UNIVERSITÀ DEGLI STUDI DI FIRENZE

Dipartimento di Statistica 'G. Parenti'

TESI DI DOTTORATO IN
STATISTICA APPLICATA – CICLO XVII
DI
MARTA BLANGIARDO

Modelling variability of gene expression data

Relatore: Clar.mo Prof. Annibale Biggeri
Coordinatore: Clar.mo Prof. Fabrizia Mealli

Gennaio 2005

CONTENTS

If you have an apple and I have an apple and we exchange apples
then you and I will still each have one apple.
But if you have an idea and I have an idea and we exchange these ideas,
then each of us will have two ideas.

George Bernard Shaw (1856 - 1950), Irish Writer

PREFACE

Microarray technology: the genesis

Proteins are the structural components of cells and tissues and perform many key functions of biological systems. The production of proteins is controlled by genes, which are coded in deoxyribonucleic acid (DNA). Protein production from genes involves two principal stages, known as transcription and translation (see figure 1).

During transcription, a single strand of messenger ribonucleic acid (mRNA) is copied from the DNA segment coding the gene. After transcription, mRNA is used as a template to assemble a chain of amino acids to form the protein.

Gene expression investigations study the amount of transcribed mRNA in a biological system. Although most proteins undergo modification after translation and before becoming functional, most changes in the state of a cell are related to changes in mRNA levels for some genes, making the transcriptome worthy of systematic measurement.

Several techniques are available to measure gene expression, including serial analysis of gene expression (SAGE), cDNA library sequencing, differential display, cDNA substraction, multiplex quantitative RT-PCR and gene expression microarray. Microarray quantify gene expression by measuring the hybridisation, or matching, of cDNA immobilised on a small glass, plastic or nylon matrix to mRNA representation from a sample under study. A separate experiment takes place in each of many individual spots, arranged as a regular pattern on the matrix, whence the name array. Arrays can currently have hundreds of thousands of spots. Such ability to measure simultaneously a large proportion of the genes on a genome opens the door to the investigation of the interactions among the genes on a large scale, the discovery of the role of the vast number of genes whose function is not adequately understood, and the characterisation of how metabolic pathways are changed under varying conditions.

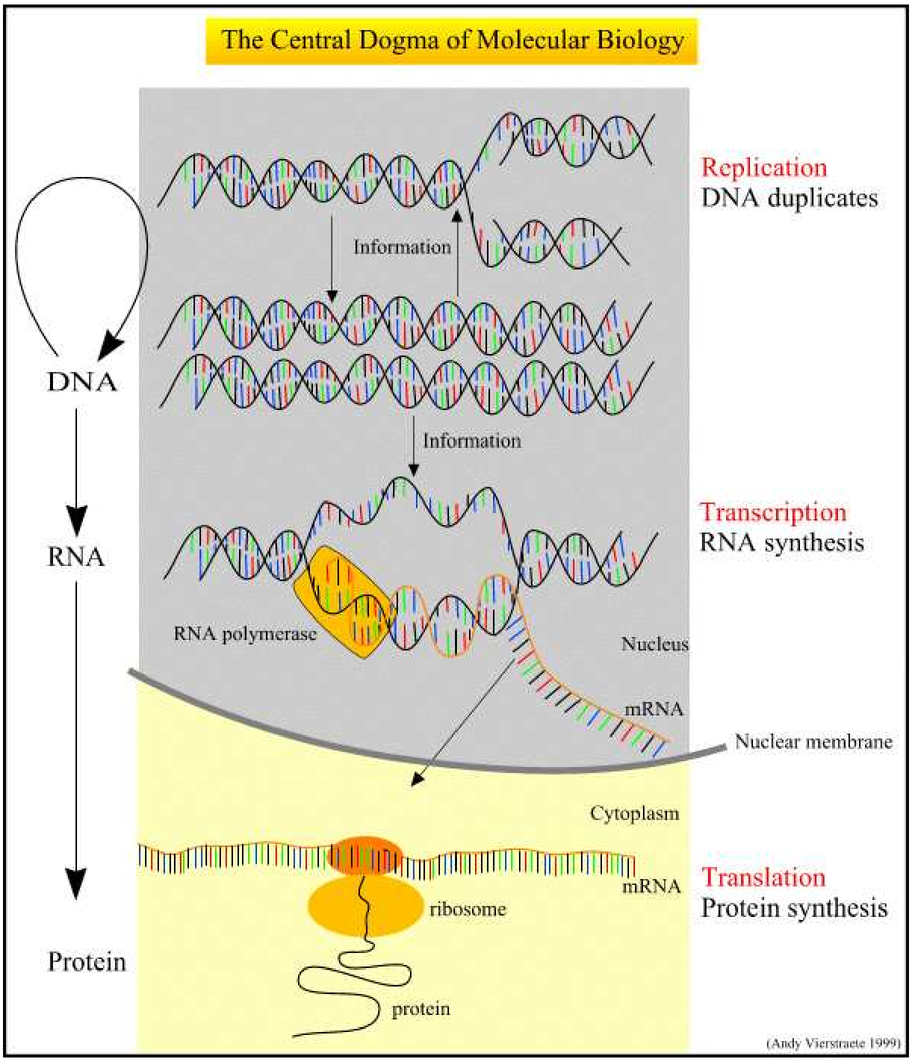

Figure 1: Central dogma of molecular biology (see color insert following page 99).

Different technologies

There are several microarray technologies. Currently two approaches are prevalent: cDNA arrays and oligonucleotide arrays. This work is based on the the first technology, but now we introduce briefly both. They both exploit hybridisation, but they differ in how DNA sequences are laid on the array and in the length of these sequences.

In spotted DNA arrays, mRNA from two different biological samples is reverse-transcribed into cDNA, labelled with dyes of different colors and hybridised to DNA sequences, each of which is spotted on a small region, or spot or glass slide. After hybridisation, a laser scanner measures dye fluorescence of each color at the fine grid of pixels. Higher fluorescence indicates higher amounts of hybridised cDNA, which in turn indicates higher gene expression in the sample. A spot typically consists of a number of pixels. Image analysis algorithms either assign pixels to a spot or not and produce summaries of fluorescence at each spot as well as summaries of fluorescence in the surrounding unspotted area (background). cDNA microarray are described in Schena *et al.* (1995) and DeRisi *et al.* (1997).

For each location on the array, a typical output consists of at least four quantities, one of each color, for both the spot and the background. Sometimes these are accompanied by measures of spot quality, to flag technical problems, or by measures of the pixel intensity variability. It is conventional to refer to the two colors as red and green. The use of two channels allows for measurements of relative gene expression across two sources of cDNA, controlling for the amount of spotted DNA, which can be variable, as well as other experimental variation. This has led to emphasis on ratios of intensities at each spot. Although this ratio is critical, there is relevant information in all four of the quantities above.

The second common approach involves the use of high density oligonucleotide arrays. This is an area of active technological development. The most widely used oligonucleotide array type is the Affymetrix GeneChip. In this array, expression of each gene is measured by comparing hybridisation of the sample mRNA to a set of probes, composed of 11-20 pairs of oligonucleotides, with a specified length. The first type of probe in each pair is known as Perfect Match and is taken from the gene sequence. The second type is known as Mismatch and is created by changing the middle base of the PM sequence to reduce the rate of specific binding of mRNA from other parts of the genome.

An RNA sample is prepared, labelled with a fluorescent dye and hybridised to an array. Unlike in two-channels arrays, a single sample is hybridised on a given array. Arrays are then scanned and images are produced and analysed to obtain a fluorescence intensity value for each probe, measuring hybridisation for the corresponding oligonucleotide. For each gene, or probe set, the typical output consists in two vectors of intensity readings, one for Perfect Match and one for Mismatch.

Variability sources

Gene expression microarray are a powerful tool, but variability arising throughout the measurement process can obscure the biological signal of interest. It is useful to classify sources of variation into five phases of data acquisition: microarray manufacturing, preparation of mRNA from biological samples, hybridisation, scanning and imaging. Each of these phases can introduce an amount of artifactual variation and/or bias that complicates the estimation of expression levels as well as the comparison of expression changes between arrays. To focus on cDNA technology, variability arises in the amplification, purification and concentration of DNA clones for spotting, in the amount of material spotted, in the ability of the spotted material to bind to the array and in the shape of the deposited spot. Systematic variation can be determined by microscopic defects in the print tip of the robotic equipment used for spotting.

During the preparation of samples, sources of variability depend on the protocol and the platform used. Important examples include labelling procedures, RNA extraction and amplification. In cDNA arrays, dye biases can arise from different ability of the dyes to incorporate into the samples and can be reduced using dye-swap, a design that provides for the replication of experiments with inverse assignation of fluorochrome at the two conditions (reference and treatment).

During hybridisation, variability arises from ambient conditions such as temperature and humidity, from edge effects (for the genes spotted near the edges of the array), from slight inhomogeneity of the hybridisation solution, from extraneous molecules or dust binding to the array, from cross-hybridisation of molecules with high sequence identity, and from washing of non hybridised materials from the array.

During scanning, natural fluorescence and binding of genetic material to the array in unspotted regions can introduce a nontrivial, spatially varying background noise. Scanning requires separating the fluorescent label from the biological material and capturing it with sensors; both phases involve randomness and re-scanned slides usually give slightly different results. Scanning intensity is an important factor, as higher intensity improves the quality of the signal but increases the risk of saturation caused by a ceiling occurring when a channel reaches maximum intensity.

In the imaging step, some technologies require human intervention for the initialisation of the imaging algorithms or the alignment of the image to a grid.

Although many of the errors are relatively small, the compounding of their effects can be significant. As a result, we can generally expect variation in the expression of a given gene across different hybridisations using the same RNA sample. This variability is involved in identification of differential expression through statistic test. In cDNA arrays many sources of noise can be quantified in the aggregate by a self-self hybridisation (calibration experiment), in which two sub samples from the same pool of RNA are labelled with different dyes and then hybridised on the same array.

Outline of the work

As described in the previous section, variability plays an important role in the analysis of microarray; the aim of this work is the variability modelling for cDNA microarray data. Starting from the analysis and the comparison of the relevant approaches in literature, we propose a Full Bayesian model taking advantage from calibration experiment.

In particular, chapter 1 presents and compares the different methodologies to model variability; chapter 2 proceeds the comparison in terms of computation aspects and of results for the approaches that give a free software. Chapter 3 presents the original modelling that takes prior information from calibration experiment. Chapter 4 analyses the results of the proposed model applied to LPS- and un-stimulated human leukocytes.

This work is structured in articles that can be read separately. However, in chapter 1-3 some references to other chapters are put in brackets, to make easy passing through the different works.

In the following pages I give some outline of what are the goals of the works.

Estimating variability in microarray experiments

In this paper, we define how to model the variability for each gene and to find a valid way to take into account gene specific information is a central point in the analysis of microarray data. Many authors have tackled this problem under different point of view and a researcher who for the first time approach this topic can find difficulties moving around. We introduce the approaches in section 1.4, clustering them as follow: modified *t* tests are in section 1.4.1, ANOVA models in section 1.4.2, error models in section 1.4.3 and Mixture models in section 1.4.4. In section 1.5 we try evaluating advantages and drawbacks for each of them.

An application to compare variability on microarray data

In this paper we present and compare the performance of five different approaches, which are based on a modification of *t* statistic.

t test is a natural choice to identify differentially expressed genes, but it is unsatisfactory for low variability level. Two problems which affect the *t* statistic in the microarray framework are considered here: the availability of only few replicates (e.g. 2,3) for each gene; the dependence between mean and standard deviation of gene log expression (the *t* statistic numerator and denominator, respectively).

In this research field the amount of data to work with is very large (usually from 1000 to 14000) and the computational aspects are as important as the methodological ones, but there is not a gold standard at the moment. R project (`www.r-project.org`) seems a good ex-

change tool for the researchers. In this perspective, we have written an add on package (*R/compvar*), that gives the results of 5 considered approaches in terms of differential expression and of variability estimate. This can be useful under two different perspective: the biologist can use the package to evaluate the power of each approach to determine differentially expressed genes; the statistician can avail of this package as a starting point, to go through and understand the different methodologies with the aim of propose an own approach to the problem.
To evaluate the library we used two public cDNA datasets, that differ for some relevant features and we discuss the results.

Using a calibration experiment to assess gene specific information:
Full Bayesian and Empirical Bayesian models for microarray data

In this paper, we propose an original approach to model the variability, taking advantage of calibration experiments (Tseng *et al.*, 2001), in which the probes hybridised on the two channels come from the same population (self-self experiment). From such an experiment, it is possible to estimate the gene-specific variance, to be incorporated in comparative experiments on the same tissue, cellular line or species.
In section 3.4 we present two different approaches to introduce prior information on gene-specific variability from a calibration experiment. The first is an Empirical Bayes model derived from Tseng *et al.*, while the second one is a full Bayesian hierarchical model. We apply the methods in the analysis of human LPS-stimulated leukocyte experiments comparing the results (section 3.5 and 3.6) and investigating the differences (section 3.7).

Calibration experiment for the analysis of microarray data: an application on un-and-LPS stimulated human leukocyte model

In this paper we present an original application of the models presented in the previous paper on peripheral blood mononuclear cells. We use the Bayesian hierarchical model to identify differentially expressed genes, taking into account the variability at gene level through calibration experiments and we compare the result to whom obtained by Tseng model, as a standard method for taking information from cali-

bration experiment. Our data on LPS-inducible gene expression profile both identified novel genes (e.g. IFI30, MLSN1, CFL2, AXIN1) suggesting new targets of study in order to better understand the pathophysiology of sepsis and inflammatory disease and confirmed the involvement of many cytokines and chemokines (IL-1b, IL-1RA, MIP-1a, -1b, -2b, -3a).

ACKNOWLEDGMENTS

First I am grateful to Annibale Biggeri, my supervisor, not only to have proposed me this topic, but also to trust me and to give me the opportunity to grow up and to take my responsibilities.
I wish to thank Simona Toti for giving me a substantial hand and for sustaining me all the time: nothing would have been the same without her; many thanks to Betti Giusti to give me the data to analyse and to carry me in the ”fabulous world” of molecular biology.
I thank Duccio Cavalieri and all the friends of *Center for Genomic Research* in Boston.
Chapter 3 was an oral presentation to the *International Workshop of Statistical Genomics*, held in Ascona (CH) and to the 19^{th} *Workshop on Statistical Modelling* held in Florence; I am grateful to Sylvia Richardson and Terry Speed for the advice they gave me in those occasion. I thank Ernst Wit and Enzo Bagnardi for listening me and reading my thesis.
Last (but not least), thanks to all my colleagues for having shared a lot of funny and stressful moments and a special thank to Gianluca.

