Bayesian Analysis (2006) **1**, Number 4, pp. 707–732

Bayesian Analysis of Comparative Microarray Experiments by Model Averaging

Paola Sebastiani[*], Hui Xie[†] and Marco F Ramoni[‡]

Abstract. A major challenge to the statistical analysis of microarray data is the small number of samples — limited by both cost and sample availability — compared to the large number of genes, now soaring into the tens of thousands per experiment. This situation is made even more difficult by the complex nature of the empirical distributions of gene expression measurements and the necessity to limit the number of false detections due to multiple comparisons. This paper introduces a novel Bayesian method for the analysis of comparative experiments performed with oligonucleotide microarrays. Our method models gene expression data by log-normal and gamma distributions with hierarchical prior distributions on the parameters of interest, and uses model averaging to compute the posterior probability of differential expression. An initial approximate Bayesian analysis is used to identify genes that have a large probability of differential expression, and this list of candidate genes is further refined by using stochastic computations. We assess the performance of this method using real data sets and show that it has an almost negligible false positive rate in small sample experiments that leads to a better detection performance.

Keywords: Microarrays, gene expression, gamma distribution, log-normal distribution, model averaging, true and false positive rates, false discovery rate

1 Introduction

One of the results of the Human Genome project is the discovery that the human genome comprises between 30,000 and 35,000 genes. Only about 50% of these genes have known functions and one of the challenges of the post-genome era is to characterize these newly discovered genes and to understand their role in cellular processes or in mechanisms leading to disease. An avenue of research focuses on gene expression: the process by which a gene transcribes the genetic code stored in its DNA sequence into molecules of mRNA that are used for producing proteins (Jacob and Monod 1961). A gene expression level is the abundance of mRNA produced during the gene expression, and the measurement of the expression levels of all the genes in a cell is nowadays made possible by the technology of microarrays.

Data analysis methods play a critical role in the successful execution and analysis of microarray experiments but notwithstanding a large body of literature (Nadon and Shoemaker 2002; Shoemaker and Lin 2005) there is no universally accepted

[*] Department of Biostatistics, Boston University, Boston MA, sebas@bu.edu

[†] Department of Biostatistics, Boston University, Boston MA,

[‡] Children's Hospital Informatics Program, Harvard Medical School, Boston MA, http://www.chip.org/~marco

statistical protocol for the analysis of microarray data
(Sebastiani et al. 2003a; The Tumor Analysis Best Practices Working Group 2004). Three main issues challenge the development of these data analytical methods: the controversial distributional nature of data generated by microarray experiments, the typically small sample size, and the problem of multiple comparisons.

In the biology literature, the standard measure of change in expression is the *fold change*, which, in the case of unpaired measures, is estimated by the ratio of the sample means (Chen et al. 1997; Newton et al. 2001). Alternative statistical techniques rely on the t statistic, or some adjusted form of it, such as the signal to noise ratio implemented in GeneCluster (Reich et al. 2004), the adjusted t statistic implemented in SAM Tusher et al. (2000), or in its empirical Bayes version (Efron et al. 2001) that are implemented in the package SIGGENES of Bioconductor (Gentleman and Carey 2002). Genes are ranked according to one of these statistics, and those genes exceeding a significance threshold are deemed to be changed across the two conditions. To reduce the number of falsely significant changes due to erroneous distributional assumptions and multiple comparisons, the choice of the significance threshold is typically distribution free, and it aims at controlling the false discovery rate — the rate of non significant changes among the detected changes — rather than the false positive rate — the rate of genes called changed among all non changes (Dudoit et al. 2001). A limitation of distribution free methods is the large sample size required to detect genes with significant change and a large true positive rate (Zien et al. 2003). By contrast, parametric approaches can be more powerful with small samples but rest on convenient distributional assumptions, such as that gene expression data follow a Gaussian distribution (Giles and Kipling 2003; Thomas et al. 2001), a log-normal distribution (Baldi and Long 2001; Ibrahim et al. 2002), or a gamma distribution with some restrictions on the shape parameters (Chen et al. 1997; Newton et al. 2001).

We investigated the adequacy of these distributional assumptions using large data sets (Sebastiani and Ramoni 2002; Sebastiani et al. 2003b, 2006), and our results suggested that a single distributional assumption is not sufficient to consistently describe gene expression data. The problem is particularly evident for gene expression data measured with synthetic oligonucleotide arrays, such as Affymetrix (Sebastiani et al. 2003a) and lead us to develop an approximate Bayesian analysis of gene expression data based on model averaging. The analysis is described in Sebastiani et al. (2006) and is implemented in the program BADGE (Bayesian Analysis of Differential Gene Expression). Briefly, BADGE implements two parallel Bayesian analyses of gene expression data, one conditional on the assumption that the expression data follow a log-normal distribution, and the other one conditional on the assumption that the expression data follow a gamma distribution. The results of the two analyses are averaged using the posterior probabilities of the two models. To make fast computations, the Bayesian inference implemented in BADGE is based on a series of numerical approximations that make it feasible to assess the evidence of differential expression for thousands of genes in a matter of seconds. Several investigations we conducted suggest that the effect of the numerical approximation is negligible on the detection of genes that change expression across two conditions but, although smaller than other popular solutions, the false positive rate of BADGE is still large. We conjecture that this effect may be due to the over-confidence of the posterior analysis caused by the numerical approximations.

To solve this issue, in this paper we propose a stochastic analysis of differential gene expression. We continue to rely on model averaging to account for model uncertainty but, rather than using numerical approximations to the posterior distribution, we use Markov Chain Monte Carlo (MCMC) methods to compute a sample from the posterior distribution of the fold change of expression for each gene, using gamma and lognormal distributions. We then average the results with MCMC estimates of the posterior weights. Beside relaxing the requirement on the minimum sample size that is necessary for comfortably using the numerical approximations implemented in BADGE, the use of stochastic computations enables us to include appropriate prior distributions on the parameters of the log-normal and gamma distribution. We describe a method to define *objective* hierarchical prior distribution that uses the expression values of genes that are not used in further analysis to model the baseline hyper-variability of gene expression measured with microarrays. The use of proper prior distributions makes the analysis very robust to outliers and leads to a significant decrease of the false positive rate without loss of power.

The next section describes the Bayesian models and the specification of prior distributions that account for the variability of gene expression data and Section 3 shows the effect of using stochastic computations in the reduction of the false positive rate in small sample experiments. Discussion and suggestions for further work are in Section Appendix.

2 Method

We first give a brief description of microarray data, followed by a general overview of our approach and then provide details of the model parameterizations and implementation in Winbugs 1.4 Thomas et al. (1992).

2.1 Microarray Data

Technically a microarray is a platform containing copies of functional DNAs of genes. There are different technologies for microarrays and we remind to the review in Sebastiani et al. (2003a) for a general overview. In this paper, we focus on synthetic oligonucleotide microarrays. A synthetic oligonucleotide microarray is a platform gridded in such a way that each location of the grid corresponds to a gene and contains several copies of a short specific DNA segment that is characteristic of the gene (Duggan et al. 1999). The short specific segments are known as *synthetic oligonucleotides* and the copies of synthetic oligonucleotides that are fixed on the platform are called the *probes*.

Each gene is associated with a number of *probe pairs* ranging from 11 in the Human Genome U133 set, to 16 in the Murine Genome U74v2 set and the Human Genome

U95v2. A probe pair consists of a perfect match probe and a mismatch probe. Each perfect match probe is chosen on the basis of uniqueness criteria and proprietary, empirical rules designed to improve the odds that probes will hybridize — that is bind — to mRNA molecules with high specificity. The mismatch probe is identical to the corresponding perfect match probe except for the nucleotide in the central position, which is replaced with its complementary nucleotide. The inversion of the central nucleotide makes the mismatch probe a further specificity control because, by design, hybridization of the mismatch probe can be attributed to either non specific hybridization or background signal caused by the hybridization of cell debris and salts to the probes (Lockhart et al. 1996). Each cell of an Affymetrix oligonucleotide microarray consists of millions of samples of a perfect match or mismatch probe, and the probes are scattered across the microarray in a random order to avoid systematic bias. To measure the expression level of the genes in a cell, investigators prepare the *target* by extracting the mRNA from the cell and making a fluorescence-tagged copy. This tagged copy is then hybridized to the probes in the microarray. During the hybridization, if a gene is expressed in the target cells, its mRNA representation will bind to the probes on the microarray, and its fluorescence tagging will make the corresponding probe brighter. Therefore, the measure of each probe intensity is taken as a proxy of the mRNA abundance for the corresponding gene in the sample, and a robust average of the intensities of the probe set determines a relative expression for the corresponding gene. Full details are in the Affymetrix document describing the statistical algorithm that is available from www.affymetrix.com/support/technical/whitepapers, and a summary is in Sebastiani et al. (2003a). Figure 11 in the supplementary material sketches the three steps of a microarray experiment.

