Exploiting blank spots for model-based background correction in discovering genes with DNA array data

Serena Arima, **Luca Tardella,**
Dipartimento di statistica, probabilità e statistiche applicate
Sapienza Università di Roma
Brunero Liseo
Dipartimento di Studi geoeconomici, linguistici,
statistici e storici per l'analisi regionale
Sapienza Università di Roma
Francesca Mariani
Istituto di Neurobiologia e Medicina Molecolare, CNR, Roma

Abstract

Motivated by a real data set arising from a study on the genetic determinants of the behavior of *Mycobacterium tuberculosis* (MTB) hosted in macrophage, we take advantages of the presence of blank spots and illustrate modeling issues for background correction and ensuing empirical findings resulting from a Bayesian hierarchical approach to the problem of detecting differentially expressed genes. We build up on the approach proposed in [23] where gene classification is based on posterior probabilities corresponding to subsets of appropriate partitions of the parameter space.

We prove the usefulness of a fully integrated approach where background correction and normalization are embedded in a single model-based framework, deriving a new tailored model to account for peculiar features of DNA array data where null expressions (blank spots) are planned by design. Also we advocate the use of an alternative normalization device resulting from a suitable reparameterization. The new model is validated using our MTB data and results are compared with the approach of [23] and with SAM method [41].

1 Introduction

In the last years microarrays technologies have dramatically changed biomedical research. They can be considered the first general genomic device capable to simultaneously measure the activity of thousands of genes at a particular moment in a particular tissue or biological sample. There are nowadays several technologies which allow to measure gene level expressions in some experimental tissue and the high density oligonucleotide arrays, or *microarrays*, are currently among the most popular ones. For a review of the microarray technology one can see, for example, [30] and [32]. In the usual context of microarray, data are summarized in a matrix whose rows correspond to gene spots and columns to replicates in different experimental conditions: each cell contains gene expression measured as absolute (or relative) intensity and detected by a scanner reading the impression (radioactive or fluorescence) measured from the array after a specified amount of time. A substantial effort has been recently made in order to address the problem of the statistical analysis of gene expression data and relevant literature starting from the pioneering works of [36], [21] and [32] can be gathered from some recent overviews [26] and [7] and books [13], [33] and [42]. Most of the difficulties analyzing gene expression data coming from array experiments arise from the presence of a large amount of measurements with comparatively little replications, the necessity of pre-processing data to comply with standard assumptions of the most popular models and also from the multiplicity issues. A recent advance in addressing all these issues in a single framework has been put forward with the Bayesian model strategy proposed in [23]. We take their approach as our starting point. However, we we will argue in the next sections that DNA array data can have some peculiar features which cannot be easily accommodated into the Lewin et al.'s framework, such as the presence of control spots which are designed to carry no expression, possibly yielding a non negligible fraction of exact null measurements. To circumvent this problem, we build upon the above mentioned model; we will show through §2 and §3 that it is possible to carry out and actually improve the inferential procedures proposed in [23] by appropriately handling data sets where gene expressions are originally detected as exact null expressions and by suitably defining a background correction and normalization device for the observed raw data. Furthermore, the proposed model takes into account the presence of more than two experimental conditions, which biologists do not necessarily consider on the same footing.

The paper is organized as follows: §2 is devoted to the description of the problem with the help of the motivating case study (MTB data) and the proposed solution. §2.2 is devoted to

review, in some details, the hierarchical Bayesian approach of [23], pointing out innovations of our tailored new model in order to adapt it to our specific context in §2.3 through §2.6; §3 illustrates the results obtained with MTB data and discusses strategies for selecting those genes which deserve further investigation.

2 Modeling and Inferential Issues

2.1 Background calibration issues

Before introducing our method, we first highlight some peculiar features of DNA data which suggested us to investigate alternative solutions with respect to currently available approaches. In gene expression data analysis, the first step consists of pre-processing data, taking logarithmic transformation of the original expression and then making suitable *normalization* steps in order to account for some external experimental variability within the same experimental condition due to the so-called *array effect*. Many procedures are discussed for example in [42] and [33] and there is still ongoing research on these topics (see e.g. [8], [12] and [35]). However, normalization procedure is applied to expression levels resulting from the array image analysis software which are usually provided as already background corrected. This means that a single (or multiple) background value is calculated for each array and is then subtracted from the foreground intensities in order to partially remove non-biological biases. Genes falling below the background values are taken as having an expression too low to be accurately detected and they are given an exact null expression; this fact is not unusual in the literature and the frequency of exact null expressions may be substantial (see [12]). To avoid that sometimes the calibration device can be manually tuned by the biologists/researchers on a case-by-case basis. It can be argued that the bias in the background intensity detection, possibly by manual intervention, may have large effect on the signal intensity automatic evaluation and its miscalibration can have severe consequences on the statistical analysis [22]. The debate on the effects of background correction is still open [37]: background correction can have the negative effect of increasing the variability, especially at low intensities and it can lead to negative values for which the logarithmic transformation is not directly applicable. On the other hand, with no-background corrected data, one increases all intensity measurements and tends to reduce the estimates of their ratios, biasing them downward. Our final solution, fully explained in the next Section, seems to be a fair compromise in terms of both logarithmic transformation feasibility and variance inflation

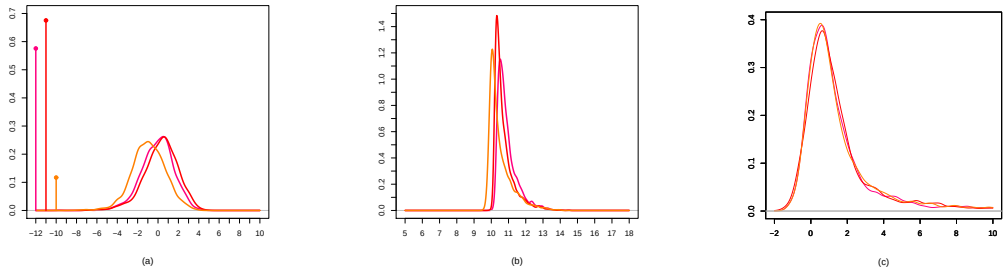

Figure 1: MACR$_H^+$ condition: automatically background corrected data (vertical lines are proportional to the amount of zeros), raw data and model-based background corrected data. All gene intensities are on a logarithmic scale.

