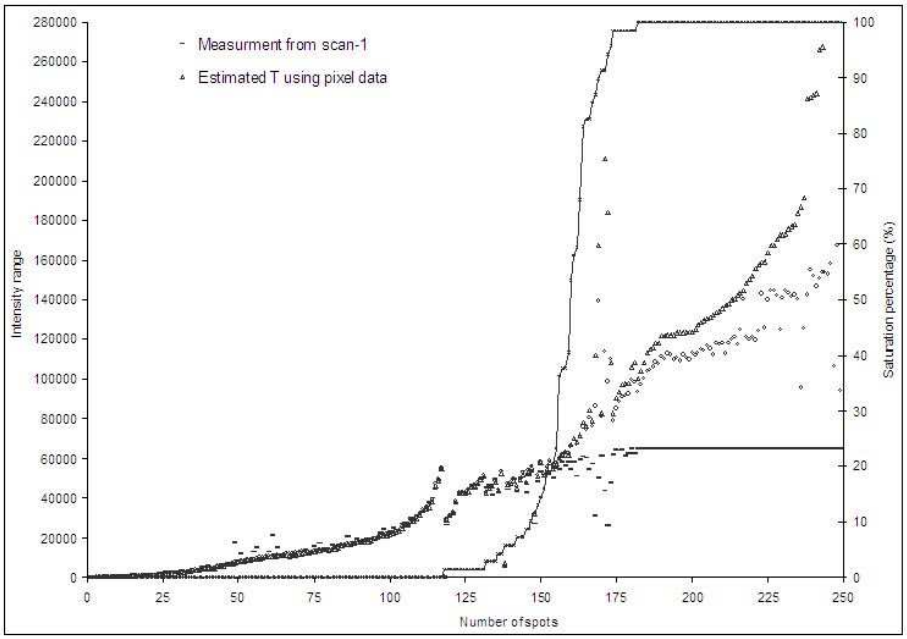

Figure 4: The figure presents two sets of comparisons. For 250 spots, the measured signal from scan-1 along with their percentage of saturation in scan-1 and the estimated (posterior median) intensity using pixel intensities from the three scans can be seen. In the same plot the estimated latent intensities obtained using spot summaries (applying the method of Gupta et al., (2006)) from three scans can also be seen. The percentage of saturation in scan-1 for these 250 spots is plotted on the secondary Y axis. The data is from Cy3 dye and intensities are sorted in ascending order with respect to the saturation percentage.

spot intensities (applying the method of Gupta *et al.,* (2006)). It can be seen from the figure that the corrected signal of spot (T_i) obtained by using spot summary information is generally lower than when estimated using the pixel intensities. The reason is that, when using the spot summaries, we cannot explain the saturation process precisely: saturation does not happen to a spot, but to the pixels belonging to it. When the percentage of saturated pixels is less than 15-20%, both models propose similar values for the latent variable T. However, the difference between the two models becomes obvious when considering spots containing 25% saturation or more. In this case, the latent T estimated using pixel information from 250 spots is much higher than when using the corresponding spot summaries.

When scan-1 is completely saturated, T is effectively estimated using the data from scan-2 and scan-3. When dealing with pixel data, it can be observed that not only the scans at higher sensitivities but also the scans made at low sensitivity can have saturated pixels. In that case, the inference regarding the corresponding T is made by borrowing strength from the neighboring unsaturated pixels (roughly in the same expression range). In Figure 5, we have plotted 65 spots with 100% saturation in scan-1 along with their saturation percentages in scan-2 and scan-3, and also the estimated signal (posterior median) for the latent variable T. It can be seen that when scan-1 has 100% saturated pixels, the profile formed by the estimated median values of T follows the same pattern as the measurements from scan-2 and scan-3.

The posterior median and the standard deviation estimates of log-transformed ϵ_{sk} errors and η_{sk} errors corresponding to the three scans and over the four intervals are displayed in Table 2 and 3 respectively. Small values of the errors over all intervals implies good accuracy in the estimation of the latent variable T_i.

In our work the correction for two dyes was done separately as the hybridization efficiencies of the two dyes are different. The existence of systematic differences between the two dyes has been documented in the literature, and they have been treated separately *i.e.* by Nava *et al.,* (2004), Dodd *et al.,* (2004) and Khondoker *et al.,* (2005) while correcting for signal saturation. A single scan of an array usually involves a lot of variability in the measurements which gets reduced when inferring T from three measurements of the same array. These precise estimates inferred seperately for the two dyes can then be used to calculate the log-ratios. Figure 6 displays the log-ratios calculated using estimated T's from the two channels. We compare it with the log-ratios calculated using the measurements from the two channels corresponding to scan-1 which is usually used to make all higher level analysis and inferences. Since for this data set, the true log-ratios are unknown, it is difficult to conclude anything further. However, intutively using precise estimates to calculate the log-ratios provide better results. Our aim here was not to focus on the differentially expressed genes but to roughly present how they could be evaluated using the estimated T's.