1 Estimating variability in microarray experiments

1.1 Abstract

Microarray technology is a powerful tool to analyse and classify thousands of genes at the same time. It is commonly used in many biological fields to compare mRNA levels between two or more conditions. A central point in the analysis of microarray data is how to model the variability for each gene and to find a valid way to take into account gene specific information. Many authors have tackled this problem under different point of view. In this paper we consider 4 types of approaches: modified *t* test, ANOVA model, two components error model and mixture model. We try evaluating advantages and drawbacks for each of them.

1.2 Introduction

Experiments based on microarray technology has assumed a very important role in the biological research field. Statistical aspects are relevant in each phase of the experimental procedure, especially in the design of experiment, quality control and normalisation, to the aim of taking under control variability sources. Under this aspect, if the choice of experimental design permits to control the sources of variability before performing the experiment, quality control and to a greater extent normalisation, take into account and try eliminating the non biological variability, that is the noise one. After normalisation, there are two main goals: one is the identification of differentially expressed genes among several varieties (class comparisons), while the other is the discovery of clusters within a collection of samples (class discovery). Class comparison is related to exposure or treatment (i.e. comparison of gene expression for a population of smokers and non smokers) and the comparison between two (or more) varieties are performed directly (i.e. loop design) or indirectly (i.e. reference design). Conversely, class discovery is based on distances between gene profiles of pairs of samples (Dobbin *et al.*, 2002). To the aim of class comparison, from a very naïf point of view two methods are mostly used: the first simply compares the expression in experimental condition versus control condition and selects the genes for which the expression values are very different in the two conditions (considering as significant the genes with at least 2 or 3 fold) (DeRisi *et al.*, 1996). The second method consists in selecting genes for which the ratio between experimental and control condition is far from the mean of experimental/control ratio (Tao *et al.*, 1999). In this framework, the variability within conditions is not taken into account making such approaches not very refined. In fact, they tend to be too much conservative at high expression level and too less conservative at low expression level.

In a classical statistical approach, it seems extremely natural to use t tests where for each gene at the numerator there is the difference between gene expression levels in two conditions to be tested as differentially or not and at the denominator there is the variance divided by the square root of the replicates number. In this context a crucial point is how to obtain a valid estimate of the denominator. Actually, when the number of replicates is small (i.e. 2,3) the sampling distribution of variance is very asymmetric, with higher probability for small values, producing an underestimate of variability and a consequent overprice of differentially effect between the two conditions. For this reason in literature many authors try stabling the measure of variability to be used. One way is to consider a global variance for the whole set of genes, or try calculating a function of the variance common for all the genes. However, this causes a loss of power, because tends to be very conservative and to increase the number of false negative. In the present work we describe the principal methods on microarray experiments that take into account the variability at gene-level. The paper follows this structure: in section 1.3 we present a brief introduction on microarray experiment and preprocessing techniques, in section 1.4 we take an excursus on the approaches we intend to treat and introduce the notation used thereafter; then subsections 1.4.1- 1.4.4 treat different methodologies in details; the Discussion (section 1.5) gives some focal differences within the treated approaches.

1.3 Background on microarray

DNA microarray analysis has become the most widely used technique for the study of gene expression patterns on a genomic scale (Schena *et al.*, 1995), (Schena *et al.*, 1996). There are many microarray technologies and the more used are two: cDNA microarray and oligonucleotide arrays. They differ in how the sequences of DNA are processed to produce the array and in the length of the sequences used. For more details see the preface; thereafter we refer to the first technology.

1.3.1 Data Preprocessing

During the complex procedure of microarray fabrication, many sources of variability arise and can obscure the biological signal of interest (see the preface). In each phase of data acquisition an amount of systematic variation can be introduced and it should be taken under control. To this aim, quality controls and normalisation procedures are used to detect and eliminate the artifacts and the systematic variations, both within a single array and across arrays.

Quality control Simon *et al.* (2004) suggest some rules to identify low quality spots. They are both at the single channel intensity level and at the relative intensity level. The first includes the number of pixels used to calculate the intensity and the strength of the intensity signal; the second comprises the ratio between the average foreground intensity and the median background intensity and spots with a large signal for one channel and low signal for the other channel (not to be eliminated, but modified to be analysed).

Normalisation After quality control data are commonly normalised. We do not treat the different aspects of this procedure, but refer to the complete review on microarray normalisation methods that can be founded in Yang *et al.* (2002). In this paper we only want to point out it is possible to normalise the data externally or treating normalisation as part of the modelling. Actually some of the approaches we refer to consider normalisation as a preliminary step in the analysis of the variability and the identification of differentially expressed genes. They generally use a local A-dependent normalisation (loess) calculated for each slide or for single print tip. On the other side, some approaches analyse directly raw data and perform a normalisation as a part of the analysis. In chapter 2 the effect of normalisation on two dataset is pointed out and in chapter 3 the importance of evaluating the differences in normalisation is a focal point.

1.4 Approaches to variability

Many authors consider a gene-specific variance at the denominator of the *t* test, and add a global constant for the whole set of genes to stable the variance. The aim of this constant is to take under control the relation between low intensity and low variability, that increases the number of false positives. The different methods treated lie on a parametric Empirical Bayesian framework (Lönnstedt and Speed, 2002), non parametric frequentist (Tusher *et al.*, 2001), non parametric Empirical Bayesian (Efron *et al.*, 2001) or parametric fully Bayesian (Baldi and Long, 2001)
Similarly it is possible to specify linear models on log expression including terms for slide, gene, treatment and dye, a subset of the interactions between terms and a random error. These terms can be treated all as fixed effects in a non parametric optic (Kerr *et al.*, 2000), or assuming some effects as fixed and some other as random. In the latter case the variance is decomposed in a sum of components and the model is framed in a parametric approach (Wolfinger *et al.*, 2001).
Instead of considering different component of variance for each gene, it is possible to decompose it accordingly to the expression values. This approach starts from the idea that the standard deviation for expression measure increases exponentially to the level of expression for high levels and linearly for low levels (not expressed genes). Mean and variance for response variable are parameterized differently for high or low values in a parametric perspective (Rocke and Durbin, 2001).
Not far from this point of view, Delmar *et al.* propose a mixture model

on gene specific variance distribution, identifying cluster of genes with equal variance under a non parametric point of view.
For our purposes, we consider M as the matrix of log ratio

$$m_{gi} = log\frac{\text{(expression level in sample 1)}_{gi}}{\text{(expression level in sample 2)}_{gi}} \tag{1.1}$$

with G rows and I column, where m_{gi} represents the logratio for g^{th} gene in the i^{th} array. When we consider the absolute expression level we work with $\mathbf{X}$, matrix of log intensity with G rows and 2I columns, where x_{gik} denotes the log intensity level for g^{th} gene, i^{th} array and k^{th} condition. We refer to the natural logarithmic scale, even if some authors model the base 2 logarithm.

1.4.1 Modified t test

Adding a constant to the denominator of *t-test* is the simplest method to try finding a valid variability measure. It permits both to eliminate the problem of asymmetric distribution of variance that arises for small number of replicates, and to reduce the loss of power that occurs when a global estimate of variability is used.

Non parametric frequentist SAM Tusher *et al.* (2001) introduce a non parametric approach that considers a modification of *t* test. For each gene a score is assigned on the basis of the change in its normalised expression divided by the ratio of standard deviation of repeated measures for that gene and the square root of replicates number. The *t* statistic proposed by the authors has a global component of variability and a gene specific one:

$$t_g = \frac{m_{g.}}{s_g/\sqrt{I} + s_0} \tag{1.2}$$

where $m_{g.}$ is the mean log ratio for the g^{th} gene over the I arrays. The component s_g is the standard deviation for the g^{th} gene coming from repeated measurements:

$$s_g = \sqrt{\frac{1}{I-1}\left\{\sum_i [m_{gi} - m_{g.}]^2\right\}} \tag{1.3}$$

For the choice of s_0 let s^α be the α percentile of s_g distribution and $t_g^\alpha = \frac{m_{g.}}{s_g/\sqrt{I}+s^\alpha}$ The authors compute the 100 quantiles of $s_g(q_1, .., q_{100})$ and for $\alpha \in (0, .05, \ldots, 1)$ they compute the absolute deviation from the median of t_g and its coefficient of variation. The chosen $\hat{\alpha}$ is which minimise the coefficient of variation and $s_0 = s^{\hat{\alpha}}$.
To find differentially expressed genes, this approach identifies a threshold and estimates the number of false positives through an approach based on permutations. The genes are ranked according to the t_g score in descending order and a large number of permutations of the labels of I samples are calculated. The relative differences are computed as t_{gp} statistic for each permutation and the expected relative difference is $t_{Eg} = \frac{\sum t_g p}{np}$ where np is the total number of permutations considered. The plot of t_g vs t_{Eg} shows a percentage of genes that lie far from the $t_g = t_{Eg}$ line by a distance greater than a specified threshold Δ and that are potentially differentially expressed. To determine the number of false positive genes, horizontal cutoffs are drawn corresponding to the smallest positive t_g and the biggest negative t_g. At each permutation the number of false positive is calculated as the number of genes that pass these thresholds and the average over all the permutations is the estimated number of false positive.
A modification of the previous approach is carried out by Efron *et al.* (2003). They start from the same statistic in (1.2) for normalised data and the variability estimate is equivalent to that in (1.3). But, differently from the previous approach, the estimate of the "fudge factor" s_0 is the 90^{th} percentile of the distribution of s_g. After the identification of the *t* statistic, this approach varies from the previous one in the identification of differentially expressed genes: the p is the probability that a gene is affected by the considered state (treatment) and a complementar probability $1-p$ is that the same gene is unaffected. Both the t_g for affected or unaffected genes have a density f_p and f_{1-p} that are estimated empirically. From these estimates, applying Bayes rule the posterior probabilities p_1 and p_0 that a gene with score t_g was affected or unaffected by treatment are obtained. For an application of these two methodologies see chapter 2.

Parametric Empirical Bayesian Lönnstedt and Speed (2002) suggest calculating a Bayesian log posterior odds statistic that originates from the classical t test, but has different assumptions. In particular, normalised m_{gi} are treated as realizations from Gaussian random variables, whose parameters $\mu_g \mid \sigma_g^2, \sigma_g^2$ have prior normal and inverse gamma distribution respectively. The indicator variable $I_g = 0$ identifies that a gene is unchanged ($\mu_g = 0$) with $1-p$ prior probability, while $I_g = 1$ shows that a gene changes ($\mu_g \neq 0$) with p prior probability.
The log posterior odds for a gene g to be expressed is the following:

$$b_g \quad = \quad log \frac{Pr(I_g = 1 \mid m_{gi})}{Pr(I_g = 0 \mid m_{gi})}$$

The authors propose an explicit formula for b_g, that is a function of p and of μ_g, σ_g^2 hyperparameters:

$$b_g \quad = \quad log \frac{p}{1-p} \frac{1}{1+nc} \left[\frac{a + s_g^2 + m_g.^2}{a + s_g^2 + \frac{m_g.^2}{1+nc}} \right]^{\nu + \frac{n}{2}}$$

where a and ν are hyperparameters in the inverse gamma prior for the variances and c is a hyperparameter in the normal prior. The p parameter is externally set (a grid of different values is suggested), while the σ_g^2 and μ_g hyperparameters are empirically estimated.
For an application of this methodology see chapter 2.

Parametric Full Bayesian Baldi and Long (2001) use a fully Bayesian hierarchical model for the raw log-expression. They consider the independence for the observations and models $x_{gik} \sim N({}_x\mu_g, {}_x\sigma_g^2)$.
Prior distribution on $({}_x\mu_g, {}_x\sigma_g^2)$ is considered jointly, avoiding to assume independence between ${}_x\mu_g$ and ${}_x\sigma_g^2$. In particular, the distribution on parameters is conjugate as follow:

$$\begin{aligned} {}_x\mu_g \mid {}_x\sigma_g^2 &\sim N(\mu_0, {}_x\sigma_g^2/\lambda_0) \\ P({}_x\sigma_g^2) &\sim \Gamma^{-1}(\nu_0, {}_x\sigma_{0g}^2) \end{aligned}$$

where μ_0 and ${}_x\sigma_g^2/\lambda_0$ can be interpreted as the location and scale of ${}_x\mu_g$, while ν_0 and ${}_x\sigma_{0g}^2$ as the degrees of freedom and scale of ${}_x\sigma_g^2$ and Γ^{-1}

indicates the inverse Gamma distribution.
To estimate ${}_x\sigma^2_{0g}$ it is possible to use the entire set of genes, but the authors suggest a flexible implementation in which the genes are ranked accordingly to their expression and for each gene ${}_x\sigma^2_{0g}$ is the result of pooling together all the neighboring genes contained in a window of size w. Using the Bayes theorem, the authors give the posterior joint distribution of interesting parameters ${}_x\mu_g$, ${}_x\sigma_g$. The posterior estimate of ${}_x\mu_g$ (after integration over ${}_x\sigma_g$) has a t distribution and gives a measure of expression for the g^{th} gene.
For an application of this methodology see chapter 2.

1.4.2 ANOVA models

Linear models and ANOVA models seem to be very useful in microarray experiments, when replicates are available, because they allow to decompose the variability distinguishing its different sources. They were applied in the microarray area firstly by Kerr *et al.* (2000).

Non parametric ANOVA The principal model presented in the first paper by Kerr *et al.* is identified by 4 principal factor and some interactions as follow:

$$x_{gikj} = \mu + A_i + D_j + V_k + G_g + GA_{gi} + GV_{gk} + \epsilon_{gikj} \qquad (1.4)$$

where A_i identifies the array effect $(i = 1, \ldots, A)$, D_j identifies the dye effect $(j = 1, 2)$, V_k identifies the condition effect ($k = 1, 2$) and G_g identifies the gene effect $(g = 1, \ldots, G)$. The normalisation is part of the model.
This is the basic formulation, but more recent complex specifications have been proposed (Kerr *et al.*, 2002). In this last version to account for non linear effects of the dyes the authors suggest using the loess adjustment (Yang *et al.*, 2002), array-by-array. The authors perform an adjustment on the log expressions instead of the logratios that are then incorporated in the linear model.
The term of interest is $(GV_{g2} - GV_{g1})$ (equivalent to m_{gi} in section 1.4.1) that estimates the differential expression between the two conditions for the g^{th} gene. To infer about differential expression they use the bootstrapping technique.

