Statistical analysis of microarray data: a Bayesian approach

RAPHAEL GOTTARDO*
University of Washington, Department of Statistics, Box 354322, Seattle, WA 98195-4322, USA
raph@stat.washington.edu

JAMES A. PANNUCCI, CHERYL R. KUSKE, THOMAS BRETTIN
Bioscience division, Mail Stop M888, Los Alamos National Laboratory, NM 87545, USA

SUMMARY
The potential of microarray data is enormous. It allows us to monitor the expression of thousands of genes simultaneously. A common task with microarray is to determine which genes are differentially expressed between two samples obtained under two different conditions. Recently, several statistical methods have been proposed to perform such a task when there are replicate samples under each condition. Two major problems arise with microarray data. The first one is that the number of replicates is very small (usually 2–10), leading to noisy point estimates. As a consequence, traditional statistics that are based on the means and standard deviations, e.g. t-statistic, are not suitable. The second problem is that the number of genes is usually very large (~10 000), and one is faced with an extreme multiple testing problem. Most multiple testing adjustments are relatively conservative, especially when the number of replicates is small. In this paper we present an empirical Bayes analysis that handles both problems very well. Using different parametrizations, we develop four statistics that can be used to test hypotheses about the means and/or variances of the gene expression levels in both one- and two-sample problems. The methods are illustrated using experimental data with prior knowledge. In addition, we present the result of a simulation comparing our methods to well-known statistics and multiple testing adjustments.

Keywords: Differential gene expression; Hierarchical Bayes; Multiple testing; Posterior probabilities; Replicated cDNA microarrays.

1. INTRODUCTION
One of the most important applications of arrays so far is the monitoring of gene expression (mRNA abundance). In terms of understanding the function of genes, knowing when, where and to what extent a gene is expressed is central to understanding both morphological and phenotypic differences. In the context of human health and treatment, the knowledge gained from these types of measurements can help determining causes and consequences of diseases, how drugs work, what genes might have therapeutic uses, etc.

In this paper, using a hierarchical Bayesian model, we derive four statistics that can be used to detect differential expression using cDNA microarrays. The identification of differentially expressed genes is a
question that arises in numerous experiments, when one wishes to compare a treatment sample to a control sample.

1.1 Background on cDNA microarrays

cDNA microarrays consist of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic arrayer. See Lockhart et al. (2000) for further details. A microarray works by exploiting the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. By using an array containing many DNA samples, scientists can measure—in a single experiment—the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array. In a typical microarray experiment, two samples, from control and treatment situations, are compared for gene expression. Treatment is of course taken in a broad sense to mean any condition different from the control. Both mRNA samples or targets are reverse-transcribed into cDNA, labeled using different fluorescent dyes (red and green dyes), then mixed and hybridized with the arrayed DNA sequences. Then, the red and green intensities on each spot are detected using a laser scanner. The ratio of the fluorescence intensities for each spot will be the starting point of a statistical analysis. Throughout this paper we will actually use the log ratio intensities, e.g. the log ratio of the red and green intensities \( \log(R/G) \). Many reasons exist to favor log ratios over raw intensities or log transformed intensities. For example, it is known that the red and green measurements tend to be correlated. Also the log transform makes the intensities more symmetric and stabilizes the variance. We refer the reader to Dudoit et al. (2002) and Speed et al. (2002) for further discussions on the subject. In theory, for a given gene, if the two samples are the same then we would expect the corresponding log ratio to be equal to zero. However, this is not quite the case in practice. It turns out that the behavior of the ratios depends on several external parameters, e.g. global intensities. To avoid this problem, people sometimes compare log ratios to control log ratios: see Dudoit et al. (2002), where control log ratios are obtained from slides where both control samples have been hybridized.

1.2 Statistical issues and notation

In this paper, we focus on the identification of differentially expressed genes in replicated cDNA microarray experiments. From now on, we refer to the treatment data as the log ratios of the treatment and control intensities, where the treatment and control intensities result from the intensities obtained from the red and green measurements, respectively. Similarly, we refer to the control group as the log ratios of the control and control intensities, where the two control intensities were obtained from different measurements, i.e. red and green intensities. We denote by \( X = (\log([\text{control from red}]_{ij}/[\text{control from green}]_{ij}))_{ij} \) the matrix of log ratio intensities from the control group, with \( g \) rows corresponding to the number of genes being studied and \( n_1 \) columns corresponding to the \( n_1 \) control hybridizations. We denote by \( Y = (\log([\text{treatment}]_{ij}/[\text{control}]_{ij}))_{ij} \) the matrix of log ratio intensities from the treatment group, with \( g \) rows corresponding to the number of genes being studied and \( n_2 \) columns corresponding to the \( n_2 \) treatment hybridizations.

If both control and treatment data are available, for each gene \( j, j = 1, \ldots, g \), we will test the null hypothesis \( H_j \) of equal treatment and control mean and/or variance of the expression levels. If only the treatment data are available, we will test the null hypothesis \( H_j \) of treatment mean expression levels equal to zero. That is, depending on the data available, we will either have a one-sample problem or a two-sample problem. Several methods have been proposed for both situations, e.g. \( t \)-tests. However, several problems arise with traditional statistics. Most of them being due to the lack of replications and the large amount of genes (\( \approx 10000 \)). For example, the number of replicates can be as few as two, leading to noisy estimates.
of the mean gene expression levels and standard deviation of the gene expression levels. Conversely, the number of genes is usually very large, and therefore one is faced with an extreme multiple testing problem.

The paper is organized as follows. Section 2 describes the bayesian computations that result in the posterior probabilities of a gene being differentially expressed. Section 3 introduces an algorithm to estimate the proportion of differentially expressed genes required in the bayesian computation. The data sets are presented in Section 4. In Section 5, we illustrate the use of each statistic using our experimental data. Section 6 presents a simulation study that compares our statistics to popular methods used with microarray data. Finally, Section 7 discusses our findings.

2. STATISTICAL INFERENCE

In this section, we introduce four statistics based on empirical Bayes modeling of microarray data. The first one represents the posterior probability of a given gene having a mean log ratio equal to zero in the one-sample case. The second, represents the posterior probability of a given gene having different mean log ratios in the control and treatment groups, respectively. The third represents the posterior probability of a given gene having different variance log ratios in the control and treatment groups, respectively. Finally, the last one represents the posterior probability of a given gene having different mean log ratios and/or different variance of the log ratios in the control and treatment groups. Even if the mean expression levels are the same, a change in the variance of the expression ratios could be the response to a biological event. In many biological experiments, and in gene expression experiments in particular, the control state can be more homogeneous when compared to the activated state(s). Consider for example, a collection of human cells grown under standard conditions. The cells are sampled at $t = 0$ (control state) and then again at regular intervals post exposure to a pathogen (activated states). The process of infection involves at the simplest level the number of human cells and the number of pathogenic microbes. This alone is difficult to replicate. Thus, while we can replicate the control state with little variation, replicating the activated state is much more difficult due to the uncontrolled variability in the infection process. An indication of a change in variance can point to those genes that are actively changing. A detailed description and the application of the B3 and B4 statistic in this very setting is the topic of work in progress. Generally speaking, it could be helpful in identifying additional sources of variation due to the experimental process. In each case, we combine data from all genes to estimate the parameters of the prior distributions. As a consequence, the sample mean and the sample standard deviation are regularized by the estimates of the hyperparameters of the prior distribution. Also, by specifying a prior probability of a gene being differentially expressed, we diminish the issue of multiple testing.

We adopt a conjugate prior approach (Robert, 2001, page 114), so that the posterior probabilities can be exactly computed (see the Appendix). In each case the data are assumed to be normally distributed with normal priors for the means, inverse gamma priors for the variances and the genes being independent. We would like to emphasize the fact that one could use a more elaborate, and perhaps more realistic, model. The parameters could then be estimated as part of the model using Markov Chain Monte Carlo (MCMC) methods, and inference could be drawn from the posterior distribution, e.g. posterior means. However, our approach allows us to derive analytical formulae that can be used directly without the use of MCMC. They are computationally cheaper, easier to understand and easier to implement. See Section 7 for further criticisms on the model and the choice of the prior.

From now on, we denote by $\Gamma(a)$, the gamma function with parameter $a$. The proportion of expressed genes is denoted by $p$. In practice the proportion of differentially expressed genes is unknown. In this section we assume that $p$ is known and therefore fixed. In the next section, we introduce a simple and very efficient algorithm that allows us to estimate $p$ (Section 3). However, in a more fully bayesian analysis, the prior for $p$ would be specified and the resulting posterior would capture uncertainty regarding $p$. In
this section, we only provide the main steps in the derivation of the statistics. The reader should refer to the Appendix for further details.

