Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 2, 167–189
DOI: 10.1214/12-BJPS200

A predictive Bayes factor approach to identify genes differentially expressed: An application to Escherichia coli bacterium data

Francisco Louzada[a], **Erlandson F. Saraiva**[b], **Luis Milan**[c] **and Juliana Cobre**[a]

[a]*Universidade de São Paulo*
[b]*Universidade Federal da Grande Dourados*
[c]*Universidade Federal de São Carlos*

Abstract. Identifying genes differentially expressed between a treatment and a control experimental condition is a common task for gene expression data analysts. Standard existing methods are the two-sample t-test, the regularized t-test (Cyber-T) and the Bayesian t-test. In this paper, we propose a Bayesian approach to identify genes differentially expressed based on the posterior probability of the difference calculated via the Bayes factor. In order to calculate the Bayes factor, we use the predictive density that is constructed by using the previously observed gene expression levels. We perform a simulation study with small sample sizes, which is usual in gene expression data analysis, to verify the performance of the proposed method and compare it with the standard ones. The results revel a better performance of the proposed methodology in identification of difference of means and/or variance. The methodology is also illustrated on the *Escherichia coli* bacterium dataset.

1 Introduction

In the last decade, DNA array technology has become an important tool for genomic research due its capacity to measuring simultaneously the expression levels of a great number of genes (or fragments of genes) under different experimental conditions. A major goal for the gene expression data analysis is to identify genes differentially expressed between a treatment and a control experimental condition. The identification of these genes is important because it may allow biologists and geneticists to study possible relationships among genes, among genes and proteins, which genes may be involved in the origin and/or evolution of same disease with genetic origin, or which genes react to a drug stimulus, and so on. For further discussion and additional references on DNA array technology, see Schena et al. (1995), DeRisi, Iyer and Brown (1997), Arfin et al. (2000), Lonnstedt and Speed (2001), Wu (2001), Hatfield, Hung and Baldi (2003).

According to Baldi and Long (2001) gene expression data can be analyzed on at least three levels of increasing complexity. In the first level, each gene is analyzed

Key words and phrases. Gene expression, t-test, modified t-test, Bayesian Inference, Bayes factor, predictive density.

Received November 2011; accepted June 2011.

separately, where the objective is verify whether the observed expression in treatment experimental condition is significantly different from observed expression in control experimental condition. In the second level, clusters of genes are identified and analyzed in terms of patterns, common functionalities and interactions. In the third level, the objective is to infer and understand the relationship among genes and proteins.

In this paper, we focus on the first level of analysis. Under this level of analysis, one of the first approaches proposed to identify genes differentially expressed was the fold-change approach (Schena et al., 1995; Allison et al., 1995). In this approach, a gene is considered differentially expressed if the average of the logarithm of the observed expression levels in treatment and control varies more than a cutoff point, R_c, which is previously prefixed. This approach however is not adequate to yield good results, once a cutoff value R_c may have different significance for different observed expression levels. Besides, this approach does not consider the variability of observed expression levels for each gene in each experimental condition.

Other method commonly used for gene expression data analysis is the so called two-sample t-test (TT) for the log transformed data (Baldi and Long, 2001; Hatfield, Hung and Baldi, 2003). The advantage of the t-test in relation to fold-change approach is that t-test consider the variability of measures from each experimental condition in the process of identification of genes differentially expressed. However, the problem with the application of t-test to this kind of data is the usual small size of treatment and control samples, which may lead to underestimated variances and small power of the test. To avoid such limitations some TT modifications were proposed, such as the Cyber-t (CT) proposed by Baldi and Long (2001) and the Bayesian t-test (BTT) proposed by Fox and Dimmic (2006). Basically, the main idea is to consider modifications of the standard error estimate of the two sample difference present in the denominator of the standard t statistics. But, one can argue that an increasing in variance would difficult the detection the changes in mean by the TT, CT and BTT because we may have a small statistic value t_g. In the other hand, if we have small variances we may have high statistics values t_g and the gene may be wrongly identified as differentially expressed.

In this paper, we propose a Bayesian approach to identify genes differentially expressed based on the posterior probability of the difference which is calculated using the Bayes factor. The advantage of using the Bayes factor is that it allows to compare the observed expression levels from the treatment and the control as well as the treatment and control distributions. In this way, a change in mean and/or variance would help to detect the distribution changes and identify the cases differentially expressed. In order to verify whether the observed data from a treatment experimental condition is supported by a model fitted by an observed data from a control experimental condition, we propose to calculate the Bayes factor via predictive density that is constructed using the previously observed gene expression levels.

A simulation study is performed to verify the performance of the proposed method and compare it with the TT, CT and BTT. The simulation study revels a better performance of the proposed method in identification of difference of means and/or variance in small sized samples, usually present in gene expression data analysis. Then, the main advantage of the proposed method is that it is easy to use like fold-change and two-sample t-test but presents better performance in situations with small sample size. We also apply the methods to a real data set, extracted from the experiment carried through with *Escherichia coli* bacterium, described in details by Arfin et al. (2000).

The paper is organized as follows. In Section 2, we model gene expression data and review the methods TT, CT and BTT. In Section 3, we develop our Bayesian approach calculating the posterior probability of difference via Bayes factor and predictive density. The performance of methods is verified and compared using artificial datasets and a real dataset in Section 4. In Section 5, we conclude the paper with final remarks on the proposed method.

2 Models for gene expression data analysis

Consider a DNA array experiment with n genes and two experimental conditions which we name by control (c) and treatment (t). Suppose that control and treatment are replicated n_c and n_t times, respectively. Denote by x_{ig_h} the ith observed expression level (or its logarithm) for gene g in experimental condition h, $h \in \{c, t\}$ and $g = 1, \ldots, n$. Let $\mathbf{x}_{g_h} = \{x_{1g_h}, \ldots, x_{n_h g_h}\}$ be realizations of independent random variables $\mathbf{X}_{g_h} = \{X_{1g_h}, \ldots, X_{n_h g_h}\}$, for $g = 1, \ldots, n$ and $h \in \{c, t\}$.

Assume that data have already been preprocessed with appropriate normalization. As it is usual in gene expression data analysis, consider the logarithm of the observed gene expression levels in control and treatment are generated from normal distributions with mean μ_{g_h} and variance $\sigma^2_{g_h}$, $X_{ig_h} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{g_h}, \sigma^2_{g_h})$, for $i = 1, \ldots, n_h$, $h \in \{c, t\}$ and $g = 1, \ldots, n$ (Baldi and Long, 2001; Fox and Dimmic, 2006; Hatfield, Hung and Baldi, 2003; Medvedovic and Sivaganesan, 2002). Denote parameters by $\theta_{g_h} = (\mu_{g_h}, \sigma^2_{g_h})$, for $g = 1, \ldots, n$ and $h \in \{c, t\}$.

The interest here is to verify whether gene g presents different gene expression levels between treatment and control experimental conditions, that is, if $\theta_{g_t} = \theta_{g_c}$ or $\theta_{g_t} \neq \theta_{g_c}$, for $g = 1, \ldots, n$. More explicitly, the null hypothesis of interest is $H_0 : \mu_{gc} = \mu_{gt}$ and $\sigma^2_{gc} = \sigma^2_{gt}$, with an overall size α. However, usually in the standard hypothesis tests, such the t-test describe below, we are testing $H_0 : \mu_1 = \mu_2$ with equal or unequal variances. Here we are also testing the equality of variances.

2.1 t-test

Under normality assumption for the logarithm of observed gene expression levels an usual statistic test used to identify genes differentially expressed is the two-sample t-test (Arfin et al., 2000; Baldi and Long, 2001; Hatfield, Hung and Baldi, 2003). The hypothesis test is based on the statistics,

$$t_g = \frac{\bar{x}_{g_t} - \bar{x}_{g_c}}{\sqrt{s_{g_t}^2/n_t + s_{g_c}^2/n_c}}, \tag{2.1}$$

which follows a Student's t distribution with $df = [s_{g_c}^2/n_c + s_{g_t}^2/n_t]^2/[(s_{g_c}^2/n_c)^2/(n_c - 1) + (s_{g_t}^2/n_t)^2/(n_t - 1)]$, degrees of freedom, where $\overline{x}_{g_h}$ and $s_{g_h}^2$ are the sample mean and variance for gene g in experimental condition $h = \{c, t\}$. Fixed a significance level α, if $|t_g|$ is greater than a threshold $t_{1-\alpha/2,df}$ (quantile $1 - \frac{\alpha}{2}$ of Student's t distribution with df degrees of freedom) then the test conclude for difference of expression levels.

A fundamental problem with the application of the t-test for gene expression data analysis are the sample sizes n_c and n_t from control and treatment that are often small, due to the realization of the experiment to be mostly expensive or tedious to be repeated (Baldi and Long, 2001). These small sample sizes may lead to underestimates of the variances and a small power of the hypothesis test.

2.2 Regularized t-test

Baldi and Long (2001) proposed a two-sample t-test replacing the denominator of (2.1) by a pooled variance estimated via a Bayesian approach.

In order to develop the Bayesian approach Baldi and Long (2001) consider for parameters of the experimental condition $h \in \{c, t\}$, $(\mu_{g_h}, \sigma_{g_h}^2)$, the conjugated prior distributions

$$\mu_{g_h}|\sigma_{g_h}^2, \mu_0, \lambda_0 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_{g_h}^2}{\lambda_0}\right) \quad \text{and} \quad \sigma_{g_h}^2|\nu_0, \sigma_0^2 \sim \mathcal{IG}(\nu_0, \sigma_0^2),$$

where $\mathcal{IG}(\cdot)$ denotes the inverse gamma distribution with mean and variance given, respectively, by $\sigma_0^2/(\nu_0 - 1)$ and $\sigma_0^4/(\nu_0 - 1)^2(\nu_0 - 2)$; μ_0, λ_0, ν_0 and σ_0^2 are hyperparameters.