2.2 Model Averaging

A typical microarray experiment produces the expression level of thousands of genes in two or more biological conditions. We denote by y_{kji} the expression level of gene k, $(k = 1, ..., p)$ measured in sample i of condition j, $(j = A, B)$, and by y_k the overall expression profile of gene k. We also denote by n_j the number of samples measured in condition j so that $i = 1, \ldots, n_j$. We measure the differential expression of each gene k by the fold change θ_k that is defined as

$$\theta_k = \frac{\mu_{kA}}{\mu_{kB}}, \quad k = 1, \ldots, p,$$

where μ_{kj} denotes the average expression level of the gene k in condition j $(j = A, B)$. A fold change $\theta_k = 1$ denotes no change of expression, while $\theta_k < 1$ and $\theta_k > 1$ indicate, respectively, over and under expression in condition 2 compared to condition 1. Therefore, we measure the evidence of differential expression by the posterior probability $1 - p(\frac{1}{\tau} \leq \theta_k \leq \tau | y_k)$ for a pre-specified threshold change $\tau > 1$. More specifically, values of $p(\theta_k < \frac{1}{\tau} | y_k)$ near 1 identify those genes that are τ-fold more expressed in condition 2, while values $p(\theta_k > \tau | y_k)$ close to 1 identify those genes that are τ-fold more expressed in condition 1. The threshold τ on the fold change of expression is determined by the sample size. Experiments with less than 20 samples per condition usually leads to the discovery of genes that change by at least 2-fold expression, while the detection of smaller effects requires larger sample size (Zien et al. 2003; Sebastiani et al. 2006).

We compute the posterior probability by assuming that the gene expression data follow either a gamma or a log-normal distribution and average the results of the two analyses. If we let M_{lk} and M_{gk} denote, respectively, the log-normal and gamma models for the expression data of gene k, the posterior probability $p(\theta_k > \tau|y_k)$ is computed as the weighted average

$$p(\theta_k > \tau|y_k) = p(\theta_k > \tau|M_{lk}, y_k)p(M_{lk}|y_k) + p(\theta_k > \tau|M_{gk}, y_k)p(M_{gk}|y_k) \qquad (1)$$

where $p(\theta_k > \tau|M_{lk}, y_k)$ and $p(\theta_k > \tau|M_{gk}, y_k)$ are the posterior probabilities of differential expression assuming a log-normal and a gamma model, and the weights $p(M_{lk}|y_k)$ and $p(M_{gk}|y_k) = 1 - p(M_{lk}|y_k)$ are the posterior probabilities of the two models. Similarly, we compute the Bayesian point estimate of the fold change by the posterior expectation of θ_k, say $E(\theta_k|y_k)$, and therefore we average the conditional point estimates

$$E(\theta_k|y_k) = E(\theta_k|M_{lk}, y_k)p(M_{lk}|y_k) + E(\theta_k|M_{gk}, y_k)p(M_{gk}|y_k). \qquad (2)$$

We compute an approximate $(1-\alpha)\%$ credible interval by averaging the credible limits under the two models. This averaging technique is known as *Bayesian model averaging* and is reviewed in Hoeting et al. (1999).

2.3 Log-normal Model

Suppose the expression data y_{kji} follow a log-normal distribution with parameters η_{kj} and $\tau_{kj} = 1/\sigma^2_{kj}$ defining the mean $\mu_{kj} = e^{\eta_{kj}+\sigma^2_{kj}/2}$, the variance $\mu^2_{kj}(e^{\sigma^2_{kj}} - 1)$, and the density function

$$f_l(y_{kji}|M_{lk}, \eta_{kj}, \sigma^2_{kj}) = \frac{1}{y_{kji}\sqrt{2\pi\sigma^2_{kj}}} e^{-\frac{1}{2\sigma^2_{kj}}(\log(y_{kji})-\eta_{kj})^2}, \quad y_{kji}, \sigma^2_{kj} > 0, j = A, B. \qquad (3)$$

With this parameterization, the fold change of expression conditional on the log-normal model is:

$$\theta_k|M_{lk} = \frac{\mu_{kA}}{\mu_{kB}} = e^{\eta_{kA}-\eta_{kB}+(\sigma^2_{kA}-\sigma^2_{kB})/2}$$

and this is our parameter of interest. Because η_{kj} and σ^2_{kj} define the mean and variance of the log-transformed data, we assume Gamma priors on the parameters τ_{kA} and τ_{kB}, and normal-gamma priors on the parameters η_{kA} and η_{kB} that are independent of τ_{kA} and τ_{kB}. The graphical model in Figure 1 summarizes the Bayesian model, conditional on the assumed log-normal distribution for the expression data.

We choose the hyper-parameters to represent the fact that a gene is not expressed in the target. Because of background noise, non expressed genes may result in noisy measurements and our prior specification describes this variability of non-expressed genes and, in particular, of non differentially expressed genes. Therefore, the prior distributions for the parameters η_{kj} and τ_{kj} are the same for $j = A, B$. Furthermore, the hierarchical prior on the parameters η_{kA} and η_{kB} allows us to model the wide range of variability of non-expressed genes.

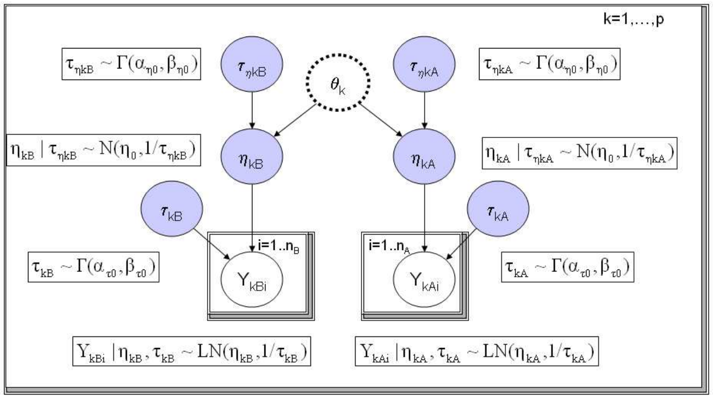

k=1,...,p
$\tau_{\eta kB} \sim \Gamma(\alpha_{\eta 0}, \beta_{\eta 0})$
$\tau_{\eta kB}$
θ_k
$\tau_{\eta kA}$
$\tau_{\eta kA} \sim \Gamma(\alpha_{\eta 0}, \beta_{\eta 0})$
$\eta_{kB} \mid \tau_{\eta kB} \sim N(\eta_0, 1/\tau_{\eta kB})$
η_{kB}
η_{kA}
$\eta_{kA} \mid \tau_{\eta kA} \sim N(\eta_0, 1/\tau_{\eta kA})$
τ_{kB}
τ_{kA}
$i=1..n_B$
$i=1..n_A$
$\tau_{kB} \sim \Gamma(\alpha_{\tau 0}, \beta_{\tau 0})$
Y_{kBi}
Y_{kAi}
$\tau_{kA} \sim \Gamma(\alpha_{\tau 0}, \beta_{\tau 0})$
$Y_{kBi} \mid \eta_{kB}, \tau_{kB} \sim LN(\eta_{kB}, 1/\tau_{kB})$
$Y_{kAi} \mid \eta_{kA}, \tau_{kA} \sim LN(\eta_{kA}, 1/\tau_{kA})$

Figure 1: Graphical model describing the parameterization for gene expression data modeled by a log-normal distribution. The subindex A and B denote the two biological conditons. Nodes in blue are the parameters that need a prior distribution. Nodes with border in dotted lines are functional nodes. Plates represents repeated parts of the graph.

We use the following strategy to elicit the prior distributions when the expression data are measured by Affymetrix chips and processed with the statistical software MAS 5.0. The software labels genes by a detection call and, specifically, genes are labelled as either "absent" or "marginally present" when the expression is non detectable, and they are labelled as "present" when there is evidence of detectable expression (We remind to Sebastiani et al. (2003a) for full details on how these detection calls are determined.) It is common practice to disregard from the analysis those genes that are systematically labelled as absent in all the samples, because they are either never expressed in the biological samples, or their expression measure is not reliable. Typically they amount to about 25–50% of the total number of genes. These data however contain information about the variability of non expressed genes and therefore we use them to build our prior distributions.