Let \( D_g \sim B(1, p) \) denote the Bernoulli random variable indicating whether the gene \( g \) is differentially expressed. In the one-sample case, expressed will always refer to a mean log ratio intensities different from zero. In the two-sample case, expressed could refer to different means and/or different variances in the control and treatment groups. For each gene \( g \) we are interested in the posterior probability of differential expression given the data, i.e. \( \text{Prob}(D_g = 1|X, Y, p) \), where \( X \) and \( Y \) are the control and treatment data as defined in the previous section. We assume that the measurements of a given gene \( g \), \( X_{gj} \) and \( Y_{gj} \) are normally distributed with means \( \mu_{1g} \), \( \mu_{2g} \) and variances \( \tau_{1g} \), \( \tau_{2g} \), respectively. Furthermore, we let the genes be independent. We will write \( (X_{gj}|\mu_{1g}, \tau_{1g}) \sim N(\mu_{1g}, \tau_{1g}) \) and \( (Y_{gj}|\mu_{2g}, \tau_{2g}) \sim N(\mu_{2g}, \tau_{2g}) \). As we mentioned previously, we let the prior means be normal and the prior variances be inverse gamma, with hyperparameters depending on the fact that the gene is differentially expressed or not. In other words, the prior means and prior variances will have different distributions when conditioning on \( (D_g = 0) \) and \( (D_g = 1) \). The genes will be separated into two groups, expressed \((D_g = 0)\) and non-expressed \((D_g = 1)\), of respective proportion \( p\% \) and \( (1 - p)\% \). If expressed is related to the mean, the separation into two groups is based on \( t \)-statistics. If expressed is related to the variance, the separation will be based on sample variance ratios between the control and the treatment groups (see the Appendix for further details).

Then, the hyperparameters are estimated by the method of moments individually in each group (expressed and non-expressed). Replacing the value of the hyperparameters in the formulae presented in this section will lead to the numerical values of the posterior probabilities \( \text{Prob}(D_g = 1|X, Y, p) \). Then, one can declare a gene differentially expressed if its posterior probabilities is greater than some predefined cutoff. Throughout this paper, we will use 0.95 as cutoff value.

### 2.1 One-sample case

Let \( D_g \sim B(1, p) \), be the Bernoulli random variable indicating whether the gene \( g \) is differentially expressed \((\mu_g \neq 0)\), i.e. \( \text{Prob}(D_g = 1) = p \), where

\[
D_g = \begin{cases} 
0 & \text{if } \mu_g = 0, \\
1 & \text{if } \mu_g \neq 0 
\end{cases}
\]  

(2.1)

and \( \mu_g \) denotes the mean expression level for the gene \( g \). For each gene \( g \) we are interested in knowing if the gene is differentially expressed given the data, i.e. \( \text{Prob}(D_g = 1|Y, p) \). The posterior probability can be computed exactly (see the Appendix on \( B_1 \)),

\[
B_1 = \text{Prob}(D_g = 1|Y, p) = \left( 1 + \frac{1 - p}{p} \sqrt{2} \frac{\Gamma(v_a)\Gamma(v_0 + n_2/2)}{\Gamma(v_0)\Gamma(v_a + n_2/2)} \frac{\tau_{0a}^n}{\tau_{a}^n} (\frac{n_a - 2}{2})^{v_0/2} \left( \frac{n_a - 2}{2} \right)^{v_0/2} \right)^{-1}
\]

(2.2)

where \( S_g^2 = \sum_{i=1}^{n_2} (Y_{g_i} - \bar{Y}_g)^2/(n_2 - 1) \) and \( S_{g0}^2 = \sum_{i=1}^{n_2} (Y_{g_i} - 0)^2/n_2 \). \( v_0 \) and \( \tau_0 \) are the hyperparameters in the inverse gamma prior for the variances of the genes that are not differentially expressed. Similarly, \( v_a \) and \( \tau_a \) are the hyperparameters in the inverse gamma prior for the variances of the genes that are differentially expressed (last section in the Appendix).
2.2 Two-sample case, difference in mean expression

Here, $D_g \sim B(1, p)$ denotes the Bernoulli random variable indicating whether the gene $g$ is differentially expressed between the control and treatment conditions ($\mu_{g1} \neq \mu_{g2}$), i.e. $\text{Prob}(D_g = 1) = p$, where

$$D_g = \begin{cases} 0 & \text{if } \mu_{g1} = \mu_{g2}, \\ 1 & \text{if } \mu_{g1} \neq \mu_{g2}. \end{cases}$$

(2.3)

$\mu_{g1}$ and $\mu_{g2}$ denote the mean expression levels for the gene $g$ in the control and treatment groups, respectively. Again, for each gene $g$ we are interested in knowing if the gene is differentially expressed, i.e. $\text{Prob}(D_g = 1 | \mathbf{X, Y, p})$. The posterior probability can be computed exactly (see the Appendix on $B_2$),

$$B_2 = \text{Prob}(D_g = 1 | \mathbf{X, Y, p}) = \begin{pmatrix} 1 + \frac{1 - p}{p} \sqrt{\frac{\Gamma(\nu_0 + \frac{n_1 + n_2}{2}) \Gamma(\nu_0) \Gamma(\nu_a + \frac{n_1 + n_2}{2}) \Gamma(\nu_a)}{\nu_0 \nu_a}} \\ \nu_0 \nu_a \end{pmatrix}^{-1} \left[ \nu_0 + \frac{1}{2}(n_1 + n_2) \right]^{\nu_0 + \frac{n_1 + n_2}{2}} \left[ \nu_0 + \frac{1}{2}(n_1 + n_2 - 1) \right]^{\nu_a + \frac{n_1 + n_2}{2}}$$

(2.4)

where $S_{g1}^2 = \frac{\sum_{i=1}^{n_1} (X_{gi} - \bar{X}_g)^2}{(n_1 - 1)}$, $S_{g2}^2 = \frac{\sum_{j=1}^{n_2} (Y_{gj} - \bar{Y}_g)^2}{(n_2 - 1)}$, and $S_g^2 = \frac{\sum_{i=1}^{n_1} (X_{gi} - (n_1 \bar{X}_g + n_2 \bar{Y}_g)/(n_1 + n_2))^2 + \sum_{j=1}^{n_2} (Y_{gj} - (n_1 \bar{X}_g + n_2 \bar{Y}_g)/(n_1 + n_2))^2}{(n_1 + n_2 - 1)}$. $v_0$ and $\tau_0$ are the hyperparameters in the inverse gamma prior for the variances of the genes that are not differentially expressed. Similarly, $v_a$ and $\tau_a$ are the hyperparameters in the inverse gamma prior for the variances of the genes that are differentially expressed (see last section in the Appendix).

2.3 Two-sample case, difference in variance of the expression levels

Here, $D_g \sim B(1, p)$ denotes the Bernoulli random variable indicating whether the gene $g$ is differentially expressed between the control and treatment conditions ($\tau_{g1} \neq \tau_{g2}$), i.e. $\text{Prob}(D_g = 1) = p$, where

$$D_g = \begin{cases} 0 & \text{if } \tau_{g1} = \tau_{g2}, \\ 1 & \text{if } \tau_{g1} \neq \tau_{g2}. \end{cases}$$

(2.5)

$\tau_{g1}$ and $\tau_{g2}$ denote the variance of the expression levels for the gene $g$ in the control and treatment groups, respectively. Again, for each gene $g$ we are interested in knowing if the gene is differentially expressed, i.e. $\text{Prob}(D_g = 1 | \mathbf{X, Y, p})$. The posterior probability can be computed exactly (see the Appendix on $B_3$),

$$B_3 = \text{Prob}(D_g = 1 | \mathbf{X, Y, p}) = \begin{pmatrix} 1 + \frac{1 - p}{p} \Gamma(\nu_0 + n_1 \tau_a + \frac{n_1 + n_2}{2}) \Gamma(\nu_0 + n_2 \tau_a + \frac{n_1 + n_2}{2}) \Gamma(\nu_0) \Gamma(\nu_a + \frac{n_1 + n_2}{2}) \Gamma(\nu_a) \\ \nu_0 \nu_a \end{pmatrix}^{-1} \left[ \nu_0 + \frac{1}{2}(n_1 + n_2 - 1) \right]^{\nu_0 + \frac{n_1 + n_2}{2}} \left[ \nu_0 + \frac{1}{2}(n_1 - 1) \right]^{\nu_0 + \frac{n_1 + n_2}{2}}$$

(2.6)

where $S_{g1}^2$ and $S_{g2}^2$, are as defined in Section 2.2. $\nu_0$ and $\tau_0$ are the hyperparameters in the inverse gamma prior for the variances of the genes that are not differentially expressed. Similarly, $\nu_{a1}$, $\tau_{a1}$ and $\nu_{a2}$, $\tau_{a2}$ are the hyperparameters in the inverse gamma prior for the variances of the genes that are differentially expressed in the control and treatment groups, respectively (see the Appendix).
Two-sample case for general difference in expression

Here, \( D_g \sim B(1, p) \) denotes the Bernoulli random variable indicating whether the gene \( g \) is differentially expressed between the control and treatment condition (\( \mu_{g1} \neq \mu_{g2} \) or \( \tau_{g1} \neq \tau_{g2} \)), i.e. \( P(D_g = 1) = p \), where

\[
D_g = \begin{cases} 
0 & \text{if } \mu_{g1} = \mu_{g2} \text{ and } \tau_{g1} = \tau_{g2}, \\
1 & \text{if } \mu_{g1} \neq \mu_{g2} \text{ or } \tau_{g1} \neq \tau_{g2}.
\end{cases}
\]

(2.7)

\( \mu_{g1} \) and \( \mu_{g2} \) denote the mean expression levels for the gene \( g \) in the control and treatment group, respectively. Similarly, \( \tau_{g1} \) and \( \tau_{g2} \) denote the variance expression levels for the gene \( g \) in the control and treatment group, respectively. Again, for each gene \( g \), we are interested in knowing if the gene is differentially expressed, i.e. \( \text{Prob}(D_g = 1 | X, Y, p) \). The probability can be computed exactly (see the Appendix on \( B_4 \)).