Using the Bayes theorem Baldi and Long (2001) obtain the posterior distributions

$$\mu_{g_h}|\mathbf{x}_{g_h}, \sigma_{g_h}^2, \mu_{n_h} \sim \mathcal{N}\left(\mu_{n_h}, \frac{\sigma_{g_h}^2}{\lambda_0 + n_{g_h}}\right) \quad \text{and} \quad \sigma_{g_h}^2|\mathbf{x}_{g_h}, \nu_{n_h}, \sigma_{n_h}^2 \sim \mathcal{IG}(\nu_{n_h}, \sigma_{n_h}^2),$$

where $\mu_{n_h} = \frac{n_{g_h}\overline{x}_{g_h}}{\lambda_0 + n_{g_h}} + \frac{\lambda_0\mu_0}{\lambda_0 + n_{g_h}}$, $\nu_{n_h} = \nu_0 + n_{g_h}$ and $\nu_{n_h}\sigma_{n_h}^2 = \nu_0\sigma_0^2 + (n_{g_h} - 1)s_{g_h}^2 + \frac{\lambda_0 n_{g_h}}{\lambda_0 + n_{g_h}}(\overline{x}_{g_h} - \mu_0)^2$ for $h \in \{c, t\}$ and $g = 1, \ldots, n$.

Then, we may calculate the posterior mean estimate for the variance as $\tilde{\sigma}_{g_h}^2 = \frac{\nu_{n_h}}{\nu_{n_h} - 2}\sigma_{n_h}^2$, $h = \{c, t\}$, and implement the Cyber-T software (CT), where the statistic t_g in (2.1) is replaced by the statistics

$$t_g = \frac{\bar{x}_{g_t} - \bar{x}_{g_c}}{\sqrt{\tilde{\sigma}_{g_t}^2/n_{g_t} + \tilde{\sigma}_{g_c}^2/n_{g_c}}} \tag{2.2}$$

and the degrees of freedom is given by $df = \nu_0 + n_{g_c} + n_{g_t} - 2$.

2.3 Bayesian t-test

Fox and Dimmic (2006) assume $X_{ig_c} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{g_c}, \sigma_g^2)$ and $X_{ig_t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{g_c} + \Delta\mu_{g_c}, \sigma_g^2)$, for $i = 1, \dots, n_h$, $h \in \{c, t\}$ and $g = 1, \dots, n$.

Based on Baldi and Long (2001), the authors develop a Bayesian approach and show that

$$\frac{\Delta\mu - \Delta\overline{x}}{\sigma_n\sqrt{1/n_{g_t} + 1/n_{g_c}}}\Big|\mathbf{x}_{g_c}, \mathbf{x}_{g_t} \sim t_{\nu_n}, \tag{2.3}$$

where $\Delta\overline{x} = \overline{x}_{g_t} - \overline{x}_{g_c}$, $\nu_n = \nu_0 + n_{g_c} + n_{g_t} - 2$, $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n_{g_c} - 1)s_{g_c}^2 + (n_{g_t} - 1)s_{g_t}^2$ and t_{ν_n} represent the Student's t distribution with ν_n degrees of freedom, for $g = 1, \dots, n$. As written in Fox and Dimmic (2006), "a hypothesis test is performed by asserting a null hypothesis that the true difference in expression levels is zero, that is, $\Delta\mu = 0$." Thus, this procedure defines the Bayesian t-test (BTT).

3 Predictive Bayes factor approach

Our predictive Bayes factor approach is given as follows. In order to represent situations with and without difference between treatment and control for a gene g, consider models M_0 and M_1, such that,

1. Under M_0 there is no difference between treatment and control, that is, $\theta_{g_t} = \theta_{g_c}$. For this situation, consider that $(\mu_{g_t}, \sigma_{g_t}^2) = (\mu_{g_c}, \sigma_{g_c}^2) = (\mu_g, \sigma_g^2) = \theta_g$. The likelihood function is

$$L_{M_0}(\theta_g|\mathbf{x}_g) \propto (\sigma_g^2)^{-n_g/2} \exp\left\{-\frac{1}{2\sigma_g^2}\sum_{i=1}^{n_g}(x_{ig} - \mu_g)^2\right\}, \tag{3.1}$$

   where $\mathbf{x}_g = \{\mathbf{x}_{g_c}, \mathbf{x}_{g_t}\}$, $n_g = n_{g_c} + n_{g_t}$, $x_{ig} = x_{ig_c}$ for $i = 1, \dots, n_c$, $x_{ig} = x_{ig_t}$ for $i > n_c$, $i = 1, \dots, n_g$ and $g = 1, \dots, n$;
2. Under M_1 there is difference, that is, $\theta_{g_t} \neq \theta_{g_c}$. For this model, the likelihood function is

$$\begin{aligned} &L_{M_1}(\theta_{g_c}, \theta_{g_t}|\mathbf{x}_{g_c}, \mathbf{x}_{g_t}) \\ &\quad = L_{M_1}^c \cdot L_{M_1}^t \\ &\quad \propto \prod_{h\in\{c,t\}} (\sigma_{g_h}^2)^{-n_{g_h}/2} \exp\left\{-\frac{1}{2\sigma_{g_h}^2}\sum_{i=1}^{n_{g_h}}(x_{ig_h} - \mu_{g_h})^2\right\} \end{aligned} \tag{3.2}$$

   for $g = 1, \dots, n$.

Now we can identify genes differentially expressed by choosing between models M_0 and M_1 which is the most supported by the data, which can be made by considering the Bayes factor (Kass and Raftery, 1995; Aitkin, 1991; Beger and Pericchi, 1996; Lavine and Schervish, 1999).

Available expert opinions may be expressed in terms of a prior distributions for parameters, since the parameters of the models M_0 and M_1 have a direct interpretation in the context of the gene expression data analysis. In order to explore the fully conjugation, consider that joint prior distribution for parameters of model M_0 is given by

$$\pi(\theta_g) = \pi(\mu_g, \sigma_g^2) = \pi(\mu_g|\sigma_g^2)\pi(\sigma_g^2) \tag{3.3}$$

and the joint prior distribution for parameters of model M_1 is given by

$$\begin{aligned} \pi(\theta_{g_c}, \theta_{g_t}) &= \pi(\theta_{g_c})\pi(\theta_{g_t}) \\ &= \pi(\mu_{g_c}, \sigma_{g_c}^2)\pi(\mu_{g_t}, \sigma_{g_t}^2) \\ &= \pi(\mu_{g_c}|\sigma_{g_c}^2)\pi(\sigma_{g_c}^2)\pi(\mu_{g_t}|\sigma_{g_t}^2)\pi(\sigma_{g_t}^2) \end{aligned} \tag{3.4}$$

in which $\pi(\mu_d|\sigma_d^2)$ and $\pi(\sigma_d^2)$, for $d = g$ under M_0 and $d = \{g_c, g_t\}$ under M_1, represent the density of the distributions

$$\mu_d|\sigma_d^2, \mu_0 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_d^2}{\lambda}\right) \quad \text{and} \quad \sigma_d^2|\tau, \beta \sim \mathcal{IG}\left(\frac{\tau}{2}, \frac{\beta}{2}\right),$$

where μ_0, λ, τ and β are known hyperparameters.

Updating the prior distributions in (3.3) and (3.4) via likelihood function in (3.1) and (3.2), respectively, the joint posterior distribution is given by

$$\mu_d, \sigma_d^2|\mathbf{x}_{g_c}, \mathbf{x}_{g_t} \sim \mathcal{N}\left(\mu_d^*, \frac{\sigma^2}{\lambda + n_d}\right)\mathcal{IG}\left(\frac{\tau + n_d + 1}{2}, \frac{\beta^*}{2}\right), \tag{3.5}$$

where $\mu_d^* = \frac{1}{n+\lambda}\sum_{i=1}^{n_d} x_{id} + \frac{\lambda}{n+\lambda}\mu_0$ and $\beta_d^* = \beta + \sum_{i=1}^{n_d} x_{id}^2 + \lambda\mu_0 - \frac{(\sum_{i=1}^{n_d} x_{id} - \lambda\mu_0)^2}{\lambda+n}$, for $d = g$ under M_0 and $d = \{g_c, g_t\}$ under M_1, $g = 1, \ldots, n$.

Fixing the prior probabilities for models $\pi(M_0) = \pi(M_1) = \frac{1}{2}$, the Bayes factor for gene g, $B_{10}(g) = \frac{I_1}{I_0}$, can be analytically calculated, where

$$\begin{aligned} I_1 &= \int\int L_{M_1}(\theta_{g_c}, \theta_{g_t}|\mathbf{x}_{g_c}, \mathbf{x}_{g_c})\pi(\theta_{g_c}\theta_{g_t})\, d\theta_{g_c}\, d\theta_{g_t} \\ &= \left[\int L_{M_1}^c(\theta_{g_c}|\mathbf{x}_{g_c})\pi(\theta_{g_c})\, d\theta_{g_c}\right] \cdot \left[\int L_{M_1}^t(\theta_{g_c}|\mathbf{x}_{g_c})\pi(\theta_{g_t})\, d\theta_{g_t}\right] \end{aligned} \tag{3.6}$$

and

$$I_0 = \int L_{M_0}(\theta_g|\mathbf{x}_g)\pi(\theta_g)\, d\theta_g \tag{3.7}$$

for $g = 1, \ldots, n$.

