Using a calibration experiment to assess gene-specific information: full Bayesian and empirical Bayesian models for two-channel

BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 1 2006, pages 50–57
doi:10.1093/bioinformatics/bti750

Gene expression

Using a calibration experiment to assess gene-specific information: full Bayesian and empirical Bayesian models for two-channel microarray data

Marta Blangiardo[1,*], Simona Toti[1], Betti Giusti[2], Rosanna Abbate[2], Alberto Magi[2,3], Filippo Poggi[2], Luciana Rossi[2], Francesca Torricelli[3] and Annibale Biggeri[1]

[1]Department of Statistics 'G.Parenti', University of Florence & Biostatistic Unit, CSPO, Florence, [2]Department 'Area Critica Medico Chirurgica' University of Florence and [3]Cytogenetic and Genetic Unit, Careggi Hospital, AOC, Florence, Italy

Received on January 14, 2005; revised on August 4, 2005; accepted on October 27, 2005
Advance Access publication November 2, 2005
Associate Editor: Alvis Brazma

ABSTRACT

Motivation: Microarray studies permit to quantify expression levels on a global scale by measuring transcript abundance of thousands of genes simultaneously. A difficulty when analysing expression measures is how to model variability for the whole set of genes. It is usually unrealistic to assume a common variance for each gene. Several approaches to model gene-specific variances are proposed. We take advantage of calibration experiments, in which the probes hybridized on the two channels come from the same population (self–self experiment). In this case it is possible to estimate the gene-specific variance, to be incorporated in comparative experiments on the same tissue, cellular line or species.

Results: We present two approaches to introduce prior information on gene-specific variability from a calibration experiment: an empirical Bayes model and a full Bayesian hierarchical model. We apply the methods in the analysis of human lipopolysaccharide-stimulated leukocyte experiments.

Availability: The calculations are implemented in WinBugs. The codes are available on request from the authors.

Contact: m.blangiardo@imperial.ac.uk

1 INTRODUCTION

In the framework of microarray analysis there are two main research goals: one is the identification of differentially expressed genes among several varieties (class comparison), while the other is the discovery of clusters within a collection of samples (class discovery) (Simon *et al*., 2003). Class comparison is related to the assessment of exposure or treatment effects (i.e. comparison of gene expression for a population of smokers and non-smokers) and the comparison can be performed directly (i.e. loop design) or indirectly (i.e. reference design). Class discovery is based on distances between gene expression profiles of pairs of samples (Dobbin and Simon, 2002) and can be absolute or relative. To the aim of class comparison the classical statistical approach is based on modified Student *t*-test procedures where, for each gene, at the numerator there is the difference between gene expression levels in two conditions to be tested and at the denominator there is the square root of the variance, divided by the number of replicates (Wit and McClure, 2004, p. 183 and followings). In this context a crucial point is how to obtain a suitable estimate of the variance. Actually, when the number of replicates is very small the sampling distribution of the variance is very asymmetric, with higher probability for small values and a strong instability of the pivotal *t*-value. For this reason in the literature many authors proposed several procedures to stabilize the variability measure (Speed, 2003, p. 51 and followings). One possibility is to consider a unique variance estimate for the whole set of genes or a function of the variance for all the genes. This approach could be used for single array inference (e.g. the Bayesian approach of Newton *et al*., 2001). Generally speaking it implies a loss of power, because it tends to be very conservative and to increase the number of false negative results. A better way to proceed can be found in a parametric or not parametric framework.

In a parametric context, many authors consider gene-specific variance estimates for the denominator of the *t*-test, but add a stabilizing constant for the whole set of genes. Baldi and Long (2001) use a full Bayesian hierarchical model for the log-expression. They discuss point estimates for the parameters and hyperparameters values. Regularized expressions for the variance of each gene are derived combining the empirical variance with a prior variance σ_{g0}^2. Several choices for the prior are proposed and among them the variance of the neighboring genes contained in a window of predefined size w (i.e. ranking the genes on the base of their expression measure, the 50 genes immediately above or below the gene under consideration). An additional hyperparameter ν_0 (prior degrees of freedom) is necessary to determine the weight assigned to the prior variance. It is tuned so that its sum is equal to a given constant ($\nu_0 + n = K$).

Lönnstedt and Speed (2002) propose a method that can be classified as empirical Bayesian: differently from a full Bayesian approach, they do not define prior distributions on hyperparameters, but substitute them by a frequentist estimate based on the marginal distribution. In particular, the authors present a B_g-statistic (a Bayes posterior logodds) instead of the classical *t*-statistic used to classify the differentially expressed genes. Following the same philosophy,

*To whom correspondence should be addressed at Imperal College School of Medicine, Norfolk Place W2 IPG. London, UK.

the variance has a gene-specific component s_g^2 and a constant term a_0. Values of B_g are explicitly calculated assuming conjugate prior on the gene expression mean and variance.

Other authors have worked on specific parametric models for the errors, starting from the idea that the standard deviation for expression measure increases proportionally to the level of expression (Newton *et al*., 2001), but does not tend to 0 for not expressed genes. From this assumption Rocke and Durbin (2001) develop an error model including a gene-specific additive component and a gene-specific multiplicative one and propose several ways to estimate the models, based on negative controls, or replicates.

In a non-parametric framework Tusher *et al*. (2001) work on *t*-tests and assign a score t_g to each gene on the basis of its change in gene expression and relative to standard deviation calculated on repeated measures. Permutations are used to identify significantly altered genes and to estimate the false discovery rate. They introduce a 'fudge factor' s_0 to the denominator of *t*-test to avoid low expression genes dominate the results. It is chosen to minimize the coefficient of variation. This method is framed in a frequentist approach, does not assume any distribution on the parameters.