We first note the following relationships that will be used to define the hyper-parameters of the prior distributions:

$$\begin{aligned}
E(\eta_{kj}) &= \eta_0 & (4)\\
E\{\log(Y_{kj})\} &= E(E\{\log(Y_{kj})|\eta_{kj},\tau_{kj}\}) = \eta_0 & (5)\\
V(\eta_{kj}) &= \beta_{\eta 0}/(\alpha_{\eta 0}-1) & (6)\\
V\{\log(Y_{kj})\} &= V(E\{\log(Y_{kj})|\eta_{kj},\tau_{kj}\}) + E(V\{\log(Y_{kj})|\eta_{kj},\tau_{kj}\}) & (7)\\
&= V(\eta_{kj}) + E(1/\tau_{kj})\\
&= \beta_{\eta 0}/(\alpha_{\eta 0}-1) + \beta_{\tau 0}/(\alpha_{\tau 0}-1)
\end{aligned}$$

To have $V(\eta_{kj}) > 0$ and $V\{\log(Y_{kj})\} > 0$ we specify $\alpha_{\eta 0} > 1$ and $\alpha_{\tau 0} > 1$, and because the marginal variances decrease with $\alpha_{\eta 0}$ and $\alpha_{\tau 0}$, we specify

$$\alpha_{\eta 0} = 2 \quad \alpha_{\tau 0} = 2.$$

We use the expression values of the genes labelled as absent to specify the hyper-parameters $\eta_0 = \bar{x}_a$, where $\bar{x}_a$ is the average of the expression values labelled as absent, in logarithmic scale. To fix the hyper-parameter $\beta_{\eta 0}$, we note that this parameter determines the variance of the marginal distribution of η_{kj}. Therefore, we compute the sample mean of the log-transformed expression values for each gene labelled as absent throughout the samples, say $\bar{x}_g$ and then compute the variance of the distribution of the sample means $\sigma_l^2 = \sum_g (\bar{x}_g - \bar{x}_a)^2/(n_a - 1)$, where n_a is the number of absent genes. We then set $\beta_{\eta 0}$ equal to this variance. To compute the last hyper-parameter $\beta_{\tau 0}$, we solve the equation $S_{xa}^2 = \sigma_l^2 + \beta_{\tau 0}$ where S_{xa}^2 is the variance of the expression values labelled as absent, in logarithmic scale. Note that $S_{xa}^2 > \sigma_l^2$ so that $\beta_{\tau 0}$ is always positive.

2.4 Gamma Model

Suppose now that the gene expression data follow a gamma distribution with parameters α_{kj}, β_{kj} that specify the mean and the variance of the distribution as $\mu_{kj} = \alpha_{kj}/\beta_{kj}$ and μ_{kj}^2/α_{kj}, and the density function

$$f_g(y_{kji}|\alpha_{kj},\beta_{kj}) = \frac{\beta_{kj}^{\alpha_{kj}}}{\Gamma(\alpha_{kj})} y_{kji}^{\alpha_{kj}-1} e^{-y_{kji}\beta_{kj}}, \quad y_{kji}, \alpha_{kj}, \beta_{kj} > 0, j = A, B. \quad (8)$$

The parameter of interest is the fold change of expression of each gene k:

$$\theta_k = \frac{\mu_{kA}}{\mu_{kB}} = \frac{\alpha_{kA}\beta_{kB}}{\alpha_{kB}\beta_{kA}}$$

so that we specify our Bayesian model by defining the conditional distribution of the gene expression data y_{kj} as a function of α_{kj} and β_{kj}, where β_{kj} is a function of α_{kj} and μ_{kj}, for $j = A, B$. We then model the prior distribution of α_{kj} and μ_{kj} by gamma distributions. The graphical model in Figure 2 summarizes the Bayesian model, conditional on the assumed gamma distribution for the expression data.

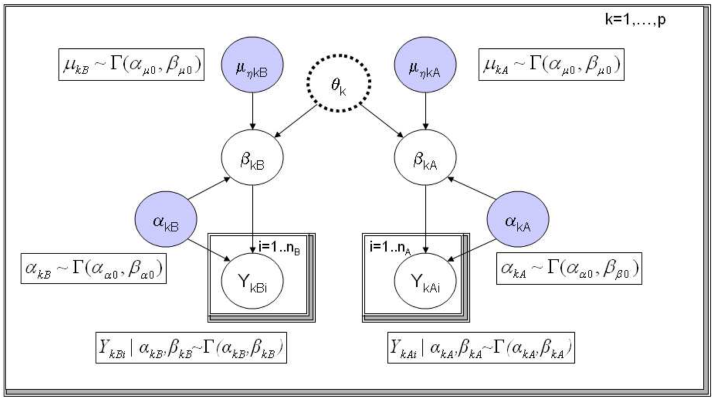

k=1,…,p
$\mu_{kB} \sim \Gamma(\alpha_{\mu 0}, \beta_{\mu 0})$
$\mu_{\eta kB}$
θ_k
$\mu_{\eta kA}$
$\mu_{kA} \sim \Gamma(\alpha_{\mu 0}, \beta_{\mu 0})$
β_{kB}
β_{kA}
α_{kB}
α_{kA}
i=1..n$_B$
i=1..n$_A$
$\alpha_{kB} \sim \Gamma(\alpha_{\alpha 0}, \beta_{\alpha 0})$
Y_{kBi}
Y_{kAi}
$\alpha_{kA} \sim \Gamma(\alpha_{\alpha 0}, \beta_{\beta 0})$
$Y_{kBi} \mid \alpha_{kB}, \beta_{kB} \sim \Gamma(\alpha_{kB}, \beta_{kB})$
$Y_{kAi} \mid \alpha_{kA}, \beta_{kA} \sim \Gamma(\alpha_{kA}, \beta_{kA})$

Figure 2: *Graphical model describing the parameterization for gene expression data modeled by a gamma distribution. The subindex A and B denote the two biological conditions. As in Figure 1, nodes in blue need a prior distribution; nodes with border in dotted lines are functional nodes, and plates represents repeated parts of the graph.*

We assume the same gamma prior distribution $\Gamma(\alpha_{a0}, \beta_{a0})$ on the parameters α_{kj} for $j = A, B$, and the same gamma prior distribution $\Gamma(\alpha_{\mu 0}, \beta_{\mu 0})$ on the parameters μ_{kj} for $j = A, B$ to represent the assumption that, a priori, each gene k is not expressed in either conditions A and B, and therefore it is also not differentially expressed. The parameters μ_{kj} and α_{kj} are marginally independent, and become conditionally dependent given β_{kj}. We note the following relationships that will be used to define the hyper-parameters of the prior distributions:

$$\begin{aligned}
E(\mu_{kj}) &= \alpha_{\mu 0}/\beta_{\mu 0} && (9)\\
E(Y_{kj}) &= E\{E(Y_{kj}|\mu_{kj})\} = \alpha_{\mu 0}/\beta_{\mu 0} && (10)\\
V(\mu_{kj}) &= \alpha_{\mu 0}/\beta_{\mu 0}^2 && (11)\\
V(Y_{kj}) &= E\{V(Y_{kj}|\mu_{kj}, \alpha_{kj})\} + V\{E(Y_{kj}|\mu_{kj})\} && (12)\\
&= E\{\mu_{kj}^2/\alpha_{kj}\} + V(\mu_{kj}) && (13)\\
&= E(\mu_{kj}^2) \times \beta_{a0}/(\alpha_{a0} - 1) + V(\mu_{kj}) && (14)
\end{aligned}$$

We first specify the parameters $\alpha_{\mu 0}$ and $\beta_{\mu 0}$ such that they match the distributions of the sample means of the genes that are always absent. To do this, we compute the average expression $\bar{y}_g$ of each gene g that is always absent, and then we compute the mean and variance of the distribution of these values:

$$\bar{y} = \frac{\sum_g \bar{y}_g}{n_a} \quad \sigma^2 = \frac{\sum_g (\bar{y}_g - \bar{y})^2}{n_a - 1}$$

where n_a is the number of absent genes. We then set

$$\alpha_{\mu 0} = \bar{y}^2/\sigma^2 \quad \beta_{\mu 0} = \bar{y}/\sigma^2$$

to ensure that $\alpha_{\mu 0}/\beta_{\mu 0} = \bar{y}$ and $\alpha_{\mu 0}/\beta_{\mu 0}^2 = \sigma^2$. We specify the last hyper-parameter α_{a0} and β_{a0} so that the marginal variance of Y_{kj} matches the variance of the absent genes, for all k and j. Therefore we need to find α_{a0} and β_{a0} so that

$$S_{ya}^2 = E(\mu_{kj}^2) \times \beta_{a0}/(\alpha_{a0} - 1) + V(\mu_{kj})$$

where S_{ya}^2 is the overall sample variance of the expression values labelled as absent. By the prior definition for μ_{kj}, we know that $E(\mu_{kj}^2) = V(\mu_{kj}) + E(\mu_{kj})^2 = \sigma^2 + \bar{y}^2$. Therefore,

$$\beta_{a0}/(\alpha_{a0} - 1) = (S_{ya}^2 - \sigma^2)/(\sigma^2 + \bar{y}^2)$$

We fix $\alpha_{a0} = 2$ to have the least informative proper prior, and $\beta_{a0} = (S_{ya}^2 - \sigma^2)/(\sigma^2 + \bar{y}^2)$.