3.1 Bayes factor via predictive distribution

In order to calculate the Bayes factor considering the observations from treatment group as future observations in relation to model fitted by observed data from a control experimental condition, we rewritten the integrals above in the follow way. Note that (3.6) can be rewritten as

$$I_1 = I_1^c \cdot I_1^t \tag{3.8}$$

in which

$$I_1^c = \left[\prod_{i=1}^{n_{g_c}} \int f(x_{ig_c}|\theta_{g_c}) \pi(\theta_{g_c}|x_{1g_c}, \dots, x_{i-1g_c})\, d\theta_{g_c} \right]$$

and

$$I_1^t = \left[\prod_{i=1}^{n_{g_t}} \int f(x_{ig_t}|\theta_{g_t}) \pi(\theta_{g_t}|x_{1g_t}, \dots, x_{i-1g_t})\, d\theta_{g_t} \right],$$

where $\pi(\theta_{g_h}|x_{1g_h}, \dots, x_{i-1g_h})$ is the posterior distribution for θ_{g_h}, $h = \{c, t\}$, given the $i - 1$ first observed data from experimental condition h, $h \in \{c, t\}$.

Analogously to (3.8), we can rewrite (3.7) as

$$I_0 = I_0^c \cdot I_0^t, \tag{3.9}$$

where $I_0^c = I_1^c$ and

$$I_0^t = \left[\prod_{i=1}^{n_{g_t}} \int f(x_{ig_t}|\theta_{g_t}) \pi\left(\theta_{g_t}|\mathbf{x}_{g_c} \cup \{x_{1g_t}, \dots, x_{i-1g_t}\}\right) d\theta_{g_t} \right].$$

Considering (3.6), (3.7) and the results above, the Bayes factor $B_{10}(g)$ is given by

$$\begin{aligned} B_{10}(g) &= \frac{I_1}{I_0} = \frac{I_1^c \cdot I_1^t}{I_0^c \cdot I_0^t} = \frac{I_1^t}{I_0^t} \\ &= \frac{\prod_{i=1}^{n_{g_t}} \int f(x_{ig_t}|\theta_{g_t}) \pi(\theta_{g_t}|x_{1g_t}, \dots, x_{i-1g_t})\, d\theta_{g_t}}{\prod_{i=1}^{n_{g_t}} \int f(x_{ig_t}|\theta_{g_t}) \pi(\theta_{g_t}|\mathbf{x}_{g_c} \cup \{x_{1g_t}, \dots, x_{i-1g_t}\})\, d\theta_{g_t}}. \end{aligned} \tag{3.10}$$

From (3.10), the integral of the numerator can be rewritten as

$$\begin{aligned} \int f(x_{ig_t}|\theta_{g_t}) \pi(\theta_{g_t}|x_{1g_t}, \dots, x_{i-1g_t})\, d\theta_{g_t} &= \frac{\int L(\theta_{g_t}|x_{1g_t}, \dots, x_{ig_t}) \pi(\theta_{g_t})\, d\theta_{g_t}}{\int L(\theta_{g_t}|x_{1g_t}, \dots, x_{i-1g_t}) \pi(\theta_{g_t})\, d\theta_{g_t}} \\ &= \frac{I_1^t(\mathbf{x}_i)}{I_1^t(\mathbf{x}_{i-1})} \end{aligned}$$

in which, $\mathbf{x}_a = (x_{1g_t}, \ldots, x_{ag_t})$ for $a = \{i, i-1\}$ and

$$I_1^t(\mathbf{x}_a) = \left[\frac{1}{\beta\pi}\right]^{n_{\mathbf{x}_a}/2} \left[\frac{\lambda}{\lambda + n_{\mathbf{x}_a}}\right]^{1/2} \frac{\Gamma((\tau + n_{\mathbf{x}_a})/2)}{\Gamma(\tau/2)} \times \left[1 + \frac{\sum_{\mathbf{x}} x + \lambda\mu_0^2}{\beta} - \frac{(\sum_{\mathbf{x}} x + \lambda\mu_0)^2}{\beta(\lambda + n_{\mathbf{x}_a})}\right]^{-(\tau + n_{\mathbf{x}_a})/2}, \tag{3.11}$$

where $n_{\mathbf{x}_a}$ is the number of observations in $\mathbf{x}_a$. Note that $\frac{I_1^t(\mathbf{x}_i)}{I_1^t(\mathbf{x}_{i-1})}$ is the Bayes factor of the model fitted using the observations $\mathbf{x}_{i-1} = (x_{1g_t}, \ldots, x_{i-1g_t})$ augmented with x_{ig_t}, $\mathbf{x}_i = (x_{1g_t}, \ldots, x_{ig_t})$, in relation to model fitted only using $\mathbf{x}_{i-1}$.

In a similar way, the integral of the denominator in (3.10) is given by

$$\int f(x_{ig_t}|\theta_{g_t})\pi\left(\theta_{g_t}|\mathbf{x}_{g_c} \cup \{x_{1g_t}, \ldots, x_{i-1g_t}\}\right) d\theta_{g_t} = \frac{I_1^t(\mathbf{x}_{g_c} \cup \mathbf{x}_i)}{I_1^t(\mathbf{x}_{g_c} \cup \mathbf{x}_{i-1})},$$

where $I_1^t(\mathbf{x}_a)$ is given in (3.11) for $\mathbf{x}_a = \{\mathbf{x}_{g_c} \cup \mathbf{x}_i, \mathbf{x}_{g_c} \cup \mathbf{x}_{i-1}\}$.

The $B_{10}(g)$ in (3.10) is calculated using

$$B_{10}(g) = \prod_{i=1}^{n_{g_t}} \frac{I_1^t(\mathbf{x}_i)}{I_1^t(\mathbf{x}_{i-1})} \cdot \frac{I_1^t(\mathbf{x}_{g_c} \cup \mathbf{x}_{i-1})}{I_1^t(\mathbf{x}_{g_c} \cup \mathbf{x}_i)}. \tag{3.12}$$

3.2 Posterior probabilities for the models

Using the Bayes theorem, the posterior probabilities for the models are given by

$$P(M_0|\mathbf{y}_g) = \frac{1}{1 + B_{10}} \quad \text{and} \quad P(M_1|\mathbf{y}_g) = \frac{B_{10}}{1 + B_{10}}. \tag{3.13}$$

Thus, if $P(M_1|\mathbf{y}_g) > P_{\text{ref}}$, where $P_{\text{ref}} \in [0.5, 1)$ is a cutoff value, we choose M_1 and the gene g presents evidence for difference between treatment and control. Otherwise, we choose M_0 and the gene g have no evidence for difference, $g = 1, \ldots, n$. Adapting the decision rule discussed in Kass and Raftery (1995) to choose between models using $P(M_1|\mathbf{y}_g)$, we consider as cutoff value $P_{\text{ref}} = 0.5$, which indicates positive evidence against M_0. We denote the Bayes factor B_{10} calculated by (3.12) as the predictive Bayes factor (PBF). One advantage of calculate B_{10} as (3.12) is that we can verify the influence of each observation from treatment group in decision on differentially ($P(M_1|\cdot) > 0.5$) or nondifferentially expressed ($P(M_1|\cdot) \leq 0.5$), as will be illustrated in Section 4.2.

4 Data analysis

In this section, the proposed PBF is applied to artificial data sets and a real data set. The artificial data sets were generated as a mix of both differentially and nondifferentially expressed genes where the fraction of differentially expressed genes is small.

In order to evaluate the performance of the PBF and to compare with TT, CT and BTT, we consider (i) the true positive rate (number of genes correctly found differentially expressed in the analysis divided by the number of genes that are differentially expressed in the underlying problem), (ii) the false positive rate (number of genes incorrectly found differentially expressed in the analysis divided by the number of genes that are nondifferentially expressed in the underlying problem) and (iii) the true discovery rate (number of genes correctly found differentially expressed divided by the number of genes found differentially expressed).

The real data set was extracted from the site www.jbc.org and refers to an experiment realized with *Escherichia coli* bacterium using nylon membranes, described in details by Arfin et al. (2000).

4.1 Artficial data sets

To generate the artificial data set, we fix $\mu_{g_c} = -14$ and $\sigma^2_{g_c} = 0.8$. These values denote the average of the observed mean and variance of the expression levels (log transformed) from control group of the *Escherichia coli* bacterium dataset. We fix $n = 1000$ and the sample sizes n_{g_c} and n_{g_t} were fixed at 4 and 8.

To verify how the method behaves when $\theta_{g_t} = (\mu_{g_t}, \sigma^2_{g_t})$ moves away from $\theta_{g_c} = (\mu_{g_c}, \sigma^2_{g_c})$, we simulate its values using $\mu_{g_t} = \mu_{g_c} \pm \delta\sigma_{g_c}$ and $\sigma_{g_t} = \gamma\sigma_{g_c}$, for $\delta = \{0.0, 0.25, 0.50, 0.75, 1, 1.25, 1.50, 1.75, 2\}$ and $\gamma = \{1, 2, 3, 4\}$, where the signal $+$ and $-$ in expression μ_{g_t} represent the situation over and under expressed, respectively.