4 Discussion

Saturation has been a prevailing problem in microarray data and it appears that no fully satisfactory ways to treat saturation, and thereby to improve the

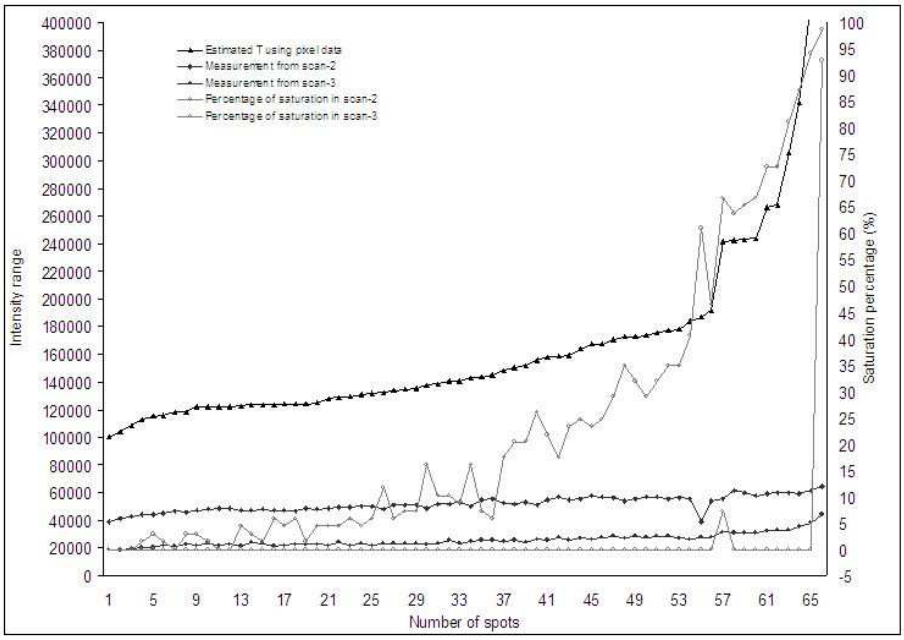

Figure 5: The figure displays scan-2 and scan-3 measurements from 65 randomly selected genes with 100 % saturated pixels in scan-1 along with their percentage of saturation in scan-2 and scan-3 (on the secondary Y axis). Posterior medians for the latent variable T corresponding to these 65 spots are connected by a dotted line. The pattern observed is similar to the pattern depicted in scan-2 and scan-3 measurements but, as can be expected, the estimates increase with the increasing percentage of saturation in scan-2.

overall signal of genes, have been developed so far. Possible approaches have been proposed for instance by Dundley *et al.,* (2002), Dodd *et al.,* (2004), Nava *et al.,* (2004) and Khondoker *et al.,* (2005). However, to our knowledge no one has used the pixel intensities, apparently as it was considered computationally too cumbersome.

In this work we have proposed a way to estimate the true signal corresponding to the two channels separately, using the Bayesian framework. Several authors have already suggested Bayesian methods for analyzing microarray data, see e.g. Baldi *et al.*(2001); Dror *et al.*(2002); Ibrahim *et al.*(2002); Parmigiani *et al.*(2002); Ramoni *et al.*(2003); Bhattacharjee *et al.*(2004); Frigessi *et al.*(2005); Lewin *et al.* (2006). We use hierarchical modeling under the Bayesian paradigm. Once the model is defined, inference follows automatically and the posterior distributions give direct answers, in terms of the conditional

Intensity-range		**Posterior median estimate of standard-deviation of $log\epsilon$ for data from Cy5 dye**		
Lower limit	**Upper limit**	**Scan-1**	**Scan-2**	**Scan-3**
200	2000	0.471±0.018	0.379±0.019	0.425±0.018
2001	5000	0.225±0.032	0.451±0.023	0.346±0.025
5001	11000	0.257±0.020	0.295±0.019	0.487±0.018
11000	-	0.339±0.006	0.027±0.017	0.315±0.006

Table 2: Posterior median estimates of the standard-deviation of $log\epsilon$ in scan-1, scan-2 and scan-3, over the four intervals of data from Cy5 dye. Similar observations were made for Cy3 (data not shown).