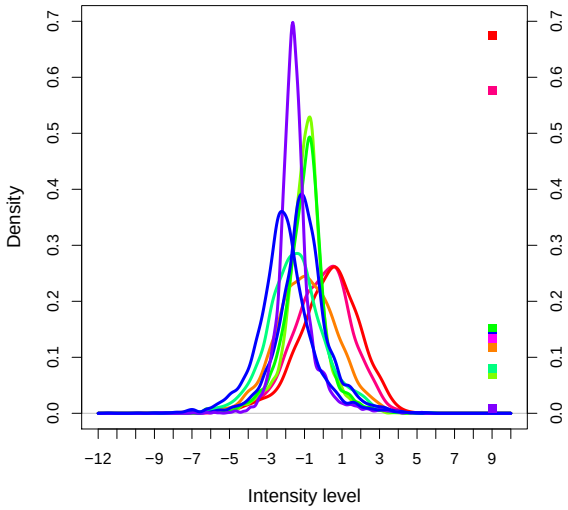

Figure 2: Logarithmic transformation of background corrected expression levels for each experimental condition; squares represent the percentage of zero's for each array and experimental condition.

control (see 1). An instance of substantial exact null expressions, is displayed in Figure 2 with colored squares of the left side showing the frequency of exact null expressions together with the distribution of the logarithmic transformation of positive background corrected expressions of MTB data in different arrays. More details on MTB data are provided in §3. A first look at

Figure 2 is sufficient to understand the extent of the array variability within the same experimental condition. The density curves do not substantially overlap and it is apparent a strong need of further normalization. Colored squares show that the fraction of exact null expressions ranges from 0.8% to almost 70%. For these expressions the logarithmic transformation step is unavailable. To overcome this problem, one can consider the following possible solutions: 1) ignoring the information coming from those spots removing them from the analysis, 2) adding a fixed constant value $\overline{c}$ to allow the logarithmic transformation [18, 6], 3) modifying the statistical model in order to account for a discrete component of the expression distributions and 4) appropriately smoothing the exact null signal in some very low but positive expression determining a more convenient continuous left tail of the distribution [14].

Of course the first solution must be discarded in those cases, as in our MTB data, where it leads to waste too much of the precious information coming from the DNA array experiments, which are time consuming, very expensive and sometimes use clinical samples not easily available. This can be a viable solution only in those experiments where the frequency of exact null expressions is very low and even in that case important biological information might be neglected. The second solution, though recently ingeniously calibrated [18] in order to avoid somehow the arbitrary determination of the constant $\overline{c}$, does not overcome the difficulty of having a relevant fraction of data for which a continuous density model can obviously result in a poor fitting. The third solution would be seemingly the more natural and faithful in this context; however it is less frequently adopted in the statistical literature and it also implies as a drawback that one must abandon the logarithmic scale evaluation which is in fact the most standard and favorite scale for biologists.

It might well happen that exact null expressions can be a consequence of automatic background calibration: in fact, this is the case in our MTB data. In Figure 2 one can notice that Array 1 and Array 3 (in the same experimental condition) show frequencies of exact null expressions as different as 0.1 and 0.7, which can give the impression that they could be the result of some imperfect functioning of the detection device. In order to better deal exact null expressions we had to go back to the original numerical outputs of the software integrated to the scanning system. In fact, one can gain insight into the structure of the arrays and find out that there are different contexts where null expressions can be detected within the available spots. It often happens that blank spots, with no DNA spotted in, are added as control devices. In those spots one should expect to detect a null expression even though there would be always the chance that a positive low expression is detected by the scanner for some experimental error. However, if an additional