The generation of the simulated data sets follow the steps:

(i) For $g = 1, \ldots, n$ generate $\mathbf{x}_{g_c}$ from $\mathcal{N}(\mu_{g_c}, \sigma^2_{g_c})$, that is, $x_{1g_c}, \ldots, x_{n_c g_c} \sim \mathcal{N}(\mu_{g_c}, \sigma^2_{g_c})$;

(ii) From index $\{1, \ldots, n\}$ choose randomly $p\%$ of these index to indicate the cases generated with difference, $p \in \{5, 10\}$. We use $p = p_{\text{over}} + p_{\text{under}}$, for $p_{\text{over}} = \{3, 5\}$ and $p_{\text{under}} = \{2, 5\}$, respectively. For example, $p = 5$ is composite by $p_{\text{over}} = 3$ plus $p_{\text{under}} = 2$;

(iii) If the index $g \in \{1, \ldots, n\}$ is chosen, then consider an indicator variable $\mathbb{I}_g = 1$ and generate $X_{ig_t} \sim \mathcal{N}(\mu_{g_t}, \sigma^2_{g_t})$, for $i = 1, \ldots, n_t$;

(iv) If the index $g \in \{1, \ldots, n\}$ is not chosen, then set up $\mathbb{I}_g = 0$ and generate $X_{ig_t} \sim \mathcal{N}(\mu_c, \sigma^2_c)$, for $i = 1, \ldots, n_t$.

For PBF application the hyperparameters were fixed in order to obtain weakly informative priors (see Appendix A). We then set up: (i) τ and β in a way that $E[\sigma^2_d] = (\beta/2)/[(\tau/2) - 1] = R$, where $R = \max(\mathbf{x}_g) - \min(\mathbf{x}_g)$ is the length of the interval of variation of the observed data $\mathbf{x}_g = \{\mathbf{x}_{g_c} \cup \mathbf{x}_{g_t}\}$, $d \in \{g, g_c, g_t\}$ and $g = 1, \ldots, n$. Thus, we obtain $\beta = (\tau - 2)R$; and we fix $\tau = 3$; (ii) μ_0 as being the middle point of the interval of variation of $\mathbf{x}_g$, $\mu_0 = [\min(\mathbf{x}_g) + \max(\mathbf{x}_g)]/2$; (iii) λ at 10^{-2}, $\lambda = 10^{-2}$.

To record the cases identified with difference by PBF we consider a indicator variable $\mathbb{I}_g^{\text{PBF}} = 1$ for cases, so that, $P(M_1|\mathbf{y}_g) > 0.5$. Otherwise, $\mathbb{I}_g^{\text{PBF}} = 0$. Analogously, for TT, CT and BTT we consider $\mathbb{I}_g^{\text{method}} = 1$ (method $= \{\text{TT}, \text{CT}, \text{BTT}\}$) for cases with p-value$_g < 0.05$ and $\mathbb{I}_g^{\text{method}} = 0$ otherwise. So, we calculate the true positive rate given by

$$\text{TPR}_{\text{method}} = \frac{\sum_{g=1}^{n} \mathbb{I}_g \cdot \mathbb{I}_g^{\text{method}}}{\sum_{g=1}^{n} \mathbb{I}_g},$$

where method $= \{\text{PA}, \text{TT}, \text{CT}, \text{BTT}\}$.

For a pair (δ, γ) and values p, n and $n_{g_c} = n_{g_t}$ fixed, we generate $L = 100$ different artificial data sets according to steps (i) to (iv) described above and we present the results using the mean of the true positive rate, that is given by $\overline{\text{TPR}}_{\text{method}} = \frac{\sum_{l=1}^{L} \text{TPR}_{\text{method}}^{(l)}}{L}$, where $\text{TPR}_{\text{method}}^{(l)}$ is the true positive rate calculated for lth generated dataset by method $= \{\text{PA}, \text{TT}, \text{CT}, \text{BTT}\}$.

Tables 1–4 present $\overline{\text{TPR}}_{\text{method}}$ for $p = \{5, 10\}$ and $n_{g_c} = n_{g_t} = 4$ and $n_{g_c} = n_{g_t} = 8$, respectively, for method $= \{\text{PBF}, \text{TT}, \text{CT}, \text{BTT}\}$. As we move from left to right side of the tables, in each line we have the distances between control and treatment means, which are increasing. As we move from top to down in columns of the tables, we have the distance between the treatment and control variances, which are increasing.

Table 1 *True positive rate*, $n_{g_c} = n_{g_t} = 4$ *and* $p = 5\%$ ($p_{\text{over}} = 3\%$, $p_{\text{under}} = 2\%$)

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.019	0.023	0.044	0.074	0.134	0.231	0.330	0.444	0.581
	TT	0.040	0.049	0.073	0.115	0.185	0.294	0.394	0.492	0.612
	CT	0.029	0.035	0.055	0.088	0.148	0.235	0.329	0.426	0.543
	BTT	0.040	0.048	0.075	0.116	0.186	0.290	0.394	0.492	0.611
2	PBF	0.061	0.061	0.082	0.113	0.159	0.195	0.260	0.346	0.412
	TT	0.050	0.049	0.063	0.086	0.113	0.142	0.183	0.245	0.286
	CT	0.037	0.039	0.052	0.070	0.095	0.122	0.156	0.219	0.257
	BTT	0.051	0.054	0.067	0.090	0.117	0.152	0.198	0.265	0.310
3	PBF	0.145	0.153	0.155	0.183	0.207	0.258	0.289	0.338	0.390
	TT	0.052	0.056	0.065	0.069	0.082	0.105	0.120	0.140	0.173
	CT	0.043	0.050	0.060	0.060	0.073	0.099	0.111	0.130	0.162
	BTT	0.061	0.062	0.074	0.075	0.092	0.124	0.139	0.158	0.198
4	PBF	0.271	0.280	0.288	0.320	0.320	0.349	0.376	0.415	0.429
	TT	0.054	0.059	0.056	0.067	0.068	0.083	0.090	0.109	0.118
	CT	0.052	0.055	0.053	0.064	0.066	0.082	0.093	0.106	0.117
	BTT	0.066	0.070	0.069	0.081	0.084	0.100	0.110	0.130	0.144

Table 2 *True positive rate, $n_{g_c} = n_{g_t} = 4$ and $p = 10\%$ ($p_{\text{over}} = 5\%$, $p_{\text{under}} = 5\%$)*

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.016	0.026	0.041	0.083	0.138	0.229	0.334	0.453	0.578
	TT	0.040	0.050	0.073	0.122	0.188	0.285	0.385	0.491	0.603
	CT	0.030	0.037	0.056	0.099	0.152	0.232	0.324	0.429	0.533
	BTT	0.039	0.049	0.073	0.123	0.191	0.288	0.388	0.499	0.612
2	PBF	0.058	0.066	0.076	0.112	0.148	0.200	0.263	0.337	0.415
	TT	0.049	0.053	0.059	0.081	0.103	0.144	0.186	0.235	0.286
	CT	0.037	0.041	0.046	0.067	0.086	0.121	0.161	0.207	0.256
	BTT	0.051	0.055	0.062	0.086	0.111	0.152	0.198	0.250	0.308
3	PBF	0.143	0.138	0.159	0.185	0.204	0.240	0.293	0.337	0.395
	TT	0.055	0.053	0.062	0.075	0.082	0.099	0.122	0.140	0.173
	CT	0.047	0.047	0.055	0.067	0.074	0.090	0.113	0.133	0.163
	BTT	0.062	0.062	0.070	0.084	0.095	0.112	0.142	0.162	0.199
4	PBF	0.262	0.281	0.285	0.293	0.323	0.340	0.382	0.417	0.454
	TT	0.051	0.061	0.059	0.062	0.073	0.082	0.092	0.106	0.125
	CT	0.047	0.057	0.056	0.058	0.070	0.079	0.092	0.103	0.127
	BTT	0.063	0.073	0.073	0.073	0.088	0.097	0.112	0.129	0.154

Table 3 *True positive rate, $n_{g_c} = n_{g_t} = 8$ and $p = 5\%$ ($p_{\text{over}} = 3\%$, $p_{\text{under}} = 2\%$)*

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.013	0.024	0.047	0.116	0.241	0.415	0.584	0.744	0.873
	TT	0.048	0.075	0.145	0.282	0.466	0.637	0.778	0.892	0.959
	CT	0.039	0.061	0.122	0.243	0.415	0.588	0.743	0.868	0.947
	BTT	0.048	0.074	0.144	0.282	0.467	0.636	0.777	0.891	0.959
2	PBF	0.094	0.112	0.142	0.190	0.282	0.381	0.480	0.576	0.687
	TT	0.047	0.058	0.091	0.140	0.214	0.305	0.404	0.514	0.622
	CT	0.039	0.049	0.081	0.125	0.194	0.284	0.379	0.490	0.596
	BTT	0.050	0.062	0.099	0.146	0.223	0.319	0.422	0.533	0.639
3	PBF	0.441	0.457	0.472	0.507	0.544	0.593	0.658	0.713	0.754
	TT	0.047	0.063	0.074	0.092	0.123	0.169	0.221	0.303	0.351
	CT	0.044	0.059	0.069	0.088	0.117	0.160	0.214	0.297	0.347
	BTT	0.053	0.071	0.084	0.104	0.140	0.192	0.245	0.327	0.382
4	PBF	0.755	0.755	0.764	0.776	0.787	0.820	0.828	0.845	0.876
	TT	0.054	0.052	0.061	0.069	0.096	0.122	0.155	0.171	0.232
	CT	0.053	0.051	0.059	0.069	0.094	0.120	0.155	0.172	0.232
	BTT	0.062	0.061	0.071	0.084	0.115	0.142	0.174	0.198	0.261

Table 4 *True positive rate*, $n_{g_c} = n_{g_t} = 8$ *and* $p = 10\%$ ($p_{\text{over}} = 5\%$, $p_{\text{under}} = 5\%$)