Intensity-range		**Posterior median estimate of standard-deviation of η for data from Cy5 dye**		
Lower limit	**Upper limit**	**Scan-1**	**Scan-2**	**Scan-3**
200	2000	2.498±0.005	2.507±0.005	2.453±0.005
2001	5000	1.164±0.004	1.912±0.007	2.231±0.008
5001	11000	0.659±0.002	1.262±0.004	1.806±0.006
11000	-	0.406±0.007	0.661±0.001	0.997±0.001

Table 3: Posterior median estimates of the standard-deviation of η in scan-1, scan-2 and scan-3, over the four intervals of data from Cy5 dye. Similar observations were made for Cy3 (data not shown).

probabilities conditioned on the observed data. Our model allows for the estimation of missing data by sampling randomly from the corresponding posterior predictive distribution, and thereby enables joint estimation of a large number of model parameters along with the latent variables. Although we have treated and corrected signal saturation separately for each dye, this is not the usual practice in microarrays where the two channels are dealt with together to estimate the differentially expressed genes. We are currently working on extending our work towards building an integrated model that aims at both correcting signal saturation and identifying differentially expressed genes. To do this, we need to add one or more layers to the present hierarchical model, to then account for both between and within array variations, as well as dye swap. Our current model was able to handle large data sets with relative ease within a reasonable time frame.

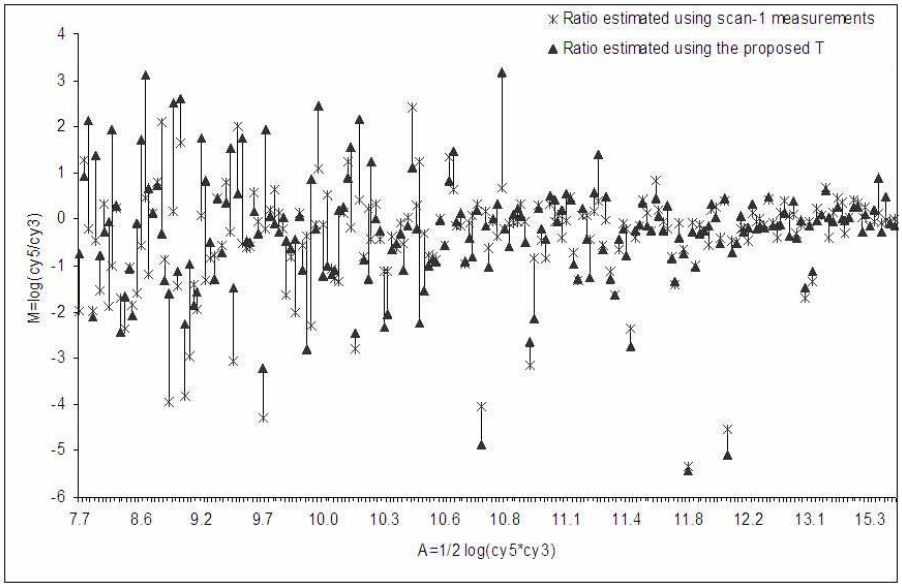

Figure 6: Plot displays the log ratio (M=log (Cy5/Cy3)) for 250 spots calculated using measurements from the two channels corresponding to scan-1 and log ratio calculated using the estimated $T's$ corresponding to the two channels. The spots are arranged in ascending order of A=1/2 log(cy5 * cy3). The differences between the log-ratios calculated using Ts and using scan-1 are both positive and negative. Since the true log-ratios are not known for this data set, it is difficult to conclude which genes are significant. The figure presents for a single array how the estimated Ts shall be used in future to find significant genes.

References

[1] Baldi, P., and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509-519.

[2] Bhattacharjee, M., Pritchard, C.C., Nelson, P.S., and Arjas, E. (2004) Bayesian integrated functional analysis of microarray data. *Bioinformatics,* **20**: 2943-2953.

[3] Dodd, L.E., Korn, E.L., McShane, L.M., Chandramouli, G.V., and Chuang, E.Y. (2004) Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics,* **20** (16): 2685-93.

[4] Dror, R.O., Murnick, J.G., and Rinaldi, N.A.(2002) A Bayesian approach to transcript estimation from gene array data: the BEAM technique. *In RECOMB 2002: Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology.* ACM PRESS.

[5] Dudoit, S., Yang, Y.H., Luu, P., and Speed, T.P. (2001) Normalization for cDNA microarray data. *Proceedings of SPIE*, **4266**, 19.