The authors specify fixed effect models where the residual $\epsilon_{gikj} \sim N(0,_\epsilon \sigma^2)$ are homoscedastic. However, they observe that even after the loess normalisation it seems a source of heteroscedasticity remains in the data. To model this component they consider two possible methods:

- heteroscedasticity for ϵ_{gikj}
- Intensity-dependent distribution of ϵ_{gikj}

The first method does not seem to work well, apart for the experiments with a very large number of replicates.
The second method starts from the observed relation between errors and intensity. The procedure adopted standardises the residuals to make independent from the intensity values.

Parametric ANOVA Wolfinger *et al.* (2001) present a mixed ANOVA model distinguishing between a normalisation model and a gene model. He does not model the dye effect, that is confused with the condition effect, working in absence of dye-swap. The normalisation model includes the same normalisation terms as in Kerr model (1.4) (μ, V_k, A_i, AV_{ik} and a residual ϵ_{gik}).
The gene model is the following:

$$r_{gik} \quad = \quad G_g + (GV)_{gk} + (GA)_{gi} + \gamma_{gik}$$

where r_{gik} is the residual obtained subtracting the fitted values $\widehat{x}_{gik}$ from x_{gik} and the other effects have the same meaning as in Kerr *et al.* model. This is a mixed model, assuming random and fixed effects. In particular, $A_i, \epsilon_{gik}, (GV)_{gk}$, GA_{gi}, γ_{gik} are random effects assumed normally distributed with 0 mean. The variance is decomposable as follow:

$$Var(x) \quad = \quad \sigma^2_A + \sigma^2_\gamma + \sigma^2_{GV} + \sigma^2_{GA_g} + \sigma^2_{\epsilon_g}$$

The random effect are assumed to be independent across index and with each other; $\sigma^2_{(GA_g)}$ and $\sigma^2_{\epsilon_g}$ have also a gene specific component (heterogeneity for genes). To estimate the effect and the variance components Restricted Maximum Likelihood is used (**?**). The interesting measure is $(GV)_{gk}$ and t tests for mixed models are used to test for differences in expression within each gene.

1.4.3 Two-components error model

A different point of view on estimating variability is carried out by Rocke and Durbin (2001), who start from a two-components model for the response y at concentration μ:

$$
\begin{aligned}
y &= \alpha + \mu \exp(\eta) + \epsilon \\
\eta &\sim N(0, \sigma_\eta) \\
\epsilon &\sim N(0, \sigma_\epsilon)
\end{aligned}
$$

where $(y - \alpha)$ is the background corrected intensity and α is estimated from replicates of not expressed genes (usually considered as 10% of lowest expression genes) or from blank spots; ϵ is the residual effect for the genes that are not expressed and η is the residual effect for high expressed genes. Actually, the model is differently specified if we distinguish between high and low intensity genes; in fact, for medium-high expression the logarithmic transformation stabilises the variance, but produces high variability for low expression levels. Then, for the g^{th} gene, the i^{th} array and the k^{th} condition:

$$
\begin{cases}
y_{gik} \approx \alpha_{ik} + \epsilon_{gi} + \epsilon_{gik} & \text{(Low)} \\
log(y_{gik}) \approx log(\alpha_{ik} + \mu_{gk}) + \eta_{gi} + \eta_{gik} & \text{(High)}
\end{cases}
\tag{1.5}
$$

Then the variability is decomposed accordingly to the previous specification:

$$
\begin{cases}
\text{Var}(y_{gik}) \approx \sigma^2_{\epsilon_{gi}} + \sigma^2_{\epsilon_{gk}} & \text{(Low)} \\
\text{Var}(log(y_{gik})) \approx \sigma^2_{\eta_{gi}} + \sigma^2_{\eta_{gk}} & \text{(High)}
\end{cases}
\tag{1.6}
$$

The interesting quantities are the differences between the two intensities on natural scale (low intensity) and on logarithmic scale (high intensity). Then the variance components to be estimate are only $\sigma^2_{\epsilon_{gi}}$ and $\sigma^2_{\epsilon_{gk}}$, $\sigma^2_{\eta_{gi}}$ and $\sigma^2_{\eta_{gk}}$.

1.4.4 Mixture models

Some authors discuss the too stringent assumption of the homoscedastic assumption on variance and the overparametrised model using a

gene specific variance. An intermediate approach can be found in mixture models. For statistical methodology we refer to McLachlan and Basford (1999), while there are several applications on microarray (Allison *et al.*, 2002), (Pan *et al.*, 2003), (Parmigiani *et al.*, 2002). These applications consider a mixture on the expression levels and try identifying groups of genes with homogeneous measure of expression assigning a variance for each group. Under a different point of view, Delmar *et al.* (2004) work to the identification of cluster of genes with equal variance. They consider a mixture models on variance distribution and assign each gene to specific groups according to the largest posterior probability to belong to.
The normalised log ratio for gene and replication is specified as follow:

$$m_{gi} \quad = \quad \mu_g + \epsilon_{gi}$$

where ϵ_{gi} is normally distributed centered on 0. The measure of differential expression is $m_{g.}$ as defined in equation (1.1) and the variability measure is the same as (1.3). The differential score is:

$$t_g \quad = \quad \frac{m_{g.}}{\sqrt{\hat{s}^2_{E_g}/I}}$$

(without any correcting global factor). If the degrees of freedom are $\nu = I - 1$, the distribution of $X_g = S^2_{E_g}$ is $\Gamma(\sigma^2_g, \nu)$; however, the true value of σ^2_g is unknown. The authors propose a mixture of gamma distributions in which each component represents a group of genes with an homogeneous variance. Let h be the number of components, $p_j, j \in [1, ..., h]$ be the probability that a gene belongs to the j^{th} group and σ^2_j be the variance for the j^{th} group. The model on variance is the following:

$$X_g \quad \sim \quad \sum_{j=1}^{h} p_j \Gamma(\sigma^2_j, \nu)$$

The parameters of the model are estimated according to the maximum likelihood principle; maximization of the log likelihood function is carried out by EM algorithm. To identify the number of variability groups (mixture components) the BIC statistic is used (McLahan and Peel, 2000). After estimating parameters $(p_1, \dots, p_h)$ and $(\sigma_1^2, \dots, \sigma_h^2)$ each gene g is assigned to a group j according to the highest posterior probability τ_{gj}. The gene variance σ_g^2 can be assigned in two different ways. The first is attributing to each gene the variability of the group it belongs to:

$$\hat{\sigma}_g^2 = \hat{\sigma}_j^2$$

The second is calculating for each gene a sum of h variances, weighted by the estimated posterior probability of a gene to belong to the j^{th} group. Identification of differentially expressed genes is performed through the *t* statistic:

$$t_g = \frac{m_{g.}}{\frac{\hat{\sigma}_g}{\sqrt{I}}}$$

For an application of this methodology see chapter 2.

1.5 Discussion

Considering a scale for the different levels of variability estimate, the modified *t* test lies on the first step. It is the simplest method that permits to stable the gene specific variability measure. It can be seen as the most intuitive approach and the easiest to implement, considering only the "global" variability component to be estimate in addition to the gene specific ones. Choosing of the s_0 to minimize the coefficient of variation (Tusher *et al.*, 2001), as well as the 90% percentile (Efron *et al.*, 2001) or estimated by data (Lönnstedt and Speed, 2002), takes under control the genes with low expression (and low variance) and avoid to call them significant. Moreover, using a full Bayesian approach (Baldi and Long, 2001), the problem is decomposed in different levels of conditional distributions and the entire distribution of all the parameters can be studied, instead of the summary statistic (i.e. mean, median). In addition, the observed relation between mean and variance is taken into account modelling jointly the distribution of (μ_g, σ_g) parameters. That issue is particularly important in chapter 2. The authors point out the possibility to pass from the full Bayesian to the empirical Bayesian approach, setting the estimate of μ_0 empirically.

On an upper step, fixed ANOVA models (subsection 1.4.2) distinguish between variability of the effects included in the model and residual variability. The approach carried out by Kerr *et al.* (2000, 2002), has the advantage to be non parametric, avoiding distributional assumption. On the other side, to estimate parameters needs bootstrap technique, that presumes homoscedasticity on residuals quantities and that is not observed on the data. The gene specific alternative does not seem to work well, apart for the experiments with a very large number of replicates. In fact, when the replicates are few, what is in most of experiments, this procedure originates narrow confidence intervals for some genes, that are called significant even if the difference in log expression is very small between the two channels. For this reason the authors suggest using the intensity-dependent distribution for residuals. In addition, they consider all the effect as fixed, even if for some effects (as AG) randomness could be very appropriate, but not used for the heavy tailed residuals.

Random effects are used by Wolfinger *et al.* (2001) under a perspective that permits to decompose the variance in several components. In particular, the heterogeneity in gene model ($r_{gik} = G_g + (GV)_{gk} + (GA)_{gi} + \gamma_{gik}$) for GV gives different degrees of variability to the genes, that have backlashes to the identification of confidence intervals resulting in differentially expressed genes.

On the same step of decomposition level, but under a different perspective, Rocke and Durbin try assigning different variability for the two intensity levels (low level versus high level). It can be seen as a different way to take into account the relation between intensity level and variability: under a frequentist point of view it does not model jointly the mean and variance parameters (as Baldi e Long do in a Bayesian perspective), but specify separately the two models for the two different levels of intensity, developing a methodology that can be considered a naïf mixture model.

On the highest step and quite close to the previous approach, Delmar *et al.* focus the attention on a cluster strategy for genes on the basis of equal variability. It limits the number of parameters and estimating variability gains strength from the closer genes (in terms of variance). However, it does not consider the relation between variance and intensity level making this approach incomplete for some aspects.

Bibliography

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(6)**, 5009-5019.

Delmar, P., Robin, S. and Daudin, J.J. (2004) Efficient variance modelling for differential analysis of replicated gene expression data, *Bioinformatics*, to be published.

DeRisi, J.L. Penland, L., Brown, P.O., Bittner, M.L. *et al* (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nature Genetics*, **14**, 457-460.

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

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2003) Empirical Bayes Analysis of a Microarray Experiment., *JASA* **96**, 1151-1160.

Kerr, M.K. and Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**, 817-837.

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2002) Statistical analysis of gene expression microarray experiment with replication., *Statistica Sinica*, **12**, 203-217.

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

Rocke, D.M. and Durbin, B. (2001) A model for Measurement Error for Gene Expression Data, *Journal of Computational Biology*, **8(6)**, 557-569.

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

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. and Davis, R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes, *PNAS*, **96**, 10614-10619.

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

Tao, H., Bausch, C., Richmond, C., Blattner, F.R. and Conway, T. (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media, *Journal of Bacteriology*, **181**, 6425-6440.

Tusher, V., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarray applied to the ionizing radiation response, *PNAS*, **98(9)**, 5116-5121.

Wolfinger, R.D., Gibsin, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology*, **8(6)**, 625-637.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalisation for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, **30(4)**, e15.

2 An application to compare variability on microarray data [*]

2.1 Abstract

Microarray experiments are a new tool that permits to consider and treat thousands of genes at the same time. They are used to compare the levels of expressions for genes in different biological samples. A natural choice to identify differentially expressed genes is perform a classical *t* test. This approach is in general unsatisfactory. Two problems which affect the *t* statistic in the microarray framework are considered here: the availability of only few replicates (e.g. 2, 3) for each gene; the dependence between mean and standard deviation of gene log expression (the *t* statistic numerator and denominator, respectively). Several methods are propose to avoid these problems. We present and compare the performance of five different approaches, which are based on average and standard deviation values of each gene, analogously of naïf *t* statistic.
Finally we consider the computational performances offered by freely available codes proposed by the authors.
The methods are applied to two public cDNA datasets, that differ for some relevant features and we discuss the results.

2.2 Introduction

Experiments based on microarray technology have assumed a very important role in the biological research field. Statistical aspects are relevant in many phases of the experimental procedure, as the design of

[*] An extract from this chapter was submitted to *Computational Statistics* titled: **"Estimating variability in microarray experiments: a comparative review of software"**, in collaboration with Simona Toti and Annibale Biggeri

experiment, quality control and normalisation, to the aim of taking under control sources of non biological variability.

After normalisation, a statistical analysis is performed to determine whether for each gene, observed difference in expression levels between different experimental conditions (e.g. cells from case versus control subjects) is significant or not.

In the classical statistical approach, a *t* test is achieved to characterise the differences between standardized means from two populations. For the microarray data, a *t* statistic is performed for each gene: the difference between gene log-expression levels in two conditions to be tested is at the numerator and the variance of this difference divided by the number of replicates is at the denominator.

However, a crucial issue that emerges using this statistic in the microarray framework is the limited number of repetition due to the high cost (and lack of material) to perform an experiment. Typically for thousands genes considered in the analysis only 2, 3 replicates for each gene are available causing underestimates of variances. Small variances lead the *t* values to be large even if its numerator is small carrying out false positives genes.

Many authors have tackled the problem under different perspectives. In the present work we describe the principal methods arising from classical *t* tests that try taking under control the issue explained above. The considered methods are: SAM (Tusher *et al.*, 2001), Non Parametric Empirical Bayes (Efron *et al.*, 2001), Parametric Empirical Bayes (Lönnstedt and Speed, 2002), Full Bayesian (Baldi and Long, 2001) and mixture model (Delmar *et al.*, 2004).

In the microarray literature, with the aim of identifying differentially express genes, also approaches different from the *t* tests were presented (ANOVA by Kerr *et al.*, error models by Rocke and Durbin), but for the sake of comparability we do not treat them.

The performance of a specific methodology considered is evaluates on its computational aspects too. Actually, for this research field characterised by a large amount of data, specific computational tools are required. So we focus the attention on the methodologies that present a freely available software.

The paper follows this structure: in section 2.3 we introduce the notation used thereafter and the approaches to identify differentially expressed genes, focussing the attention on variability modelling. Section 2.4 describes the dataset we use and the preprocessing we perform; section 2.5 presents the results in terms of differentially expressed genes and of variability estimate; finally we explain some discussion points in section 2.6.

2.3 Estimating Variability

In this section we introduce the notation used thereafter. The data matrix can be presented in two different forms: an absolute expression matrix **X** in which each row identifies a gene and each column corresponds to a channel $(k = 1, 2)$ for each array $(i = 1, \ldots, I)$ and x_{gik} denotes the log intensity for g^{th} gene, i^{th} array and k^{th} channel. Alternatively, the data matrix can be considered as a relative expression matrix M, in which each row corresponds to a gene and each column identifies an array: m_{gi} denotes the log ratio (as the difference between x_{gi1} and x_{gi2}) for g^{th} gene and i^{th} array. We also define A as the matrix of mean intensities, the arithmetic mean of the two channels log intensities and its elements as a_{gi}
A classical tool for descriptive analysis of data is the M versus A plot.