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.010	0.020	0.053	0.116	0.241	0.395	0.591	0.747	0.868
	TT	0.049	0.072	0.155	0.272	0.454	0.627	0.799	0.894	0.952
	CT	0.037	0.058	0.129	0.235	0.410	0.581	0.765	0.871	0.941
	BTT	0.049	0.072	0.154	0.270	0.454	0.626	0.798	0.894	0.953
2	PBF	0.093	0.103	0.140	0.203	0.282	0.377	0.474	0.588	0.695
	TT	0.052	0.058	0.091	0.138	0.215	0.308	0.400	0.511	0.622
	CT	0.043	0.050	0.079	0.123	0.197	0.281	0.373	0.490	0.598
	BTT	0.055	0.061	0.095	0.147	0.228	0.321	0.417	0.529	0.642
3	PBF	0.443	0.442	0.467	0.509	0.543	0.596	0.646	0.705	0.759
	TT	0.054	0.057	0.069	0.099	0.131	0.169	0.228	0.286	0.361
	CT	0.051	0.053	0.064	0.094	0.125	0.163	0.220	0.277	0.350
	BTT	0.061	0.064	0.078	0.112	0.146	0.188	0.254	0.317	0.391
4	PBF	0.763	0.756	0.759	0.779	0.784	0.806	0.829	0.848	0.873
	TT	0.055	0.053	0.060	0.072	0.094	0.119	0.148	0.193	0.243
	CT	0.053	0.052	0.057	0.072	0.093	0.120	0.147	0.193	0.242
	BTT	0.063	0.062	0.072	0.087	0.111	0.138	0.170	0.222	0.272

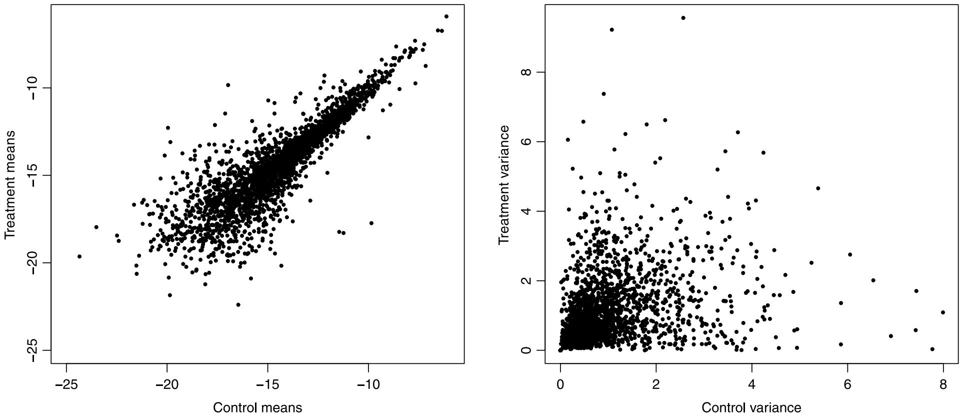
The PBF present better performance than other methods, except, for $\gamma = 1$ fixed where t-tests present greater $\overline{\text{TPR}}$. For the four methods the $\overline{\text{TPR}}$ increases as the value of δ increases, that is, when the mean of the treatment distribution moves away from the mean of the control distribution. Increasing the variance of the treatment ($\gamma = \{2, 3, 4\}$) the PBF present higher $\overline{\text{TPR}}$ than TT, CT and BTT for all cases simulated. Besides that, we can note that among the t-tests (TT, CT and BTT), the BTT present better performance.

Tables 5–8 in Appendix B show the mean of the false positive rates by method. Tables 9–12 in Appendix C show the mean of the true discovery rates by method. For cases simulated, the PBF present smaller false positive rate and greater true discovery rate than t-tests.

These results show a better performance of the PBF in relation to t-tests in identification of difference of means and variance. From the biological practical point of view, the PBF may identify gene differences which are not identified by TT, CT and BTT, specially, genes with differences in means and variances.

4.2 *Escherichia coli* bacterium data set

In this section, consider the gene expression data set on *Escherichia coli* bacterium, composed by $n = 4290$ genes (Arfin et al., 2000). Each gene g have four measure from the control and four measure from the treatment. Figure 1 shows the treatment and control observed means and variances for all genes of this dataset.



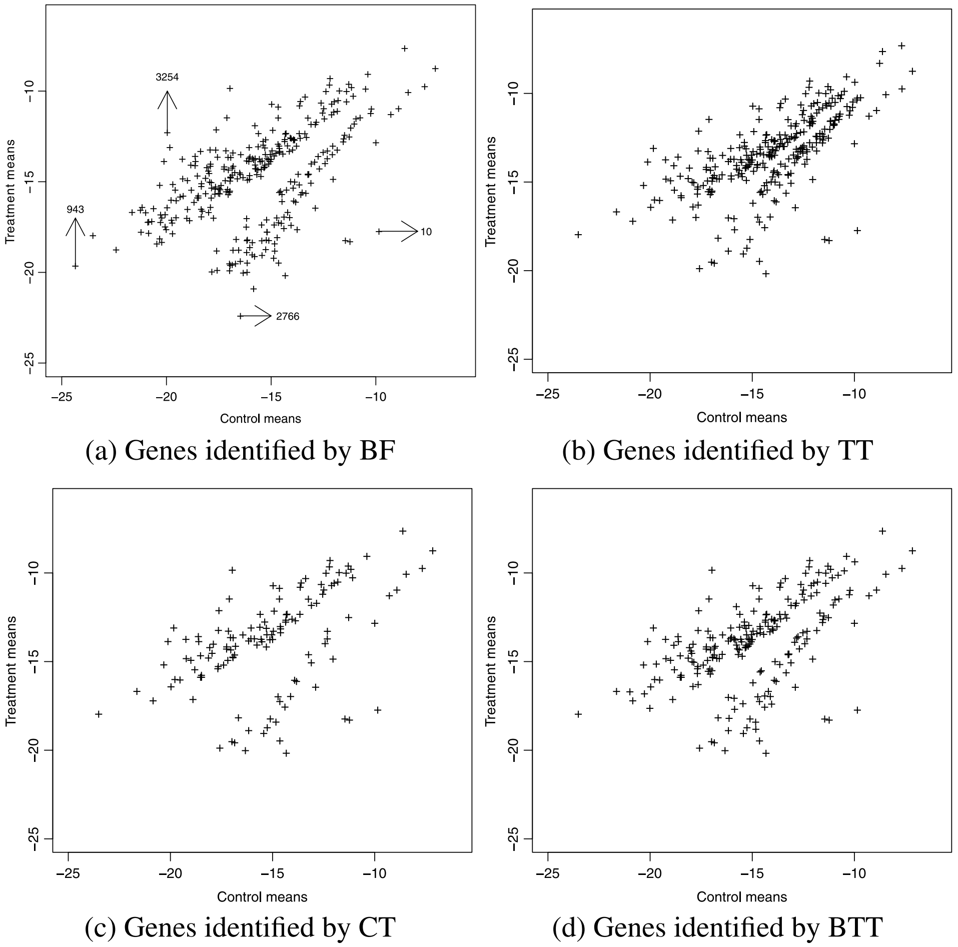
(a) Observed means (b) Observed variances

Figure 1 *Treatment and control observed means and variances.*

Results for the four methods are presented in Figures 2 and 3, respectively. In the figures, the signal "+" indicate genes identified with evidence for difference. Figure 2 show the observed treatment and control means of genes identified with evidence for difference by PBF, TT, CT and BTT, respectively. Figure 3 shows the results in relation to the observed treatment and control variances. The PBF identifies 236 genes with evidences for difference, while TT identifies 287, CT 131 and BTT 217 genes.

Genes with means well apart are identified by the four methods, as can be noted in Figure 2. An example is the gene 10 (hdeB) that is highlighted in Figures 2(a) and 3(a). But, cases with mean and variances well apart are not identified by TT, CT and BTT and are identified by PBF, as can be noted in Figure 3. Examples are genes 943 (b0562 (f143)), 2766 (b1326 (f262)) and 3254 (dbpA) that are highlighted in Figures 2(a) and 3(a). One possible reason is the low performance of TT, CT and BTT in situations with differences in means and variances, as observed in the artificial data sets. Besides, it show us that PBF is capable of identify differentially expressed genes which are not identified by TT, CT and BTT, specially, genes with differences in means and/or variances.

However, from probability $P(M_1|\mathbf{y})$ the PBF does not differentiate among the situations differing only in means, only in variances, and in both parameters. For example, for gene 943 (see Figure 2(a)), we have $P(M_1|\mathbf{y}_{943}) = 0.9446$ and the conclusion is that this gene present evidences for difference. But, we can not say if this difference is in relation to mean, variance or both parameters only using $P(M_1|\mathbf{y}_{943}) = 0.9446$. It happens because PBF is a method developed to identify genes with different expression based on changes in distribution, not only with different means or variances of the expression in separately. Nevertheless, from observed mean and variance and Figures 2(a) and 3(a) we can see that gene 943 present mean and variance higher in treatment in relation to control.