Very similar to the previous, Efron *et al*. (2001) propose a simple empirical Bayes model in which the fudge factor to be added at the denominator is the *90*th percentile of the standard deviation for all the genes. Delmar *et al*. (2004) develop a finite mixture model for the marginal gene-specific distribution (which can be classified as non-parametric maximum likelihood). In particular, estimating gene-specific variance can be seen as a classification problem, where the number of components and the gene belonging are estimated. Since the number of groups is much lower than the number of genes, the estimates of group variance are very stable.

Heuristically, Comander *et al*. (2004) pooled genes to calculate more reliable variance estimates by average of minimum intensity values. There is no parametric statistical modelling of variance as function of intensity, but instead a loess smoothed estimate of variance is derived. Uncertainty in this procedure is not considered and a *Z*-test is used.

All the previous approaches work with a classical comparative experiment (with replications), where samples from two populations are compared. A different approach is introduced by Tseng *et al*. (2001) who propose calibration experiments in which the probes hybridized on the two channels come from the same population (self–self experiment). Such experiments make possible to incorporate the gene-specific variability information in comparative experiments on the same tissue, cellular line or species, with a prior ignorance on the remaining parameters and represent an alternative way to face the problem of variance estimate.

We followed the Tseng's approach and performed a calibration experiment before doing the comparative one. We built a full Bayesian model and a simpler Empirical Bayesian model. We analysed data on lipopolysaccharide (LPS) stimulated and un-stimulated human leukocyte, obtaining prior knowledge on variability from self–self experiment.

The structure of the paper is as follows. In Section 2 we describe the calibration and comparative experiments (Subsection 2.1) and the data preprocessing phase (Subsection 2.2); in Section 3 we present the normalization procedure used, and then focus the attention on the full Bayesian model and on the Empirical Bayesian one; model graphs and details on implementation follow; in Section 4 we describe the results in terms of differentially expressed genes; In Section 5 a sensitivity analysis is reported and in Section 6 we discuss the differences between the two models.

2 MATERIALS

2.1 LPS microarray experiment

2.1.1 Calibration experiment Mononuclear cells were obtained from peripheral blood (PMBC) of 10 healthy subjects by density gradient centrifugation on Ficoll-Hypaque. Cells from each subjects were incubated in RPMI 1640 at 37° in a humidified atmosphere with 5% CO_2 for 3 h in standard conditions (absence of lipopolysaccharide). Total RNA was extracted and equal amount of total RNA from different subjects was pooled. Total RNAs were split into six aliquots and then retro-transcribed with amino-allyl-dUTP, hydrolysed, purified and labelled with NHS-Cyanine dyes (three aliquots with Cy3, probe A and three aliquots with Cy5, probe B). Then, three arrays were produced having the two probes purified, mixed and hybridized on the arrays. After incubation, the three arrays were scanned by the 4000B scanner (Axon). Image analysis was performed by GenePix 4.1 software.

2.1.2 Comparative experiment Mononuclear cells were obtained from peripheral blood (PMBC) of the same 10 healthy subjects used in calibration experiment by density gradient centrifugation on Ficoll-Hypaque. Cells from each subjects were divided into two aliquots; the first was incubated in RPMI 1640 at 37° in a humidified atmosphere with 5% CO_2 for 3 h in the presence of LPS (10 μg/ml, stimulated cells). The second was incubated in the same conditions but in the absence of LPS (un-stimulated cells). Total RNA was extracted and equal amount of total RNA separately, from stimulated or un-stimulated cells, was pooled. Total RNAs were retro-transcribed with amino-allyl-dUTP, hydrolysed, purified and labelled with NHS-Cyanine dyes following th dye-swap design (Cy3 and Cy5, coupled, to un-stimulated and stimulated specimens). The two probes were purified, mixed and hybridized on the arrays. After incubation, arrays were scanned by the 4000B scanner (Axon). Image analysis was performed by GenePix 4.1 software. For the comparative experiment, two arrays finally were printed according to the dye-swap design.

Therefore, the complete experiment consists in 5 arrays made up 22×21 spots grid, for a total of 14 784 spots. The 14 784 spots included 13 971 oligonucleotides representing each one different gene, 29 negative controls (mixtures of oligonucleotide of other organisms), 2 positive controls (a mixture of all the human oligonucleotides) and 872 blanks (only printing solution). Out of 14 784, 1502 (10.2%) spots were absent because of a failure during the printing procedure.

2.2 Microarray data preprocessing

2.2.1 Quality control The process of microarray fabrication is subjected to many sources of variability and could contain a large amount of noise. In particular, it is possible that the noise dominates the signal for some spots. We applied the quality control present in GenePix Pro 4.1, with the aim of evaluating the presence of artefacts (bubbles, hair, fibres). After GenePix Pro 4.1 quality control and the visual inspection, the analysable spots resulted 80, 87 and 90% as concerned the 3 self–self experiments, and 83 and 87%, for the 2 arrays of the comparative experiment.

2.2.2 Spots selection for the analysis of gene-specific variances To the purpose of the present paper, we restricted our attention to a subset of genes for which extraneous sources of variability can be excluded. To select these spots all the five arrays were screened following the criteria suggested by Simon *et al*. (2003). In particular, we excluded a spot if the number of pixels used to calculate the intensity was less than 25 for the foreground intensity in either channel, if the signal was lower than 200 for both the channels or if the ratio between the average foreground intensity and the median background intensity was smaller than 1.5 in either channel. Spots with a large signal for one channel and low, undetectable signal for the other were not eliminated, but modified to become analysable, forcing the low intensity signal (defined as less than 200) to 200. In this paper we considered 2887 genes represented in all the 5 arrays (3 calibration arrays and 2 comparative arrays).