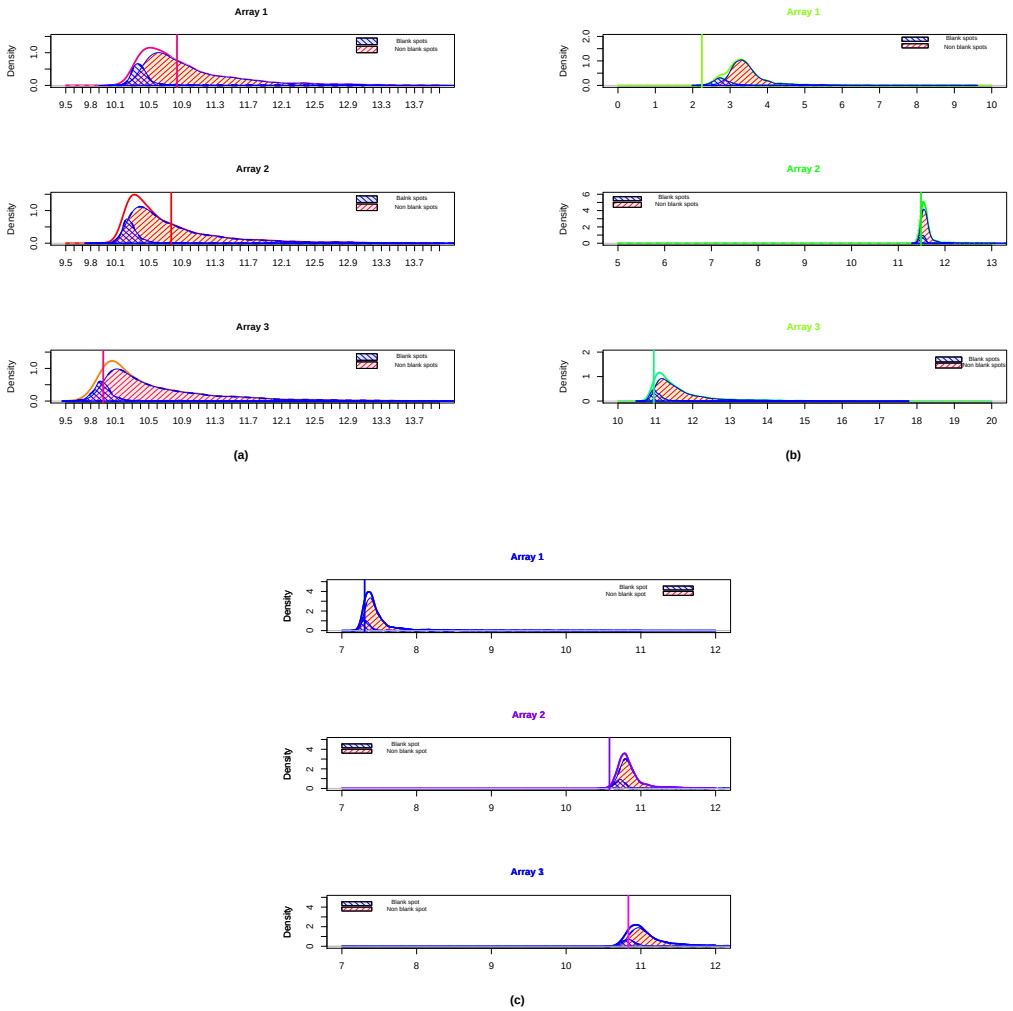

Figure 3: Logarithmic transformation of raw gene expression for MACR_H^+ (Panel (a)), MACR_V^- (Panel (b)) and MACR_H^- (Panel (c)) experimental conditions: blank spots and gene of interest distributions are highlighted. Vertical bars indicate log background level calculated by the image software

unexpected incidence of null expressions for non blank spots is experienced, one may suspect that something is going wrong with the background calibration. Panel (a) of Figure 3 shows the distribution of raw gene expressions, on a logarithmic scale, for one experimental condition (MACR_H^+ , more details in §3). Raw gene expressions are a measure of intensity of grey-level of each spot obtained by black pixel counts; they are typically measured as foreground intensities average. The distribution of blank spots and non blank spots are highlighted with different colors and shadow for each array: vertical lines indicate the background value calculated by the scanning software. As it can be noted, for the same experimental condition, the raw gene expression distributions show similar shapes and features in the different arrays but the vertical lines do not always match them. This troublesome feature is present also for replicates of other experimental conditions (Panel (b) and (c) of Figure 3). As a consequence, the corresponding background corrected values do not reflect faithfully the variability of the raw data and risk to introduce an arbitrary bias in the experimental results. Note also that all genes with raw intensities below the vertical threshold will be transformed into exact null expressions, giving account of the squares of the left side of Figure 2.

Several approaches have been proposed for decreasing the background noise in image analysis [43] and there is still ongoing research ([4] and [24]); the consequences of a wrong background calibration have been deeply analyzed in, for example, [22]. [14] proposed an alternative method for background correction in a Bayesian framework and [44] proposed a two-stage normalization method to adjust for the effect of background intensities. In this paper, we propose a different solution to adjust the background relying on a model-based procedure fully exploiting the information coming from blank spots, avoiding to leave room to recalibration steps made through ad-hoc clerical interaction of biologists with the scanner reading software. In a Bayesian framework, we develop a hierarchical model embedding background correction, normalization and differentially expressed gene selections.

2.2 The start up model

We base our hierarchical model on the approach proposed in [23], hereafter named LMRGA model. We first briefly review the LMRGA model ingredients using a slightly different notation from the original paper. In a basic DNA array experiment, G gene expressions are observed under two (or more) different experimental conditions and, for each combination of genes and conditions, r_c replications are available: in real applications r_c's can be very small. The key

points of the LMRGA model are: 1) there is an ANOVA structure for the log-expressions which accounts for the experimental condition and also a specific *gene effect*, 2) gene expressions are affected by an *array effect* which causes a systematic difference in expression levels between different arrays and 3) the array effect, for a single gene is not restricted to be constant but it is assumed to be a smooth function of an underlying gene expression level. These features are integrated in LMRGA with the following statistical model: for $g = 1, \cdots, G$; $c = 1, 2$ and $r = 1, \cdots, r_c$,

$$Y_{gcr} \sim N\left(\alpha_g + 0.5(-1)^c \delta_g + \beta_{gcr}, \sigma^2_{gc}\right) \tag{1}$$

where α_g is the overall effect due to gene g, δ_g is the *differential* effect between the two conditions; β_{gcr} is assumed to capture the array effect and it is modeled as a function of the overall gene effect parameter α_g, that is $\beta_{gcr} = f_{cr}(\alpha_g)$.
The rest of the model can be described as a standard hierarchical Bayesian model, with appropriate prior distributions over the parameters and the hyperparameters. Also identifiability constraints are imposed through the relations$\sum_{r=1}^{r_c} \beta_{gcr} = 0$, for all g and c. In the following subsections we will introduce suitable modifications of the former model structure to account for peculiarities of our data set and eventually yielding improved inference.

2.3 A new model embedding blank spots

In §2.1 we got insight into some possible consequences of background corrected data on the statistical analysis when background calibration problems occur. In order to reduce the bias that can affect background corrected data, we suggest to start modeling raw gene expressions as it is originally read by the scanning software with no background correction or other preliminary transformation whatsoever. Hence we propose a fully model-based background correction, taking advantages from the presence of blank spots. Several other techniques are available in the literature which take into account the information coming from expressions of blank or housekeeping genes but, to our knowledge, none of them include a formal probabilistic model component to be used for either background correction and/or normalization of the remaining genes. Pre-processing through housekeeping genes has also been widely dealt with (see [33] and [20]) by subtracting the mean expression of multiple housekeeping genes; here we propose to embed that step using a model which extracts from blank spots their array specific mean effect and use it to discount the global background nuisance. More formally we start with log raw gene expression Y_{gcr} for gene g, condition c and replicate r. They are positive numbers for

which the transformation on logarithmic scale does not need further adjustments. We adopt here the following model-based approach relying on the presence of blank spots for which the corresponding underlying mean expression level is known to be null.