\[
B_4 = \text{Prob}(D = 1 | X, Y, p)
\]

\[
= \left( 1 + \frac{1 - p}{p} \sqrt{\frac{2}{\nu_1 + \frac{n_1 + \nu_2}{2}}} \Gamma(v_a1 + \frac{\nu_1}{2}) \Gamma(v_a2 + \frac{\nu_2}{2}) \right)^{-1} \left[ \frac{\nu_1 + a_1 - 1}{\nu_1} S_g^2 \right]^{\nu_1 + a_1 + 1} \left[ \frac{\nu_2 + a_2 - 1}{\nu_2} S_g^2 \right]^{\nu_2 + a_2 + 1} - 1 \left[ \frac{\nu_1 + a_1 + 1}{\nu_1} S_g^2 \right]^{\nu_1 + a_1 - 1} \left[ \frac{\nu_2 + a_2 + 1}{\nu_2} S_g^2 \right]^{\nu_2 + a_2 - 1}
\]

(2.8)

where \( S_g^2 \) and \( S_g^2 \) are as defined in Section 2.2. \( \nu_0 \) and \( \tau_0 \) are the hyperparameters in the inverse gamma prior for the variances of the genes that are not differentially expressed. Similarly, \( \nu_{a1} \) and \( \nu_{a2} \), \( \tau_{a1} \) and \( \tau_{a2} \) are the hyperparameters in the inverse gamma prior for the variances of the genes that are differentially expressed in the control and treatment group, respectively (see last section in the Appendix).

The underlying hierarchical Gaussian model we introduced here is similar to models developed by Baldi et al. (2001) and Lonnstedt et al. (2002). Baldi et al. model log-expression values by independent normal distributions, parametrized by corresponding means and variances with hierarchical prior distributions. However, they recommend, alternatively to a full bayesian treatment, an intermediate solution. They recommend using a regularized t-test, where the sample variance is replaced by the posterior mean of the variance in the hierarchical model. The result is a weighted average between the prior variance and the sample variance. This version of the t-test is implemented in a web-server called Cyber-T (Baldi et al., 2001). The sample variance is replaced by the following point estimate:

\[
\sigma^2 = \frac{\nu_0 \sigma_0^2 + (n - 1)s^2}{\nu_0 + n - 2}
\]

(2.9)

where \( \nu_0 + n \geq 2 \), \( \sigma_0 \) is the prior standard deviation and \( s \) is the sample standard deviation. For further details on how to compute \( \nu_0 \) and \( \sigma_0 \) we refer the reader to the original paper (Baldi et al., 2001).

Lonnstedt et al. model log ratio expression values as a mixture of normal distributions. Then, they base their analysis on the log odds ratio of a gene being differentially expressed. However, they do not propose a rule for deciding if a gene is differentially expressed, but they regard their method as a way of ranking genes. Moreover, they only derived a statistic for the one-sample problem. In Sections 5 and 6 we compare \( B_1 \) and \( B_2 \) to the regularized t-test of Baldi et al. on both experimental and synthetic data. Since Lonnstedt et al. recommend their method for ranking, we do not feel confident in deciding on a cutoff value. As a consequence, we decided not to apply it to the experimental data. However, using a ROC (receiving operating characteristic) curve, it is still possible to compare their statistic to \( B_1 \) on the synthetic data (Section 6).
3. Estimating \( p \), the portion of differentially expressed genes

In this section, we present a very simple and efficient algorithm that can be used to estimate \( p \) for each of the \( B \) statistics. One of the strong points of the statistics described in Section 2 is that they are relatively robust to the prior value of \( p \). Let us consider an example where the number of genes is \( N = 20\,000 \) and the true value of \( p \) is 0.01. If one chooses a ‘bad’ value of \( p \), let us say 0.1, the number of differentially expressed genes detected (e.g. posterior probability greater than 0.95), namely \( d \), will be inflated. However, \( d \) will not be \( 20\,000 \times 0.1 = 2000 \). It should lie between 100 and 2000. As a consequence, \( \hat{p} = d/N \) would be a better estimate for \( p \) than 0.1. Then, one can use an iterative process to get closer to the true value of \( p \). Algorithm 1 describes the procedure with further details.

\[ \text{Algorithm 1 Estimate the proportion } p \text{ of differentially expressed genes} \]

```
1: Start with an initial value \( p^{(0)} > 0 \) (initial guess)
2: while \( p^{(m+1)} \neq p^{(m)} \) and \( p^{(m)} > 0 \) do
3: \quad d \leftarrow \#\{\text{Prob}(D_i = 1|X, Y, p) > 0.95 : 1 \leq i \leq N\} \text{ (number of differentially expressed genes detected)}
4: \quad \text{Up-date } p^{(m)} \text{ to } p^{(m+1)} = d/N \text{ (proportion of differentially expressed genes detected)}
5: end while
```

Let us define \( S(p) = \#\{\text{Prob}(D_i = 1|X, Y, p) > 0.95 : 1 \leq i \leq N\}/N \), \( S(p) \) is a mapping from \([0, 1]\) onto \([0, 1]\). If the value of \( p \) is nearly correct, \( S(p) \) should be close to \( p \). If possible, we would like to have \( S(p) = p \), i.e. we would like to find the fixed point(s) of \( S \). Since \([\{0, 1\},|.|]\) is a complete metric space, if \( S \) satisfies the conditions of the contraction mapping principle (McDonald et al., 1999, page 500), Algorithm 1 would converge to the unique fixed point of \( S \) no matter what the starting value is. The conditions are clearly not satisfied since \( S \) is not continuous and has at least two fixed points (0 and 1). However, the converse of the contraction mapping principle is false. That is, even if the conditions are not satisfied, the algorithm might converge. In theory, if the function \( \delta \) only depends on \( p \) through \((1 - p)/p \), then it should be increasing in \( p \). Therefore the sequence defined by \( p^{(0)} = p_0 \) and \( p^{(m+1)} = S(p^{(m)}) \) should be monotonic bounded above by \( 1 \) and below by \( 0 \) and then convergent. By robustness, if the starting value \( p^{(0)} \) is smaller than the true value, then \( p^{(1)} \) should be larger than \( p^{(0)} \). Similarly, if \( p^{(0)} \) is larger than the true value, then \( p^{(1)} \) should be smaller than \( p^{(0)} \). Then there should be at most one \( \hat{p} \) such that \( S(\hat{p}) = \hat{p} \) and \( 0 < \hat{p} < 1 \), see Figure 1. Consequently, the algorithm should converge to the unique limit independently of the starting value. Figure 1 shows the graph of \( S(p) \) for the experimental data described in Section 4, using \( B_2 \). It shows two different sequences converging to the limit using different starting values. In practice, \( S(p) \) depends on \( p \) through the hyperparameters as well (see the Appendix). As a consequence, when the number of differentially expressed genes is very small (\( p \ll 0.01 \)), the estimation might be sensitive to changes in \( p \). In some rare cases, Algorithm 1 will not converge. It will oscillate between a finite number of values close to the true value of \( p \). In most situations, the convergence of Algorithm 1 is independent of the starting value. One can run the algorithm with different starting values and check that the limits are the same. Even though the limits might not be the same, they should all be very close to the true value of \( p \). Note that the starting value should not be too small or too large, to avoid getting too close to the fixed points 0 or 1. For example, a good starting value would be 0.2. The convergence of Algorithm 1 is rather intuitive and can be understood from Figure 1. The algorithm depends on the estimation of the hyperparameters which is different for each dataset. Therefore, it is impossible to give general conditions for convergence. We have applied our algorithm to a large number of datasets, and so far it has always converged. However, it was possible to
generate synthetic data with a very small number of differentially expressed genes and a small number of replicates where the convergence failed.

4. DESCRIPTION OF THE EXPERIMENTAL DATA

The experiment was designed to explore the effects of CO$_2$ on gene expression. About 1700+ *Bacillus anthracis* chromosomal and plasmid genes were represented on the microarray. *Bacillus anthracis* is the causative agent of anthrax. Of particular interest was the transcriptional response of the bacteria to CO$_2$. Earlier work (Dai et al., 1995; Koehler et al., 1994; Pezard et al., 1991; Leppla, 1982) has shown that CO$_2$ concentrations above the atmospheric level induces the toxin genes responsible for the onset of the disease state of anthrax. This experiment could potentially produce new information regarding the effects of CO$_2$ on gene expression. The replicates analyzed in this study represent technical replication; each replicate originated from the same culture and was spotted on a different array.

From now on, we will refer to the control data as the log ratios from the control measurements (absence of CO$_2$), i.e. the matrix $(\log(\text{control from red}_{ij})/\log(\text{control from green}_{ij}))_{ij}$, with 1710 rows corresponding to the number of genes being studied and 12 columns corresponding to the 12 control hybridizations. Similarly, we will refer to the treatment data as the log ratios from the treatment (presence of CO$_2$) and control (absence of CO$_2$) measurements, i.e. the matrix $(\log(\text{treatment}_{ij})/\log(\text{control}_{ij}))_{ij}$, with 1710 rows corresponding to the number of genes being studied and 12 columns corresponding to the 12 treatment hybridizations.