2.5 Implementation and Analysis

We have implemented the model averaging procedure into a set of functions that run under the R package (version 1.9) and are interfaced to Winbugs 1.4 through the package R2WinBUGS[1]. We first identify the set of genes that are consistently labelled as absent throughout all samples and use their expression values to define the prior hyper-parameters. Then, for each gene and each log-normal and gamma model, we use WinBugs 1.4 to generate a sample from the posterior distribution of θ_k that we use to estimate the posterior probability $p(\theta_k > \tau|M_m)$ $(M_m = M_g, M_l)$ as

$$p(\theta_k > \tau|M_m) = \frac{\sum_h (\theta_{kh} > \tau)}{n}$$

where $\{\theta_{kh}\}$ $(h = 1, ..., n)$ is the MCMC sample of n values for the parameter θ_k. In practice, an initial burn-in of 1,000 iteration followed by a sample of 1,000 iterations seem to be sufficient to reach the convergence when the gene expression data are assumed to follow a log-normal distribution. When we assume a gamma distribution for the gene expression data, convergence may be slower so we use initial values for the parameters that match the empirical moments of the data. (See some discussion in the supplementary material, section Diagnostic). For each gene, we also estimate the mixing weights by computing a stochastic estimate of the marginal likelihood, in log-scale. We use these estimates to approximate the Bayes factor by $B_k = exp(\log(p(M_{lk}|y_k)) - \log(p(M_{gk}|y_k)))$ and then derive the mixing weights of the log-normal and gamma models as $B_k/(1+B_k)$ and $1/(1 + B_k)$.

[1] Available for download at `http://cran.r-project.org/src/contrib/Descriptions/R2WinBUGS.html`

Our analysis scores each gene by the posterior probability of differential expression that is then used to rank genes. To select the genes differentially expressed across the two conditions, we fix a small threshold $\alpha(= 5\%)$ on the posterior probability of differential expression, and select the genes for which $p(1/\tau \leq \theta_k \leq \tau | y_k) < \alpha$.

Although MCMC methods can be time consuming, in our experience we have successfully analyzed sets comprising up to 5,000 genes in at most 15 minutes, using an Intel Pentium 1.7 GHz, and 1.0 GB of RAM.

3 Evaluation

The objective of this evaluation is to assess the performance of our method in real microarray experiments of different sample sizes, in which there is a known number of genes that change expression between two biological conditions.

3.1 Material and Methods

We created test sets from a spiked-in microarray study prepared by Affymetrix using the U133 array A. The data set consists of three technical replicates of 14 separate microarray experiments. In each experiment, the target consisted of a complex human background, and 42 genes whose expression was artificially amplified at concentrations ranging from 0pM to 512pM. While the expression profiles of the non-spiked genes should simply be random noise, the expression values of the spiked-in genes increase proportionally to the concentrations. As an example, Figure 12 in the supplementary material shows the expression profiles of four of these genes, for increasing concentrations. Note that the data were scaled to a target value of 1, so that the expression values range between 0 and 6968.0, with a median expression 31.2.

These 42 spiked-in genes were assigned to each of the 14 experiments using the Latin square design described in Figure 3. One of the 42 samples did not pass the quality control steps and we removed it from any subsequent analysis. Furthermore, we found that the expression profiles of 27 extra genes matched the expression profiles of some of the spiked-in genes: four had the expression profiles matching those of three spiked-in genes assigned to group G1 in the original Latin square experiment; two expression profiles matched those of the spiked-in genes assigned to group G8; five expression profiles matched those of the spiked-in genes in group G13; and 15 profiles matched those of the spiked-in genes in group G14 (See Table 2 in the supplementary material for full details). We therefore used the expression profiles of the 42+27 spiked-in genes and the background noise to create test sets for differential analysis in which we varied the sample size, and the number of true changes represented by the artificially spiked-in genes.

To create these test sets, we proceeded as follows. We first note that by the Latin square arrangement, each group of low concentrations ranging between 0pM and 4pM is matched by a group of high concentrations ranging between 8pM and 512pM: For

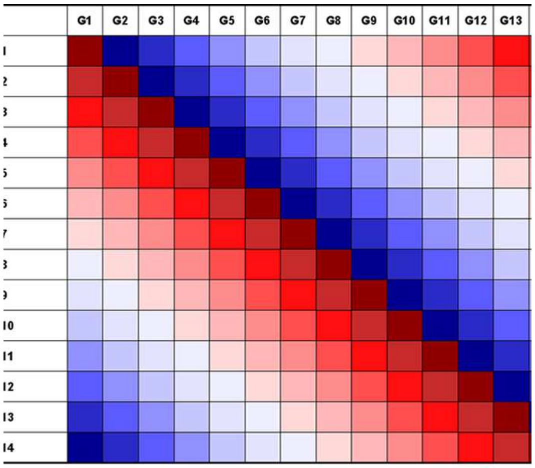


Figure 3: *Latin square design used to generate the artificial test sets. Each row represents one of the 14 distinct experiments, and each column represents the groups of genes that were artificially spiked in. Each cell (i, j) represents the concentration of the spiked-in genes in run i and group j. The color code is red to denote concentrations from 0pM to 4pM, or low concentrations, and blue to denote concentrations from 8pM to 512pM, or high concentrations. For example, in run 1 the genes in group G1 are all spiked-in at concentration 0pM, while the genes in group G2 are all spiked-in at concentration 512pM and all genes in group G8 are spiked-in at concentration 4pM.*

example the first seven low concentrations of the spiked-in genes in group G1 (red cells in Column 1, Figure 3) are matched by the first seven high concentrations of the spiked-in genes in group G8 (blue cells in Column 8, Figure 3). Similarly the low concentrations in run 2–8 of the spiked-in genes in group G2 are matched by the high concentrations of the spiked-in genes in group G9, and so on. Because our preliminary analysis showed that the expression of each gene spiked at low concentrations 0–4pM was much smaller than the expression of the same gene spiked at higher concentrations 8–512pM (See Figure 12 in the supplementary material), we dichotomized the concentrations as low (0–4pM) and high (8–512pM) and used the genes that were spiked-in with either low or high concentration to create test sets of different sample sizes as illustrated in Figure 4.

We started by creating three initial sets of 41 samples each: the first set comprised all 12 spiked-in genes in groups G1 and G8, as well as all 22,244 genes that were not spiked-in any of the 41 samples; the second set comprised all 11 spiked-in genes in groups G6 and G13 and the 22,244 genes that were not spiked-in any of the 41 samples; the third set comprised the 21 spiked-in genes in groups G7 and G14, and the remaining 22,244 genes that were not spiked-in any of the 41 samples. We then divided the 41 samples in each set into two biological conditions A and B comprising all experiments at low or high concentrations. For example, in the first set we assigned the triplicated experiments 1–7 to condition A, and the triplicated experiments 8–14 to condition B,