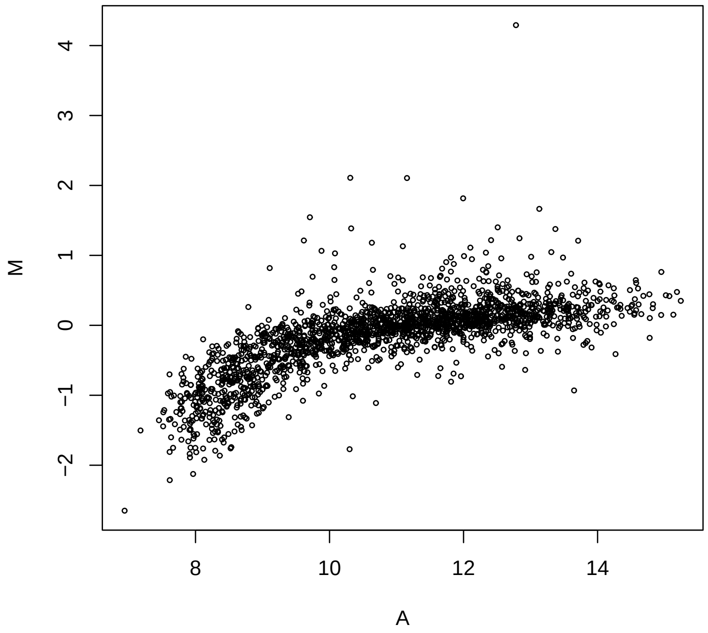

Figure 2.1: M vs A plot for raw data (TCDD dataset).

From figure 2.1 the relation between intensity level (a_{gi}) and variability in the sense of difference between channel intensities (m_{gi}) emerges. In particular, the association between the two variables is characterised by high variability associated with extremely intensity values.
An other explorative tool is the histogram for the observed variances of raw data $s_g = \sqrt{\frac{1}{I-1}\left\{\sum_i \left[m_{gi} - m_{g.}\right]^2\right\}}$, where $m_{g.}$ is the mean log ratio for the g^{th} gene over the I arrays.

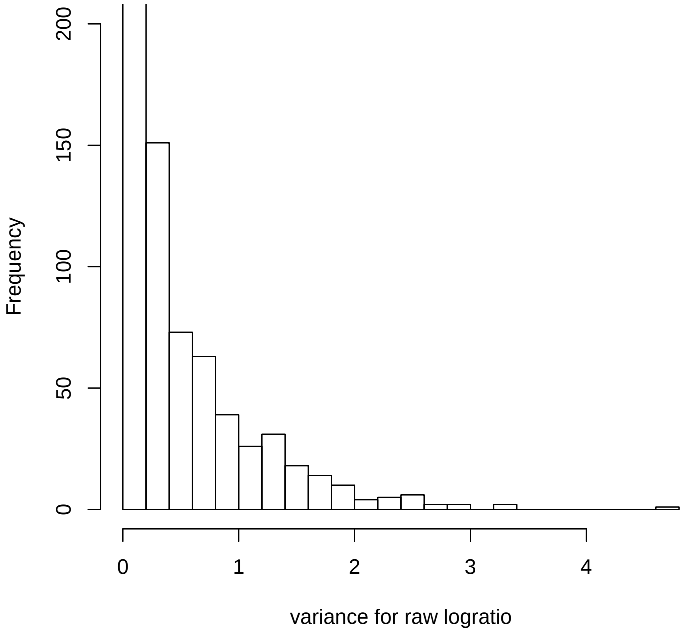

Figure 2.2: Histogram for variance of raw logratio. (The frequency of the first bin is truncated at 200, but it is 1400).

In figure 2.2 the hypothesis of homoscedasticity appears reliable for the genes belonging to the first bin (the frequency is 1400) but for the other bins heteroscedasticity is checked.
The issues of relation between M and A and of heteroscedasticity cover a central role in the analysis of microarray and can originate substantial differences in the results. On the basis of them, we try evaluating 5 approaches to the analysis: all of them take into account the heteroscedasticity problem; some of them take into account also the relation between M and A.

2.3.1 SAM and Non Parametric Empirical Bayes

The classical statistic for *t* test in microarray analysis is defined as $t_g = m_{g.} \backslash \frac{s_g}{\sqrt{I}}$.
Adding a global constant to the denominator of *t test* is the simplest method to try finding a robust variability measure. It permits both to eliminate the problem of asymmetric distribution of variability and to reduce the loss of power, when small number of replicates is available.

Variability For each gene a score is assigned on the basis of the change in its normalised expression divided by the standard deviation of repeated measures of that gene. The *t* statistic has a global component of variability and a gene specific one:

$$t_g = \frac{m_{g.}}{\frac{s_g}{\sqrt{I}} + s_0} \tag{2.1}$$

Different choices of s_0 are available: Tusher *et al.* (2001), assign s_0 to the percentiles of the s_g distribution, for each percentile compute the median absolute deviation from the median of t_g, the coefficient of variation for these values and choose the percentile that minimise the coefficient of variation as s_0.
Efron assigned s_0 to the 90% percentile of the s_g distribution (Efron *et al.*, 2001).

Differential Expression Both the methods are based on the non parametric theory, but their approach to the identification of differentially expressed genes is quite different.

SAM Tusher *et al.* identifies a threshold (Δ) for a gene to be differentially expressed and estimate the number of false positives through an approach based on permutations. This approach does not assign a p-value, but uses the q-value. It is defined as the minimum pFDR that can be achieved over all the rejection regions that contain the observed statistic, where pFDR is:

$$pFDR(\Delta) = E\left[\frac{V(\Delta)}{R(\Delta)} | R(\Delta) > 0\right] \tag{2.2}$$

for V false positive over R rejected hypothesis for Δ. The q-value gives the strength that a gene is differentially expressed.

Non Parametric Empirical Bayesian Efron *et al.* consider the probability that a gene is affected by the considered state (treatment) and the complementary probability that the same gene is unaffected. The densities for the two probabilities are empirically obtained and through the Bayes rule the posterior probability to be expressed or not is calculated.

As for the relation between M and A, these two methods do not propose any particular solutions. However, they suggest performing a loess normalisation usually considered able to eliminate this relation.

2.3.2 *Parametric Empirical Bayesian*

Lönnstedt and Speed (2002) treat m_{gi} as realizations from Gaussian random variables:

$$m_{gi} \quad \sim \quad N(\mu_g, \sigma_g^2) \tag{2.3}$$

where μ_g and σ_g^2 have prior normal and inverse gamma distribution respectively.

Variability The hyperparameters of σ_g^2 prior distribution are empirically estimated. The relation between mean and variance is taken into account assuming the conditional dependence of μ_g from σ_g^2. So, $\mu_g \sim N(0, c\sigma_g^2)$ where c is a positive scale parameter.

Differential Expression The indicator variable $I_g = 0$ identifies that a gene is unchanged ($\mu_g = 0$) with $1 - p$ prior probability, while $I_g = 1$ shows that a gene changes ($\mu_g \neq 0$) with p prior probability.
The authors suggest the test statistic b_g:

$$b_g \quad = \quad log\frac{Pr(I_g = 1 \mid m_{gi})}{Pr(I_g = 0 \mid m_{gi})} \tag{2.4}$$

which is the log posterior odds for the g gene. For a specific threshold of probability to be differentially expressed (p^*) a gene is retained differentially expressed if $b_g > log\frac{p^*}{1-p^*}$. The authors propose an explicit formula for b_g, that has a gene specific component and a global one; the first is function of $m_{g\cdot}$ and σ_g^2, while the second is a function of p and of

μ_g, σ_g^2 hyperparameters; the p parameter is externally set (a grid of different values is suggested),while μ_g hyperparameters are empirically estimated.

2.3.3 *Parametric Full Bayesian*

Baldi and Long (2001) use a fully Bayesian hierarchical model for the raw log expression. They model

$$x_{gik} \sim N({}_x\mu_g, {}_x\sigma_g^2) \tag{2.5}$$

Variability Analogously to the previous model, the authors take into account the relation between ${}_x\mu_g$ and ${}_x\sigma^2$ through a conditional dependence of the former to the latter. The distribution on parameters is normal and inverse gamma respectively. The original point of this approach is in the specification of shape hyperparameter for inverse gamma distribution (${}_x\sigma_{0g}^2$). To estimate it the genes are ranked accordingly to their expression and for each gene ${}_x\sigma_{0g}^2$ is the result of pooling together all the neighboring genes contained in a window of size w. In this way the correction factor of gene-specific variance is not global, but calculated grouping genes of similar expression.

Differential Expression Using the Bayes theorem, the authors give the posterior joint distribution of ${}_x\mu_g$ and ${}_x\sigma_g$. Integrating over the last, the marginal distribution of ${}_x\mu_g$ is a Student *t* . The hypothesis of differential expression for g gene is tested by the *t* statistic p-value.

2.3.4 *Mixture Models*

In the previous paragraphes we have presented several models for a global correction of gene specific variance. An alternative can be found in the clustering of variances. There are several applications of mixture models on microarray data (Allison *et al.*, 2002), (Pan *et al.*, 2003), (Parmigiani *et al.*, 2002). These applications consider a mixture on the expression levels and try identifying groups of genes with homogeneous measure of expression assigning a variance for each group. Under a different point of view, Delmar *et al.* (2004) work to the identification of cluster of genes with equal variance. They consider a mixture models on variance distribution and assign each gene to specific groups according to the largest posterior probability to belong to.

Variability The authors propose to model the variability as a mixture of gamma distribution in which each component is computed from a group of genes with an homogeneous variance. Let h be the number of components, $p_j, j \in [1, ..., h]$ be the probability that a gene belongs to the j^{th} group and σ_j^2 be the variance for the j^{th} group. Let S_g^2 be the variance of logratio and ν the degrees of freedom $\nu = I - 1$, then $X_g = \nu S_g^2$ is a gamma mixture:

$$X_g \sim \sum_{j=1}^{h} p_j \Gamma(\sigma_j^2, \nu) \tag{2.6}$$

The parameters of the model are estimated according to the maximum likelihood principle; maximization of the log likelihood function is carried out by EM algorithm. To identify the number of variability groups (mixture components) the BIC statistic is used (McLahan and Peel, 2000).

Differential Expression Identification of differentially expressed genes is performed through the *t* statistic:

$$t_g = \frac{m_{g.}}{\hat{\sigma}_g}\sqrt{I} \tag{2.7}$$

where $\hat{\sigma}_g$ can be calculated in two different ways. The first one is $\hat{\sigma}_g^2 = \hat{\sigma}_j^2$. The second is a sum of the h variances weighted by the estimated posterior probabilities of a gene to belong to the j^{th} group.
The distribution of t_g is calculated taking into account the uncertainty of gene assignment to a specific group.
Data are normalised externally before analysis; the relation between mean and variance is considered only through normalisation.

2.4 Data and Preprocessing

2.4.1 TCDD Dataset

The data presented here were published in Kerr *et al*. (2002) and are an experiment to study tetrachlordibenzo- p-dioxin (TCDD). It is known that this compound induce several biological and biochemical responses, including gene induction. The experiment used the human hepatoma cell line HepG2 as an in vitro model to study TCDD. HepG2 is an established cell line for which metabolic enzymes are known to be inducible (Kikuchi *et al.*, 1998), (Li *et al.*, 2003). Thus it can be considered a prototype of the TCDD response. The experimental design included replication to control the noise that is associated with microarray data, obtained using six arrays. A separate labelling reaction was performed for each hybridisation. Each gene is singly spotted. Each array has 1900 genes. Table 2.1 summarizes the triple dye-swap experimental design.

	Cy3	Cy5
Array 1	Variety 2	Variety 1
Array 2	Variety 2	Variety 1
Array 3	Variety 1	Variety 2
Array 4	Variety 2	Variety 1
Array 5	Variety 1	Variety 2
Array 6	Variety 1	Variety 2

Table 2.1: Design of TCDD experiment.

2.4.2 E-Coli Dataset

This dataset was presented in Tseng *et al*. (2001) and it is relative to a calibration experiment on Escherichia coli cells grown in glucose. Calibration experiments used the same mRNA pool divided into two aliquots and labelled separately with two different dyes in order to investigate variations in this technology. Two array were performed, in which genes are singly spotted on each of them. The number of genes considered is 4129. This dataset is particularly interesting, because the calibration design permits to exclude the presence of any variety effect that can originate systematic differences in the expression for the two dyes (the probability of a false negative is 0). For this reason all the variability is the non biological one and it seems very useful to the aim of comparing the performance of different methodologies in which variability estimate play a valuable role. In table 2.2 we present the experimental design for this dataset.

	Cy3	Cy5
Array 1	Variety 1	Variety 1
Array 2	Variety 1	Variety 1

Table 2.2: Design of E-Coli experiment.

2.4.3 Quality control and normalisation

We perform quality control according to Simon criteria (Simon *et al.*, 2003). In particular, we excluded a spot if the number of pixels used to calculate the intensity is less than 25 for the foreground intensity in either channel, if the signal is lower than 200 for both the channels or if the ratio between the average foreground intensity and the median background intensity is smaller than 1.5 in either channel. Viceversa, spots with a large signal for one channel and low signal for the other are not eliminated, but modified to become analysable, forcing the low intensity signal (defined as less than 200) to 200.

Moreover, when the method requires normalisation, we normalise the data through a local A-dependent normalisation (loess) globally calculated for each slide (Yang *et al.*, 2002). It is pointed out that a loess for each print tip is better than the one for the entire array, but we are not able to apply the first, due to the absence of information about the grid definition for the two datasets.

The number of genes remained after quality controls is 1887 for TCDD dataset and 3880 for E-Coli dataset.