5. ILLUSTRATION OF THE STATISTICS

In this section we illustrate the statistics $B_1$, $B_2$, $B_3$ and $B_4$ on the data presented in Section 4. This data set is particularly interesting as we know four genes that should be differentially expressed (see Section 4). We refer to those genes as the toxin genes. In the one-sample case, those genes should have a mean expression level significantly different from zero. In the two-sample case, the mean expression level in the treatment and control group should be significantly different. We also compare our statistics to classical methods such as one- and two-sample t-tests. For those statistics, the $p$-values will be adjusted using Benjamini’s (Benjamini et al., 1995), and Holm’s (Holm, 1979) multiple testing adjustments. Holm’s $p$-
value adjustment controls the Family Wise Error Rate (FWER), which is the probability of making one or more Type I errors among the hypothesis. Benjamini et al. showed that the FWER approach can be really conservative, especially when the number of hypotheses tested is large. Alternatively, they proposed a more powerful sequential procedure that controls the false discovery rate (FDR), which is the proportion of false positives among all rejected hypotheses. We also illustrate the regularized $t$-test of Baldi et al. (2001). The corresponding $p$-values will be adjusted using Benjamini’s method. We used (2.9) to regularize the $t$-test. In their paper, Baldi et al. recommended keeping $n + v_0 = 10$ where $n$ is the number of replicates. Since here we have 12 replicates, we chose $v_0$ such that $n + v_0 = 20$, i.e. $v_0 = 8$. We estimated $\sigma_0$ from the standard deviations of all the genes. As mentioned previously, we did not use the Lonnstedt et al. (2002) statistic because we did not feel confident in choosing a cutoff to declare a gene differentially expressed. In this section, using our bayesian statistics we declare a gene differentially expressed if its posterior probability is greater than 0.95. Using other statistics, we declare a gene differentially expressed if the associated $p$-value is less than 0.05. Before performing any analysis, the data were normalized using a non-linear normalization technique, first introduced by Yang et al. (2002), to allow between-slide comparisons.

### 5.1 The $B_1$ statistic

In order to use the one-sample formula, we only used the treatment data. When using the bayesian statistic $B_1$, we estimated $p$, the prior probability of a gene being differentially expressed, to 0.02. We used Algorithm 1 introduced in Section 3 to estimate $p$. To be consistent, when using Benjamini’s false discovery rate (FDR) approach, we fixed the proportion of true null hypotheses to 0.98 and the expected FDR to 0.05. See Benjamini et al. (1995), Storey (2001a,b) and Efron et al. (2001) for further details on the application of the FDR to microarray experiments. For each method, we graphed the absolute sample means against the sample standard deviations for all genes, then highlighted the differentially expressed genes detected with each method. This representation allows us to observe if a gene was declared differentially expressed because of a large mean, small standard deviation, or both. Figure 2 shows the genes detected with each method. As expected, the method based on the $t$-statistics and Holm’s adjustment is very conservative and only detects five genes. Moreover, it does not detect any of the toxin genes (orange circles), although they have large mean expression levels. The FDR is clearly more powerful with 126 genes detected. It detects all the toxin genes. However, many of the genes detected seem to be false positive, especially when they have small standard deviations and small means. Overall, the expected FDR, fixed to 0.05 in the procedure, seems to be inflated, perhaps because many of the genes have very small standard deviations, though more probably because Benjamini’s adjustment assumes that the genes are independent. It is well known that the group of genes might be correlated. As a consequence, the FDR might not be controlled at the 0.05 level. Storey et al. (2001) introduced a procedure to estimate the FDR under dependence. However, the technique is computationally intensive and the dependence is not easy to estimate. The regularized $t$-test with Benjamini’s adjustment declares 86 genes differentially expressed. It seems to reduce the number of false positive genes, e.g. genes with small standard deviations. Alternatively to $t$-statistics, we also tried one-sample Wilcoxon statistics and obtained similar results (not shown). Finally, the statistic $B_1$, with $p$ estimated to 0.02, detects 34 genes, which is close to what we expected ($0.02 \times 1700$). The multiple testing issue is taken care of. Moreover, because of the prior distribution put on the variances, we avoid detecting genes with very small standard deviations and small means, i.e. potentially false positives. Figure 2 shows that the toxin genes have large mean expression levels, as we expected. However, they also have large associated standard deviations. In a similar fashion, genes that have large mean expression levels, also have large variances. Therefore, using a $t$-statistic, those genes will be more penalized than others. Using our statistic, $B_1$, we have to estimate the hyperparameters for the variance of the genes that are differentially expressed (see the Appendix). Since we do not know
which ones are differentially expressed, we use the top 2% of the genes that have greatest absolute $t$-statistic. Even though large $t$-statistics might occur because of small standard deviations, some are led by large sample means. Therefore, the top 2% should contain a large number of genes with large sample means. Using this process, $B_1$ will incorporate the fact that potentially expressed genes have greater standard deviations. As a consequence, we detect two genes (Figure 2) that have large mean expression levels and associated larger standard deviations that are not detected by any other methods.

5.2 The $B_2$ statistic

When using the bayesian statistic $B_2$, we estimated $p$, the prior probability of a gene being differentially expressed, to 0.008. Again the estimation was done using Algorithm 1, Section 3. To be consistent, when using Benjamini’s FDR approach, we fixed the proportion of true null hypotheses to 0.992 and the expected FDR to 0.05. This time, we graphed the numerators against the denominators of the two-sample $t$-statistics for all genes, then highlighted the differentially expressed genes detected. The absolute numerator of the two-sample $t$-statistic is the difference in mean expression levels and the denominator a measure of the variance expression levels. Consequently, using this representation allows us to observe if a gene was declared differentially expressed because of a large mean difference, small standard deviations, or both. Figure 3 shows the genes detected with each method. The method based on the $t$-statistics and Holm’s adjustment is less conservative with 11 genes detected. It also detects three of the toxin genes. Again, one of them is not detected because of the large associated standard deviation. The adjustment based on the FDR is clearly more powerful with 27 genes detected and detection of all the toxin genes. Again, because of the dependence structure between the genes, we cannot be certain that the FDR is controlled at the 0.05 level. This time, since the number of genes detected by the regularized $t$-test is the same, we did not highlight them. In the two-sample case, the denominator of the $t$-statistic depends on both variances from the control and the treatment. Therefore, it is less likely to have a very small denominator, and the regularization is not as effective. This observation will be confirmed in the next section, when applied to the synthetic data. We also compared $B_2$ to the non-parametric technique described in Dudoit et al. (2002). They estimate the $p$-values associated with the two-sample $t$-statistics by permutation and adjust the $p$-values using the Westfall and Young algorithm, (Westfall et al., 1993). Since their procedure controls the FWER, and the number of replicates is large, the results (not shown here) are almost identical.
to the ones obtained using $t$-tests and adjusting the $p$-values using Holm’s. Again, we tried two-sample Wilcoxon statistics and obtained similar results (not shown). Finally, the statistic $B_2$, with $p$ estimated to 0.008, detects 13 genes, which includes the four toxin genes. Because of the prior distribution put on the variances, the genes detected by the bayesian approach have large differences in mean and associated larger variances. Again, we detect one gene with large mean difference, not detected by any other methods. The behavior of this gene is similar to the behavior of the toxin genes. It could be a true differentially expressed gene.

5.3 The $B_3$ and $B_4$ statistics

Now, we use the statistics $B_3$ and $B_4$ to detect change in mean expression levels and/or variance expression levels between the control and treatment group. To the best of our knowledge, there are no other statistics to perform similar tasks in the context of microarray data. It is hard to evaluate their performances. We still compare the method based on $B_3$ to a simple $F$ test, since they both detect changes in variances. In this case, to be able to visualize both changes in mean expressions and variance of the expressions we plotted the absolute mean differences against the ratio of the variances. When forming the ratios, we always put the largest variance on top, i.e. all ratios are greater than 1. Again, we estimated the true value of $p$ using Algorithm 1. The proportion of differentially expressed genes for $B_3$ and $B_4$ were estimated to 0.012 and 0.07 respectively, using Algorithm 1. Figure 4 shows the genes differentially expressed detected by $B_2$, $B_3$ and $B_4$ and the $F$ statistic. It demonstrates the potential of the statistics $B_3$ and $B_4$. $B_2$ only detects genes with large mean difference whereas $B_3$ detects genes with standard deviation ratios. Conversely, $B_4$ detects both. Note that the genes detected by $B_4$ are not the union of the ones detected by $B_2$ and $B_3$. A moderate change in the mean combined to a moderate change in the variance might be significant for $B_4$ even though each change separately is not. The toxin genes (Figure 4) are detected by both $B_2$ and $B_3$ since they show great change in mean expression level. However, they do not show changes in the variances of the expression levels. This does not mean that interesting genes do not show changes in the variance of the expression levels. We are currently investigating some of the genes that showed great changes in the variance of the expression levels. Finally, the $F$ test statistic clearly detects the genes with large ratios, i.e. potentially different variances. The adjustment based on the FDR is clearly more powerful.
that the adjustment based on the FWER. Because of the dependence between genes, the number might be inflated. Moreover, it is well known (Miller, 1997) that the $F$ statistic is very sensitive to departure from normality. Of course, the $B_i$ statistics also depend on the fact that the data are sampled from a normal distribution, but they are more robust: see the discussion at the end of the next section.