3 METHODS

In this section we present the two methods we used to analyse the data. The first model, is a full Bayesian hierarchical model while the second, originally proposed by Tseng *et al*. (2001), is an instance of the empirical Bayes approach.

3.1 Normalization

We performed two different types of normalization (Yang *et al*., 2002): for each slide a local A-dependent normalization (loess), considering all the genes present on the array, is used for empirical Bayes model. For Bayesian hierarchical model, the normalization step was part of the modelling phase.

3.2 Models

3.2.1 Bayesian hierarchical model The model is split into two parts.

Calibration model. The first submodel is used to estimate gene-specific variances from the calibration experiment. To this purpose we specified the following model, which is in the same philosophy of Lewin *et al*., 2005, for the unnormalized log-intensity

$$x_{igc} \sim N(\mu_{igc}, {}_x\sigma_g), \tag{1}$$

where i denotes array ($i = 1, 2, 3$), g denotes gene $g = 1, \ldots, 2887$ and c denotes channel $c = 1, 2$, where as usual $c = 1$ denotes Cy3 dye and $c = 2$ denotes Cy5 dye. For notation simplicity we refer to ${}_x\sigma_g$ as the variance.

The normalization procedure was achieved by an ANOVA model (see Kerr *et al*., 2002 for a general introduction to the analysis of variance approach to microarray data)

$$\mu_{igc} = \alpha_{ig} + \delta_c + \gamma_g, \tag{2}$$

where α_{ig} denotes the gene-specific array–gene interactions, δ_c the dye-effects and γ_g the normalized gene effects. $\gamma_g \sim N(\mu_\gamma, \sigma_\gamma)$ are exchangeable, with μ_γ non-informative Gaussian and $1/\sigma_\gamma$ non-informative Gamma hyperpriors. All the other normalization parameters were fixed effects modelled with non-informative Gaussian hyperpriors. The gene-specific variances were assumed to follow a Lognormal distribution ${}_x\sigma_g \sim \log N(\mu_\sigma, \sigma_\sigma)$ with $\mu_\sigma \sim N(0, 10\,000)$ and $1/\sigma_\sigma \sim Ga(0.001, 0.001)$ non-informative hyperpriors. This assumption of a skewed distribution for variance is standard and flexible enough to allow high variances for few genes.

Comparative model. The second submodel is specified for the comparative experiment and incorporates relevant information from the calibration experiment. The kernel likelihood is the same as for the calibration model. For the i-th array ($i = 1, 2$) the unnormalized log-intensity

$$x_{igc} \sim N(\mu_{igc}, {}_x\sigma_g) \tag{3}$$

was modelled as Gaussian for gene g and channel $c = 1, 2$. The gene-specific variances were modelled as lognormal variables ${}_x\sigma_g \sim \log N(\mu_\sigma, \sigma_\sigma)$ with informative parameters values obtained from the self–self experiment. In particular, we assumed μ_σ equal to the mean of the appropriate posterior distribution on the self–self data:

$$\begin{aligned} E[\mu_\sigma \mid x^{\text{self}}] &= \frac{\int \mu_\sigma f(x^{\text{self}} \mid \mu_\sigma)\pi(\mu_\sigma)d\mu_\sigma}{\int f(x^{\text{self}}|\mu_\sigma)\pi(\mu_\sigma)d\mu_\sigma} \\ &= \frac{\int \mu_\sigma \int f(x^{\text{self}} \mid \mu_\sigma, \sigma_\sigma)\pi(\mu_\sigma, \sigma_\sigma)d\sigma_\sigma d\mu_\sigma}{\text{const}(x^{\text{self}})}, \end{aligned} \tag{4}$$

Where x^{self} are the self–self expression data and const(x^{self}) is a normalizing constant depending only on data. Analogously, for σ_σ we plugged in the posterior mean of the corresponding posterior distribution $f(\sigma_\sigma \mid x^{\text{self}})$.

A linear model was assumed for μ_{igc} as follows:

$$\mu_{igc} = \alpha_{ig} + \tau_g + \delta_c + \gamma_g. \tag{5}$$

Here the model terms τ_g can be interpreted as a normalized log-ratio and quantify the treatment (LPS) effects. Their distribution was assumed Gaussian with gene-specific mean μ_{τ_g} and variance σ_{τ_g}. Summarizing, the prior distributions for τ_g, μ_{τ_g} and σ_{τ_g} were assumed as follows:

$$\tau_g \sim N(\mu_{\tau_g}, \sigma_{\tau_g}) \tag{6}$$

$$\mu_{\tau_g} \sim N(\mu_\tau, \sigma_\tau), \quad 1/\sigma_{\tau_g} \sim Ga(\nu_\tau, \beta_\tau), \tag{7}$$

with informative hyperparameters μ_τ, σ_τ, ν_τ, β_τ.

This formulation is sensible since a Gaussian distributed effect parameter τ_g, on the log scale, is justified by most of the literature on generalized linear mixed models (see Clayton in Markov Chain Monte Carlo in Practice, 1996). The conjugate hyperpriors [Equation (6)] are standard and assume an exchangeable structure, i.e. same ignorance about the status of the gene (differentially or not differentially expressed). More sophisticated mixture models could be introduced (see Parmigiani *et al*., 2002).