2.5 Results

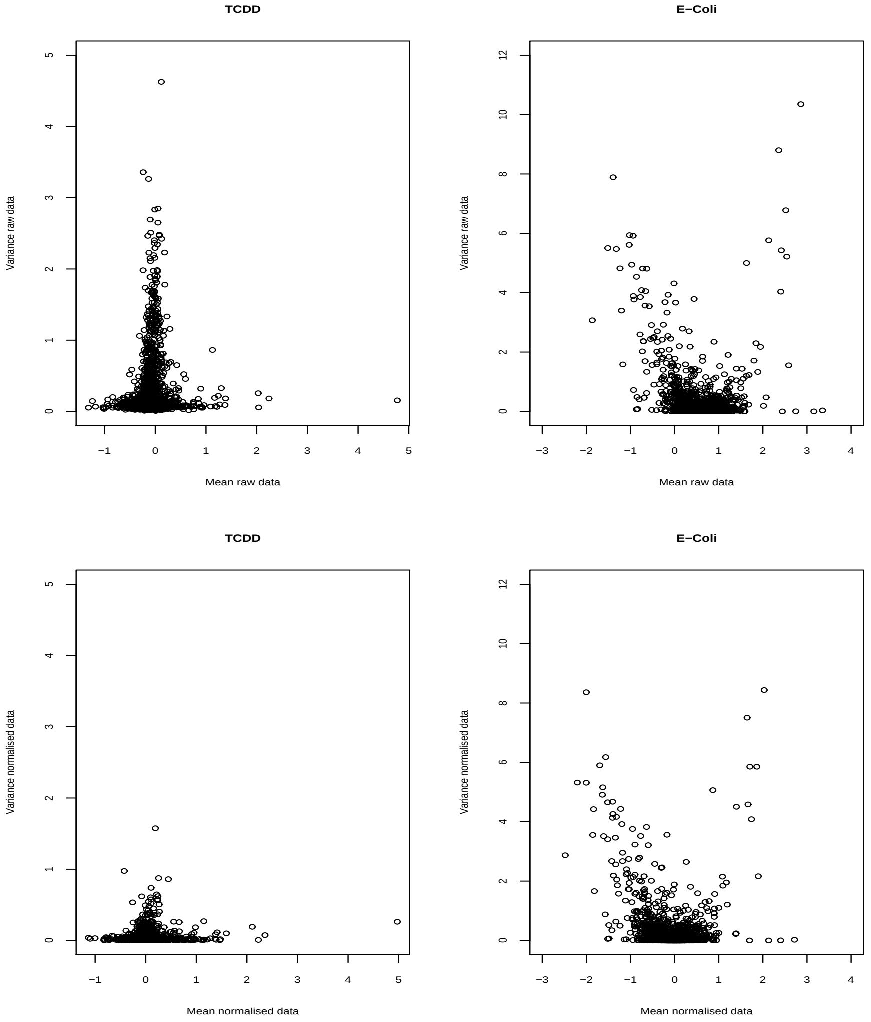

Figure 2.3: Mean vs variance plot.

Figure 2.3 visualises the relation between mean of logratio intensity and variability for each gene.
The TCDD dataset shows the high capability of normalisation in reducing variability. It points out the effect of loess normalisation in correct-

ing the relation between M and A (see figure 2.1). On the other hand, the E-Coli M vs A plot (not reported) does not show a systematic relation. The visible normalisation effect is the translation of the scatterplot to be centered in 0 (correction of dye effect).

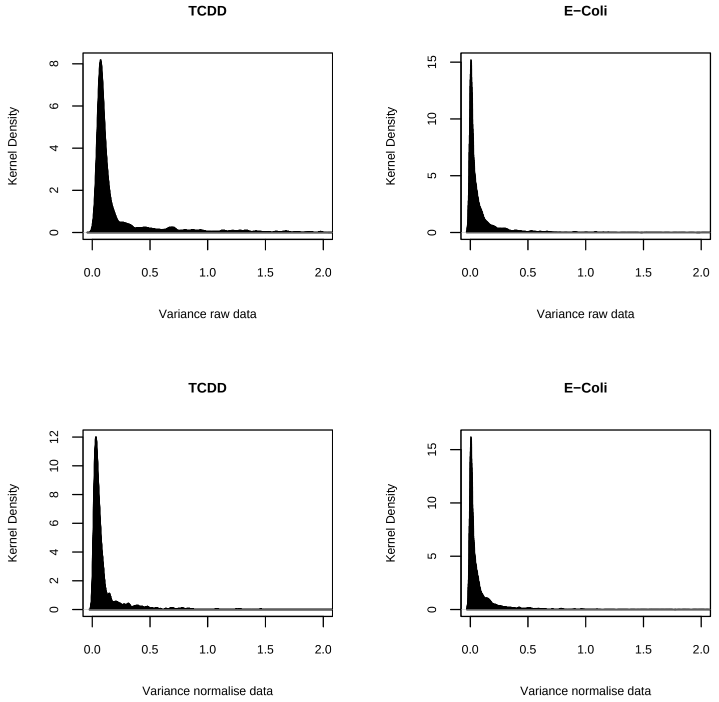

Figure 2.4: Observed variance kernel density.

Figure 2.4 points out the observed variability distribution for the two datasets and the effect of normalisation in variance distribution. Actually, the loess normalisation tends to reduce the extreme values to the centre of the distribution and its effect is more visible on a comparative experiment (on the left) than on a calibration one (on the right). However, a non negligible heteroscedasticity remains even after normalisation.

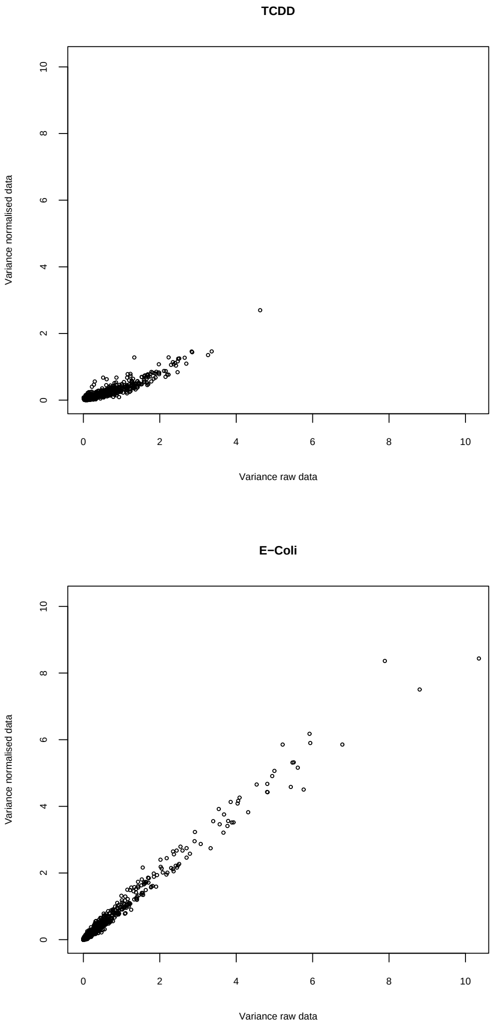

Figure 2.5: Comparison between raw and normalised variability.

In figure 2.5 there are the confirms of what described above: the plot of raw logratio variability versus normalised logratio variability shows a reduction for the comparative dataset (TCDD) and a substantial unchange for the E-Coli dataset.

2.5.1 *Comparison within methods*

Variability comparison First we point out the comparison of different methodologies in terms of effects on the variability estimate.

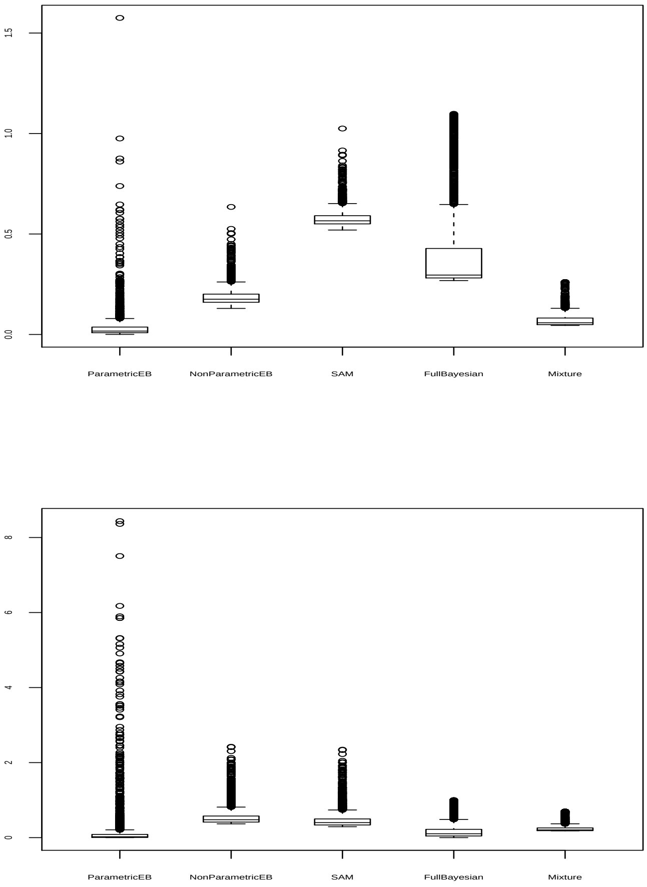

Figure 2.6: Boxplot for variability measures.

In particular, figure 2.6 shows that parametric EB is the method that reproduces the highest dispersion. Viceversa, mixture model presents the most concentrate distribution of variance. The other approaches display an intermediate dispersion, with an evident shift due to the global variability component at denominator of t statistic. In particular, non parametric EB and SAM present a very similar distribution apart from the strength of the shift, for the different way of computing the global component of variability.

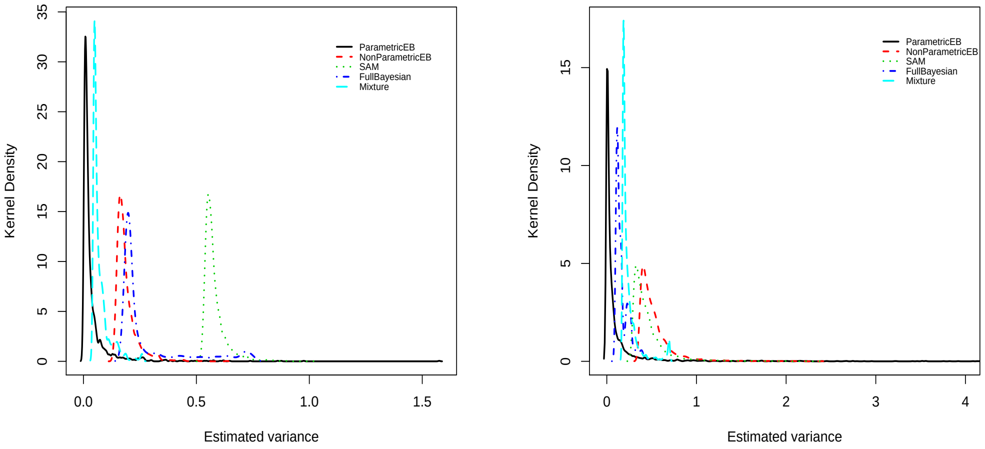

Figure 2.7: Kernel density for estimated variability (see color insert following page 99).

Figure 2.7 shows the kernel density of variability for the 5 approaches. In terms of relative comparisons, let EB be the reference approach as the method with the density distribution of variability closest to 0: it has a narrow distribution, but with a long right tail. The mixture model is the method with the highest and narrowest distribution, but it is shifted to be centered around 0.25. For the TCDD dataset there is a strong distinction between the other three distribution: Non Parametric Empirical Bayesian and SAM show a similarity, but are centered on different values, while Full Bayesian approach has a larger distribution and a longer tail. For E-Coli dataset the densities for the other three approaches are more confused: the full Bayesian model presents a lower variability and a narrower distribution than SAM and non parametric empirical Bayesian one.

Differential expression comparison The way to estimate variability carried out above influences the number of differentially expressed genes.

	SAM	NonParEB	Mixture	FullBay	ParEB
SAM	264				
NonParEB	79	79			
Mixture	72	56	72		
FullBay	62	55	63	63	
ParEB	43	33	39	38	43

Table 2.3: Table of differentially expressed genes for TCDD dataset.

Table 2.3 shows the number of genes differentially expressed with the 5 different methodologies for TCDD dataset. SAM analysis returns the largest number of differentially expressed genes (264), including the results of all the other analysis. Also there is the complete overlapping between mixture and Empirical Bayesian approach, that is the approach finding the minimum number of genes as differentially expressed. For the other pair of methods a large but not complete overlapping is observed.

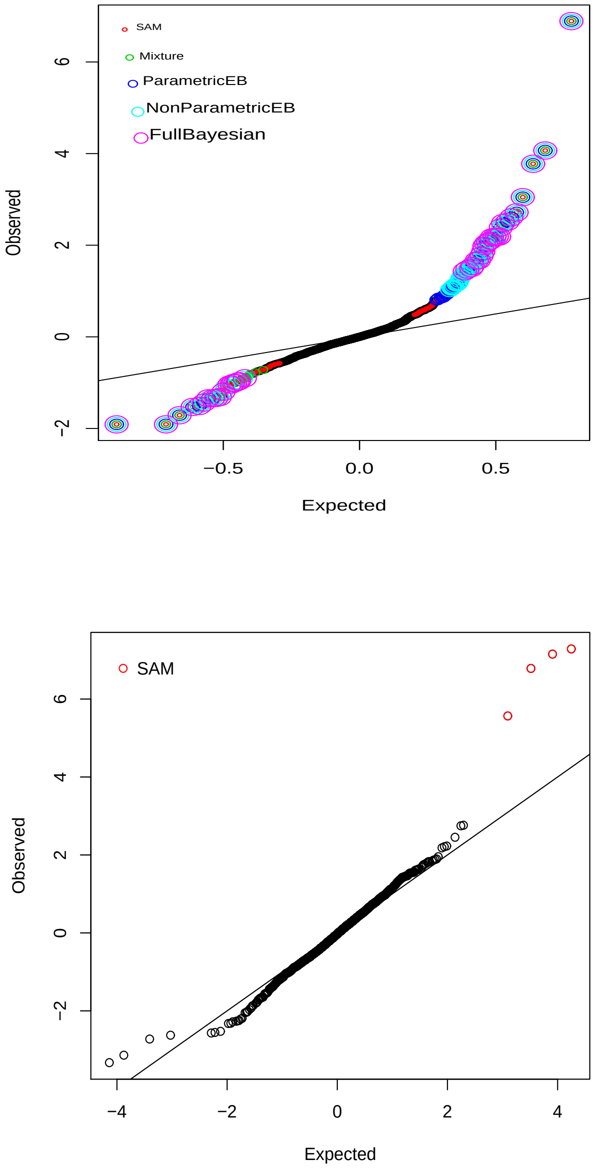

Figure 2.8: Differentially expressed genes by SAM plot (see color insert following page 99).

In figure 2.8 the observed relative difference (t_g) are plotted versus expected relative difference under H_0 hypothesis (mean of t_g calculated on permutated data). The solid line is the $45°$ line (observed equal to expected). Genes highlighted with different color and size are the differentially expressed by the 5 approaches. It better characterises the overlapping within the different methodologies for TCDD dataset (top): moving to the extreme values, the agreement within the methodologies increases. If the coordinates of a gene lie on the bisector, the observed value is equal to the expected one. In fact, in E-Coli plot (bottom) all the points are spread on the bisector and only 4 are identified as significant by SAM analysis.