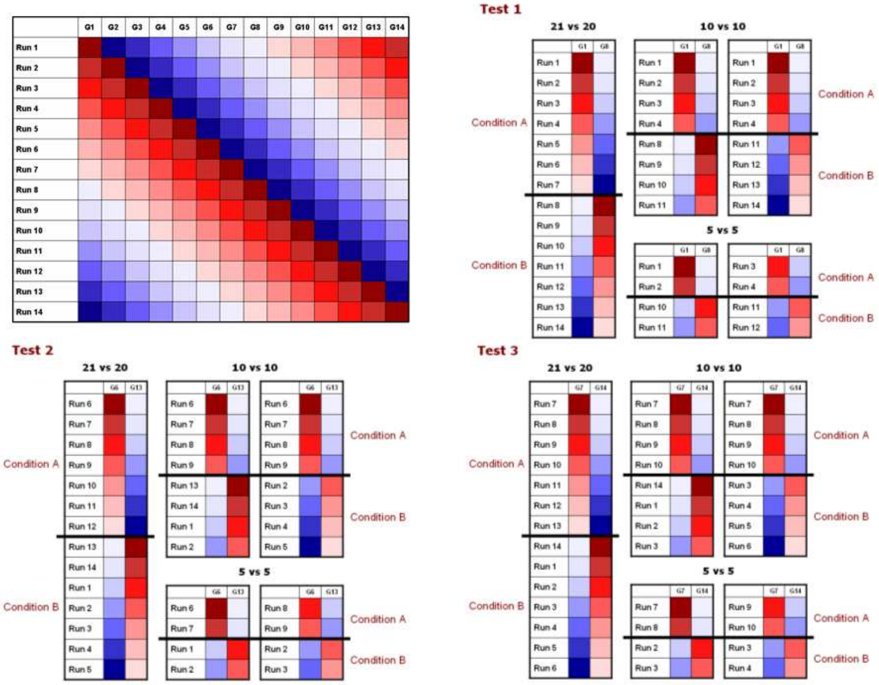


Figure 4: *Test sets used for the evaluation that were created from the original microarray study. As in Figure 3, each row represents one of the 14 distinct experiments, and each column represents the groups of genes that were artificially spiked-in at concentrations represented by the color intensity. The color code is red to denote concentrations from 0pM to 4pM, and blue to denote concentrations from 8pM to 512pM.*

thus providing a test set for comparative analysis with 12 known changes, given by the seven known spikes-in genes in group G1 and the five spike-in genes in group G8, and 20 and 21 samples in conditions A and B. From this set, we selected two further subsets of 10 samples per condition, and two subsets of 5 samples per condition. We created data sets of different sample size from the other two sets in a similar way, as shown in Figure 4. Note that the data sets of different sizes comprise genes that were spiked-in at different concentrations, thus providing test sets for differential analysis of different complexity with median fold change ranging from 1.04 to 88 between the two conditions.

We analyzed each of the 15 data sets by first removing the genes labelled as absent, whose expression values were used to build the prior distributions as described in Section 2. The genes with at least one expression value labelled as present were analyzed with the original version of BADGE (Sebastiani et al. 2006), which uses numerical approximations for fast computations. The effect of these numerical approximations is to yield a posterior probability of differential expression that is often too extreme, with the consequence of a large power and a large false positive rate. We used this first approximate analysis as a filter to only remove those genes in which there is no evidence of differential expression and we filtered out all those genes with a posterior probability of differential expression $0.025 < p(\theta_k > 1|y_k) < 0.975$. We used the stochastic analysis on the remaining genes and we selected those genes with $p(\theta_k > 2|y_k)$ or $p(\theta_k < 1/2|y_k)$ exceeding 95% as differentially expressed. This choice for the threshold τ will be discussed in the next section. For comparisons, we also detected the differentially expressed genes by controlling the false discovery rate using SAM (Tusher et al. 2000) that is implemented in the package *siggenes* in Bioconductor[2]. We limited the analysis with SAM to the default search ranges for the smoothing parameters, and transformed the raw expression data using the cubic root transformation as suggested in Tusher et al. (2000).

3.2 Results

Table 1 reports the results of the evaluation. The selected number of genes that were analyzed after removal of the absent genes was on average 14,000. In this initial selection, BADGE identified a number of significant genes ranging from 31, with 20 falsely significant genes, to 1106 with 1085 falsely significant genes and a false discovery rate ranging from 56% to 99%. The behavior of SAM was mixed: in five tests SAM selected an unacceptable number of significant genes, with almost 100% FDR; in the remaining ten tests SAM usually selected a small number of different genes with a large true positive rate, although the performance is almost always worse than the stochastic Bayesian analysis. The stochastic Bayesian analysis is remarkably accurate, with a number of false detection that is almost negligible and a false discovery rate that is at most 27%. The large accuracy is also paired by a large power that is above 85% in the 15 sets. These results confirm our preliminary power analysis that relatively small samples allow for the estimation of large effects defined by at least a 2-fold change of expression (Sebastiani et al. 2006), and hence $\tau = 2$. Furthermore, the positive results are stable across the 15 test sets.

We believe the higher sensitivity and specificity of the stochastic analysis is induced by both the use of a prior distribution that models the background noise and a more conservative threshold τ on the fold change. This is confirmed by the fact that, if we restrict the initial selection in BADGE to those genes that change expression by at least 2-fold, the average number of gene selected by BADGE decreases but remains larger than the number of genes selected with the stochastic analysis. For example the number of selected genes is 82 and 37 in Test 1, column 1 of Table 1 with five samples per group, compared to 14 and 13 genes selected by the stochastic analysis. These results confirm that the approximate analysis of the original implementation of BADGE has high power at the price of a large false positive rate. On the other hand, the stochastic analysis maintains the high power but reduces the false positive rate.

[2] http://bioconductor.org/

	Test 1			Test 2			Test 3		
	12 True Change			11 True Changes			21 True Changes		
$n_1 = 20, n_2 = 21$	Selected	FDR	TP	Selected	FDR	TP	Selected	FDR	TP
SBADGE	11	0.00	11	11	0.00	11	21	0.00	21
BADGE	76	0.84	12	31	0.64	11	58	0.64	21
SAM	6989	1.00	10	19	0.42	11	6906	1.00	21
	9 True Changes			11 True Changes			21 True Changes		
$n_1 = n_2 = 10$	Selected	FDR	TP	Selected	FDR	TP	Selected	FDR	TP
SBADGE	11	0.27	8	13	0.15	11	24	0.13	21
BADGE	256	0.96	9	573	0.98	11	663	0.97	21
SAM	10260	1.00	7	11710	1.00	11	11168	1.00	21
	12 True Changes			11 True Changes			21 True Changes		
$n_1 = n_2 = 10$	Selected	FDR	TP	Selected	FDR	TP	Selected	FDR	TP
SBADGE	11	0.00	11	12	0.08	11	21	0.00	21
BADGE	38	0.68	12	30	0.63	11	48	0.56	21
SAM	10	0.33	8	10	0.09	10	18	0.00	18
	9 True Changes			11 True Changes			21 True Changes		
$n_1 = n_2 = 5$	Selected	FDR	TP	Selected	FDR	TP	Selected	FDR	TP
SBADGE	14	0.43	8	15	0.27	11	23	0.09	21
BADGE	1104	0.99	9	496	0.98	11	1106	0.98	21
SAM	8	0.13	7	14	0.21	11	229	0.91	21
	12 True Changes			11 True Changes			21 True Changes		
$n_1 = n_2 = 5$	Selected	FDR	TP	Selected	FDR	TP	Selected	FDR	TP
SBADGE	13	0.17	10	14	0.21	11	22	0.05	21
BADGE	251	0.01	12	101	0.89	11	248	0.92	21
SAM	171	0.95	9	12	0.08	11	23	0.09	21

Table 1: *Detection accuracy of* BADGE, *the extension with MCMC estimation* (SBADGE), *and significant analysis of microarray* (SAM) *in the 15 data sets generated from the spiked-in experiment. In each test, we identified the significant genes by bounding the posterior probability of differential expression to 0.95. The false discovery rate (FDR) is the ratio of falsely significant genes and the total number of significant genes. The true positives (TP) is the total number of detected true changes. In* SAM *we used 18% nominal FDR.*

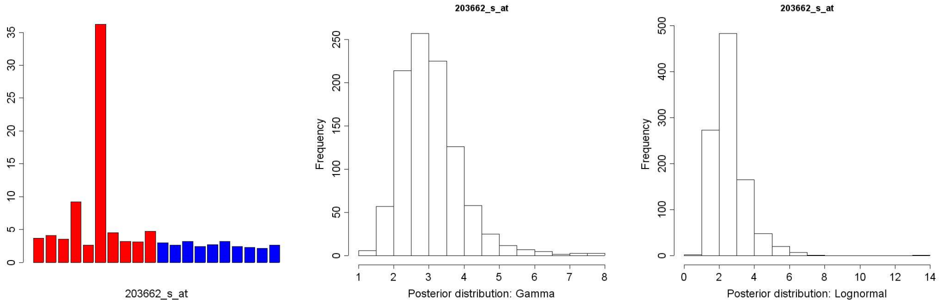


Figure 5: *Left: Expression data of the gene 203662_s_at (y-axis) in 20 microarray samples comprising 10 samples in condition A (red) and 10 samples in condition B (blue) (x-axis). This gene was selected as differentially expressed by the approximate analysis in* BADGE*. Middle: posterior distribution of the fold change using a gamma distribution for the expression values. Right: posterior distribution of the fold change using a lognormal distribution for the expression values. The two distributions lead to* $p(\theta > 2|M_g) = 0.937$ *with weight 0.94 and* $p(\theta > 2|M_l) = 0.725$ *with weight 0.06, so that the overall posterior probability is 0.92 and the gene is not selected as differentially expressed.*