Let Y_{g_0cr} be the logarithm of the raw intensity of a control gene g_0 with $g_0 \in G_0$ where G_0 denotes the set of blank spots. Since they are *known* to be not differentially expressed, their raw intensities should reflect only the background noise and can then be used as a building block for a more reliable built-in background correction.

Taking advantage of the presence of such blank spots, we propose to model Y_{g_0cr} as the distribution of the background bias with mean μ^0_{cr} and precision τ^0_{cr}. We then use the following piece of hierarchial model including these background values as explicit parameters:

$$Y_{g_0cr_{g_0c}} \sim N(\mu^0_{cr}, \tau^0_{cr}) \tag{2}$$

$$\mu^0_{cr} \sim N(\mu^{'}_c, \tau^{'}_c) \tag{3}$$

$$\tau^0_{cr} \sim N(\mu^{''}_c, \tau^{''}_c) I(0, +\infty); \tag{4}$$

we standardize the control gene ($g_0 \in G_0$) raw intensities and denote $Z_{g_0cr} = (Y_{g_0cr} - \mu^0_{cr})\sqrt{\tau^0_{cr}}$. It follows from Equation (2)-(4) that $Z_{g_0cr} \sim N(0, 1)$. As we will see in the next Section the distribution of the standardized Z_{g_0cr} will be also exploited for defining reliable automatically calibrated benchmarks for declaring over/under expressed genes.

The above model concerns only those genes in the control group with labels in G_0 but the background level parameter is shared with the remaining part of our model concerned with genes $g \in G_1$ of real interest, where G_1 denotes those genes which are not blank. Hence, if Y_{gcr} denotes the logarithm of the raw gene expression for gene of interest $g \in G_1$, we model a standardized version of it as follows:

$$Z_{gcr} = (Y_{gcr} - \mu^0_{cr})\sqrt{\tau^0_{cr}}. \tag{5}$$

The transformation in Equation (5) can be considered as an *embedded first stage background correction*, in which raw data are de-noised with respect to the background bias. The beneficial effect of embedding the normalization step and the expression modeling into a single model framework represents one of the key achievement in [23]. Here we believe it is possible to enhance their approach by using the useful information coming from blank spots through (2)-(4) and also adopting the following alternative parameterizations

$$Z_{gcr} = \lambda_{gcr} + \varepsilon_{gcr} \qquad g \in G_1 \tag{6}$$

with $\lambda_{gcr} = \mu_{gc} + \nu_{gcr}$, $g \in G_1$, $r = 1, .., r_c$ and $c = 1, 2, ..., C$.
Here μ_{gc} is the underlying mean expression level of each gene $g \in G_1$ in the condition c depurated by the array effect. The differences $\mu_{gc} - \mu_{gc'}$ can be interpreted in terms of log-fold change due to the differential effect between the first two experimental conditions c and $c^{'}$. In order to account for non-biological array-specific biases we are encouraged by the arguments and findings of [23] to embed the array effect in the model as a function of the gene mean effect. However we highlight that in our model, differently from [23], the specification of the new array effect ν_{gcr} is $\nu_{gcr} = f_{cr}(\mu_{gc})$, which depends on gene g through its entire underlying expression μ_{gc} and not an overall gene effect as in Equation (1). Not only this choice seems rather natural but it turned out to be justified on additional evidence on simulated data not reported here. In fact, with two or more experimental conditions the baseline α_g is a quantity which may depend on the chosen ANOVA parametrization and on the number of experimental conditions. Moreover in the presence of pronounced differential expression, its meaning is at least ambiguous.
Normalized expressions for genes of interest can be obtained as

$$W_{gcr} = Z_{gcr} - \nu_{gcr}. \tag{7}$$

We will refer to Equation (7) as *second stage normalization* component. The error term in Equation (6), ε_{gcr}, is modeled as normally distributed with mean 0 and a *gene specific* variance in the c-th condition, denoted as σ^2_{gc}. These variances have been modeled as exchangeable within each condition with log-normal prior distribution with mean μ_c and precision η_c. The model is identified by constraining $\sum_{r=1}^{r_c} \nu_{gcr} = 0 \quad \forall g, c$.

2.4 Prior Distribution and Implementation

A fully Bayesian inference is carried out completing the model proposed in the previous section with prior distributions on the corresponding unknown parameters. For our analysis we elicited vaguely informative priors. In Equations (3)-(4), we define prior distributions for mean and variance of blank spots. Flat normal and truncated normal hyperpriors are assigned respectively to mean and variance hyperparameters $\mu_c^{'}$, $\mu_c^{''}$ and $\tau_c^{'}$, τ''_c. Polynomial coefficients specifying the array effect have independent $N(0, 10^2)$ priors. Mean gene effect μ_{gc} is modeled, once again differently from LMRGA , using a two components normal mixture:

$$\mu_{gc} \sim \omega_c N(\phi 1_c, \eta 1_c) + (1 - \omega_c) N(\phi 2_c, \eta 2_c) \tag{8}$$

where $\phi 1_c \sim N(0, 10^3)$, $\phi 2_c \sim N(0, 10^3)$ and $\eta 1_c \sim IG(10^2, 10^2)$, $\eta 2_c \sim IG(10^2, 10^2)$. The weight of the mixture component is modeled through a uniform prior.