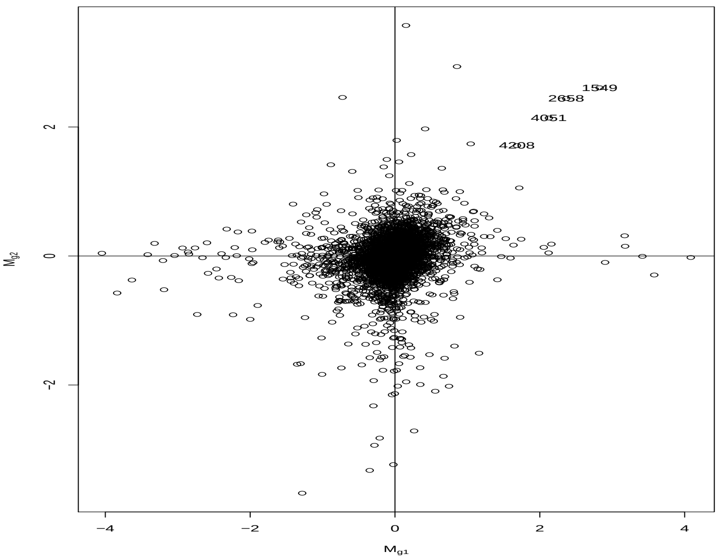

Figure 2.9: m_{1g} vs m_{2g} plot for normalised data.

Figure 2.9 presents the plot of normalised logratio for the two array of E-Coli. The four genes with highlighted ID are founded differentially expressed by SAM method. They lie on the first quadrant of the plot, characterised by positive logratio for both the arrays. However, the q-values associated with the four genes is quite high (0.36), indicating a not negligible probability that the gene is a false positive.

2.6 Discussion

For a comparative purpose we used calibration experiment as an alternative to simulation study: like simulation, it permits to validate a methodology, but has the advantage of being a real experiment. As such, it presents all the problem linked to the RNA extraction, array fabrication, scanning and fluorescence calculation. Due to the absence of a hypothetical differentially effect of treatments, the genes individuated as differentially expressed are false positive for sure, while false negative are not present at all.

From E-Coli experiment it seems that all the methodologies are able to identify the absolute absence of differentially expressed genes. The only exception is SAM analysis, that identifies 4 genes as differentially expressed. Actually, figure 2.9 shows that for the 4 genes found significant with SAM the logratio of the two arrays are very far from 0 (where they should lie for the absence of whatever difference between the two channels). However, the q-values determined by SAM, inform the researcher that when the 4 genes are significant, the minimum probability to find a false positive is 0.36. The q-value is defined as the probability that the null hypothesis is true given a statistic as extreme or more extreme than the observed. It is the minimum positive false discovery rate for calling a gene significant. As a reference, let us consider the q-values (table 2.4) for differentially expressed genes of TCDD experiment (real comparative experiment): they range between 0.01508 and 0.06893, showing a high strength of being differentially expressed. In comparison, for the E-Coli dataset, a minimum expected number of 36 genes false positives between 100 individuates as differentially expressed, is an index of low reliability.

qvalue	SAM	EB	NonParEB	FullBay	Mixture
0.01508	21	21	21	21	21
0.01552	2	2	2	1	2
0.01562	5	4	5	2	4
0.01626	6	6	6	4	6
0.01648	2	1	2	1	1
0.01659	4	2	4	2	2
0.01704	1	1	1	0	1
0.01739	9	7	6	4	7
0.01755	1	1	1	0	1
0.01875	1	1	1	0	1
0.01901	3	3	2	1	3
0.01957	1	1	1	0	1
0.02018	13	8	11	0	8
0.02046	2	1	0	2	1
0.02078	4	2	2	1	2
0.02084	1	0	0	1	0
0.02096	1	0	1	0	0
0.02185	1	0	1	0	0
0.02189	1	1	0	1	1
0.02250	1	0	0	1	1
0.02414	1	0	0	0	0
0.02458	1	0	0	0	0
0.02568	1	0	1	0	0
0.02627	1	0	1	0	0
0.02636	2	0	2	0	0
0.02685	2	0	0	0	0
0.02692	2	1	0	0	1
0.02705	2	0	0	1	1
0.02708	1	0	1	0	0
0.02727	1	0	1	0	0
0.02799	1	0	0	0	1
0.02980	1	0	1	0	0
0.03042	1	0	0	0	0
0.03068	1	0	1	0	0
0.03077	2	0	0	0	1
0.03101	1	0	0	0	0
0.03116	5	0	1	0	0
0.03128	2	0	1	0	1
0.03150	1	0	0	0	1
0.03353	9	0	2	0	1
0.03377	2	0	0	0	0
0.03573	1	0	0	0	0
0.03663	1	0	0	0	0
0.03682	1	0	0	0	1
0.03728	2	0	0	0	0
0.03741	2	0	0	0	0
0.03829	1	0	0	0	0
0.03850	1	0	0	0	0
0.03879	5	0	0	0	0
0.04047	4	0	0	0	0
0.04087	1	0	0	0	1
0.04136-0.06893	125	0	0	0	0

Table 2.4: Comparison of differentially expressed genes: this table presents the ranked q-values for the 264 genes emerged by SAM analysis.

In table 2.4 the frequency distribution of differentially expressed genes on the basis of SAM q-values is performed for the 5 methodologies. The other columns report the frequencies for the other methods. We remark that differentially expressed genes with q-values from 0.01508 to 0.01659 are indicated as differentially expressed by all the methodologies; then the different power to find differentially expressed gene emerges, up to the 125 genes with q-value from 0.04136 to 0.06893, called differentially expressed only by SAM.

The issue of heteroscedasticity is tackled in terms of global correction of gene specific variance (SAM and Non Parametric Empirical Bayes) on one hand, and in terms of mixture model on gene specific variability distribution on the other hand. In between, the Empirical Bayesian approach is closer to the first, considering an hyperparameter of variance distribution as the global variability component; more close to mixture model, the full Bayesian approach, introduces the idea of gene grouping through the sliding window on which the correction term for gene specific variance is calculated.

The other issue of the relation between M and A is taken into account properly only by EB and full Bayesian approach, while the other approaches consider it through the normalisation: the first are the most conservative methods, but eliminate an important problem, that can originate false positive (negative).

In conclusion the 5 approaches treated are comparable methodologies of different complexity level that rise from the Student *t* and propose some modifications lying under very different point of view, but taking into account the problem of heterogeneity. They present a different power of identification of differentially expressed genes, but all seem to give a good capability to recognise false positive.

2.7 R Library

This is a note to introduce *R/compvar*, a library written in R. It analyses and compares the different methodologies described in section 2.3 and provides the output presented and discussed in the previous sections.

2.7.1 Introduction

R/compvar is an add-on package for the freely available statistical language R (`www.r-project.org`) to analyse microarray data. It can be obtained on request (`blangiar@ds.unifi.it`).
The name *compvar* stands for comparison of variability. It performs a complete analysis of microarray including:

- Creation of data matrix from raw log intensity and normalise data
- Visualisation of the data focussing the attention on heteroscedasticity and on relation between M and A
- The following analysis
    1. SAM (Tusher *et al.*, 2001)
    2. Non Parametric Empirical Bayesian (Efron *et al.*, 2001)
    3. Parametric Empirical Bayesian (Lönnstedt and Speed, 2002)
    4. Full Bayesian (Baldi and Long, 2001)
    5. Mixture model (Delmar *et al.*, 2004)
- Comparison of different methodologies in terms of variability and differential expression

The dataset used for example are the two described in section 2.4.

2.7.2 Functions

We present the single functions that form the R/compvar library following the order in which they should be called. We better describe each of them, focusing on input and output variables and giving some additional notes when needed.

`CreateData`

This function takes the raw intensities and returns log_2 intensities, raw and normalised logratio to be used to perform the analysis.

Input

Data Matrix of raw intensity (*n genes, 2x n array*)

ID Array of ID

norm Normalisation method to be performed

The normalisation is performed array by array. The type of normalisation allowed are: "n" for no normalisation; "m" for median normalisation "l" for global loess normalisation "p" for print tip normalisation "s" for print tip resized normalisation.

Output
A list object that contains:

log.intensity A matrix of log intensities (*n genes, 2x n array*)

raw.logratio A matrix of raw logratio (*n genes, n array*)

norm A matrix of normalised logratio (*n genes, n array*)

`preprocessing`

This function uses the raw and normalised logratio and returns plots of kernel densities, M vs A and variance of raw logratio vs variance of normalised logratio. It is very useful for understanding and evaluating the relation between intensity and variability as well as the heteroscedasticity.

Input

log.intensity A matrix of log intensities (*n genes, 2 x n array*)

raw.logratio A matrix of raw logratio (*n genes, n array*)

norm A matrix of normalised logratio (*n genes, n array*)

directory.out The directory to store the output

Output
Three plots:

MvsAAllData.ps a M vs A plot that explain the relation between these two quantities.

DensityAllData.ps a Kernel Density plot for variance of raw logratio and for variance of normalised logratio (through CreateData function)

rawVSnorm.ps a plot to compare the two variances (for raw and normalised logratio) `sam.analysis`

This function performs the SAM analysis (Tusher *et al.*, 2001). It builds a modified t statistic, adding a constant (fudge factor) to the denominator to stable the variance. A plot shows the observed t statistic vs the expected values under the null hypothesis (calculated through permutations)

Input

ID Array of ID for the genes

M Matrix of normalised logratio (*n genes*, *n array*)

cl The number of classes for the analysis (by default it is one class-paired analysis)

B Number of permutation to perform

delta A gene will be called differentially expressed, if its posterior probability of being differentially expressed is large than or equal to delta.

alpha.s0 The possible values of the fudge factor s0 in terms of quantiles of the standard deviations of the genes.

thres.fdr For each value contained in thres.fdr, two lines parallel to the 45-degree line are generated in the SAM plot.

lambda.p0 Number between 0 and 1 that is used to estimate p0. If set to 1 (default), the automatic p0 selection using the natural cubic spline fit is used.

vec.lambda.p0 Vector of values for λ used in the automatical computation of p0.

na.rm If TRUE the missing values are not considered; if FALSE they are replaced by the genewise mean.

R.fold If TRUE (default), the fold change for each differentially expressed gene will be computed.

R.unlog If TRUE, 2data will be used in the computation of the R.fold. This is recommend if data contains the log2 transformed gene expression levels.

directory.out The directory to store the output

The relation between M and A is taken into account only through normalisation. The global factor to correct gene specific variance is calculated as that minimize the variation coefficient of *t* statistic distribution.

Output
Two files and the SAM plot. The files contain:

SAMvariance values of estimated variance. The variance is the sum of two component, the gene specific one and the global one.

samOutput information about all the genes.

samDiff information about the differentially expressed genes.

The SAM plot draws the observed relative difference versus the expected relative difference): the solid line in green is the 45° line (when observed is equal to expected).

`NonParEB`

This function performs the non parametric Empirical Bayesian analysis (Efron et al., 2002). builds a modified *t* statistic, adding a constant (fudge factor) to the denominator to stable the variance. To find differentially expressed genes it considers the probability that a gene is affected by the considered state (treatment) and the complementary probability that the same gene is unaffected. The densities for the two probabilities are empirically obtained and through the Bayes rule the posterior probability to be expressed or not is calculated.

Input

ID Array of ID for the genes

M Matrix of normalised logratio (*n genes*, *n array*)

cl The number of classes for the analysis (by default it is one class-paired analysis)

B Number of permutation to perform

delta A gene will be called differentially expressed, if its posterior probability of being differentially expressed is large than or equal to delta.

na.rm If TRUE the missing values are not considered; if FALSE they are replaced by the genewise mean.

R.fold If TRUE (default), the fold change for each differentially expressed gene will be computed.

R.unlog If TRUE, 2data will be used in the computation of the R.fold. This is recommend if data contains the log2 transformed gene expression levels.

directory.out The directory to store the output

The relation between M and A is taken into account only through normalisation. The global factor to correct gene specific variance is calculated as the 90^{th} percentile of the variance distribution.

Output
Two files and the volcano plot. The files contain:

NonParEBvariance values of gene specific variance. The variance is the sum of two component, the gene specific one and the global one.

NonParEBOutput information about all the genes.

NonParEBDiff information about the differentially expressed genes.

The volcano plot highlights the differentially expressed genes.

EB

This function performs the parametric Empirical Bayesian analysis (Lönnstedt and Speed, 2002). It takes the normalised logratio and observed variability and through a Bayesian model returns an odds ratio for each gene: log(Pr(the gene is differentially expressed) / Pr(the gene is not differentially expressed)).

Input

ID Array of ID for the genes

M Matrix of normalised logratio (*n genes*, *n array*)

p The threshold of probability to be not differentially expressed allowed for a gene to be called differentially expressed

directory.out The directory to store the output

This method takes into account the relation between mean (A) and variance (M), modelling at the same time, mean and variance. The authors calculate a B statistic, as log(Pr(the gene is differentially expressed) / Pr(the gene is not differentially expressed)). A closed form is found for the B statistic, that has a gene specific component and a global one.

Output
Four files and the volcano plot. The files contain:

EBvariance values of gene specific variance. The variance has invgamma distribution: $\sigma^2 \sim \Gamma^{-1}(\nu, 1)$.

EBOutput Information about all the genes (M,posterior variance, posterior log odds)

EBDiff Information about differentially expressed genes (M,posterior variance, posterior log odds)

EBVolcano values to perform the volcano plot (x is the B statistic and y the posterior log odds)

The volcano plot highlights the differentially expressed genes.

`fullbay`

This function performs the full Bayesian approach (Baldi and Long, 2001). It considers a gene specific variance and a correction calculated over a sliding window of neighbors genes in terms of expression.

Input

data Matrix of raw intensity values (*n genes, 2 x n array*)

ngenes Number of genes (number of rows of M, raw.logratio, log.intensity and data matrix)

narray Number of arrays

cs The first column contain data

ce The last column contain data (for paired analysis equal to the number of arrays)

experror The error allowed for identifying differentially expressed genes

winsize The size of the sliding window to calculate the correction factor for gene specific variance

conf The confidence on prior estimate

minrep The minimum number of non missing replicates to be able to calculate the variability estimates

directory.out The directory to store the output

The method considers the Bonferroni correction. It takes into account the relation between mean (A) and variance (M).

Output
Four files that contain:

FullBayOutput Information about all the genes, including values of *t* test and p-values.

FullBayDiff information about significant genes including values of *t* test and p-values.

FullBayVariance values (as sum of gene specific component and global component calculated through the sliding window)

FullBayVolcano values to build volcano (plot of *t* statistic (x) versus posterior log odds (y), to evaluate the differentially expressed genes) `mixture`

This is the function to perform the mixture analysis (Delmar et al., 2004) in a multi slide microarray experiment. It identifies cluster of genes with equal variance considering a mixture models on variance distribution and assign each gene to specific groups according to the largest posterior probability to belong to.