Figure 5 shows an example of the gain of accuracy induced by the use of stochastic computations. The bar plot in the left panel shows the expression values of the gene 203662_s_at[3] in 20 samples: the average expression is below 5 with the exception of one sample with expression value above 35. The presence of this outlier leads to the selection of this gene as differentially expressed in the initial approximate analysis. The plots in the middle and right panels show the posterior distributions of the fold parameter in 1,000 simulations. The two distributions lead to $p(\theta > 2|M_g) = 0.937$ with weight 0.94 and $p(\theta > 2|M_l) = 0.725$ with weight 0.06, so that the overall posterior probability is 0.92 and the gene is not selected as differentially expressed. This example suggests that the analysis conditional on the log-normal distribution is more robust to outliers. However, as shown in Figure 6, the analysis conditional on the lognormal distribution is often too cautious: the gene 213060_s_at is one of those artificially spiked in in the experiment and, although the data in the left panel suggests some evidence of differential expression, the posterior distribution of the fold parameter conditional on the lognormal distribution rules out a detectable change of expression ($p(\theta < 1/2|M_l) = 0.214$ with weight 0.26), while the analysis conditional on the gamma model points to an almost detectable change of expression ($p(\theta < 1/2|M_g) = 0.92$ with weight 0.74). The effect of averaging the two inferences is an overall posterior probability of 0.73. We note that, in both examples, the larger weights assigned to the gamma models are more consistent with the data distributions: the left top panel of Figure 13 (supplementary material) shows the histogram of the expression values in the original scale, and after the logarithmic transformation (top right panel) for the gene analyzed in Figure 5.

[3] See a description of gene annotation at
http://www.affymetrix.com/support/technical/technotes/mouse430_technote.pdf.

The bottom panel of the same figure shows the histograms of the expression values in the original (left) and logarithmic scale (right) for the gene analyzed in Figure 6. In both cases, the residual asymmetry left after the logarithmic transformation is more consistent with a gamma model rather than a log-normal model.

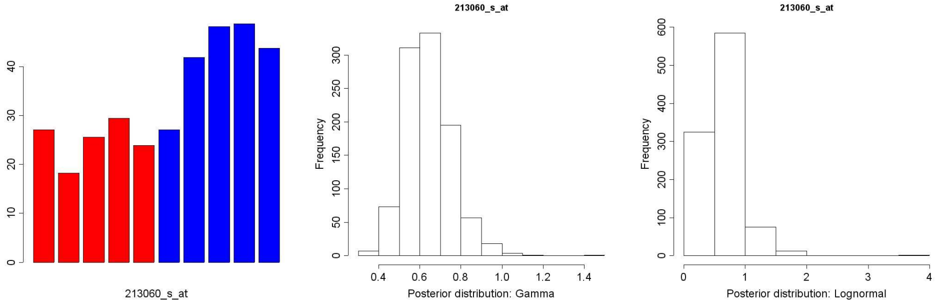


Figure 6: *Left: Expression data of the gene 213060_s_at (y-axis) in 10 microarray samples comprising 5 samples in condition A (red) and 5 samples in condition B (blue) (x-axis). This gene was selected as differentially expressed by the approximate analysis in* BADGE. *Middle: posterior distribution of the fold change using a gamma distribution for the expression values. Right: posterior distribution of the fold change using a lognormal distribution for the expression values. The two distributions lead to* $p(\theta < 1/2|M_g) = 0.92$ *with weight 0.74 and* $p(\theta < 1/2|M_l) = 0.214$ *with weight 0.26, so that the overall posterior probability is 0.73.*

4 Discussion

In this paper we proposed a Bayesian hierarchical model coupled with model averaging to identify genes that change expression between two biological conditions. The well known variability of gene expression data measured with microarrays is modelled through a set of hierarchical prior distributions, whose parameters are determined empirically by matching them with the expression levels of the genes that are not expressed in either biological condition. In this way, the inference on one gene borrows information from the other genes. This is particulary beneficial in microarray studies, where there are many genes simultaneously under study but the sample size for each gene is small.

A Bayesian model averaging technique is used to address the problem of model uncertainty for the gene expression data. Two distributions, log-normal and gamma distributions, are used to fit the individual gene expression data and infer the posterior distribution of the parameter that describes the change of expression between the two biological conditions. The two conditional inferences are then averaged with weights given by their posterior probabilities. We evaluated the proposed method in a spiked-in study produced by Affymetrix. The analysis showed that our method has advantages of computational stability and low detection error rates compared to the popular SAM procedure (Tusher et al. 2000). We are also beginning to use this method for the analysis of real microarray experiments. One successful application is the identification of the molecular profile of HIV positive women in South Africa, who transmit the virus to their newborns (Montano et al. 2006).

A further advantage of our method is that it is easily extendable to accommodate other distributions. So far we implemented the method with log-normal and gamma distributions and our experience shows that these two distributions are generally sufficient to describe gene expression data measured with Affymetrix oligonucleotide arrays. More research is needed to assess the properties of this method when applied to gene expression data measured with cDNA arrays (Sebastiani et al. 2003a) or other emerging platforms such us the Illumina Sentrix BedChip, and compared it to other methods that are more suitable for these platforms.

There are several technical issues that remain to be investigated. We are currently assessing the sensitivity of the results to the prior hyper-parameters and preliminary analysis suggests that specifying the prior hyper-parameters to match the background intensity of non-expressed genes increases specificity. However, the prior hyper-parameters need to be "experiment dependent" as different microarray experiments may lead to different ranges of variability. We are also extending this approach to incorporate a further predictive step that can help identifying genes that are necessary and sufficient to describe the molecular signatures characterizing the two biological conditions. The main intuition is to build a predictive model that uses the genes with detected differential expression to identify each of the two biological conditions. Furthermore, in this paper we focused attention to microarray data preprocessed with the statistical software MAS 5.0. This software uses a non-parametric statistical method to label gene expression as "absent", "marginally present" or "present and we use genes that are consistently labelled as absent to build our prior distributions. Alternative approaches to MAS 5.0 include RMA (Cope et al. 2004) and dChip (Li and Wong 2001) that provide different filters to identify genes with low/nonreliable expression values. Therefore, our method can be easily applied to microarray data preprocessed with either RMA or dChip.

Another issue that needs further investigation is how to determine the significant changes of expression between two conditions. Our method now uses the criterion that $p(1/\tau \leq \theta_k \leq \tau) < \alpha$ for pre-specified values of τ and small α. The choice of the threshold τ is sample size dependent and our preliminary power analysis suggests that microarray experiments comprising less than 20 samples per condition require $\tau \geq 2$ (Sebastiani et al. 2006). This value for τ appears to remove the need for a more conservative value α that would lead to a loss of power. In our evaluation, we tried for example detection rules based on a smaller value α and $\tau = 1$, but we could not optimize the trade off between sensitivity and specificity as with the results obtained when $\tau = 2$. The choice of the best threshold α is still an open question and we are investigating two approaches, one based on a decision theoretic approach to optimize the trade off between sensitivity and specificity, and the other one based on optimizing the FDR.

Software

The program BADGE is available from http://genomethods.org/badge/. The stochastic extension uses a series of scripts for the R package and Winbugs that will be made available on a dedicated web site.

Supplementary material

Diagnostics We use the traceplot of simulations from the posterior distributions of θ_k to assess the convergence of MCMC with 2,000 iterations. The traceplots for gene 24 in test 1 is shown in Figures 7 and 8 for the fitting of log-normal model and Gamma Model. The plot is representative of traceplots for all the genes. The traceplots are for 3 chains with 2000 iterations for each chain. In each chain, an initial burn-in period of 1,000 iterations are followed by a sample of 1,000 iterations. The traceplots show a pattern of convergence.

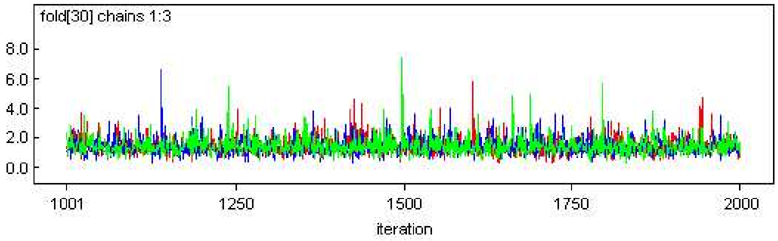


Figure 7: *Traceplot for gene 30 after 1000 burn-in period with a log-normal model.*

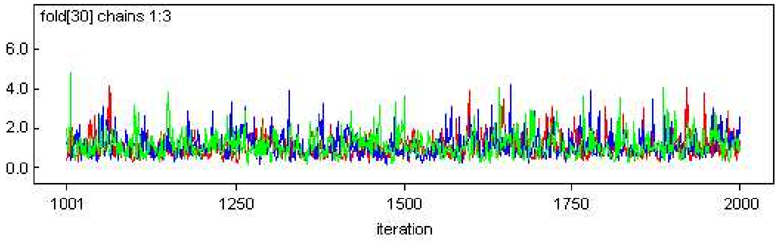


Figure 8: *Traceplot for gene 30 after 1000 burn-in period with a Gamma model.*

Gelman (1996) presents a diagnostic statistic, called R-hat that is used when multiple Markov chains are run, and recommends that values of R-hat for all the estimated parameters should be smaller than 1.2. In our analysis, R-hat values range between 1 and 1.02 for log-normal models, and between 1.003 and 1.031 when Gamma models, thus confirming that 2,000 iterations appear to be sufficient to achieve convergence.