Mixture distributions are widely used in modeling gene expression data (see, for example [11], [27]), assuming gene expressions to be a mixture of over-expressed, under-expressed and equal-expressed genes. However, in this work, we do not model Y_{gcr} as a mixture, but, relying on [23], we keep assuming gene expression comes from a normal distribution but we model expression means as a Normal mixture, which is a very flexible distribution allowing for skewness: this is an important feature since in most experiments, the majority of genes are not differentially expressed and the symmetry between up and down regulated genes is not always a reliable assumption [31]. Therefore, a skew distribution should reflect this particular feature of the data, spreading most of the gene expressions around the mean and asymmetrically allocating some of them on the tails. We achieved satisfactory results with the normal mixture distributions: in fact, when we compare the two competing models, one assuming a normal distribution for the mean gene expression and the other assuming the mixture in Equation (8) as we will show also in §3.1, the model in Equation (8) provides a better fit in terms of Bayesian p-values distribution. We are actually working on a version of the model based on a skew-normal error distributions, which automatically account for skewness [1].

We estimate the model using WinBUGS [38] and R [34], two freely available softwares which allow the user to perform Monte Carlo Markov Chain (MCMC) simulations of the posterior distributions. Code is available upon request. Using several chains with different starting points, we allow 10000 iterations for the sampler to stabilize and another 50000 for sampling the joint posterior. Absence of convergence is checked by visual inspection of the trace plots and by using several starting points.

2.5 Interesting Genes selection

Once the joint posterior distribution of all parameters is available - at least approximately - one can take advantage of the flexibility of this framework to select a list of interesting genes. We fully benefit from the original solution proposed in [23], in which several criteria, aiming at comparing the mean gene expression effects in different experimental conditions, have been developed. The appealing feature of such criteria is that they can be easily modified in order to answer to specific complex biological questions, taking into account the statistical significance of the conclusions accounting also for multiplicities. In this Section, we introduce selection criteria

and suggest a new method to specify reference values, completely based on data and model features.

The criteria in [23] can be summarized as follows: we calculate the probability of a gene to be differentially expressed in condition $c = 1$ with respect to condition $c = 2$ as

$$p_g \equiv \mathbb{P}(|\mu_{g1} - \mu_{g2}| > \delta_{cut} \cap \mu_{g1} > \alpha_{cut}| \text{ data }). \tag{9}$$

Genes for which $p_g > \pi_{cut}$ are declared differentially expressed. The choice of the π_{cut} value is driven by evaluating, as originally suggested in [23], the impact of the π_{cut} on the False Discovery Rate (FDR) [5, 15].

The values δ_{cut} and α_{cut} correspond to statements of biological interest: α_{cut} is the limit value above which a gene can be declared expressed and δ_{cut} is the difference which is needed, from a biological perspective, to declare a gene differentially expressed. By employing the above criterion, the posterior probability of a gene being differentially expressed is evaluated both from a statistical and biological standpoint.

Reference values for α_{cut} and δ_{cut} are usually suggested by biologists' experience. However, it is obvious that reference values ought to depend on the measurement scale, which is something biologists do not have full control of. If, as we suggest, one starts dealing with raw data, the normal reference system the biologists are used to is changed. We think it is then natural to calibrate these reference quantities using the data features and the specified model. We can then take advantage of the new feature introduced in Equation (5) which adjusts the data on a common scale, in the sense that the expressions of all genes (both blank and of interest) are standardized with respect to the overall array-specific blank spots distribution. Hence, the transformed blank spots expression follows a standard normal distribution, Z_{g_0cr} and it is then natural to calibrate α_{cut} and δ_{cut} on the basis of this reference distribution: in fact, since we know that blank spots are not expressed, a plausible value for α_{cut} could be chosen as an appropriate extreme quantile of the reference distributions Z_{g_0cr}. Analogously, we can calibrate the choice of the cut-off value δ_{cut}: since blank spots are known to be not differentially expressed, we can suppose that the minimum value to declare a gene as differentially expressed should be larger than an multiple of suitable extreme quantile range of the reference distribution. The criterion in Equation (9) can be flexibly extended to more complex situations where multiple experimental conditions have to be compared.

3 Results

3.1 *Mycobacterium Tuberculosis* data

Data we have analyzed, referred to MTB data, have been produced by a team of biologists of the Institute of Neurobiology and Molecular Medicine (INMM - National Research Council, Rome, Italy) using low density oligonucleotide arrays also named *macroarrays*. The INMM main project aims at developing innovative molecules strengthening the immune defenses against tuberculosis infections. In order to better understand the genetic mechanism of the bacterium behavior, after being inoculated in an immune system, the INMM biologists were able to extract expression levels originated from some tissue containing human macrophage exposed to MTB in the following three conditions: $\textsc{Macr}_V^-$ infected *in-vitro* culture of human macrophages, $\textsc{Macr}_H^+$ infected human macrophages extracted from broncho alveolar lavage from patients affected by MTB, $\textsc{Macr}_H^-$ human macrophages extracted from broncho alveolar lavage from patients affected by other pulmonary diseases. MTB data consist of three replicates (arrays) for each experimental condition for a total of 9 *macroarrays*. Each array contains 4608 gene spots; roughly 15% of them, have no DNA spotting them (blank spots). In a first screening stage, researchers investigate on which genes are possibly activated/disactivated as a reaction of the infectious disease carrier and then be deemed responsible of the success/failure of the immune system. Partial results based on a mixture model for $\textsc{Macr}_H^+$ and $\textsc{Macr}_V^-$ conditions, have been reported in [10]. In the subsequent step of the analysis, researchers will mostly focus their investigation efforts only on those genes which show a differential pattern of gene expression of macrophages in the $\textsc{Macr}_H^+$ condition simultaneously with respect to those in the $\textsc{Macr}_V^-$ and $\textsc{Macr}_H^-$ conditions. These simultaneous comparisons allow biologists to detect which genes modify their expression levels as a specific consequence of the tuberculosis infection (*TB–specific*), discarding those whose modifications can be due to other pulmonary diseases (*a-specific*). At a first stage we looked at the 4608 expression levels resulting from the array image analysis software Array Vision 7.0 (Imaging Research Inc., Canada); for each array and experimental condition, the software calculates a global median background values and expression levels are provided as already background corrected. Figure 2 shows background corrected intensities for each array and experimental condition. Exact null expressions are detected very often also in spots other than blank spots. Figure 3 shows the distribution of raw gene expression, on a logarithmic scale, for $\textsc{Macr}_H^+$ experimental condition. The distributions of blank spots are more tightly concentrated around their means, resembling some Gaussian-like distribution: their means can then be safely