6. SIMULATION

In this section we compare classical statistical methods for microarray data to our $B_i$ statistics using randomly generated data sets. The comparisons are made in terms of false positives and true positives among the genes detected differentially expressed by each method. When comparing $B_1$ to the log odds ratio of Lonnstedt et al. (2002), we use a ROC curve.

6.1 The $B_1$ statistic

The treatment data set, from our data, was used as a model for the simulated treatment data sets. Similarly, the control data were used as a model for the control data sets. For each sample size, $n = 3, \ldots, 15$, we generated pairs of data sets (control and treatment) with 10,000 genes. The log ratio intensities for the genes were simulated independently, from a normal distribution. Even though genes might be biologically related and then correlated, it is hard to evaluate the degree of correlation. Consequently, we decided to generate uncorrelated genes. The replicates for each gene were simulated as independent normal observations. For the control data sets we generated all the genes with mean zero. For each treatment data set, we fixed the number of differentially expressed genes to 100, i.e. the proportion of genes with mean different from 0 (and then different from the control) is 0.01. Among those 100, 50 were generated with a mean uniformly distributed between 0.1 and 1.5, and 50 were generated with a mean uniformly distributed between 1.5 and 3; that is, it would resemble a microarray data set with high and moderate expressed genes. The standard deviations were uniformly generated between 0.05 and 1, independently from the mean. Figure 5 shows an example of a randomly generated pair of data.

Since the data were generated from a normal distribution, we decided to compare our bayesian statistics ($B_1$ and $B_2$ only) to classical $t$-tests. The $t$-test is actually the uniformly most powerful (UMP) test for a single hypothesis (Shao, 1999). We also used the regularized $t$-test (Baldi et al., 2001). Again, when applying the regularized $t$-test we fixed $n + v_0 = 20$ where $n$ is the number of replicates (2.9).
main difficulty remains with multiple testing adjustments. As in Section 5, we used both Benjamini’s and Holm’s adjustments. Using $t$-tests, we declared a gene differentially expressed if its adjusted $p$-value is less than 0.05. Using our bayesian statistics, we declared a gene differentially expressed if its posterior probability is greater than 0.95. Since we generated the data, the true value of $p$ was known. Therefore, we fixed $p$ to 0.01; we also tried $p = 0.005$ and $p = 0.02$ to show that small changes in $p$ do not affect the analysis too much. Finally, we also used Algorithm 1 to estimate $p$ for each data set. To be consistent, when applying Benjamini’s adjustment, we fixed the proportion of true hypotheses to 0.99 and the expected FDR to 0.05. Figures 6 (a) and (b) show the result of the simulations for the one-sample problem, i.e. when we only use the treatment data. For those simulations we used 1000 data sets. As we expected, the method based on Holm’s adjustment is very conservative and the FDR approach is more powerful. However, both methods perform really badly when the number of replicates is small. They are not detecting anything for sample sizes less than five. The regularized $t$-test is somehow more powerful with a smaller false positive rate, though the improvement is not great. Conversely, the statistic $B_1$ performs much better. The method based on the statistic $B_1$ is clearly more powerful, especially when the number of replicates is small. The number of false positive is controlled very well, especially for the true value of $p$. Overall, the statistic $B_1$ performs best for the true value of $p$. However, slight changes in $p$ do not affect the analysis too much. Even when $p = 0.005$ or $p = 0.02$, the bayesian approach is preferable. Moreover, the curve of true positives (beta) for $B_1$ with estimated $p$ is very close to the curve when $p = 0.01$, meaning that Algorithm 1 (Section 3) performs well.

As mentioned previously, it is hard to compare our statistic $B_1$ to the log odds ratio statistic developed by Lonnstedt et al. (2002). However, it is still possible to do so, using a ROC curve. In their paper, Lonnstedt et al. (2002), used a ROC curve to compare their statistics to others such as $t$-statistic. Therefore, we feel that it would be a good measure. For a range of cutoff value for each statistic, the number of false positive and false negative genes are averaged over 1000 datasets for each cutoff values. Then the ROC curve is obtained by plotting the averaged false positives, against the averaged false negatives. In our case, we chose 20 cutoff values for both methods such that the number of expressed genes were the same. The log odds ratios were computed using the sma R package developed by Dudoit et al. The package can be downloaded at http://cran.r-project.org/src/contrib/PACKAGES.html#sma.

Figure 7 shows the results of the simulation for both methods and different sample sizes. For sample sizes 3 and 4, the statistic $B_1$ is more powerful than the log odds ratio of Lonnstedt et al. For sample
sizes greater than 5, $B_1$ is still more powerful, though the difference is almost indistinguishable. For a given gene, their statistic corresponds to the log odds ratio of differential expression which is in 1–1 correspondence with the posterior probability of differential expression. The difference when the sample size is 2 and 3 is most likely due to the parametrization and the way we estimate the hyperparameters. For example, we first use the $t$-statistics to form two groups, which might be more robust than using mean averages. When the number of replicates is large enough, the two statistics will be almost equivalent. However, Lonnstedt et al. recommended using their statistic for ranking genes by level of expression. Our method is more than a way of ranking genes. We introduced a very simple and efficient algorithm that can be used to estimate $p$, the proportion of differentially expressed genes. Having a relatively good estimate of $p$, one can declare a gene differentially expressed if its posterior probability is greater than 0.95 (as used in the estimation process, Section 3). This allows us to select a group of differentially expressed genes with good power while keeping a small error rate (Figure 6).

6.2 The $B_2$ statistic

When comparing $B_2$, the data were generated as for $B_1$, the difference being that now we use both the control and the treatment datasets. Figures 6 (c) and (d) show the results of the simulations. Again, as expected, the procedures that control the FWER are quite conservative. Benjamini’s adjustment based on the FDR is clearly more powerful. As for the one-sample problem, $B_2$ performs very well. However, the
improvement is not as large. Even though changes in $p$ do not modify the outcome too much, the change in power is larger than for $B_1$. For sample sizes less than 6 in each group, $B_2$ performs better than all others, even when $p = 0.05$. When $p = 0.01$ and $p = 0.02$, the improvement is present for all sample sizes. Finally, the curve of true positives (beta) for the bayesian approach with estimated $p$ is between the curves $p = 0.005$ and $p = 0.01$. The estimation seems a little more conservative than for the one-sample case. However, for sample sizes greater than 10, the curve with estimated $p$ gets very close to the curve $p = 0.01$. This time, the regularized t-test does not bring much improvement. This confirms what we observed in the experimental data: In the two-sample case, the regularization is not as effective.

Note that in the one-sample case and the two-sample case, all the beta curves in Figures 6 (a) and (c) seem to converge to 80 and not 100. This is an artifact due to the way we generated the data. In the treatment data some of the genes were generated with very small mean and relatively large variance (Figure 5). As a consequence, they are very hard to detect with a small number of replicates.

**6.3 The $B_3$ and $B_4$ statistics**

As we have already mentioned, because of lack of `concurrents` it is hard to evaluate $B_3$ and $B_4$. However, we still consider a simulation study to evaluate their false positive rate and compare them to a simple $F$-statistic. This time the data were generated a little differently. The log ratios were still generated as independent, from a normal distribution. The 100 differentially expressed genes were generated with both different means and different variances between the control and the treatment groups. For those 100, the mean were set equal to zero in the control group and uniformly generated between 0.1 and 1.5 in the treatment group. The variances in the control were uniformly generated between 0.05 and 0.3 and the variances in the treatment were uniformly generated between 0.3 and 0.8. The remaining genes were generated with identical means (equal to zero) and same variances in both the control and treatment group. The variances were uniformly generated between 0.05 and 0.3 and set equal in the treatment and control group. Figure 8 shows the results of the simulations when the data were generated from a normal distribution. The conclusion about the FDR and the FWER are still the same. The statistic $B_3$ performs well, though the improvement is not great. The false positive rate is very small even when $p = 0.02$. The $B_4$ statistic is clearly the most powerful here, which is what we expected since it detects both changes in

---

**Fig. 7.** Comparison of the Lonnstedt et al. log odds ratio and $B_1$. For a certain cutoff value, each method defines the numbers of false positive and false negative in each of the simulated datasets. The lines reflect the averages of these numbers over a range of cutoffs.
means and variances. The false positive rate for $B_4$ is relatively low, except when $n = 2$, which is not surprising since the estimation is only based on two observations.

It turns out that our $B_i$ statistics were the most powerful, when the true value of $p$ is used. In general, the true value of $p$ is unknown. Even though the results are not too sensitive to the choice of $p$, changes in $p$ do alter the analysis. Therefore, we recommend using Algorithm 1 to estimate the true value. The $B_i$ with estimated $p$ still perform well. Moreover, the $t$-statistic, and even more the $F$ statistic, are more dependent on the normal distribution of the data than our bayesian statistics. We performed simulations with different distributions such as Laplace and Student and obtained similar results for all of the $B_i$, whereas the methods based on the $t$-test and $F$ test were clearly less powerful with high false positive rate.