(a) Genes identified by BF

(b) Genes identified by TT

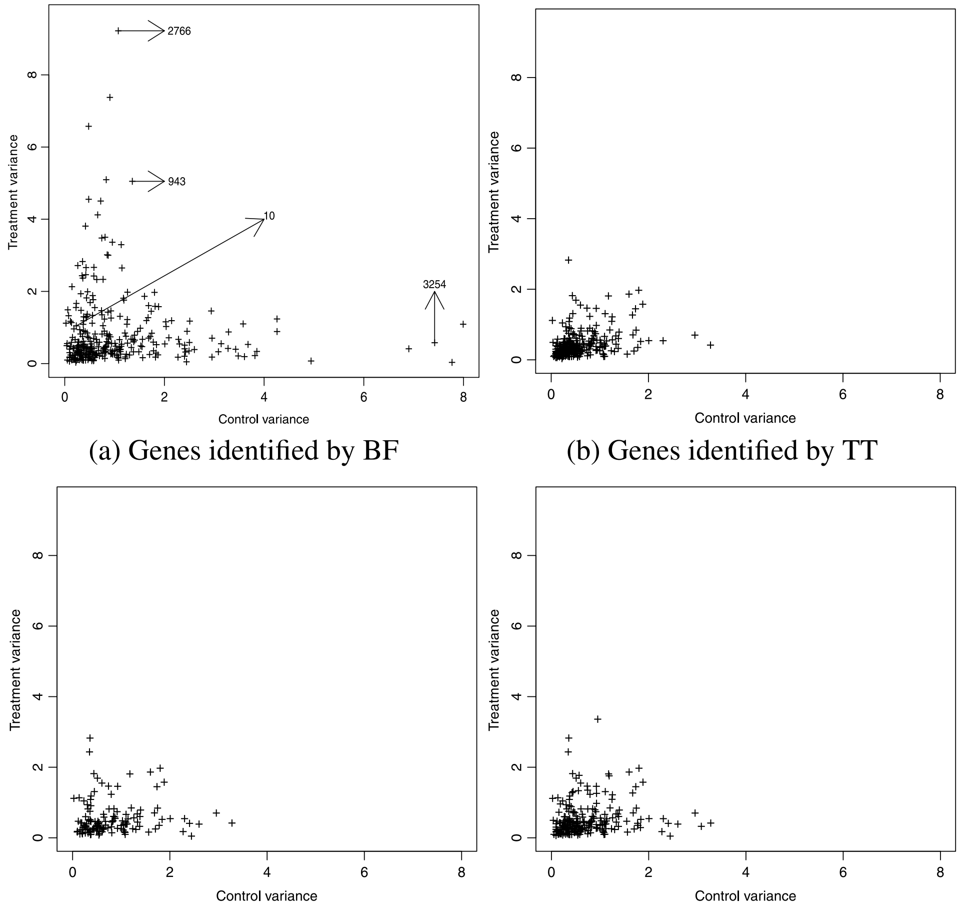
(c) Genes identified by CT

(d) Genes identified by BTT

Figure 2 *Treatment and control observed means and genes identified with evidence for difference.*

The proposed PBF is a method to identify genes with evidences for difference of gene expression levels in relation to mean and/or variances. The identification of genes differentially expressed is a first step to Biologist/Geneticist to study the genes identified in more details and answer some questions. Thus, for gene 943 cited, the PBF method concludes that this gene present evidences for difference because $P(M_1|\mathbf{y}_{943}) = 0.9446$. Identified the evidence for difference the biologist/geneticist will study this gene in details and so answer questions like: if the variation in expression measurements is due more sources of variation or whether is due the treatment condition, or if gene reacted to a drug stimulus and so on.

To illustrate the influence of each observation from treatment group in the value of $P(M_1|\cdot)$ given the value of B_{10} calculated as in (3.12), we present the Fig-



(a) Genes identified by BF

(b) Genes identified by TT

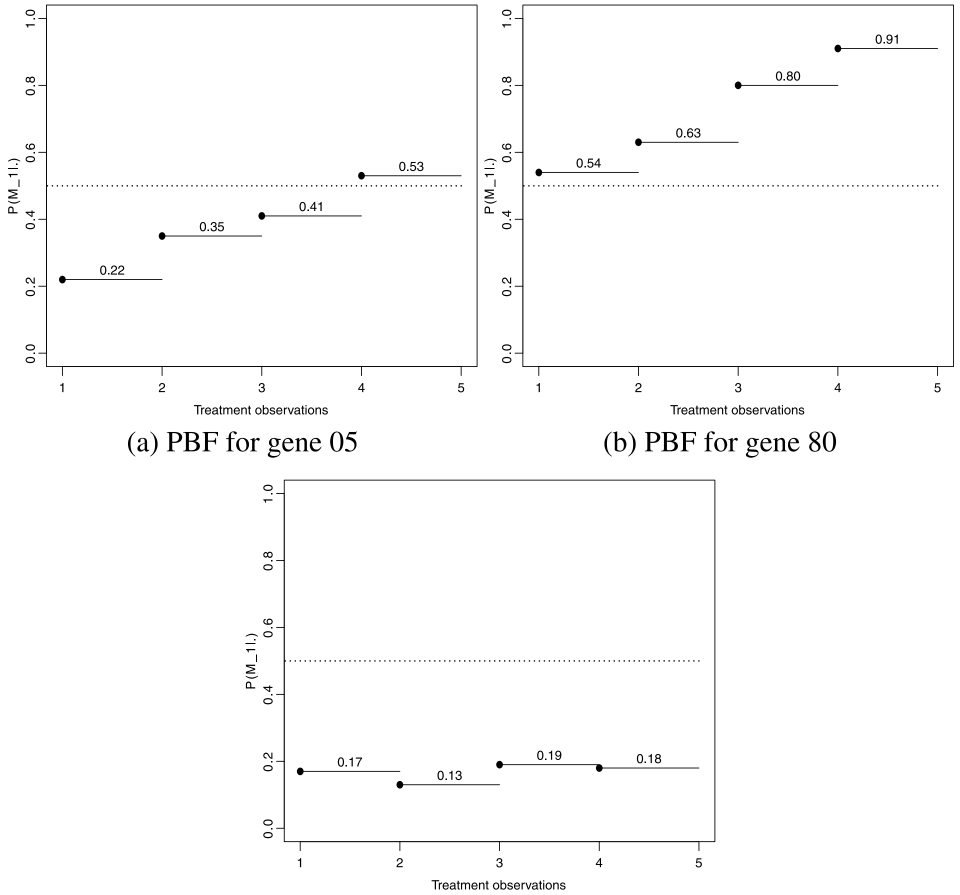
(c) Genes identified by CT

(d) Genes identified by BTT

Figure 3 *Treatment and control observed variances and genes identified with evidence for difference.*

ure 4. This figure shows the value of $P(M_1|\cdot)$ for B_{10} calculated according to PBF (equation (3.12)), for genes 05, 80 and 156. In this figure, each "•" (from left to right) represent the value of $P(M_1|\cdot)$ given the set $\mathbf{x}_i = \{x_{1g_t}, \ldots, x_{n_t g_t}\}$, for $i = 1, 2, 3, 4$. For instance, the second "•" is the value of $P(M_1|\cdot)$ given the set $\mathbf{x}_2 = \{x_{1g_t}, x_{2g_t}\}$. The hatched line represent the value $P_{\text{ref}} = 0.5$.

The gene 05 (Figure 4(a)) is considered differentially expressed by PBF give the 4 observations, $P(M_1|\cdot) = 0.53$. Excluding the fourth observation the gene would be considered nondifferentially, $P(M_1|\cdot) = 0.41$. For genes 80 and 156 the PBF always indicate the situation differentially and nondifferentially expressed, respectively.



(a) PBF for gene 05

(b) PBF for gene 80

(c) PBF for gene 156

Figure 4 *PBF for genes* 05, 80 *and* 156.

5 Discussion

We propose a Bayesian approach to identify differentially expressed genes based on posterior probability of the difference which is calculated via Bayes factor. In order to calculate the Bayes factor considering the model fitted by observed data from the control experimental condition, we use the predictive density.

The performance of the proposed PBF method as well as its comparison with the TT, CT and BTT was verified on an artificial and a real datasets. Results from the artificial and the real data sets show a better performance of PBF in relation to TT, CT and BTT in identification of difference, mainly, in situations with variance difference.

Results also suggest a possible complementarity among the methods, where, cases with difference of means and similar variances are easily identified by the t-tests, while cases with changes in variance with or without expressive changes in mean are adequately identified by PBF. The advantage of the PBF method is that it can identify genes which are not identified by the usual approaches. The biological interest in this fact is that PBF may bring to light genes that are not identified when using only t-test or modified t-test ones. Moreover, the PBF methods can be easily implemented in usual softwares such as the software R (the Comprehensive R Archive Network, http://cran.r-project.org). The source code used in data set analysis can be obtained by email the authors.

Although we have considered $P_{\text{ref}} = 0.5$ as the cutoff value to choose between models using $P(M_1|\mathbf{y}_g)$, other values of P_{ref} up to 1 may be considered. The PBF reduces the number of identification, but continues identifying genes with means well apart that are not identified by the others methods.

In this paper, we consider such first level of analysis, as also made by Bald and Long (2001) and Fox and Dimmic (2006), making a comparison gene by gene. However, considering that a challenge in gene expression data analysis is to deal with the multiple testing problem, a further development is to the proposed a method to control the false discovery rate when thousands of hypotheses are realized simultaneously. Besides, to extend the proposed method to second level of analysis in order to identify clusters of genes.

Appendix A: Some issues on hyperparameter specification

As discussed by Sinharay and Stern (2002) and Kass and Raftery (1995), the Bayes factor is sensible to choice of prior hyperparameters. Thus, as done by Baldi and Long (2001) and Fox and Dimmic (2006), we opt to describe a procedure to specify the hiperparameters than realize a sensibility analysis.