Informative prior on log-ratio. Actually values for $\mu_\tau, \sigma_\tau, \nu_\tau, \beta_\tau$ were obtained from the calibration experiment as follows. On the calibration arrays we calculated a residual effect $r_{igc} = x_{igc} - \mu_{igc}$ and reconstructed a 'normalized log-ratio' under the null hypothesis for each slide as the difference between the residual effect of $c = 1$ channel and the residual effect of $c = 2$ channel is given as

$$t_{ig} = r_{ig1} - r_{ig2}, \tag{8}$$

where r_{igc} was the residual for the c-th channel on the i-th array ($i = 1, 2, 3$).

Then for each gene we calculated the plug-in values for the μ_{τ_g} prior as:

$$\hat{\mu}_\tau = \frac{1}{G}\sum_g t_{\cdot g} \tag{9}$$

$$\hat{\sigma}_\tau = \frac{1}{G-1}\sum_g (t_{\cdot g} - \hat{\mu}_\tau)^2, \tag{10}$$

where $t_{\cdot g} = \frac{1}{3}\sum_i t_{ig}$ (Fig. 1).

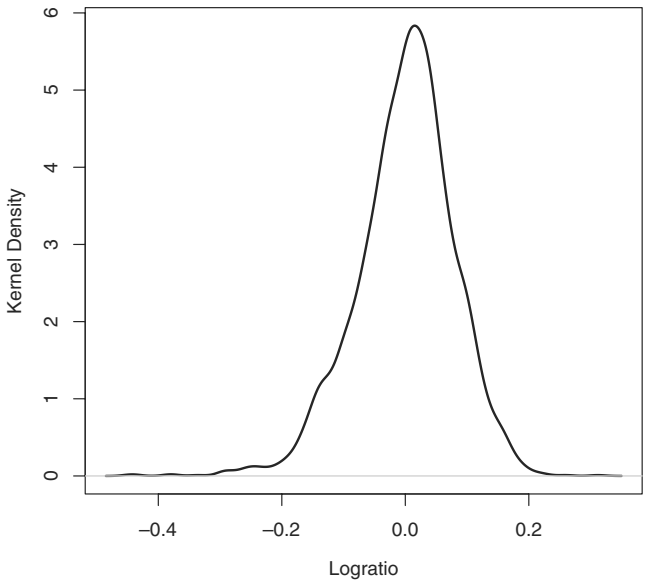

Fig. 1. Kernel density plot of normalized log-ratios $t_{\cdot g}$ for self–self experiment.

Similarly, we obtained the plug-in values for the prior Gamma parameters ν_τ and β_τ from the mean and variance of $\hat{\sigma}_{\tau_g} = [\frac{1}{2}\sum_i (t_{ig} - t_{\cdot g})^2]$:

$$\hat{\nu}_\tau = \mathrm{Ave}(\hat{\sigma}_{\tau_g}) \cdot \hat{\beta}_\tau \tag{11}$$

$$\hat{\beta}_\tau = \frac{\mathrm{Ave}(\hat{\sigma}_{\tau_g})}{\mathrm{Var}(\hat{\sigma}_{\tau_g})}, \tag{12}$$

where Ave(.) and Var(.) denote the average and variance operator.

3.2.2 Tseng's empirical Bayes model To adapt the model proposed by Tseng *et al*. (2001) we reformulated it as follow. We normalized the data externally by loess (Yang *et al*., 2002) through the MAANOVA library implemented in R (www.r-project.org) (Wu *et al*., 2003). The normalized log-ratio m_{ig} for g-th gene and i-th array were modelled as

$$m_{ig} \sim N(\tau_g, {}_m\sigma_g), \tag{13}$$

where τ_g was the mean and ${}_m\sigma_g$ was the variance of log-ratio over the replicates of the comparative experiment for the gene g. To make easy compare it with the full Bayesian model and the likelihood can be written as follows:

$$m_{ig} = \mathrm{normalized}(x_{ig1} - x_{ig2}) \tag{14}$$

$$m_{ig} \sim N(\mu_{ig}, {}_m\sigma_g), \tag{15}$$

where $\mu_{ig} = \tau_g$. The distribution of τ_g was assumed Gaussian with gene-specific parameters and all the hyperparameters had a classic Bayesian non-informative distribution [compare with Equations 6 and 7]. The information pooled from the calibration experiment was used to obtain an informative prior distribution for ${}_m\sigma_g$:

$${}_m\sigma_g \sim \frac{w_g}{\chi^2_k/k}, \tag{16}$$

where k was the number of degree of freedom of a χ^2-deviate; w_g was a weighted average of gene-specific and overall empirical variance calculated on the calibration arrays ($i = 1, \ldots, I^{\mathrm{self}}$) as follows:

$$\hat{s}_g = \frac{1}{I^{\mathrm{self}} - 1} \sum_{i=1}^{I^{\mathrm{self}}} (m_{gi}^{\mathrm{self}} - \bar{m}_{g\cdot}^{\mathrm{self}})^2 \tag{17}$$

$$\hat{s}_\cdot = \frac{1}{G} \sum_{g=1}^{G} \hat{s}_g \tag{18}$$

$$w_g = \frac{[(I^{\mathrm{self}} - 1) \cdot \hat{s}_g + \hat{s}_\cdot]}{I^{\mathrm{self}}}. \tag{19}$$

In other words, in the Tseng model the information on the gene-specific variability from the self–self experiment is utilized to derive an informative inverse Gamma prior.

However, the two variance modelling are deeply different. The empirical Bayes approach uses the information from the self–self experiment to plug in values of parameters of the gene-specific variance prior ${}_m\sigma_g \sim (w_g k)/[\Gamma(\frac{1}{2}, \frac{1}{2})]$; the full Bayes approach uses the posteriors given calibration data to obtain values for the hyperparameters of the hyperpriors governing the gene-specific variance priors $\sigma_{\tau_g} \sim 1/[\Gamma(\nu_\tau, \beta_\tau)]$.