For example, the $F$ test with $p$-values adjusted by Benjamini’s applied to Laplace distributed data led to a false positive rate of 300. Furthermore, when we estimate $p$, we use an iterative process (Algorithm 1) that depends on the rejection region and the data. Whereas Benjamini’s adjustment is based on a rule that heavily depends on independence. We have also simulated dependent genes, with blocks of 50 highly dependent genes (results not shown). The power of the bayesian approach remained about the same, whereas all the other methods were significantly less powerful with a decrease of threefold or more. This showed that the bayesian statistics are robust to departure from both normal and independence assumptions.
7. DISCUSSION

We have developed a bayesian framework for the analysis of microarray data to address a number of issues present with well known statistical techniques. We used a hierarchical bayesian model with independent Gaussian modeling. Using our model, we developed four statistics representing posterior probabilities of differential expression. We started with the easiest case, when one is constrained to a one-sample problem. We then generalized it to the two-sample case. The underlying hierarchical Gaussian model is similar to models developed by Baldi et al. (2001) and Lonnstedt et al. (2002). However, we derived more general formulae that can be used in a wide range of settings. For example, the statistics $B_3$ and $B_4$ detect genes with difference in the variance of the log expression ratios which is rather new in the context of microarray. The derivations of $B_3$ and $B_4$ were so natural from the model that we decided to introduce them in this paper. We believe that they have potential. At this point, we are not exactly sure why a change in the variance of the expression levels would be the response to a biological event. This is definitively a question we would like to answer in the future. We also introduced a very efficient algorithm to estimate the proportion of differentially expressed genes $p$. Having a good estimate of $p$, we control the false positive rate while keeping a very good power. The four statistics we introduced are computed using point estimates specific to each gene and hyperparameter estimates common to all. As a consequence, our statistics compensate for small-sample bias and allows the detection of differentially expressed genes with as few as two replicates. However, because of the high variability of microarray data across replicates we would recommend the use of at least three replicates to decrease the error rate. When developing our statistics we assumed that the observations were normally distributed and the genes were independent. While the normality assumption is nearly correct for log ratios, the independence of the genes is clearly not satisfied. However, as pointed out by Lonnstedt et al. (2002), the normality and independence assumptions are not to be taken literally, rather as a tool leading to explicit formulae. Moreover, we have shown using real data sets that our statistics lead to satisfactory results. Using simulated data, we have shown that our bayesian approach is a good alternative to classical statistics, e.g. the $t$-statistic. The algorithm introduced to estimate $p$, the proportion of differentially expressed genes, gives us strong control on the number of false positives while keeping good power.

In Sections 4 and 6 we showed that the two-sample $t$-statistics were more powerful and had a smaller error rate than the one-sample version. The regularization proposed by Baldi et al. (2001) does not seem really necessary in the two-sample case. In the bayesian framework, it seems that there is no real improvement when using control ratios ($B_2$ against $B_1$). This is a consequence of the hierarchical structure of the model. For example, the variances are shrunk together, which reduces the negative effect of very small variances.

Finally, although the methods described in this paper were developed for microarray studies, we have applied them to Affymetrix chip with success. For example, one could used the log transformed expression level estimates from the multiplicative model of Li et al. (2000).

The software producing this analysis was written in C and wrapped in the R statistical language. Both the C code (Bayes_microarray_0.1.tar.gz) and the R package (amd_0.1.tar.gz) are freely available at ftp://bpublic.lanl.gov/compbio/software/.

8. ACKNOWLEDGEMENTS

We thank David Higdon for helpful discussions on this problem. Finally, we would like to thank the referees for their comments on the manuscript. This work was performed under the auspices of the Department of Energy under contract to the University of California and was supported by the Molecular Foundations of Pathogenesis project, funded by Laboratory Directed Research and Development at Los Alamos National Laboratory.
APPENDIX A

A.1 Posterior probability of differential expression

From now on we will denote by $N(a, b)$ the normal distribution with mean $a$ and variance $b$ and the corresponding density by $N(x; a, b)$. Similarly, we will denote by $IG(a, b)$ the inverse gamma distribution with mean $b/(a - 1)$ and variance $b^2/((a - 1)^2(a - 2))$ and the corresponding density by $IG(x; a, b)$, i.e. $IG(x; a, b) \propto e^{-b/\sqrt{x}}/x^{a+1}$.

Throughout the Appendix X and Y are as defined in Section 1.2.

The $B_1$ statistic. We assume that $Y_{gj}$ is normally distributed with mean $\mu_g$ and variance $\tau_g$, i.e. $(X_{gj}|\mu_g, \tau_g) \sim N(\mu_g, \tau_g)$. Let $D_g \sim B(1, p)$ be the Bernoulli random variable indicating whether the gene $g$ is differentially expressed ($\mu_g \neq 0$), as defined by (2.1).

For each gene $g$ we are interested in knowing if the gene is differentially expressed. From Bayes theorem, the probability can be computed in the following way:

$$\Pr(D_g = 1|Y, p) = \frac{p\Pr(Y_g|D_g = 1)}{p\Pr(Y_g|D_g = 1) + (1-p)\Pr(Y_g|D_g = 0)} \quad (A.1)$$

where $Y_g$ is the vector of measurements for gene $g$ and $p$ is the proportion of differentially expressed genes, i.e. $\Pr(D_g = 1|Y)$. We replaced $Y$ by $Y_g$ since the genes are independent. We need to compute $\Pr(Y_g|D_g = 0)$ and $\Pr(Y_g|D_g = 1)$. Since $Y_g$ is assumed to be absolutely continuous, those probabilities represent densities. From now on, for simplicity’s sake, we will fix the gene $g$ and omit the index $g$. When $D = 0$, the conditional densities are

$$(Y_{j}|D = 0, \tau) \sim N(0, \tau), \quad (\tau|D = 0) \sim IG(\nu_0, \tau_0)$$

where $\nu_0$ and $\tau_0$ are fixed hyperparameters. When $D = 1$, the conditional densities are

$$(Y_{j}|D = 1, \mu, \tau) \sim N(\mu, \tau), \quad (\mu|D = 1, \tau) \sim N(\mu_a, \kappa \tau), \quad (\tau|D = 1) \sim IG(\nu_a, \tau_a)$$

where $\kappa$, $\mu_a$, $\nu_a$ and $\tau_a$ are fixed hyperparameters. Then

$$\Pr(Y|D = 0) = \int \prod_{j=1}^{n_0} N(Y_{j}; 0, \tau)IG(\tau; \nu_0, \tau_0)d\tau \quad (A.2)$$

where the integration is performed over all possible values of $\tau$. The integration of the density is easily performed by identifying the posterior inverse gamma of $\tau, IG(\nu_0 + n_2/2, \tau_0 + \sum_{j=1}^{n_0} Y_{j}^2/2)$. Similarly,

$$\Pr(Y|D = 1) = \int \prod_{j=1}^{n_2} N(Y_{j}; \mu, \tau)N(\mu; \mu_a, \kappa \tau)IG(\tau; \nu_a, \tau_a)d\mu d\tau \quad (A.3)$$

where the integration is performed over all possible values of $\tau$ and $\mu$. Again, the integration of the densities is performed by identifying the posterior normal of $\mu|\tau, N(\sum_{j=1}^{n_2} (Y_{j} + \mu_a/\kappa)/(n_2 + 1/\kappa), \tau/(n_2 + 1/\kappa)), \tau|\nu_a + n_2/2, \nu_a + \sum_{j=1}^{n_2} Y_{j}^2 + \mu_a^2/\kappa - (\sum_{j=1}^{n_2} Y_{j} + \mu_a/\kappa)^2/(n_2 + 1/\kappa))/2)$. 

Then, one can explicitly compute the posterior probability of a gene being expressed as

\[
\text{Prob}(D = 1|Y, p) = \left(1 + \frac{1 - p}{p} \sqrt{\kappa n_2 + 1} \frac{\Gamma(v_0 + \frac{n_2}{2}) \frac{v_0}{\tau_0}}{\Gamma(v_0) \frac{v_0}{\tau_0} + \frac{1}{\tau_0} \frac{v_0}{\tau_0} + \frac{n_2}{2}} \right)^{-1}
\]

Replacing \( \mu_a \) and \( \kappa \) by the estimates \( \bar{Y} \) and \( 1/n_2 \) leads to (2.2). Note that \( \kappa \) and \( \mu_a \) are hyperparameters that are common to all the genes. However, replacing them by the sample mean and the inverse of the number of observations significantly simplifies the formula. The important feature of the model, i.e. the shrinkage of the variance, is conserved.

The \( B_2 \) statistic. We assume that \( X_{gj} \) is normally distributed with mean \( \mu_g \) and variance \( \tau_g \), i.e. \( (X_{gj}|\mu_g, \tau_g) \sim N(\mu_g, \tau_g) \). Similarly \( Y_{gj} \) is assumed normally distributed with mean \( \mu_g \) and variance \( \tau_g \), i.e. \( (Y_{gj}|\mu_g, \tau_g) \sim N(\mu_g, \tau_g) \). Let \( D_g \sim B(1, p) \), be the Bernoulli random variable indicating whether the gene \( g \) is differentially expressed between the control and treatment condition \( (\mu_{g1} \neq \mu_{g2}) \), as defined by (2.3). \( \text{Prob}(D_g = 1|X, Y) \) can be computed using Bayes’ theorem. We need to compute \( \text{Prob}(X_g, Y_g|D_g = 0) \) and \( \text{Prob}(X_g, Y_g|D_g = 1) \). Again, we will fix the gene \( g \) and omit the index \( g \).