interpreted as background levels which may depend on both replication and experimental condition. Note that all genes with raw intensities below the vertical threshold will be automatically changes in null expressions. Although the gene expression distributions of the three arrays are very similar with possibly only a minor shift, the background values are so different that in the first and second array 57% and 67% of genes (almost blank spots) have density values smaller than the built-in background value while in the third array this percentage decreases to 11%. Among those roughly 45% come from blank spots. Similar features are present in $\textsc{Macr}_H^-$ and $\textsc{Macr}_V^-$ conditions, such that background corrected values do not reflect faithfully the variability of the raw data and risk to introduce an arbitrary bias in the comparison of different experimental conditions.

3.2 Model fitting and validation

We implement the Bayesian model described in §2 to these data using WinBUGS. Normalized data are estimated as in Equation (7) via posterior averages; in particular, we choose $f_{cr}(\cdot)$ to be a third degree polynomial, but also fourth degree polynomial or more complex spline functions can be used. The normalization device works fine as it is shown in Panel (c) of Figure 1 for the $\textsc{Macr}_H^+$ condition: after removing background and array effects, curves of de-noised signals for the same experimental condition look very similar with respect to the original ones.
The assumptions of a single normal distribution rather than the mixture in (8) for the mean gene expressions are compared in terms of fit. In particular, we aim at comparing, through Bayesian p-values [28], the observed mean expressions to those evaluated using predictive values under two different models: one assuming the aforementioned mixture distribution and the other assuming a normal distribution for the mean gene expression μ_{gc}. In particular, for each model we compute the following posterior predictive p-values, $\mathbb{P}(\mu_{gc} > \mu_{gc\cdot}^{obs})$: under the null hypothesis the model being true, the distribution of the p-values is almost uniform [3]. We have checked via histograms of the posterior predicted p-values that the proposed model fits the data better than the alternative one (see 4).

Therefore we are confident on the fact that this model is more appropriate for our data.

3.3 Differentially expressed genes selection

The criterion in Equation (9) can be flexibly extended to more complex situations, where more than two experimental conditions have to be compared. In particular, in our MTB data, 3

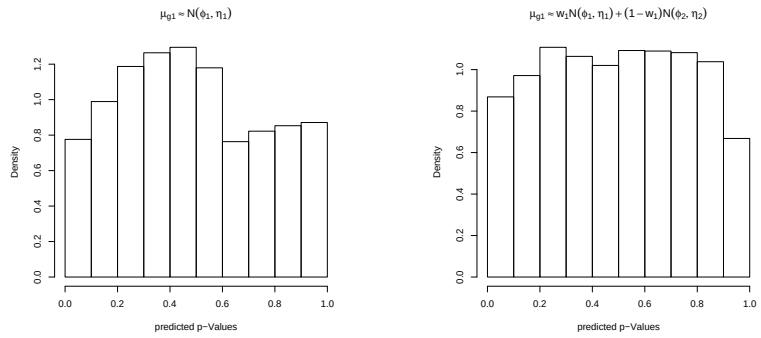

Figure 4: Posterior predicted p-values for the observed and predicted mean expressions under the model assuming normal (left panel) and mixture of normal (right panel) for α

experimental conditions are considered and biologists are interested in distinguishing *TB-specific* genes from *a-specific* genes. *TB-specific* genes are genes that are over/under expressed in the MACR_H^+ condition with respect to the MACR_V^- condition, but not differentially expressed in MACR_H^- with respect to MACR_V^- condition; *a-specific* are genes which are over/under expressed in the MACR_H^+ or MACR_H^- with respect to MACR_V^- condition. The selection of *TB-specific* gene list can then be achieved by modifying the criterion (9) in the following way:

$$p_g \equiv \mathbb{P}(|\mu_{g1} - \mu_{g2}| > \delta_{cut} \cap |\mu_{g3} - \mu_{g2}| < \delta_{cut} \cap \mu_{g1} > \alpha_{cut}| \text{ data })$$

Notice that the condition $|\mu_{g1} - \mu_{g2}| > \delta_{cut} \cap |\mu_{g3} - \mu_{g2}| < \delta_{cut}$ is not deemed sufficient to declare differential expression unless the basic condition $\mu_{g1} > \alpha_{cut}$ is simultaneously verified. This allows to combine very flexibly both biological interest and statistical significance of the conclusions when more than two experimental conditions are considered. A list of *TB-specific* and *a-specific* gene is obtained by applying the criterion in the previous Equation: in particular, according to §2.6, we choose α_{cut} and δ_{cut} equal respectively to the quantile at level $\alpha = 0.90$ and $\alpha = 0.99$ of the reference distribution Z_{g0cr}. Following [23], the threshold for $p_g = 0.8$ guarantees a FDR level equal to 0.05. Out of 3891 non blank spots, 505 genes are found *a-specific*, 12 genes are declared under expressed *TB-specific* and 143 over expressed *TB-specific*. We compare our list of *a-specific* genes with those reported in [10]: using the same MTB data, they compare only two experimental conditions (MACR_H^- and MACR_V^-), by applying a Bayesian mixture model to t scores as in [2]. *a-specific* genes resulting from our approach restricted to MACR_H^- and MACR_V^- conditions, mostly agree with those reported in [10], with some remarkable exceptions (15 genes): 3 of them are blank spots and for most of them, the mean raw expressions are as low