3.2.3 Tseng's prior with internal normalization To better address model comparison we modified the empirical Bayes model proposed by Tseng including the normalization step into the model as follows:

$$x_{igc} \sim N(\mu_{igc}, {}_x\sigma_g) \tag{20}$$

$$\mu_{igc} = \alpha_{ig} + \tau_g + \delta_c + \gamma_g \tag{21}$$

$${}_x\sigma_g \sim \log N(\mu_\sigma, \sigma_\sigma), \tag{22}$$

where the parameters of the lognormal distribution on ${}_x\sigma_g$ were informative coming from the calibration experiment (see Subsection 3.2.1), and the normalization parameters were modelled following standard ANOVA [see Equation (5)]. The hyperpriors for τ_g were modelled following Tseng's proposal ($\sigma_{\tau_g} \sim (w_g k)/[\Gamma(\frac{1}{2}, \frac{1}{2})]$ (Fig. 2).

3.2.4 Bayesian hierarchical model with loess normalization We also modified the Bayesian hierarchical model to carry out a loess normalization instead of the linear one. We performed a loess normalization through MAANOVA library and then we calculated the normalized values for the two channels as follows:

$${}_nx_{ig1} = x_{ig1} - \tfrac{1}{2} l_{ig}, \quad {}_nx_{ig2} = x_{ig2} + \tfrac{1}{2} l_{ig}, \tag{23}$$

where 1 is the red channel, 2 is the green one and l is the coefficient used to scale the log-ratio in the classical global loess normalization.

The normalized channel intensity (on log scale) are

$${}_nx_{igc} \sim N(\mu_{igc}, {}_x\sigma_g) \tag{24}$$

and we perform a further normalization in calibration experiment

$$\mu_{igc} = \alpha_{ig} + \gamma_g \tag{25}$$

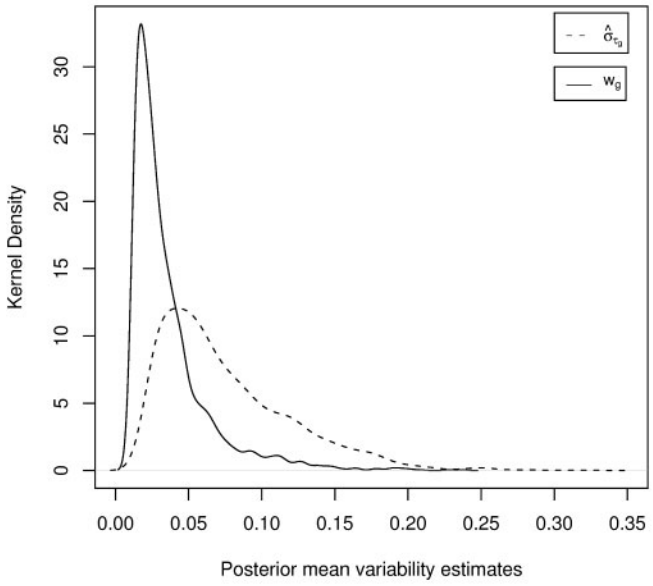

Fig. 2. Kernel density plot of estimates of gene-specific variability w_g and $\hat{\sigma}_{\tau_g}$ from self–self experiment.

as well as in comparative experiment

$$\mu_{igc} = \alpha_{ig} + \gamma_g + \tau_g \tag{26}$$

for eliminating the array effects that are not considered in the loess normalization performed separately for each slide. The model specification thereafter follows the structure defined in Equations (3)–(12).

3.3 The graph of the model

A system of conditional distributions can be often represented through the correspondent directed acyclic graph (DAG, directed for the link between each pair of nodes, acyclic for the impossibility of turning on the same node after leaving it, following the direction of the arrows) (Gilks *et al*., 1996). In a DAG the circles denote unobserved quantities, while single squares indicate observed quantities and double squares indicate a mathematical quantity; the arrows between the nodes are solid to mean a stochastic dependence, while dashed arrow denotes functional relationships; solid lines show stochastic undirected dependence. Repetitive structures (arrays, for example), are shown as stacked rectangles. Figure 3 shows the graph for the Bayesian hierarchical model presented in Section 3.2.1 while Figure 4 shows the DAG for Tseng's model presented in Section 3.2.2.

3.4 Implementation

To estimate the parameters of interest we use the marginal posterior distributions approximated by MCMC methods implemented in WinBugs 1.4 (Spiegelhalter *et al*., 2003); the Bayesian hierarchical model with ANOVA normalization as well as with loess normalization, and Tseng's model with internal normalization are estimated by Metropolis-within-Gibbs routine, a generalization of Gibbs that can be used for non-log concave sampling (Tanner, 1996); the Tseng's empirical Bayes model can also be fitted by Gibbs sampling in WinBugs. We have checked the convergence both visually by

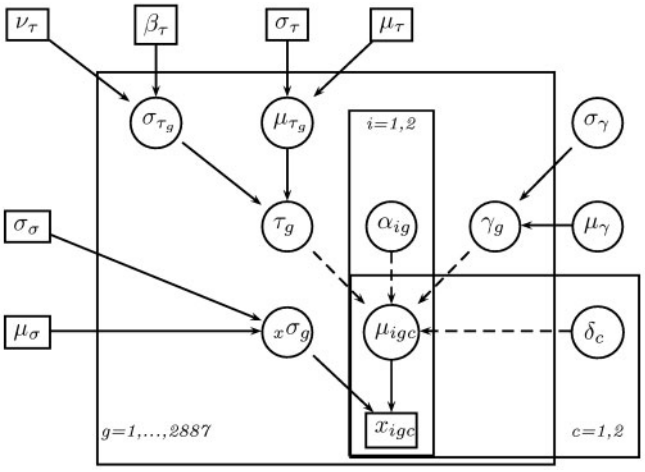

Fig. 3. Graph of hierarchical Bayesian model for treated samples (for untreated ones the τ_g effects are absent).