When \( D = 0 \), the conditional densities are

\[
(X_j|D = 0, \mu, \tau) \sim N(\mu, \tau), \quad (Y_j|D = 0, \mu, \tau) \sim N(\mu, \tau),
\]

\[
(\mu|D = 0, \tau) \sim N(\mu_0, \kappa_0 \tau), \quad (\tau|D = 0) \sim IG(v_0, \tau_0)
\]

where \( \kappa_0, \mu_0, \nu_0 \) and \( \tau_0 \) are fixed hyperparameters. When \( D = 1 \), the conditional densities are

\[
(X_j|D = 1, \mu_1, \tau) \sim N(\mu_1, \tau), \quad (\mu_1|D = 1, \tau) \sim N(\mu_{a1}, \kappa_1 \tau)
\]

\[
(Y_j|D = 1, \mu_2, \tau) \sim N(\mu_2, \tau), \quad (\mu_2|D = 1, \tau) \sim N(\mu_{a2}, \kappa_2 \tau), \quad (\tau|D = 1) \sim IG(\nu_a, \tau_a)
\]

where \( \kappa_1, \kappa_2, \mu_{a1}, \mu_{a2}, \nu_a \) and \( \tau_a \) are fixed hyperparameters. Then,

\[
\text{Prob}(X, Y|D = 0) = \int \int \prod_{j=1}^{n_1} N(X_j; \mu, \tau) \prod_{j=1}^{n_2} N(Y_j; \mu, \tau) N(\mu; \mu_0, \kappa_0 \tau) \frac{IG(\tau; v_0, \tau_0)}{d\mu d\tau} \quad (A.5)
\]

where the integration is performed over all possible values of \( \tau \) and \( \mu \). The integration of the densities is performed by identifying the posterior normal of \( \mu|\tau, N((\sum_{j=1}^{n_1} X_j + \sum_{j=1}^{n_2} Y_j + \mu_0)/\kappa_0)/(n_1 + n_2 + 1/\kappa_0), \tau/(n_1 + n_2 + 1/\kappa_0)) \) and the posterior inverse gamma of \( \tau, IG\left(v_0 + \frac{n_1 + n_2}{2}, \tau_0 + \frac{1}{2} \left( \sum_{j=1}^{n_1} X_j^2 + \sum_{j=1}^{n_2} Y_j^2 + \mu_0^2/\kappa_0 - \frac{(\sum_{j=1}^{n_1} X_j + \sum_{j=1}^{n_2} Y_j + \mu_0)^2}{n_1 + n_2 + 1/\kappa_0} \right) \right) \).
Similarly,

$$\text{Prob}(\mathbf{X}, \mathbf{Y}|D = 1) = \int_{\tau_1}^{\tau_2} \int_{\mu_1}^{\mu_2} \prod_{j=1}^{n_1} N(X_j; \mu_1, \tau) \prod_{j=1}^{n_2} N(Y_j; \mu_2, \kappa_1 \tau) \times \prod_{j=1}^{n_2} N(Y_j; \mu_1, \mu_2, \kappa_2 \tau) IG(\tau; \nu_0, \tau_0) d\mu_2 d\mu_1 d\tau$$  \hspace{1cm} (A.6)

where the integration is performed over all possible values of $\tau, \mu_1$ and $\mu_2$. Again, the integration of the density is easily performed by identifying the posterior normal of $\mu_1|\tau, N(\sum_{j=1}^{n_1} (X_j + \mu_1)/\kappa_1)/(n_1 + 1/\kappa_1), \tau/(n_1 + 1/\kappa_1)$, the posterior normal of $\mu_2|\tau, N(\sum_{j=1}^{n_2} (Y_j + \mu_2)/\kappa_2)/(n_2 + 1/\kappa_2), \tau/(n_2 + 1/\kappa_2)$, and the posterior inverse gamma of $\tau$,

$$IG \left( \nu_0 + \frac{(n_1 + n_2)}{2}, \tau_a + 1 \left\{ \frac{\mu^2_{a1}}{\kappa_1} + \frac{\sum_{j=1}^{n_1} X_j^2}{\kappa_1 n_1 + 1} + \frac{\mu^2_{a2}}{\kappa_2} + \frac{\sum_{j=1}^{n_2} Y_j^2}{\kappa_2 n_2 + 1} \right\} \right).$$

Then, one can explicitly compute the posterior probability of a gene being expressed as

$$P(D = 1|X, Y, p) = \left(1 + \frac{1 - p \sqrt{\kappa_1 n_1 + \tau \sqrt{\kappa_2 n_2}}}{p \sqrt{\kappa_0 (n_1 + n_2) + \tau \sqrt{\kappa_0 (n_1 + n_2)}}} \right)^{-1} \hspace{1cm} (A.7)$$

where

$$R = \left[ \tau_0 + \frac{1}{2} \left\{ \frac{\mu^2_{a1}}{\kappa_1} + \frac{\sum_{j=1}^{n_1} X_j^2}{\kappa_1 n_1 + 1} + \frac{\mu^2_{a2}}{\kappa_2} + \frac{\sum_{j=1}^{n_2} Y_j^2}{\kappa_2 n_2 + 1} \right\} \right]^{\nu_0 + \frac{n_1 + n_2}{2}}.$$

Replacing $\mu_0, \mu_{a1}, \mu_{a2}, \kappa_0, \kappa_1, \kappa_2, \nu_0, \tau_0$, by their respective estimates $(n_1 \bar{X} + n_2 \bar{Y})/(n_1 + n_2), \bar{X}, \bar{Y}, 1/(n_1 + n_2), 1/n_1$ and $1/n_2$, leads to (2.4). Again, those hyperparameters are common to all the genes. However, replacing them by the sample mean and the inverse of the number of observation significantly simplifies the formula.

**The $B_3$ statistic.** We assume that $X_{gj}$ is normally distributed with mean $\mu_{g1}$ and variance $\tau_{g1}$, i.e. $(X_{gj}|\mu_{g1}, \tau_{g1}) \sim N(\mu_{g1}, \tau_{g1})$. Similarly $Y_{gj}$ is assumed normally distributed with mean $\mu_{g2}$ and variance $\tau_{g2}$, i.e. $(Y_{gj}|\mu_{g2}, \tau_{g2}) \sim N(\mu_{g2}, \tau_{g2})$. Let $D_g \sim B(1, p)$ be the Bernoulli random variable indicating whether the gene $g$ is differentially expressed between the control and treatment condition ($\tau_1 \neq \tau_2$), as defined in (2.5). For each gene $g$, we are interested in knowing if the gene is differentially expressed, i.e. $\text{Prob}(D_g = 1|\mathbf{X}, \mathbf{Y})$, which can be computed using Bayes’ theorem. Again, we will fix the gene $g$ and omit the index $g$. When $D = 0$, the conditional densities are

$$(X_j|D = 0, \mu_1, \tau) \sim N(\mu_1, \tau), \hspace{1cm} (\mu_1|D = 0, \tau) \sim N(\mu_{a1}, \kappa_1 \tau),$$

$$(Y_j|D = 0, \mu_2, \tau) \sim N(\mu_2, \tau), \hspace{1cm} (\mu_2|D = 0, \tau) \sim N(\mu_{a2}, \kappa_2 \tau), \hspace{1cm} (\tau|D = 0) \sim IG(\nu_0, \tau_0).$$
where \( \kappa_1, \kappa_2, \mu_{a1}, \mu_{a2}, \nu_0 \) and \( \tau_0 \) are fixed hyperparameters. When \( D = 1 \), the conditional densities are

\[
(X_j | D = 1, \mu_1, \tau_1) \sim N(\mu_1, \tau_1), \quad (\mu_1 | D = 1, \tau_1) \sim N(\mu_{a1}, \kappa_1 \tau_1), \quad (\tau_1 | D = 1) \sim IG(\nu_{a1}, \tau_{a1})
\]

\[
(Y_j | D = 1, \mu_2, \tau_2) \sim N(\mu_2, \tau_2), \quad (\mu_2 | D = 1, \tau_2) \sim N(\mu_{a2}, \kappa_2 \tau_2), \quad (\tau_2 | D = 1) \sim IG(\nu_{a2}, \tau_{a2})
\]

where \( \kappa_1, \kappa_2, \mu_{a1}, \mu_{a2}, \nu_{a1}, \nu_{a2}, \tau_{a1} \) and \( \tau_{a2} \) are fixed hyperparameters. The calculation of Prob(\( X, Y | D = 0 \)) is performed similarly to (A.6). Similarly,

\[
\text{Prob}(X, Y | D = 1) = \int \int \int \prod_{j = 1}^{n_1} N(X_j; \mu_1, \tau_1) \prod_{j = 1}^{n_2} N(Y_j; \mu_2, \tau_2) N(\mu_1; \mu_{a1}, \kappa_1 \tau_1) N(\mu_2; \mu_{a2}, \kappa_2 \tau_2)
\]

\[
\times N(\mu_1; \mu_{a1}, \kappa_1 \tau_1) IG(\nu_{a1}, \tau_{a1}) IG(\nu_{a2}, \tau_{a2}) d\mu_1 d\mu_2 d\tau_1 d\tau_2
\]

where the integration is performed over all possible values of \( \tau_1, \tau_2, \mu_1 \) and \( \mu_2 \). Again, the integration of the density is easily performed by identifying the posterior normal of \( \mu_1 | \tau_1, N(\sum_{j=1}^{n_1} (X_j + \mu_{a1}/\kappa_1) / (n_1 + 1/\kappa_1), \tau_1 / (n_1 + 1/\kappa_1)) \), the posterior normal of \( \mu_2 | \tau_2, N(\sum_{j=1}^{n_2} (Y_j + \mu_{a2}/\kappa_2) / (n_2 + 1/\kappa_2), \tau_2 / (n_2 + 1/\kappa_2)) \), the posterior inverse gamma of \( \tau_1 \),

\[
IG(\nu_{a1} + n_1/2, \tau_{a1} + 1/2 \left( \sum_{j=1}^{n_1} X_j^2 + \mu_{a1}^2/\kappa_1 \right) - \left( \sum_{j=1}^{n_1} X_j \right)^2)
\]

and the posterior inverse gamma of \( \tau_2 \),

\[
IG(\nu_{a2} + n_2/2, \tau_{a2} + 1/2 \left( \sum_{j=1}^{n_2} Y_j^2 + \mu_{a2}^2/\kappa_2 \right) - \left( \sum_{j=1}^{n_2} Y_j \right)^2)
\]