Gene	Bckg Corr	MACR^-_V			MACR^-_H		
g		1	2	3	1	2	3
	(a)	0.21	0.00	0.00	0.255	0.055	0.00
ID 168	(b)	3.10	11.43	10.94	7.39	10.65	10.83
	(c)	0.97	-1.26	-0.38	0.64	-0.75	-0.15
	(a)	0.63	0.00	0.87	0.17	0.26	0.14
ID 1324	(b)	3.61	11.49	12.09	7.36	10.86	11.06
	(c)	2.60	-0.40	3.29	0.34	1.09	1.25
	(a)	5.84	3.13	10.60	14.78	5.45	6.02
ID 1893	(b)	5.37	12.04	14.25	9.13	12.64	13.47
	(c)	8.36	7.23	10.13	19.15	16.36	15.89

Table 1: Automatically background corrected data ((a)),raw data on log scale ((b)) and model-based normalized data ((c)) for three genes: genes are identified by their row index in the data matrix

as those of blank spots while their differences in the corresponding automatically background corrected expressions are artificially amplified so that in [10] they are declared differentially expressed. To emphasize this concepts, Table (1) shows values for automatically background corrected densities, raw and model-based background corrected signals of three genes for which the two approaches lead to different conclusions. In [10], *ID168* and *1324* are declared over-expressed while in our approach they are declared not differentially expressed; on the other hand, *ID1893* results over-expressed in our approach while not differentially expressed in [10]. It is easy to see that the disagreement of the two methods is mainly due to the fact that mixture t score model, applied to automatically background corrected data, ia strongly affected by amplification/reduction of the expression induced by miscalibration of the background values. The list of differentially expressed genes are also compared with those obtained with SAM method [41]: it requires normalized data and therefore it also suffers of the same problems of the mixture model. In particular, using automatically background corrected data and also the correction suggested in [14], applying SAM procedure, most of the blank spots are declared differentially expressed, which is in strong disagreement with biologists knowledge. Despite the fact the SAM method and our model are based on FDR control, SAM method seems to produce an excessive discovery. Same conclusions can be drawn by extending the comparisons to three

experimental conditions.
However it must be noted that selection of *TB-specific* genes is straightforward in Bayesian framework, but classical methods, such as t-test or SAM analysis, are not so manageable to adapt in order to take into account specific biologist requests.

4 Discussion

In our work, we show how, within a Bayesian hierarchical model approach, adding a model component for blank spots can help avoiding background values miscalibration due to either defective software built-in correction or subjective clerical intervention. Differently from LMRGA model, a) we start with raw data and the *first stage background correction* is carried out in the blank spot model component (2)-(4) and (5), b) we apply a different ANOVA parametrization (Equation (6)) to data standardized through Equation (5), allowing a more natural extension for the array effect in the presence of multiple experimental conditions, c) we use a different specification for the distribution of the mean gene expressions in order to account for typical asymmetry.
With our MTB data, we could fully benefit from the decision theoretic approach of [23] which allowed us to tailor new criteria to cope with the biological specific questions, such as the distinction between *TB-specific* and *a-specific*. Moreover, including the blank spot model component allows us to rely on the blank spot distribution to calibrate reference values α_{cut} and δ_{cut} to define differential expressions more objectively and completely relying on data and model features, avoiding subjective/arbitrary biologists' interpretations of the scale of measurement. This work provides a new strategy to cope with DNA data with many exact null expressions and exploit, when blank spots are available, some of the information coming from the design of the array, which is typically discarded or used to validate statistical methods. Instead, we prove that the information provided by blank spots can usefully contribute to achieve the main goals of the analysis in a more reliable way better accounting for background correction.

References

[1] Azzalini,A., Capitanio,A. (1999) Statistical applications of the multivariate skew normal distribution,*Journal of the Royal Statistical Society. Series B.*, **61**, 579-602.

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

[3] Bayarry,M.J., Berger,J. (2000) P-values for composite null models, *Journal of American Statistical Association*, **95**, 1127-1142.

[4] Baek,J., Son,Y.S., McLachlan,G.J. (2007) Segmentation and intensity estimation of microarray images using a gamma-t mixture model, *Bioinformatics*, **23**, 458-465.

[5] Benjamini,Y., Hochberg,Y. Bofelli,F. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of Royal Statistical Society, Series B: Methodological*, **57**, 289-300.

[6] Blangiardo,M., Toti,S., Giusti,B., Abbate,R., Magi,A., Poggi,F., Rossi,L., Torricelli,F., Biggeri,A. (2006) Using a calibration experiment to assess gene-specific information: full Bayesian and empirical Bayesian models for two-channel microarray data, *Bioinformatics*, **22**, 50-57.

[7] Blangiardo,M., Richardson,S. (2007) Statistical tools for synthesizing lists of differentially expressed features in related experiments, *Genome Biology*, **8**,R54.

[8] Blangiardo,M., Richardson,S. (2007) A Bayesian calibration for combining different pre-processing methods in Affymetrix chips, Technical Report, Imperial College, London.

[9] Burguengo,M., Crossa,J., Grimanelli,D., Leblanc,O., Autran,D. (2005) Spatial Analysis of cDNA Microarray Experiments, *Crop Science Society*, **45**, 748-757.

[10] Cappelli,G., Volpe,E., Grassi,M., Liseo,B., Colizzi,V., Mariani,F. (2006) Profyling of *Mycobacterium tuberculosis* gene expression during human macrophage intection: Upregulation of the alternative sigma factor G, a group of transcriptional regulators, an proteins with unknown function, *Research in Microbiology*, **157**, 445-455.

[11] Do,K., Muller,P., Tang,F. (2005) A Bayesian mixture model for differential gene expression, *J. R. Stat. Soc., Series C*, **54**, 111-139.