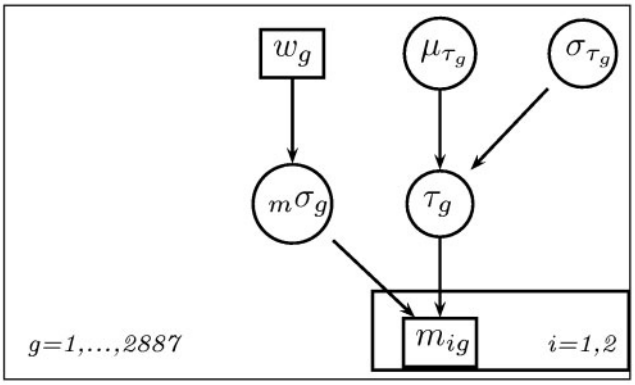

Fig. 4. Graph of Tseng's *et al*. model for normalized log ratios m_{ig}.

Gelman-Rubin statistics (Gelman and Rubin, 1992) and using different starting points. We have performed 10 000 burn-iterations followed by 4000 sampling iterations for all the models. Fitting the Bayesian hierarchical model on calibration experiment takes 1 h to do 100 iterations on a workstation HPXW6000 with 2 GbRAM and Intel Xeon CPU2. 8 GHz processor, for the large number of posterior distributions it has to store to be subsequently incorporated in the comparative experiment analysis. Performing the comparative experiment takes 380 s for 1000 iterations. Fitting Tseng's model takes 300 s to perform 1000 iterations.

4 RESULTS

We explored the posterior distribution of the treatment effects τ_g to identify the differentially expressed genes taking 95% two sides probability level. Genes found differentially expressed with at least one of the two methods are shown in Table 1. Using the Bayesian hierarchical model we found 26 differentially expressed genes. Out of 26 genes IFI30 and PRKAG2 were under-expressed in LPS stimulated leukocytes. Using the Tseng *et al*. one we found 46 differentially expressed genes. Out of 46 genes, 20 emerged

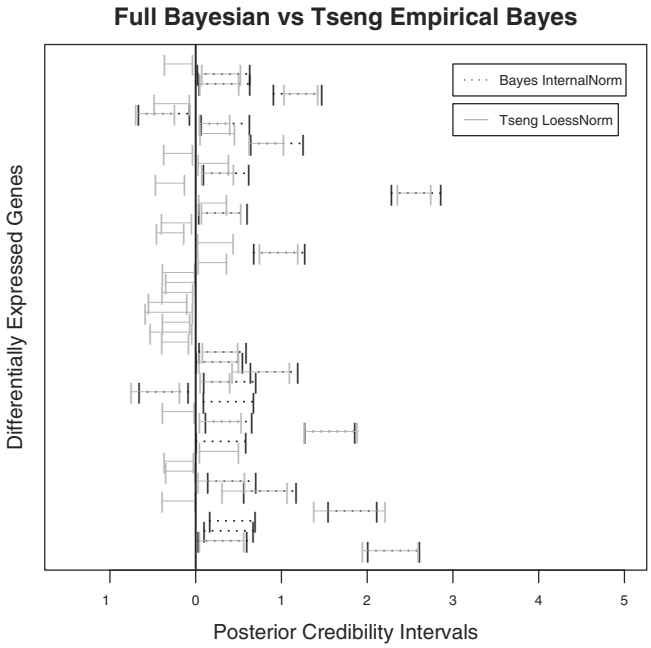

Fig. 5. Posterior credibility intervals at 95% for differentially expressed genes: full Bayesian model versus empirical Bayesian one.

downregulated in LPS stimulated leukocytes. Out of 26 genes, 22 identified by the first model were highlighted also by the Tseng *et al*. one (Fig. 5 and Table 1).

The LPS-induced transcripts identified by both models mainly consist of gene encoding protein associated with cytokines and chemokines including interleukin (IL)-1 beta, IL-1 receptor antagonist (RA), macrophage inflammatory protein (MIP)-1 alpha, MIP-1 beta, MIP-2 beta, MIP-3 alpha; cytoskeletal protein such as vimentin and cofillin 2 (Mor-Vaknini *et al*., 2003); and plasminogen activator inhibitor type 2 (PAI-2) (Pepe *et al*., 1997).

To facilitate the interpretation of our results, we reported the results obtained by a classic analysis of the comparative arrays only, without taking into account the calibration ones. The analysis of the comparative experiment by Tusher's SAM resulted in 18 significant differentially expressed genes, using a cut-off at $p = 0.01$. Fifteen were also identified by the Bayesian approaches. Due to the limited sample size, a low sensitivity is expected compared to the analysis which took into account the calibration arrays. The Bayesian approaches provided also a more stable inference on genes with small sample standard deviation, among which three were significant by SAM but were not confirmed by the Bayesian analyses. No negative log-ratios emerged as significative by SAM.

5 SENSITIVITY ANALYSIS AND MODEL COMPARISON

The results presented in the previous section are difficult to interpret comparatively because the two models use different normalization procedures. To gain insight on the behaviour of the different approaches we need to evaluate differentially expressed genes taking fixed the normalization procedure (Subsection 3.2.3).