Following the scheme used by Richardson and Green (1997) and Stephens (2000) to fix hyperparameters in the context of mixture models, we use the range of the data to define the hyperparameters of the prior distribution on σ^2. We then set up τ and β in a way that $E[\sigma_d^2] = (\beta/2)/[(\tau/2) - 1] = R$, where $R = \max(\mathbf{x}_g) - \min(\mathbf{x}_g)$ is the length of the interval of variation of the observed data $\mathbf{x}_g = \{\mathbf{x}_{g_c} \cup \mathbf{x}_{g_t}\}$, $d \in \{g, g_c, g_t\}$ and $g = 1, \dots, n$. Thus, we obtain $\beta = (\tau - 2)R$; and we fix $\tau = 3$. Then, the prior distribution on σ^2 is

$$\sigma^2|\tau, \beta \sim \mathcal{IG}\left(\frac{3}{2}, \frac{R}{2}\right).$$

For example, for gene 943 that present $P(M_1|\mathbf{y}_{943}) = 0.9446$, we have $R = -15.5120 + 26.9905 = 11.4785$. Using the software R (fixing set.seed(10)), we generate 1000 values of the distribution $\mathcal{IG}(1.5, 5.73925)$ and we obtain an average of 10.7918 and a variance of 857.4616. For gene 928 that presents $P(M_1|\mathbf{y}_{943}) = 0.1085$, we have $R = -14.0489 + 17.7174 = 3.6685$. Generating 1000 values of the distribution $\mathcal{IG}(1.5, 1.8344)$, we obtain an average of 3.4490 and a variance of 87.5842.

To define the hyperparameters of the prior distribution on μ we set up: (i) μ_0 as being the middle point of the interval of variation of $\mathbf{x}_g$, $\mu_0 = [\min(\mathbf{x}_g) + \max(\mathbf{x}_g)]/2$; (ii) λ at 10^{-2}, $\lambda = 10^{-2}$. Thus, the prior distribution on μ is

$$\mu|\lambda, \sigma^2 \sim N(\mu_0, 100 \cdot R).$$

Considering $\sigma^2 = E[\sigma^2] = R$, so for gene 943, we have $\mu|\lambda, \sigma^2 \sim \mathcal{N}(-21.2512, 1147.843)$. For gene 928, $\mu|\lambda, \sigma^2 \sim \mathcal{N}(-15.8832, 366.8518)$. The same scenario remains for the others genes of the dataset.

From examples above, we consider this procedure as a good way to define the hyperparameters due the prior distributions obtained be weakly informative, that is, with large variance.

Appendix B: False positive rate

The false positive rate is given by

$$\text{FPR}_{\text{method}} = \frac{\sum_{g=1}^{n} (1 - \mathbb{I}_g) \cdot \mathbb{I}_g^{\text{method}}}{n - \sum_{g=1}^{n} \mathbb{I}_g},$$

where method = {PBF, TT, CT, BTT}. See Tables 5–8.

Appendix C: True discovery rate

The true discovery rate is given by

$$\text{TDR}_{\text{method}} = \frac{\sum_{g=1}^{n} \mathbb{I}_g \cdot \mathbb{I}_g^{\text{method}}}{\sum_{g=1}^{n} \mathbb{I}_g^{\text{method}}},$$

where method = {PBF, TT, CT, BTT}. See Tables 9–12.

Acknowledgments

We thank the Reviews for their comments, suggestions and criticisms. The researches of Francisco Louzada is supported by the Brazilian organization CNPq.

Table 5 *False positive rate*, $n_{g_c} = n_{g_t} = 4$ *and* $p = 5\%$ $(p_{\text{over}} = 3\%,\ p_{\text{under}} = 2\%)$

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.017
	TT	0.041	0.041	0.041	0.042	0.042	0.041	0.041	0.041	0.041
	CT	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.029
	BTT	0.040	0.041	0.041	0.041	0.042	0.041	0.040	0.041	0.041
2	PBF	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018
	TT	0.040	0.042	0.041	0.041	0.041	0.041	0.042	0.042	0.041
	CT	0.029	0.031	0.030	0.030	0.030	0.030	0.030	0.030	0.030
	BTT	0.040	0.042	0.041	0.040	0.041	0.041	0.042	0.041	0.041
3	PBF	0.018	0.017	0.017	0.018	0.018	0.017	0.018	0.018	0.017
	TT	0.042	0.040	0.041	0.041	0.041	0.041	0.042	0.042	0.040
	CT	0.031	0.029	0.029	0.030	0.030	0.030	0.030	0.030	0.030
	BTT	0.042	0.040	0.041	0.041	0.041	0.042	0.042	0.041	0.041
4	PBF	0.017	0.017	0.018	0.017	0.017	0.018	0.018	0.017	0.018
	TT	0.040	0.041	0.042	0.040	0.041	0.041	0.042	0.040	0.041
	CT	0.029	0.029	0.030	0.029	0.029	0.029	0.030	0.029	0.030
	BTT	0.040	0.041	0.042	0.040	0.041	0.041	0.042	0.040	0.041

Table 6 *False positive rate*, $n_{g_c} = n_{g_t} = 4$ *and* $p = 10\%$ $(p_{\text{over}} = 5\%,\ p_{\text{under}} = 5\%)$

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.018	0.018	0.018	0.019	0.017	0.018	0.017	0.018	0.018
	TT	0.041	0.041	0.042	0.042	0.040	0.040	0.040	0.040	0.041
	CT	0.030	0.029	0.030	0.031	0.029	0.030	0.029	0.029	0.030
	BTT	0.041	0.041	0.042	0.043	0.040	0.041	0.040	0.040	0.041
2	PBF	0.018	0.018	0.017	0.018	0.018	0.018	0.017	0.018	0.018
	TT	0.041	0.041	0.041	0.041	0.042	0.041	0.041	0.041	0.042
	CT	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
	BTT	0.041	0.041	0.041	0.041	0.042	0.041	0.041	0.041	0.042
3	PBF	0.018	0.017	0.018	0.018	0.018	0.018	0.018	0.017	0.018
	TT	0.041	0.040	0.042	0.041	0.041	0.041	0.042	0.041	0.041
	CT	0.030	0.030	0.031	0.030	0.030	0.030	0.031	0.029	0.030
	BTT	0.041	0.041	0.042	0.041	0.041	0.041	0.042	0.040	0.042
4	PBF	0.018	0.017	0.018	0.018	0.017	0.018	0.017	0.019	0.017
	TT	0.040	0.041	0.040	0.041	0.040	0.042	0.040	0.042	0.041
	CT	0.029	0.030	0.030	0.030	0.028	0.030	0.029	0.031	0.030
	BTT	0.041	0.040	0.040	0.041	0.040	0.042	0.040	0.042	0.041

Table 7 *False positive rate*, $n_{g_c} = n_{g_t} = 8$ *and* $p = 5\%$ ($p_{\text{over}} = 3\%$, $p_{\text{under}} = 2\%$)

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.010	0.010	0.011	0.011	0.011	0.011	0.011	0.011	0.011
	TT	0.047	0.047	0.047	0.049	0.047	0.048	0.048	0.048	0.048
	CT	0.037	0.037	0.037	0.038	0.037	0.038	0.038	0.038	0.038
	BTT	0.047	0.047	0.047	0.049	0.047	0.048	0.048	0.048	0.048
2	PBF	0.011	0.010	0.011	0.010	0.011	0.011	0.010	0.010	0.010
	TT	0.047	0.048	0.048	0.049	0.048	0.048	0.049	0.048	0.047
	CT	0.037	0.037	0.038	0.038	0.038	0.038	0.038	0.038	0.037
	BTT	0.047	0.047	0.048	0.048	0.048	0.048	0.049	0.048	0.047
3	PBF	0.011	0.010	0.010	0.011	0.010	0.010	0.010	0.011	0.011
	TT	0.047	0.048	0.048	0.049	0.047	0.047	0.048	0.048	0.049
	CT	0.037	0.038	0.037	0.038	0.037	0.037	0.038	0.038	0.038
	BTT	0.047	0.048	0.048	0.048	0.047	0.047	0.048	0.048	0.048
4	PBF	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
	TT	0.048	0.047	0.048	0.048	0.048	0.046	0.048	0.047	0.047
	CT	0.037	0.036	0.037	0.038	0.037	0.036	0.037	0.037	0.037
	BTT	0.048	0.047	0.048	0.048	0.048	0.046	0.048	0.047	0.047

Table 8 *False positive rate*, $n_{g_c} = n_{g_t} = 8$ *and* $p = 10\%$ ($p_{\text{over}} = 5\%$, $p_{\text{under}} = 5\%$)

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.010	0.011	0.010	0.011	0.010	0.011	0.011	0.010	0.011
	TT	0.048	0.047	0.047	0.048	0.048	0.050	0.049	0.047	0.048
	CT	0.037	0.037	0.037	0.038	0.038	0.039	0.038	0.036	0.037
	BTT	0.047	0.047	0.047	0.048	0.048	0.049	0.048	0.047	0.048
2	PBF	0.011	0.011	0.011	0.010	0.010	0.011	0.010	0.010	0.010
	TT	0.049	0.048	0.047	0.047	0.048	0.050	0.048	0.047	0.048
	CT	0.038	0.038	0.037	0.037	0.037	0.039	0.038	0.037	0.038
	BTT	0.049	0.047	0.047	0.047	0.048	0.049	0.048	0.047	0.047
3	PBF	0.010	0.011	0.011	0.010	0.010	0.011	0.011	0.010	0.010
	TT	0.048	0.048	0.048	0.048	0.047	0.049	0.047	0.049	0.048
	CT	0.037	0.038	0.038	0.037	0.037	0.039	0.037	0.038	0.037
	BTT	0.048	0.048	0.048	0.048	0.047	0.049	0.047	0.049	0.047
4	PBF	0.011	0.011	0.010	0.011	0.019	0.010	0.010	0.019	0.010
	TT	0.048	0.048	0.048	0.048	0.048	0.048	0.047	0.048	0.047
	CT	0.038	0.038	0.037	0.038	0.037	0.037	0.037	0.037	0.037
	BTT	0.048	0.048	0.048	0.048	0.047	0.047	0.047	0.048	0.047