[12] Dawes,N.L., Glassey,J. (2007) Normalisation of Multicondition cDNA Macroarray Data, *Comparative and Functional Genomics*, **2007**, Article ID 90578, 12 pages, 2007. doi:10.1155/2007/90578.

[13] Do, K.A., Muller,P., Vannucci,M. (2006) Bayesian Inference for Gene Expression and Proteomics, Cambridge University Press, New York.

[14] Edwards,D. (2002) Non-linear normalization and background correction in one-channel cDNA microarray studies, *Bioinformatics*, **19**, 825-833.

[15] Farcomeni,A. (2007) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion, *Statistical Methods in Medical Research*, 1-42.

[16] Gentleman,R., Carey,V.J., Huber,W., Irizarry,R.A., Dudoit,S.(2005) Bioinformatics and Computational Biology Using R and Bioconductor, Springer.

[17] Gilks,W.R., Richardson,S., Spiegelhalter,D.G. (1996) Markov Chain Monte Carlo in practice, *Chapman and Hall*, **London**.

[18] Gottardo,R., Raftery,A.E., Yeung,K.Y., Bumgarner,R.E. (2006) Quality Control and Robust Estimation for cDNA Microarrays with Replicates, *Journal of American Statistical Association*, **101**(473), 30-40.

[19] Gottardo,R., Pannucci,J.A., Kuske,C.R., Brettin,T. (2003) Statistical analysis of microarray data: a Bayesian approach, *Biostatistics*, **4**(4), 597-630.

[20] Jaeger,J., Spang,R. (2006) Selecting normalization genes for small diagnostic microarrays, *Bioinformatics*, **7**, 388-398.

[21] Kerr,M.K., Churchill,G.A. (2001) Experimental design for gene expression microarrays, *Biostatistics*, **2**, 183-201.

[22] Kim,J.H., Kim,H.Y., Lee,Y.S. (2002) Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles, *Exp. Mol. Med.*, **34**, 224-232.

[23] Lewin,A., Richardson,S., Marshall,C., Glazier,A., Aitman,T. (2006) Bayesian Modeling of Differential Gene Expression, *Biometrics*, **62**, 1-9.

[24] Li,Q., Fraley,C., Bumgarner,R.E., Yeung,K.Y., Raftery,A. (2005) Donuts, scratches and blanks: robust model-based segmentation of microarray images, *Bioinformatics*, **21**, 2875-2882.

[25] Lonnstedt,I., Speed,T. (2002) Replicated Microarray Data, *Statistica Sinica*, **12**, 31-46.

[26] Manda OM,S., Walls,R.E., Gilthorpe,M.S. (2007) A full Bayesian hierarchical mixture model for the variance of gene differential expression, *BMC Bioinformatics*, **8**, 124.

[27] McLachlan,G.J., Bean,R.W., Peel,D. (2002) A mixture model based approach to the clustering of microarray expression data, *Bioinformatics*, **18**, 413-422.

[28] Meng,X.L. (1994) Posterior predictive p-values, *Annals of Statistics*, **2**, 1142-1160.

[29] Newton,M., Noueiry,A., Sarkar,D., Ahlquisi,P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture model, *Biostatistics*, **5**, 155-176.

[30] Nguyen,A.B., Arpat,D.V., Wang,N., Carroll,R.J. (2002) DNA microarray experiments: biological and technological aspects, *Biometrics*, **58**, 701-717.

[31] Oshlack,A., Emslie,D., Corcoran,L.M., Smyth,G.K. (2007) Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes, *Genome Biology*, **8**, R2+. doi: 10.1186/gb-2007-8-1-r2.

[32] Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics*, **18**, 546-554.

[33] Parmigiani,G., Garett,E.S., Irizarry,R.A., Zeger,S.L. (2003) The Analysis of Gene Expression Data, Springer.

[34] R Development Core Team (2006) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienne,Austria.ISBN 3-900051-07-0, URL http://www.R-project.org.

[35] Ritchie,M.E., Silver,J., Oshlack,A., Holmes,M., Diyagama,D., Holloway,A., Smyth,G.K. (2007) A comparison of background correction methods for two-color microarrays, *Bioinformatics*, **23**, 2700-2707.

[36] Schadt,E.E., Li C., Ellis B., Wong W.H. (1999) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data, Tehcnical Report, Department of Statistics, UCLA.

[37] Scharpf,R.B., Iacobuzio-Donahue,C.A., Sneddon,J.B., Parmigiani,G. (2007) When should one subtract background fluorescence in 2-color microarrays?, *Biostatistics*, **8**, 695-707.

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

[39] Spiegelhalter,D.J., Best,N., Carlin,B.P., Carlin,B.P., van der Linde,A. (2002) Bayesian measures of model complexity and fit, *Journal of Royal Statistical Society, Series B*, **64**, 479-498.

[40] Tseng,G.C., Oh,M., Rohlin,L., Liao,J.C., Wong,W.H. (2001) Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variations and Assessment of Gene Effects, *Nucleic Acid Research*, **29**, 2549-2557.

[41] Tusher,V., Tibshirani,R., Gilbert,C. (2001) Significance analysis microarray applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences USA*, **98**, 0048.1-0048.16.

[42] Wit,E., McClure,J. (2004) Statistics for Microarrays, John Wiley & Sons.

[43] Yang,M.C., Ruan,Q.G., Yang,J.J., Eckenrode,S., Wu,S., Mclndoe,R.A., She,J.X. (2001) A statistical method for flagging weak spots improves normalization and ratio estimates in microarray, *Physiol. Genomics*, **7**, 45-53.

[44] Yoon,D., Yi,S.G., Kim,J.H., Park,T. (2004) Two-stage normalization using background intensities in cDNA microarray data, *Bioinformatics*, **5**, 1-12.

[45] Zhao,Y., Li,M.C., Simon,R. (2006) An adaptive method for cDNA microarray normalization, *Bioinformatics*, **6**, 28-38.