Input

ID Array of ID for the genes

M Matrix of normalised logratio (*n genes*, *n array*)

log.intensity The matrix of log intensity (as output from CreateData) (*n genes*, *2 x n array*)

pval The p value to call a gene differentially expressed

directory.out The directory to store the output

The measure of variability at the denominator of *t* statistic is gene specific in the sense that is a sum of the group variances weighted by the estimated posterior probabilities of a gene to belong to the j^{th} group.

Output
Four files that contain:

MixtureOutput Information about all the genes, including values of *t* test and p-values.

MixtureDiff information about differentially expressed genes including values of *t* test and p-values.

MixtureVariance gene specific variance estimates as a sum of the group variances weighted by the estimated posterior probabilities of a gene to belong to the j^{th} group.

MixtureVolcano values to build volcano (plot of *t* statistic (x) versus posterior log odds (y), to evaluate the differentially expressed genes) and three plots:

- Plot of log-variance versus mean intensity (the genes belonging to different mixture components are highlighted with different colors)
- Plot of mean logratio versus mean log-intensity
- Plot *t* statistic versus denominator

`comp.var`

This function compare the different methodologies (SAM, EB, NonParEB, FullBay, Mixture) in terms of variability estimates.

Input

ngenes Number of genes (number of rows of M, raw.logratio, log.intensity and data matrix)

SAM If different from NULL means that SAM analysis was performed. It should be the variance array output from `sam.analysis`

EB If different from NULL means that EB analysis was performed. It should be the variance array output from `EB`

NonParEB If different from NULL means that NonParEB analysis was performed. It should be the variance array output from `NonParEB`

FullBay If different from NULL means that FullBay analysis was performed. It should be the variance array output from `fullBay`

Mixture If different from NULL means that Mixture analysis was performed. It should be the variance array output from `mixture`

directory.out The directory to store the output

At least two approaches are needed to compare in terms of variability.

Output
Two plots:

- Kernel density for variability distribution
- Boxplot of variances

`comp.volc`

This function plots the volcano to compare the different methodologies (EB, FullBay, Mixture) in terms of differential expression.

Input

ngenes Number of genes (number of rows of M, raw.logratio, log.intensity and data matrix)

EB If different from NULL means that EB analysis was performed. It should be the variance array output from `EB`

FullBay If different from NULL means that FullBay analysis was performed. It should be the variance array output from `fullBay`

Mixture If different from NULL means that Mixture analysis was performed. It should be the variance array output from `mixture`

directory.out The directory to store the output

You need to perform at least two methodologies to compare in terms of differential expression. SAM and NonParEB cannot be used to perform volcano.

Output
A matrix of volcano plots (depending on the number of methods performed), where the cut off identifies the threshold for differential expression and the highlighted genes are those called differentially expressed.

`comp.diff`

This function draws the typical SAM plot to compare the different methodologies (SAM, EB, NonParEB, FullBay, Mixture) in terms of differential expression.

Input

ID Array of ID

M Matrix of normalised logratio (*n genes*, *n array*)

narray Number of array

ngenes Number of genes (number of rows of M, raw.logratio, log.intensity and data matrix)

SAM Must be different from NULL to draw the plot. It should be the variance array output from `sam.analysis`

EB If different from NULL means that EB analysis was performed. It should be the variance array output from `EB`

NonParEB If different from NULL means that NonParEB analysis was performed. It should be the variance array output from `NonParEB`

FullBay If different from NULL means that FullBay analysis was performed. It should be the variance array output from `fullBay`

Mixture If different from NULL means that Mixture analysis was performed. It should be the variance array output from `mixture`

directory.out The directory to store the output

You need to perform the SAM analysis to do this plot. It uses SAM as common reference for the other approaches.

Output

The SAM plot: observed relative difference (t_g) are plotted versus expected relative difference under H_0 hypothesis (mean of t_g calculated on permutation data). The solid line is the 45° line (observed equal to expected). The genes found differentially expressed by one of the methodologies used are highlighted with different colors and size.

BIBLIOGRAPHY

Allison, D.B., Gadbury, G.L., Heo, M., Fernández, J.R.,Lee, C., Prolla, T.A. and Weindruch, R. (2002) A mixture model approach for the analysis of microarray gene expression data, *Computational Statistics and Data Analysis* , **39(1)**, 1-20.

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(6)**, 5009-5019.

Delmar, P., Robin, S. and Daudin, J.J. (2004) Efficient variance modelling for differential analysis of replicated gene expression data, *Bioinformatics*, to be published.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2003) Empirical Bayes Analysis of a Microarray Experiment., *JASA* **96**, 1151-1160.

Kerr, M.K., Martin, M., Churchill, G.A. (2000) Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**, 817-837.

Kikuchi, H., Hossain, A., Yoshida, H. and Kobayashi, S. (1998) Induction of cytochrome P-450 1A1 by omeprazole in human HepG2 cells is protein tyrosine kinase-dependent, and is not inhibited by naphthoflavone, *Archives of Biochemestry and Biophysics*, **358**,351-358.

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

Li, W., Harper, P.A, Tang, B.K., and Okey, A.B. (1998) Regulation of cytochrome P450 enzymes by aryl hydrocarbon receptor in human cells: CYP1A2 expression in the LS180 colon carcinoma cell line after treatment with 2,3,7,8-tetrachlorodibenzo-p-dioxin or 3-methylcholanthrene, *Biochemestry Pharmacology*, **56(5)**, 599-612.

McLachlan, G.J. and Basford, K.E. (1988) Mixture Models, *New York, Marcel Dekker, Inc.*

McLahan, G. and Peel, D. (2000) Finite Mixture Models, *New York, Wiley*.

Pan, W., Lin, J., and Le, C. (2003) A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data, *Functional Integrational Genomics*, **3(3)**, 117-124.

Parmigiani, G., Garrett, E.S, Anbazhagan, R. and Gabrielson, E. (2001) A model for Measurement Error for Gene Expression Data, *Journal of the Royal Statistical Society B*, **64(4)**, 717-736.

Rocke, D.M. and Durbin, B. (2001) A model for Measurement Error for Gene Expression Data, *Journal of Computational Biology*, **8(6)**, 557-569.

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

Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalisation, models of variations and assessment of gene effects, *Nucleic Acids Research*, **29**, 2549-2557.

Tusher, V., Tibshirani, R. and Chu,G. (2001) Significance analysis of microarray applied to the ionizing radiation response, *PNAS*, **98(9)**, 5116-5121.

3 Using a calibration experiment to assess gene specific information: Fully Bayesian and Empirical Bayesian models for microarray data[*]

3.1 Abstract

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 (Tseng *et al.*, 2001), in which the probes hybridised on the two channels come from the same population (self-self experiment). From such an experiment, it is possible to estimate the gene-specific variance, to be incorporated in comparative experiments on the same tissue, cellular line or species.

We propose two different approaches to introduce prior information on gene-specific variability from a calibration experiment. The first is an empirical Bayes model derived from Tseng *et al.* while the second one is an original full Bayesian hierarchical model. We apply the methods in the analysis of human lipopolysaccharide stimulated leukocyte experiments comparing the results and investigating the differences. The approaches are implemented in WinBUGS (Spiegelhalter *et al.*, 2003).

[*] An extract from this chapter with the same title was submitted to *Bionformatics*, in collaboration with Simona Toti, Annibale Biggeri and Betti Giusti.

3.2 Introduction

In the framework of microarray analysis there are two main streams: 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 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 *et al.*, 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, p. 183 and following). 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 stabilise the variability measure (Speed *et al.* (2003), p. 51 and following). 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, 2001); otherwise it causes 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 stabilising 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 empir-

ical variance with a prior variance σ^2_{g0}. 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, the variance has a gene specific component s^2_g 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 minimise the coefficient of variation. This method is framed in a frequentist approach, does not assuming any distribution on the parameters.
Very similar to the previous, but in a non parametric context, 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.* (Tseng *et al.*, 2001) who propose calibration experiments in which the probes hybridised 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 the following: in section 3.3 we describe the calibration and comparative experiments (3.3.1) and the data preprocessing phase (3.3.2); in section 3.4 we present the normalisation 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 3.5 we describe the results in terms of differentially expressed genes; in section 3.6 a sensitivity analysis is reported and in section 3.7 we discuss the differences between the two models.

3.3 Materials

3.3.1 LPS microarray experiment

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 hours 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 6 aliquots and then retro-transcribed with amino-allyl-dUTP, hydrolyzed, purified and labelled with NHS-Cyanine dyes (3 aliquots with Cy3, probe A and 3 aliquots with Cy5, probe B). Then, three arrays were produced having the two probes purified, mixed and hybridised on the arrays. After incubation, the three array were scanned by the 4000B scanner (Axon). Image analysis was performed by GenePix 4.1 software.

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 hours in presence of lipopolysaccharide (LPS, 10μ g/ml, stimulated cells). The second was incubated in the same conditions but in 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, hydrolyzed, 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 hybridised 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, 2 arrays finally were printed according to the dye-swap design.

Therefore, the complete experiment consists in 5 arrays made up 22x21 spots grid, for a total of 14784 spots. The 14784 spots included 13971 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). 1502 (10.2%) out of 14784 spots were absent because of a failure during the printing procedure.

3.3.2 *Microarray data preprocessing*

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 apply the quality control present in GenePix Pro 4.1, with the aim of evaluating the presence of artifacts (bubbles, hair, fibers).
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 $83\%, 87\%$, for the 2 arrays of the comparative experiment.

Spots selection for the analysis of gene specific variances

To the purpose of the present paper, we restrict our attention to a subset of genes for which extraneous sources of variability can be excluded. To select these spots all the 5 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 is less than 25 for the foreground intensity in either channel, if the signal is lower than 200 for both the channels or if the ratio between the average foreground intensity and the median background intensity is smaller than

1.5 in either channel. Spots with a large signal for one channel and low, undetectable signal for the other are not eliminated, but modified to become analysable, forcing the low intensity signal (defined as less than 200) to 200.
In this paper we consider 2887 genes represented in all the 5 arrays (3 calibration arrays and 2 comparative arrays).

3.4 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.

3.4.1 Normalisation

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

3.4.2 Models

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 (Lewin *et al.*, 2003) for the unnormalised log-intensity

$$x_{igc} \quad \sim \quad N\left(\mu_{igc}, {}_{x}\sigma_g\right) \tag{3.1}$$

where i denotes array ($i = 1, 2, 3$), g denotes gene $g = 1, ..., 2887$ 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 normalisation procedure was achieved by an ANOVA model (Kerr *et al.*, 2002)

$$\mu_{igc} \;=\; \alpha_{ig} + \delta_c + \gamma_g \tag{3.2}$$

where α_{ig} denotes the gene-specific array-gene interactions, δ_c the dye-effects and γ_g the normalised 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 Normalisation parameters were fixed effects modelled with non informative Gaussian hyperpriors. The gene-specific variances were assumed to follow a Lognormal distribution ${}_x\sigma_g \sim logN(\mu_\sigma, \sigma_\sigma)$ with $\mu_\sigma \sim N(0, 10000)$ and $1/\sigma_\sigma \sim Ga(0.001, 0.001)$ noninformative hyperpriors.

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 unnormalised log-intensity

$$x_{igc} \;\sim\; N\left(\mu_{igc}, {}_x\sigma_g\right) \tag{3.3}$$

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

$$\begin{aligned} E\left[\mu_\sigma \mid x^{\text{self}}\right] &= \frac{\int \mu_\sigma f(x^{\text{self}} \mid \mu_\sigma)\, \pi(\mu_\sigma)\, d\mu_\sigma}{\int f(x^{\text{self}} \mid \mu_\sigma)\, \pi(\mu_\sigma)\, d\mu_\sigma} \\ &= \frac{\int \mu_\sigma \int f(x^{\text{self}} \mid \mu_\sigma, \sigma_\sigma)\, \pi(\mu_\sigma, \sigma_\sigma)\, d\sigma_\sigma d\mu_\sigma}{const(x^{\text{self}})} \end{aligned} \tag{3.4}$$

where x^{self} are the self self expression data and $const(x^{\text{self}})$ is a normalising constant depending only on data. Analogously, for σ_σ we plug in

the posterior mean of the corresponding posterior distribution $f(\sigma_\sigma \mid x^{\text{self}})$.
A linear model was assumed for μ_{igc}:

$$\mu_{igc} \;=\; \alpha_{ig} + \tau_g + \delta_c + \gamma_g \tag{3.5}$$

Here the model terms τ_g can be interpreted as a normalised 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 follow:

$$\tau_g \;\sim\; N(\mu_{\tau_g}, \sigma_{\tau_g}) \tag{3.6}$$

$$\mu_{\tau_g} \sim N(\mu_\tau, \sigma_\tau) \quad , \quad 1/\sigma_{\tau_g} \sim Ga(\nu_\tau, \beta_\tau) \tag{3.7}$$

with informative hyperparameters $\mu_\tau, \sigma_\tau, \nu_\tau, \beta_\tau$.

Informative prior on log-ratio
Actually values for $\mu_\tau, \sigma_\tau, \nu_\tau, \beta_\tau$ were obtained from the calibration experiment as follow. On the calibration arrays we calculated a residual effect $r_{igc} = x_{igc} - \mu_{igc}$ and reconstructed a "normalised 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:

$$t_{ig} = r_{ig1} - r_{ig2} \tag{3.8}$$

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

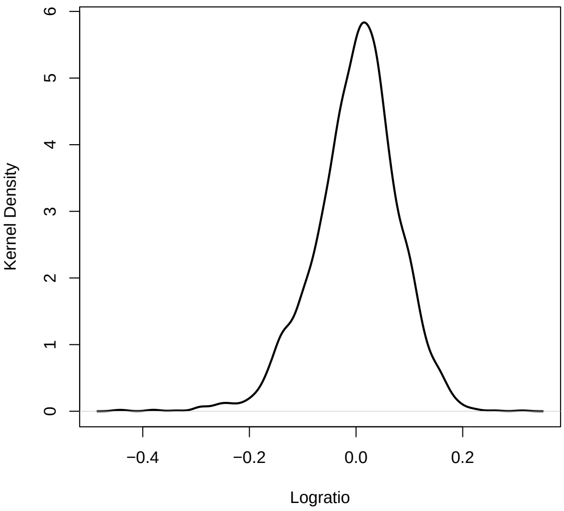

Figure 3.1: Kernel density plot of normalised log-ratios $t_{\cdot g}$ for self self experiment.