Table 9 *True discovery rate,* $n_{g_c} = n_{g_t} = 4$ *and* $p = 5\%$ ($p_{\text{over}} = 3\%$, $p_{\text{under}} = 2\%$)

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.052	0.066	0.114	0.184	0.290	0.409	0.499	0.566	0.641
	TT	0.049	0.060	0.086	0.128	0.191	0.275	0.340	0.389	0.440
	CT	0.049	0.058	0.089	0.132	0.210	0.291	0.369	0.432	0.495
	BTT	0.050	0.057	0.087	0.129	0.193	0.271	0.342	0.390	0.443
2	PBF	0.157	0.149	0.197	0.249	0.322	0.363	0.427	0.503	0.552
	TT	0.061	0.059	0.074	0.100	0.126	0.152	0.189	0.238	0.269
	CT	0.062	0.064	0.082	0.110	0.142	0.175	0.213	0.282	0.312
	BTT	0.063	0.064	0.079	0.105	0.130	0.161	0.202	0.253	0.285
3	PBF	0.296	0.320	0.321	0.353	0.381	0.442	0.455	0.502	0.544
	TT	0.061	0.069	0.078	0.082	0.095	0.117	0.133	0.150	0.185
	CT	0.068	0.083	0.099	0.097	0.114	0.148	0.163	0.186	0.225
	BTT	0.070	0.076	0.088	0.088	0.106	0.135	0.150	0.169	0.204
4	PBF	0.462	0.466	0.466	0.505	0.497	0.517	0.523	0.565	0.561
	TT	0.066	0.071	0.065	0.081	0.081	0.096	0.102	0.123	0.132
	CT	0.085	0.090	0.084	0.103	0.105	0.128	0.139	0.158	0.171
	BTT	0.079	0.081	0.079	0.097	0.097	0.112	0.121	0.143	0.156

Table 10 *True discovery rate,* $n_{g_c} = n_{g_t} = 4$ *and* $p = 10\%$ ($p_{\text{over}} = 5\%$, $p_{\text{under}} = 5\%$)

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.090	0.141	0.203	0.330	0.469	0.583	0.692	0.740	0.784
	TT	0.097	0.120	0.162	0.244	0.341	0.440	0.517	0.577	0.622
	CT	0.098	0.122	0.169	0.263	0.368	0.462	0.557	0.620	0.667
	BTT	0.095	0.118	0.162	0.244	0.345	0.439	0.519	0.581	0.625
2	PBF	0.267	0.292	0.329	0.413	0.471	0.548	0.632	0.678	0.721
	TT	0.116	0.127	0.138	0.180	0.215	0.279	0.337	0.392	0.432
	CT	0.120	0.132	0.149	0.201	0.239	0.310	0.378	0.437	0.484
	BTT	0.119	0.129	0.145	0.191	0.227	0.290	0.350	0.403	0.449
3	PBF	0.473	0.469	0.498	0.532	0.560	0.596	0.640	0.684	0.710
	TT	0.128	0.128	0.140	0.169	0.183	0.212	0.244	0.276	0.318
	CT	0.148	0.151	0.163	0.202	0.216	0.250	0.291	0.334	0.379
	BTT	0.143	0.144	0.156	0.186	0.206	0.235	0.271	0.308	0.349
4	PBF	0.623	0.643	0.646	0.640	0.687	0.683	0.711	0.714	0.749
	TT	0.122	0.143	0.142	0.143	0.172	0.179	0.203	0.220	0.252
	CT	0.149	0.178	0.177	0.177	0.216	0.223	0.260	0.269	0.321
	BTT	0.146	0.167	0.171	0.164	0.199	0.206	0.237	0.257	0.294

Table 11 *True discovery rate, $n_{g_c} = n_{g_t} = 8$ and $p = 5\%$ ($p_{\text{over}} = 3\%$, $p_{\text{under}} = 2\%$)*

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.055	0.113	0.185	0.363	0.549	0.662	0.740	0.791	0.814
	TT	0.051	0.076	0.138	0.233	0.341	0.412	0.460	0.498	0.515
	CT	0.052	0.081	0.148	0.250	0.370	0.451	0.508	0.552	0.572
	BTT	0.051	0.077	0.137	0.235	0.343	0.414	0.461	0.499	0.516
2	PBF	0.321	0.361	0.420	0.493	0.581	0.660	0.726	0.749	0.778
	TT	0.051	0.059	0.092	0.132	0.192	0.252	0.303	0.362	0.411
	CT	0.054	0.065	0.102	0.148	0.215	0.286	0.348	0.408	0.460
	BTT	0.053	0.064	0.099	0.138	0.198	0.261	0.313	0.371	0.418
3	PBF	0.690	0.701	0.708	0.714	0.742	0.757	0.774	0.785	0.786
	TT	0.050	0.064	0.074	0.091	0.121	0.160	0.195	0.248	0.277
	CT	0.059	0.075	0.089	0.109	0.142	0.186	0.232	0.291	0.326
	BTT	0.057	0.073	0.084	0.102	0.136	0.179	0.214	0.263	0.297
4	PBF	0.795	0.795	0.805	0.801	0.803	0.811	0.812	0.812	0.822
	TT	0.056	0.056	0.063	0.070	0.095	0.121	0.145	0.160	0.206
	CT	0.070	0.069	0.077	0.086	0.118	0.149	0.183	0.196	0.248
	BTT	0.064	0.065	0.072	0.084	0.112	0.139	0.161	0.181	0.227

Table 12 *True discovery rate, $n_{g_c} = n_{g_t} = 8$ and $p = 10\%$ ($p_{\text{over}} = 5\%$, $p_{\text{under}} = 5\%$)*

		δ								
γ	Method	0	0.25	0.50	0.75	1	1.25	1.5	1.75	2
1	PBF	0.095	0.172	0.358	0.547	0.723	0.803	0.860	0.891	0.902
	TT	0.102	0.143	0.266	0.387	0.512	0.586	0.646	0.679	0.689
	CT	0.099	0.146	0.278	0.409	0.547	0.625	0.694	0.728	0.738
	BTT	0.103	0.144	0.266	0.385	0.513	0.587	0.648	0.682	0.691
2	PBF	0.500	0.511	0.596	0.682	0.757	0.799	0.835	0.867	0.887
	TT	0.106	0.118	0.175	0.244	0.334	0.410	0.482	0.547	0.593
	CT	0.111	0.127	0.193	0.267	0.372	0.447	0.525	0.598	0.640
	BTT	0.112	0.124	0.184	0.256	0.347	0.421	0.494	0.557	0.602
3	PBF	0.826	0.825	0.829	0.850	0.860	0.860	0.874	0.883	0.894
	TT	0.110	0.116	0.136	0.189	0.238	0.278	0.349	0.393	0.457
	CT	0.130	0.134	0.157	0.219	0.277	0.317	0.398	0.446	0.513
	BTT	0.123	0.129	0.151	0.208	0.258	0.299	0.373	0.419	0.479
4	PBF	0.891	0.890	0.893	0.892	0.897	0.900	0.903	0.905	0.909
	TT	0.113	0.108	0.122	0.144	0.182	0.218	0.260	0.311	0.366
	CT	0.137	0.131	0.145	0.175	0.217	0.267	0.308	0.368	0.422
	BTT	0.128	0.125	0.144	0.168	0.208	0.246	0.289	0.342	0.393

References

Aitkin, M. (1991). Posterior Bayes factor. *Journal of the Royal Statistical Society, Ser. B* **53**, 111–142.

Allison, D. B., Cui, X., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W. (2000). Global gene expression profiling in Escherichia coli K12. *The Journal of Biological Chemistry* **275**, 29672–29684.

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

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122. MR1394065

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–668.

Fox, R. J. and Dimmic, M. W. (2006). A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics* **7**, 126.

Hatifield, G. W., Hung, S. and Baldi, P. (2003). Differential analysis of DNA microarray gene expression data. *Molecular Microbiology* **47**, 871–877.

Kass, R. and Raftery, A. (1995). Bayes factor. *Journal of the American Statistical Association* **90**, 773–795.

Lavine, M. and Schervish, M. J. (1999). Bayes factor: What they are and what they are not. *The American Statistician* **53**, 119–122. MR1707756

Lönnstedt, I. and Speed, T. (2001). Replicated microarray data. *Statistica Sinica* **12**, 31–46. MR1894187

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixture with unknown number of components. *Journal of the Royal Statistical Society* **59**, 731–792. MR1483213

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

Sinharay, S. and Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician* **56**, 196–201. MR1940207

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump method. *The Annals of Statistics* **28**, 40–74. MR1762903

Wu, T. D. (2001). Analyzing gene expression data from DNA microarray to identify candidates genes. *Journal of Pathology* **195**, 53–65.

F. Louzada
J. Cobre
ICMC
Universidade de São Paulo
SP
Brazil
E-mail: louzada@icmc.usp.br
jucobre@icmc.usp.br

E. F. Saraiva
FACET
Universidade Federal da Grande Dourados
Dourados
Brazil
E-mail: ErlandsonSaraiva@ufgd.edu.br

L. Milan
DEs
Universidade Federal de São Carlos
SP
Brazil
E-mail: dlam@ufscar.br