In addition, the two graphs in Figures 9 and 10 display deviance plots for each log-normal and gamma model for the gene expression data used in in test 1, and appear to confirm that a burn-in of 1,000 iteration followed by 1,000 is sufficient to reach convergence.

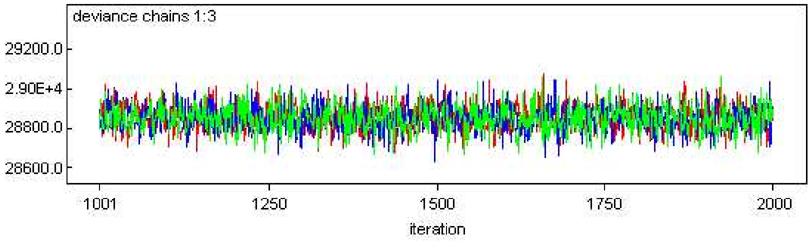


Figure 9: *Traceplot of deviance for log-normal model with data from all genes used in test 1.*

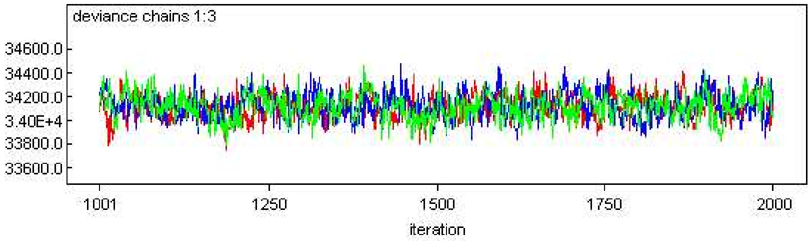


Figure 10: *Traceplot of deviance for Gamma model with data from all genes used in test 1.*

Group ID	Gene ID	Extra spiked-in genes
1	203508_at; 204563_at; 204513_s_at	204890_s_at; 204891_s_at; 203173_s_at; 213060_s_at;
2	204205_at; 204959_at; 207655_s_at	
3	204836_at; 205291_at; 209795_at	
4	205569_at; 207777_s_at; 204912_at	
5	205692_s_at; 207160_at; 212827_at*	209374_s_at*
6	204417_at; 205267_at;209606_at	
7	205398_s_at; 209354_at; 209734_at;	
8	206060_s_at*; 205790_at; 200665_s_at	205397_x_at; 208010_s_at*
9	204430_s_at; 207540_s_at; 207641_at	
10	203471_s_at; 204951_at; 207968_s_at	
11	AFFX-r2-TagA_at; AFFX-r2-TagB_at; AFFX-r2-TagC_at	
12	AFFX-r2-TagD_at; AFFX-r2-TagE_at; AFFX-r2-TagF_at	
13	AFFX-r2-TagH_at; AFFX-r2-TagG_at; AFFX-DapX-3_at	AFFX-DapX-5_at; AFFX-DapX-M_at; AFFX-r2-Bs-dap-3_at; AFFX-r2-Bs-dap-5_at; AFFX-r2-Bs-dap-M_at
14	AFFX-LysX-3_at; AFFX-PheX-3_at; AFFX-ThrX-3_at;	AFFX-LysX-5_at; AFFX-LysX-M_at; AFFX-PheX-5_at; AFFX-PheX-M_at; AFFX-ThrX-5_at; AFFX-ThrX-M_at; AFFX-r2-Bs-lys-3_at; AFFX-r2-Bs-lys-5_at; AFFX-r2-Bs-lys-M_at; AFFX-r2-Bs-phe-3_at; AFFX-r2-Bs-phe-5_at; AFFX-r2-Bs-phe-M_at; AFFX-r2-Bs-thr-3_s_at; AFFX-r2-Bs-thr-5_s_at; AFFX-r2-Bs-thr-M_s_at

Table 2: *Summary of the original 42 spiked-in genes in the latin square experiment. The last column reports the extra spikes that we identified in the data, using the program Caged (http://www.genomethods.org/caged/). The spikes marked with a star were already identified as possible targets.*

Acknowledgments

Research partially funded by the NIH/NHGRI R01:HG003354-01A2.

References

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. 708

Chen, Y., Dougherty, E., and Bittner, M. (1997). "Ratio-based Decisions and the Quantitative Analysis of cDNA Microarray Images." *Journal of Biomedical Optics*, 2: 364–374. 708

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004). "A Benchmark for Affymetrix GeneChip Expression Measures." *Bioinformatics*, 20: 323–331. 723

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2001). "Statistical Methods for Identifying Genes with Differential Expression in Replicated cDNA Microarrays Experiments." *Statistica Sinica*, 12: 111–139. 708

Duggan, J. D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). "Expression Profiling Using CDNA Microarrays." *Nature Genetics*, 21: 10–14. 709

Efron, B., Storey, J. D., and Tibshirani, R. (2001). "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association*, 96: 1151–1160. 708

Gelman, A. (1996). "Inference and monitoring convergence." In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov Chain Monto Carlo in Practice*, 131–143. Boca Raton, FL: Chapman and HallCRC. 724

Gentleman, R. and Carey, V. (2002). "Bioconductor." *R News*, 2(1): 11–16.
URL http://CRAN.R-project.org/doc/Rnews/ 708

Giles, P. J. and Kipling, D. (2003). "Normality of Oligonucleotide Microarray Data and Implications for Parametric Statistical Analyses." *Bioinformatics*, 19: 2254–2262. 708

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian Model Averaging: A Tutorial (with discussion)." *Statistical Science*, 14: 382–417. 711

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002). "Bayesian Models for Gene Expression With DNA Microarray Data." *Journal of the American Statistical Association*, 97: 88–99. 708

Jacob, F. and Monod, J. (1961). "Genetic Regulatory Mechanisms in the Synthesis of Proteins." *Journal of Molecular Biology*, 3: 318–356. 707

Li, C. and Wong, W. H. (2001). "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection." *Proceedings of the National Academy of Science, USA*, 98: 31–36. 723

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nature Biotechnology*, 14: 1675–1680. 710

Montano, M., Rarick, M., Sebastiani, P., Brinkmann, P., Russell, M., Essex, A. N. M., Wester, C., and Thior, I. (2006). "Gene-Expression Profiling of HIV-1 Infection and Perinatal Transmission in Southern Africa." *Genes and Immunity*. In press. 723

Nadon, R. and Shoemaker, J. (2002). "Statistical Issues with Microarrays: Processing and Analysis." *Trends in Genetics*, 18: 265–271. 707

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data." *Journal of Computational Biology*, 8(1): 37–52. 708

Reich, M., Ohm, K., Tamayo, P., Angelo, M., and Mesirov, J. P. (2004). "GeneCluster 2.0: An Advanced Toolset for Bioarray Analysis." *Bioinformatics*, 20(11): 1797–8. 708

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. (2003a). "Statistical challenges in functional genomics (with discussion)." *Statistical Science*, 18: 33–70. 708, 709, 710, 712, 723

Sebastiani, P., Jeneralczuk, J., and Ramoni, M. (2006). "Design and Analysis of Screening Experiments with Microarrays." In Dean, A. and Lewis, S. (eds.), *Screening*, 115–138. Springer. 708, 711, 718, 719, 723

Sebastiani, P. and Ramoni, M. (2002). "Bayesian Differential Analysis of Gene Expression Data." In *Proceedings of the Joint Statistical Meeting. Section on Bayesian Statistical Science.*. American Statistical Association. 708

Sebastiani, P., Yu, Y. H., and Ramoni, M. F. (2003b). "Bayesian Machine Learning and Its Potential Applications to the Genomic Study of Oral Oncology." *Advances in Dental Research*, 17: 104–108. 708

Shoemaker, J. S. and Lin, S. M. (2005). *Methods of Microarray Analysis IV*. New York NY: Springer. 707