Then, one can explicitly compute the posterior probability of a gene being differentially expressed as

\[
P(D = 1 | X, Y, p) = \left( 1 + \frac{1 - p}{p} \frac{\Gamma(\nu_0 + \nu_{a1} + \nu_{a2} + 2) \Gamma(\nu_{a1}) \Gamma(\nu_{a2})}{\Gamma(\nu_{a1} + 1/\kappa_1) \Gamma(\nu_{a2} + 1/\kappa_2) \Gamma(\nu_0)} \right)^{-1}
\]

\[
R = \left[ \frac{\tau_{a1} + \left( \frac{\mu_{a1}}{\kappa_1} + \sum_{j=1}^{n_1} X_j^2 \right) / \left( n_1 + 1/\kappa_1 \right) \nu_{a1} + \nu_{a1}/2}{\tau_{a2} + \left( \frac{\mu_{a2}}{\kappa_2} + \sum_{j=1}^{n_2} Y_j^2 \right) / \left( n_2 + 1/\kappa_2 \right) \nu_{a2} + \nu_{a2}/2} \right]^{\nu_{a1} + \nu_{a2}/2} 
\]

\[
\left[ \frac{\tau_{a1} + \left( \frac{\mu_{a1}}{\kappa_1} + \sum_{j=1}^{n_1} X_j^2 \right) / \left( n_1 + 1/\kappa_1 \right) \nu_{a1} + \nu_{a1}/2}{\tau_{a2} + \left( \frac{\mu_{a2}}{\kappa_2} + \sum_{j=1}^{n_2} Y_j^2 \right) / \left( n_2 + 1/\kappa_2 \right) \nu_{a2} + \nu_{a2}/2} \right]^{\nu_{a1} + \nu_{a2}/2}
\]

Replacing \( \mu_{a1}, \mu_{a2}, \kappa_1, \kappa_2 \), by their respective estimates \( \hat{X}, \hat{Y}, 1/n_1 \) and \( 1/n_2 \) leads to (2.6). Again those hyperparameters are common to all the genes. Again, replacing them by the sample mean and the inverse of the number of observations significantly simplifies the formula.
The B Statistic. We assume that $X_{ij}$ is normally distributed with mean $\mu_{ij}$ and variance $\tau_{ij}$, i.e. $(X_{ij} | \mu_{ij}, \tau_{ij}) \sim N(\mu_{ij}, \tau_{ij})$. Similarly, $Y_{ij}$ is assumed normally distributed with mean $\mu_{ij}$ and variance $\tau_{ij}$, i.e. $(Y_{ij} | \mu_{ij}, \tau_{ij}) \sim N(\mu_{ij}, \tau_{ij})$. Let $D_g \sim B(1, p)$ be the Bernoulli random variable indicating whether the gene $g$ is differentially expressed between the control and treatment conditions ($\mu_1 \neq \mu_2$ or $\tau_1 \neq \tau_2$), as defined in (2.7). For each gene $g$ we are interested in knowing if the gene is differentially expressed, i.e. $\text{Prob}(D_g = 1 | X, Y)$, which can be computed using Bayes’ theorem. Again, we fix the gene $g$ and omit the index $g$. When $D = 0$, the conditional densities are

$$(X_j | D = 0, \mu, \tau) \sim N(\mu, \tau), \quad (Y_j | D = 0, \mu, \tau) \sim N(\mu, \tau), \quad (\mu | D = 0, \tau) \sim IG(\mu_0, \tau_0)$$

where $\mu_0, \tau_0$ are fixed hyperparameters. When $D = 1$, the conditional densities are

$$(X_j | D = 1, \mu_1, \tau_1) \sim N(\mu_1, \tau_1), \quad (Y_j | D = 1, \mu_1, \tau_1) \sim N(\mu_1, \tau_1), \quad (\mu_1 | D = 1) \sim IG(\mu_{a1}, \tau_{a1})$$

$$(Y_j | D = 1, \mu_2, \tau_2) \sim N(\mu_2, \tau_2), \quad (\mu_2 | D = 1, \tau_2) \sim N(\mu_{a2}, \kappa_2 \tau_2), \quad (\tau_2 | D = 1) \sim IG(\nu_2, \tau_2)$$

where $\kappa_1, \kappa_2, \mu_{a1}, \mu_{a2}, \nu_1, \nu_2, \tau_{a1}$ and $\tau_{a2}$ are fixed hyperparameters. The calculation of $\text{Prob}(X, Y | D = 0)$ is performed similarly to (A.5). The calculation of $\text{Prob}(X, Y | D = 1)$ is performed as for (A.8).

Then, one can explicitly compute the posterior probability of a gene being differentially expressed as

$$P(D = 1 | X, Y, p) = \left(1 + \frac{1 - p \sqrt{\nu_1 \nu_2 + 1}}{p \sqrt{\nu_0 (\nu_1 + \nu_2) + 1}} \Gamma(\nu_0 + \frac{\nu_1 + \nu_2}{2})\Gamma(\nu_1)\Gamma(\nu_2) \Gamma(\tau_{a1}) \Gamma(\tau_{a2}) \right)^{-1}$$

where

$$R = \left[\begin{array}{c}
\frac{\mu^2_{a1}}{\tau_{a1}} + \sum_{j=1}^{\nu_1} X_j^2 - \frac{\sum_{j=1}^{\nu_1} X_j + \frac{\nu_1}{\nu_1 + 1}}{\nu_1 + 1} \\
\frac{\mu^2_{a2}}{\tau_{a2}} + \sum_{j=1}^{\nu_2} Y_j^2 - \frac{\sum_{j=1}^{\nu_2} Y_j + \frac{\nu_2}{\nu_2 + 1}}{\nu_2 + 1}
\end{array}\right]^{-1}$$

(A.12)

Replacing $\mu_0, \mu_{a1}, \mu_{a2}, \nu_0, \nu_1, \nu_2, \tau_{a1}, \tau_{a2}$ by their respective estimates $(n_1 \bar{X} + n_2 \bar{Y})/(n_1 + n_2), \bar{X}, \bar{Y}, 1/(n_1 + n_2), 1/n_1$ and $1/n_2$, and using Bayes’ theorem leads to (2.8). Again, these hyperparameters are common to all the genes. Again, replacing them by the sample mean and the inverse of the number of observation significantly simplifies the formula. Moreover, the shrinkage of the variance is conserved.

### A.2 Estimation of hyperparameters

In each case we assumed that we knew $p$ a priori. In a typical microarray experiment, one usually only expects a small proportion of genes to be differentially expressed. A reasonable assumption would be to assume that $p$ is between 0.001 and 0.2. In Section 3, we have also introduced a simple and efficient algorithm to estimate $p$. The other hyperparameters were estimated by the method of moments. The means of the prior distribution of the means were set equal to the sample means. The scale parameter for the variance of the prior mean was always set equal to the inverse of the number of finite observations.

The hyperparameters for the inverse gamma distribution were estimated using the observed variances.
• For $B_1$, $(\nu_0, \tau_0)$ were estimated using the top proportion $p$ of genes with greatest absolute $t$-statistic. $(\nu_0, \tau_0)$ were estimated using the remaining $1 - p$ proportion. One could also use the absolute expression; however, the $t$-statistic is a better measure.

• For $B_2$, $(\nu_0, \tau_0)$ were estimated using the top proportion $p$ of genes with greatest absolute two-sample $t$-statistic. $(\nu_0, \tau_0)$ were estimated using the remaining $1 - p$ proportion.

• For $B_3$, $(\nu_1, \tau_1)$ and $(\nu_2, \tau_2)$ were estimated using the top proportion $p$ of genes with greatest variance of the expression ratios. When forming the ratios, the greatest variance was always put on top. $(\nu_0, \tau_0)$ were estimated using the remaining proportion.

• For $B_4$, $(\nu_1, \tau_1)$ and $(\nu_2, \tau_2)$ were estimated using the top proportion $p/2$ of genes with greatest absolute two-sample $t$-statistic, and the top proportion $p/2$ of genes with greatest variance of the expression ratios. When forming the ratios, the greatest variance was always put on top. $(\nu_0, \tau_0)$ were estimated using the remaining proportion.

**REFERENCES**


[Received 2002; revised March 25, 2003; accepted for publication March 17, 2003]