Table 1. Differentially expressed genes: posterior mean and posterior credibility interval at 95%

ID	Symbol	Bayesian flierarchical model		Empirical Bayesian model	
		Post mean	Post CrI	Post mean	Post CrI
2064	VIM	0.32	(0.04,0.63)	0.28	(0.05,0.50)
2563	TAC1			0.20	(0.04,0.36)
2890	PRKCG	0.41	(0.14,0.70)	0.29	(0.03,0.57)
12183	KIAA0935			−0.20	(−0.37,−0.04)
14623	IFI30	−0.36	(−0.66,−0.09)	−0.46	(−0.75,−0.19)
23672	LRP6			−0.29	(−0.53,−0.04)
42500	ARL5			0.26	(0.05,0.45)
43265	MLSN1	0.39	(0.10,0.70)	0.22	(0.05,0.40)
68879	BPM4			−0.20	(−0.36,−0.04)
73817	SCYA3	2.30	(2.01,2.61)	2.28	(1.95,2.59)
75356	TCF4			0.22	(0.02,0.44)
75498	SCYA20	0.92	(0.64,1.19)	0.75	(0.43,1.10)
75703	SCYA4	1.56	(1.27,1.86)	1.57	(1.27,1.88)
75716	SERPINB2	1.19	(0.91,1.47)	1.22	(1.03,1.42)
76095	IER3	0.87	(0.56,1.17)	0.68	(0.31,1.07)
78452	SLC20A1			−0.17	(−0.35,0)
81134	IL1RN	0.96	(0.64,1.26)	0.82	(0.62,1.02)
89690	GRO3	0.98	(0.68,1.27)	0.97	(0.74,1.19)
92381	—			−0.17	(−0.35,−0.01)
99508	—			−0.20	(−0.39,−0.02)
100015	HAB1			−0.33	(−0.55,−0.1)
103839	KIAA0987			−0.19	(−0.39,−0.01)
103931	DKF2P434B	0.28	(0.01,0.55)	0.27	(0.04,0.50)
118463	TTS-2.2			−0.23	(−0.39,−0.08)
126256	IL1B	2.57	(2.28,2.86)	2.55	(2.36,2.74)
129727	KIAA0464			−0.20	(−0.39,−0.01)
138263	—	0.31	(0.01,0.59)		
166204	PHF1			0.19	(0.03,0.36)
169301	—	0.40	(0.11,0.65)	0.29	(0.04,0.53)
171185	P38IP	0.31	(0.04,0.60)	0.30	(0.07,0.53)
178078	GRM4			−0.28	(−0.48,−0.07)
179657	PLAUR	0.37	(0.09,0.67)		
180141	CFL2	0.43	(0.16,0.69)		
184434	AXIN1	0.41	(0.1,0.67)		
184711	—			−0.30	(−0.47,−0.13)
184776	RPL23A			−0.30	(−0.59,−0.04)
195453	RPS27	0.32	(0.02,0.63)	0.30	(0.07,0.52)
198951	JUNB			0.27	(0.05,0.50)
240122	CDC14B	0.35	(0.06,0.63)	0.22	(0.04,0.40)
251928	NPIP			−0.20	(−0.37,−0.03)
259842	PRKAG2	−0.37	(−0.67,−0.07)	−0.47	(−0.7,−0.25)
266902	NTF5	0.32	(0.03,0.60)	0.31	(0.04,0.56)
270062	—			−0.29	(−0.45,−0.14)
272205	FLJ10034			−0.23	(−0.39,−0.07)
272801	FLJ20464	0.36	(0.09,0.62)	0.25	(0.07,0.44)
272802	FLJ20499			0.21	(0.03,0.38)
274431	—	0.33	(0.04,0.59)	0.29	(0.08,0.49)
274535	SCYA3LI	1.82	(1.55,2.11)	1.80	(1.38,2.21)
278976	—			−0.22	(−0.4,−0.05)
279886	RANBP9			−0.21	(−0.39,−0.03)

The largest differences were observed in the downregulated genes. The full Bayesian models found two negative genes and three negative genes. On the other side, by Tseng model 20 genes emerged as downregulated, but using the internal linear

ANOVA normalization it found only 2 negative genes. Generally speaking, as theoretically expected, the full Bayesian model seems more conservative and robust with regard to the choice of normalization procedure. The Tseng model seems less conservative and more sensitive to the normalization procedure adopted. Since this results is based on the analysis of only one dataset, we do not know if one particular normalization procedure has to be recommended. The reader should note that theoretically the EB model is more sensible to normalization procedures. A full comparison among different normalization approaches to be used in the EB approach is outside the scope of the present paper.

6 DISCUSSION

The observed differences in number of differentially expressed genes between the Bayesian hierarchical model and the Tseng empirical Bayesian one are related to different factors, namely normalization method and specification of prior information. In the Bayesian Hierarchical method, the normalization step is performed inside the model through a multi-slide linear normalization (ANOVA). In the empirical Bayesian approach, data are normalized outside the model, through a loess normalization performed separately for each array. When incorporating the normalization into the model, the likelihood is based on single channel expression measures over replicates, while with an external normalization, the likelihood is based on an empirical measure of relative expression.

This is a very important point in modelling gene-specific variances. In fact, 'many ratios with high variances result from spots that have a medium or high intensity in one channel and a very low intensity in the other' (Comander *et al*., 2004, p. 4) and building a model with single channel intensity can be much more sensitive than modelling the empirical log-ratio. Coherently, using the Tseng prior with the normalization step into the model (Subsection 3.2.3) all the genes emerged downregulated in the previous analysis were no more differentially expressed.