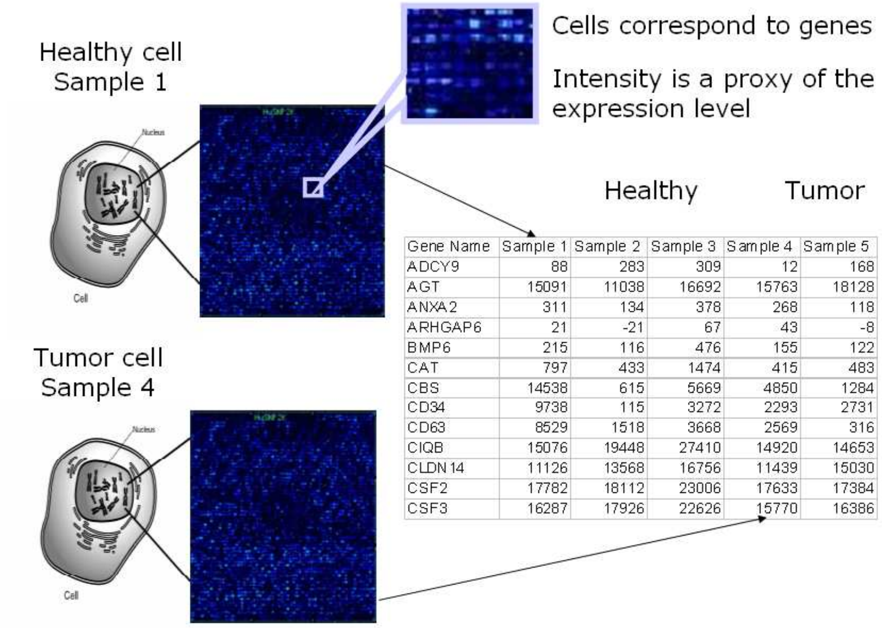
The Tumor Analysis Best Practices Working Group (2004). "Expression Profiling – Best Practices for Data Generation and Interpretation of Clinical Trials." *Nature Genetics*, 5: 229–237. 708

Thomas, A., Spiegelhalter, D. J., and Gilks, W. R. (1992). "BUGS: A program to perform Bayesian inference using Gibbs Sampling." In Bernardo, J., Berger, J., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 837–42. Oxford, UK: Oxford University Press. 709

Thomas, J. G., Olson, J. M., Tapscott, S. J., and Zhao, L. P. (2001). "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles." *Genome Research*, 11: 1227–1236. 708

Tusher, V. G., Tibshirani, R., and Chu, G. (2000). "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response." *Proceedings of the National Academy of Science, USA*, 98: 5116–5121. 708, 719, 722

Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2003). "Microarrays: How Many Do You Need?" *Journal of Computational Biology*, 10(3–4): 653–67. 708, 711



Gene Name	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
ADCY9	88	283	309	12	168
AGT	15091	11038	16692	15763	18128
ANXA2	311	134	378	268	118
ARHGAP6	21	-21	67	43	-8
BMP6	215	116	476	155	122
CAT	797	433	1474	415	483
CBS	14538	615	5669	4850	1284
CD34	9738	115	3272	2293	2731
CD63	8529	1518	3668	2569	316
CIQB	15076	19448	27410	14920	14653
CLDN14	11126	13568	16756	11439	15030
CSF2	17782	18112	23006	17633	17384
CSF3	16287	17926	22626	15770	16386

Figure 11: *A sketch of a microarray experiment. The mRNA in a cell is fluorescently labelled and hybridized to the microarray. After the hybridization, the intensity of each probe is captured into an image that is then processed to produce a proxy of the expression level of each gene in the target. Each microarray measures the molecular profile of a cell, and several microarray samples are needed to be able to detect the genes that have differential expression. In this figure, five microarrays were used to measure the molecular profiles of three healthy cells (Samples 1–3) and two tumor cells (Samples 4 and 5).*

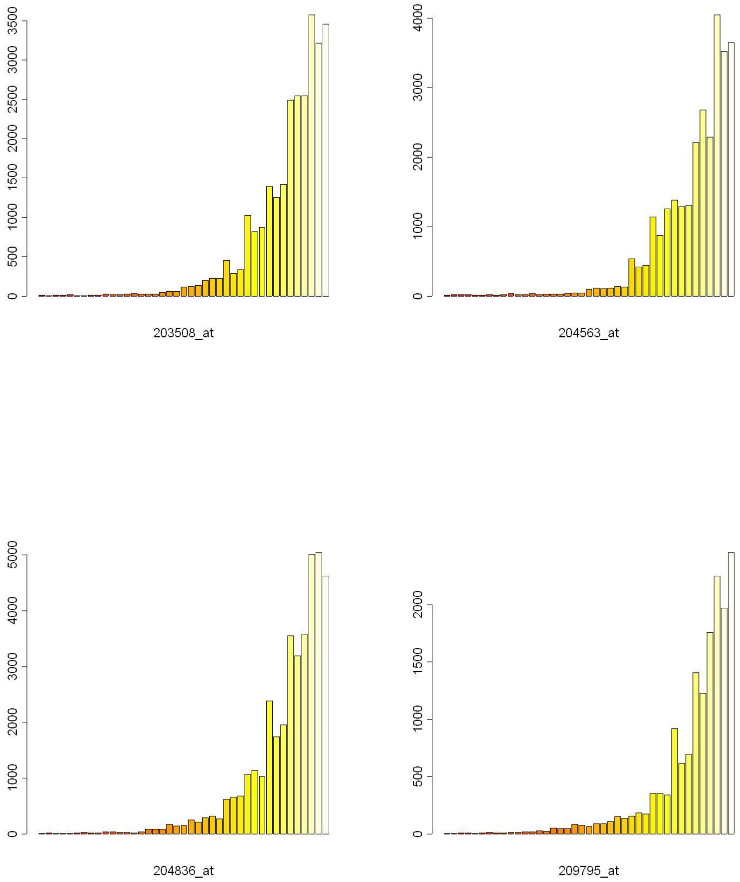


Figure 12: *Barplots of four spiked-in genes. The x-axis reports the concentrations, in increasing order from 0pM to 512pM, and the y-axis reports the intensity values (unscaled-unnormalized measurements). Each concentration was measured in three technical replicates.*

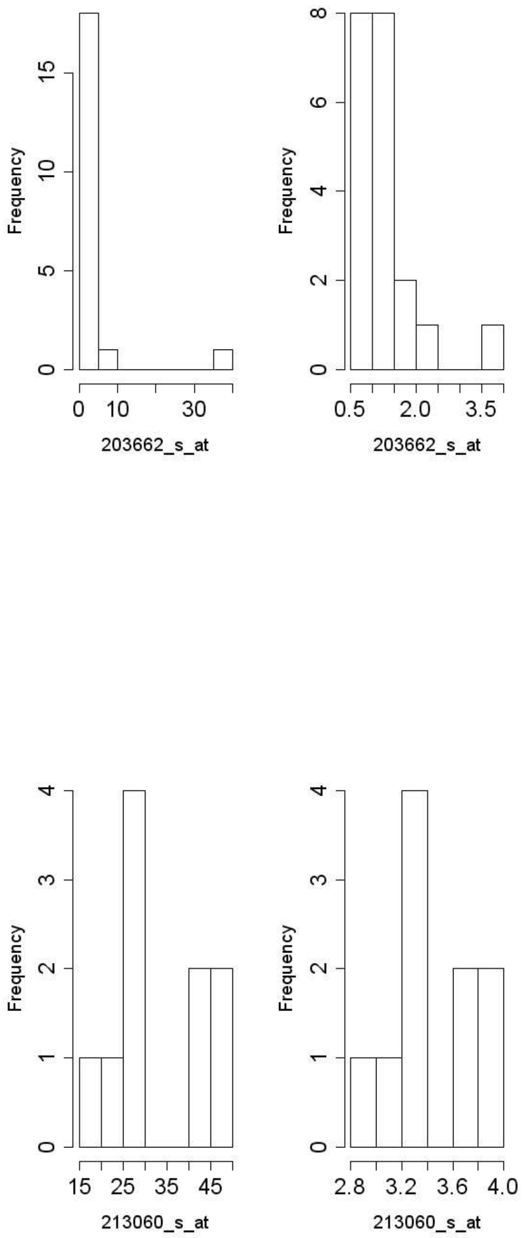


Figure 13: *Histograms of the raw expression data for the genes displayed in Figures* 5 *(top panel) and* 6 *(bottom panel). In each panel, the left histogram displays the original expression values, and the right histogram displays the expression values after the logarithmic transformation.*