[6] Dudley, A., Aach, J., Steffen, M., and Church, G. (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl Acad. Sci., USA,* **99,** 7754-7759.

[7] Ekstrom CT, Bak S, Kristensen C and Rudemo M. (2004) Spot shape modeling and data transformations for microarrays. *Bioinformatics.* Sep 22; **20(14)**:2270-8.

[8] Finkelstein, D.B., Gollub, J., and Cherry, J.M. (2002) Normalization and systematic measurement error in cDNA microarray data. *Joint Statistical Meeting.*

[9] Glasbey, C.A. and Khondoker, M.R. (2005) Correction for pixel censoring in cDNA microarrays. In "Statistical Solutions to Modern Problems: *Proceedings of the 20th International Workshop on Statistical Modelling*", eds. A.R. Francis, K.M. Matawie, A. Oshlack, G.K. Smyth, **17-31**.

[10] Gupta, R., Arjas, E., Kulathinal, S., and Auvinen, P. (2006) A latent variable model for estimating gene expression intensities from multiple scanned DNA microarray. *Technical report* (available on request).

[11] Ibrahim, J., Chen, M.H., and Gray, R. (2002) Bayesian models for gene expression with DNA microarray data. *J. Am. Stat. Assoc.*, **97**, 88-99.

[12] Khondoker, M.R., Glasbey, C.A. and Worton, B.J. (2006) Statistical estimation of gene expression using multiple laser scans of microarrays. *Bioinformatics*, **22**, 215-219.

[13] Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2006) Bayesian modeling of differential gene expression. *Biometrics*. **62**: 1-9

[14] Nava, J.G., Hijum, S., and Trelles, O.(2004) Saturation and Quantization Reduction in Microarray Experiments using Two Scans at Different Sensitivities. *Statistical Applications in Genetics and Molecular Biology,* **3,** No. 1, Article 11.

[15] Parmigiani, G., Garrett, E.S., Anbazhagan, R., and Gabrielson, E. (2002) A statistical framework for expression-based molecular classification in cancer. *J. R. Stat. Soc.,* **64**, 717-736.

[16] Ramoni, M.F., and Sebastiani, P. (2003) Bayesian methods for microarray data analysis. IMA Workshop 1: Statistical Methods for Gene Expression: Microarrays and Proteomics, Minneapolis, USA.

[17] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-70.

[18] Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1999) WinBUGS, Version 1.2. User Manual, 1999, MRC Biostatistics Unit.

[19] Wit, E., and McClure, J. (2003) Statistical adjustment of signal censoring in gene expression experiments. *Bioinformatics,* **19,** 1055-1060.

[20] Yang, Y., Buckley, M., Dudoit, S., and Speed, T. (2001) Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Stat.,* **11,** 108-136.

[21] Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., 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 Acid Research*, **30**, E15.

5 Appendix

5.1 Code in WinBugs (without interval specification)

model[
for(i in 1 : N)[
LT[i] ~ dunif(0,15)
LQe1[i] <- LT[i]
LQe2[i] <- LT[i] - a
LQe3[i] <- LT[i] - a - b
for(j in 1 : P)[
Lye1[i , j] ~ dnorm(mue1[i],tau.etae1)
Lye2[i , j] ~ dnorm(mue2[i],tau.etae2)
Lye3[i , j] ~ dnorm(mue3[i],tau.etae3)]
mue1[i] ~ dnorm(LQe1[i],tau.epie1)
mue2[i] ~ dnorm(LQe2[i],tau.epie2)
mue3[i] ~ dnorm(LQe3[i],tau.epie3)]
a ~ dunif(0,5)
b ~ dunif(0,5)
tau.etae3 ~ dgamma(1.0E-5,1.0E-5)
tau.etae2 ~ dgamma(1.0E-5,1.0E-5)
tau.etae1 ~ dgamma(1.0E-5,1.0E-5)
sigma.etae1 <- 1 / sqrt(tau.etae1)
sigma.etae2 <- 1 / sqrt(tau.etae2)
sigma.etae3 <- 1 / sqrt(tau.etae3)
tau.epie3 <- 1 / sqrt(sigma.epie3)
tau.epie2 <- 1 / sqrt(sigma.epie2)
tau.epie1 <- 1 / sqrt(sigma.epie1)
sigma.epie1 ~ dunif(0,2)
sigma.epie2 ~ dunif(0,2)
sigma.epie3 ~ dunif(0,2)]
where N is the number of genes and P is the number of pixels.

5.2 Directed acyclic graph (DAG) for the model