Using the Bayesian hierarchical modelling with loess normalization (Subseection B.2.4) 27 genes were found differentially expressed; 18 out of 27 overlap those obtained by the empirical Bayes model and only 2 out of them were downregulated.

The full Bayesian model originates likely more conservative estimates of relative expression with respect to the empirical Bayes one. The sensitivity analysis performed in the previous section shows that the Bayesian model is more robust to the different normalization procedures adopted.

The empirical Bayesian model and the full Bayesian one insert prior information on variability from the calibration experiment in different ways. In the first the prior distribution for the variance of the normalized gene log-ratio ($_m\sigma_g$) is a function of a weighted average between the observed gene-specific variances (s_g) and their average among the set of genes ($s.$) on the calibration arrays [Equation (16)]. It is not assumed a hyperprior distribution on the prior parameters, but instead an estimate is plugged in, following the empirical Bayesian approach. The proposed estimate in Tseng model lies on the theory of the generalized estimator of James-Stein (Efron and Morris, 1972) and has optimality properties under a frequentist point of view.

The full Bayesian hierarchical model inserts information from self–self experiment at the normalized log-ratio level for each gene, as well as at the single channel intensity level (Fig. 3).

The gene-specific log-ratio (τ_g) probability density has informative distribution on its parameters $\mu_{\tau_g}, \sigma_{\tau_g}$ [Equation (7)]. The single channel intensity likelihood has a gene-specific prior distribution for the variance with parameters $\mu_\sigma, \sigma_\sigma$ estimated from the self–self experiment [Equation (4)]. An alternative would be to consider the whole posterior distribution of μ_σ and σ_σ from the calibration experiment. The hierarchical structure of the model is a robust answer to the problem of putting in prior knowledge. The introduction of a supplementary layer in the model permits to filter the available previous information in a sensible way.

As showed in Figure 2, in our data Bayesian posterior estimates of gene-specific variances tend to be larger than the empirical Bayes estimates. The reader can also appreciate that the distribution of log-ratios (Fig. 1) from calibration experiment has a heavier tail for negative values and a positive mode. Coherently, our Bayesian analysis for the comparative experiment is more conservative and gives more penalty to negative log-ratios.

Both models reveal a shrinkage effect: additional materials to illustrate this point can be requested to the authors.

In conclusion, we showed how information from calibration experiments can be utilized to improve inference on differentially expressed genes in comparative experiments.

The approach presented is specific for two-channel arrays. However, our modelling is based on absolute gene expression level, the log-ratio being a model parameter to be estimated. Therefore it can be adapted to Affymetrix platforms.

We can point out that the calibration experiment is a good answer to the problem of gene-specific variability estimate and allows us to include prior information both working in a full Bayesian framework and in an Empirical Bayesian one. It naturally extends to a sequence of experiments (e.g. time course experiments): it permits to update prior information and to take under control sources of variations that can be introduced between different experiments. Moreover, a calibration experiment can be used as baseline for future experiments on the same tissue, cellular line or species.

Conflict of Interest: none declared.

REFERENCES

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 5009–5019.

Comander,J. *et al.* (2004) Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, **5**, 1–21.

Delmar,P., Robin,S. and Daudin,J.J. (2005) Efficient variance modelling for differential analysis of replicated gene expression data. *Bioinformatics*, **21**, 502–508.

Dobbin,K. and Simon,R. (2002) Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, **18**, 1438–1445.

Efron,B. and Morris,C. (1972) Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika*, **59**, 335–347.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*., **96**, 1151–1160.

Gelman,A. and Rubin,D.B. (1992) Inference from iterative simulations using multiple sequences. *Stat. Sci*., **7**, 457–511.

Gilks,W.R., Richardson,S. and Spiegelhalter,D.J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Kerr,M.K. *et al.* (2002) Statistical analysis of gene expression microarray experiment with replication. *Stat. Sin*., **12**, 203–217.

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

Lewin,A., Richardson,S., Marshall,C., Glazier,A. and Aitman,T. (2005) Bayesian Modelling of Differential Gene Expression. *Biometrics*, in press.

Mor-Vaknini *et al.* (2003) Vimentin is secreted by activated macrophages. *Nat. Cell Biol.*, **5**, 59–63.

Newton,M.A. *et al.* (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Parmigiani,G. *et al.* (2002) A statistical framework for expression based molecular classification in cancer. *J. R. Stat. Soc.*, **64**, 717–736.

Pepe,G. *et al.* (1997) Tissue factor and plasminogen activator inhibitor type 2 expression in human stimulated monocytes is inhibited by heparin. *Semin. Thrombosis Hemostasis*, **23**, 135–141.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression Data. *J. Comput. Biol.*, **8**, 557–569.

Simon,R.M., Korn,E.L and McShane,L.M. (2003) *Design and Analysis of DNA Microarray Investigations*. Springer-Verlag, New York.

Speed,T. (ed.) (2003) *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall, New York, NY.

Spiegelhalter,D., Thomas,A., Best,N. and Lunn,D. (2003) WinBUGS, version 1.4. *User manual MRC Biostatistics Unit*, Cambridge, UK.

Tanner,M.A. (1996) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.*Springer, New York.

Tseng,G.C. *et al.* (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.

Tusher,V.G. *et al.* (2001) Significance analysis of microarray applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wit,E. and McClure,J. (2004) *Statistics for Microarrays* John Wiley and Son, Chichester, UK.

Wu,H., Kerr,M.K., Cui,X. and Churchill,G.A. (2003) MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In Parmigiani, G., Garett, E., Irizarry, R. and Zeger, S. (eds). *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York, NY, pp. 313–341.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.