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

$$\widehat{\mu}_\tau = \frac{1}{\mathrm{G}} \sum_g t_{\cdot g} \tag{3.9}$$

$$\widehat{\sigma}_\tau = \frac{1}{\text{G-1}} \sum_g \left(t_{\cdot g} - \widehat{\mu}_\tau\right)^2 \tag{3.10}$$

where $t_{\cdot g} = \frac{1}{3}\sum_i t_{ig}$ (see figure 3.1).
Similarly, we obtained the plug-in values for the prior Gamma parameters ν_τ and β_τ from the mean and variance of $\{\widehat{\sigma}_{\tau_g}\} = \left(\frac{1}{2}\sum_i \left(t_{ig} - t_{\cdot g}\right)^2\right)$:

$$\widehat{\nu}_\tau = Ave(\widehat{\sigma}_{\tau_g}) \cdot \widehat{\beta}_\tau \tag{3.11}$$

$$\widehat{\beta}_\tau = \frac{Ave(\widehat{\sigma}_{\tau_g})}{Var(\widehat{\sigma}_{\tau_g})} \tag{3.12}$$

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

Tseng's Empirical Bayes model

To adapt the model proposed by Tseng *et al.* (2001), first we normalised 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 normalised log-ratio m_{ig} for g^{th} gene and i^{th} array were modelled as:

$$m_{ig} \sim N(\tau_{g},{}_{m}\sigma_{g}) \tag{3.13}$$

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 to the full Bayesian model the likelihood can be written as follow:

$$\begin{aligned} m_{ig} &= \textit{normalised}(x_{ig1} - x_{ig2}) && (3.14) \\ m_{ig} &\sim N(\mu_{ig},{}_{m}\sigma_{g}) && (3.15) \end{aligned}$$

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 to equation 3.6 and 3.7). The information pooled from the calibration experiment was used to obtain an informative prior distribution for ${}_m\sigma_g$:

$${}_{m}\sigma_{g} \sim \frac{w_g}{\frac{\chi^2_k}{k}} \tag{3.16}$$

where k was the number of degree of freedom of a chi squared deviate; w_g was a weighted average of gene-specific and overall empirical variance calculated on the calibration arrays $(i = 1, ..., I^{\text{self}})$:

$$\begin{aligned} \widehat{s}_g &= \frac{1}{\mathrm{I}^{\text{self}} - 1} \sum_{i=1}^{\mathrm{I}^{\text{self}}} (m^{\text{self}}_{gi} - \overline{m}^{\text{self}}_{g\cdot})^2 && (3.17) \\ \widehat{s}. &= \frac{1}{\mathrm{G}} \sum_{g=1}^{G} \widehat{s}_g && (3.18) \\ w_g &= \frac{[(\mathrm{I}^{\text{self}} - 1) \cdot \widehat{s}_g + \widehat{s}.]}{\mathrm{I}^{\text{self}}} && (3.19) \end{aligned}$$

In other words, in the Tseng model the information on the gene-specific variability from the self self experiment is utilised 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 ${}_m\sigma_g \sim \frac{w_g k}{\Gamma(\frac{1}{2},\frac{1}{2})}$; 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 $\sigma_{\tau_g} \sim \frac{1}{\Gamma(\nu_\tau,\beta_\tau)}$.

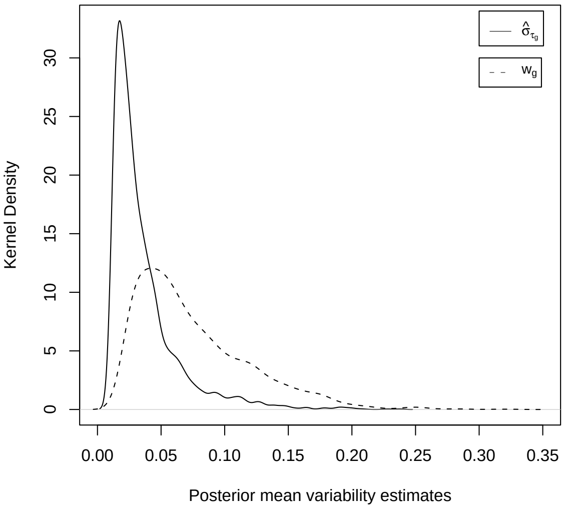

Figure 3.2: Kernel density plot of mean posterior estimates of variability $\hat{\sigma}_{\tau_g}$ from self-self experiment for full Bayesian model.

Tseng's prior with internal normalisation

To better address model comparison we modified the Empirical Bayes model proposed by Tseng including the normalisation step into the model:

$$\begin{aligned} x_{igc} &\sim N\left(\mu_{igc},{}_x\sigma_g\right) & (3.20) \\ \mu_{igc} &= \alpha_{ig} + \tau_g + \delta_c + \gamma_g & (3.21) \\ {}_x\sigma_g &\sim logN(\mu_\sigma, \sigma_\sigma) & (3.22) \end{aligned}$$

where the parameters of the lognormal distribution on ${}_x\sigma_g$ were informative coming from the calibration experiment (see paragraph 3.4.2), and the normalisation parameters were modelled following standard ANOVA (see equation 3.5). The hyperpriors for τ_g were modelled following Tseng's proposal $\left(\sigma_{\tau_g} \sim \frac{w_g k}{\Gamma(\frac{1}{2},\frac{1}{2})}\right)$ (see figure 3.2).

Bayesian hierarchical model with loess normalisation

We also modified the Bayesian hierarchical model to carry out a loess normalisation instead of the linear one. We performed a loess normalisation through MAANOVA library and then we calculated the normalised values for the two channels as follow:

$$ {}_nx_{ig1} = x_{ig1} - \frac{1}{2}l_{ig} \quad , \quad {}_nx_{ig2} = x_{ig2} + \frac{1}{2}l_{ig} \qquad (3.23)$$

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 normalisation. The normalised channel intensity (on log scale) are

$$ {}_nx_{igc} \sim N\left(\mu_{igc},{}_x\sigma_g\right) \qquad (3.24)$$

and we perform a further normalisation in calibration experiment

$$\mu_{igc} = \alpha_{ig} + \gamma_g \qquad (3.25)$$

as well as in comparative experiment

$$\mu_{igc} = \alpha_{ig} + \gamma_g + \tau_g \qquad (3.26)$$

for eliminating the array effects that are not considered in the loess normalisation performed separately for each slide. The model specification thereafter follow the structure defined in equations 3.3-3.12.

3.4.3 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 denote functional relationships; solid lines show stochastic undirected dependence. Repetitive structures (arrays, for example), are shown as stacked rectangles. Figure 3.3 shows the graph for the Bayesian hierarchical model presented in section 3.4.2 while figure 3.4 shows the DAG for Tseng's model presented in section 3.4.2.

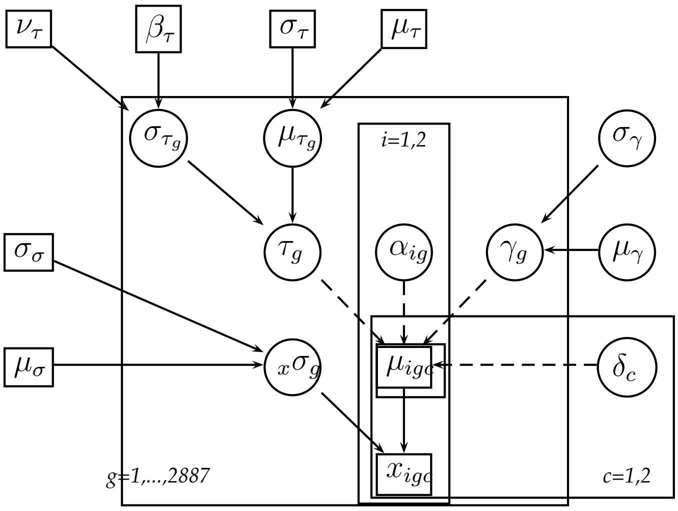

Figure 3.3: Graph of Hierarchical Bayesian model for treated samples (for untreated ones the τ_g effects are absent).

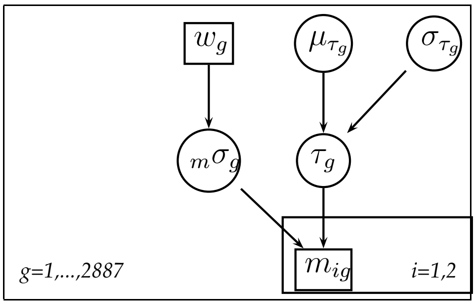

Figure 3.4: Graph of Tseng's et al. model for normalised log ratios m_{ig}.

3.4.4 Implementation

To estimate the parameters of interest we use the marginal posterior distributions approximated by MCMC methods implemented in WinBugs 1.4; the Bayesian hierarchical model with ANOVA normalisation as well as with loess normalisation, and Tseng's model with internal normalisation 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 Gelman-Rubin statistics (Gelman *et al.*, 1992) and using different starting points. We have performed 10000 burn in iterations followed by 4000 sampling iterations for all the models. Fitting the Bayesian hierarchical model on calibration experiment takes 1 hour to do 100 iterations on a workstation $HPXW6000$ with $2GbRAM$ and Intel Xeon $CPU2.8GHz$ 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 sec. for 1000 iterations. Fitting Tseng's model takes 300 seconds to perform 1000 iterations.

3.5 Results

We explored the posterior distribution of the treatment effects τ_g t 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 3.1. Using the Bayesian hierarchical model we found 26 differentially expressed genes. 2 (IFI30 and PRKAG2) out of 26 genes were underexpressed in LPS stimulated leukocytes. Using Tseng *et al.* one we found 46 differentially expressed genes. 20 out of 46 genes emerged downregulated in LPS stimulated leukocytes. 22 out of 26 genes identified by the first model were highlighted also by the Tseng *et al.* one (figure 3.5 and table 3.1).

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).

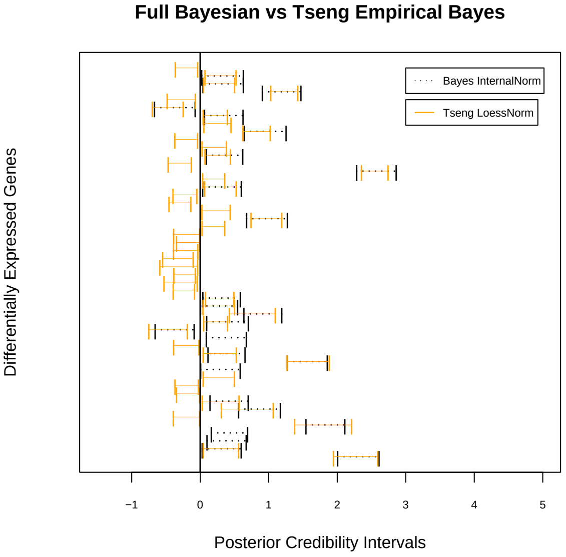

Figure 3.5: 95% posterior credibility intervals for differentially expressed genes: Full Bayesian model vs Empirical Bayesian one (see color insert following page 99).

		Bayesian Hierarchical Model		Empirical Bayesian Model	
ID	Symbol	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)

Table 3.1: Differentially expressed genes: posterior mean and posterior credibility interval at 95%.

3.6 Sensitivity Analysis and Model Comparison

The results presented in the previous section are difficult to interpret comparatively because the two models use different normalisation procedures. To gain insight on the behavior of the different approaches we need to evaluate differentially expressed genes taking fixed the normalisation procedure (subsection 3.4.2).

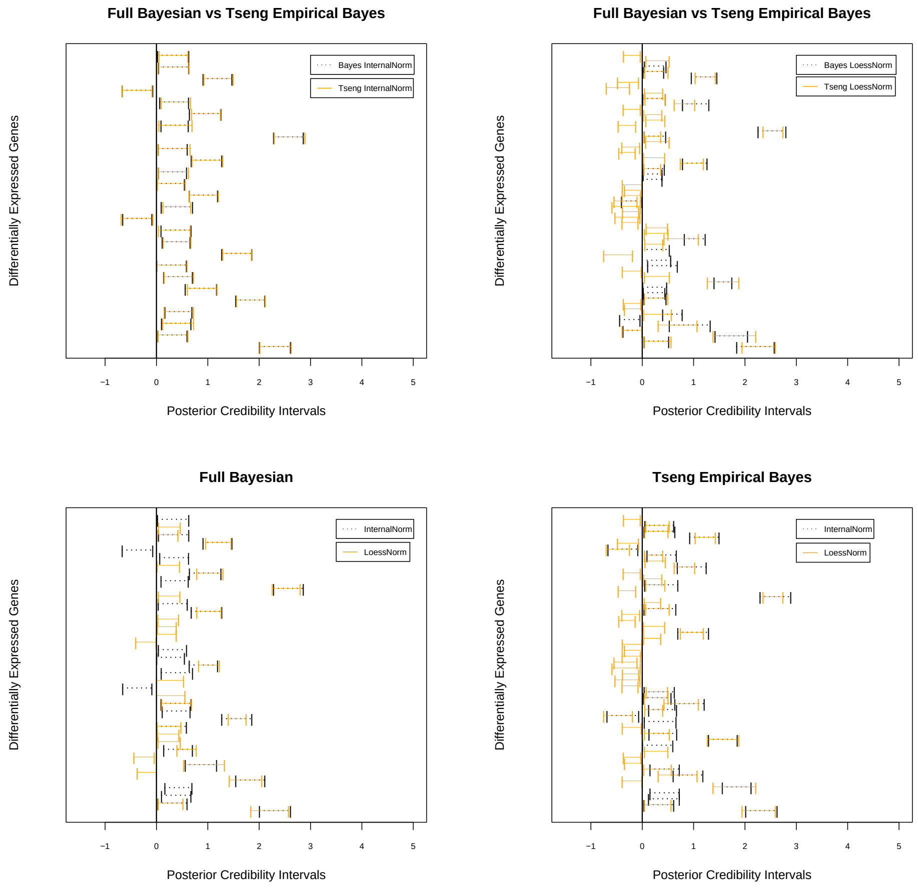

Figure 3.6: 95% posterior credibility intervals for differentially expressed genes (see color insert following page 99).

Figure 3.6 plots differentially expressed genes and their 95% posterior credibility intervals for: (up, left) the full Bayesian model and that with Tseng prior specification with internal linear ANOVA normalisation; (up, right) the Tseng model and the full Bayesian one with loess normalisation; (bottom, left) the full Bayesian model with both normalisation procedures; (bottom, right) the Tseng model with both normalisation procedures; . The largest differences were observed in the down regulated genes. The full Bayesian models found 2 negative genes and 3 negative genes. On the other side, by Tseng model 20 genes emerge as down regulated, but using the internal linear ANOVA normalisation 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 normalisation procedure. The Tseng model seems less conservative and more sensitive to the normalisation procedure adopted.

3.7 Discussion

The observed differences in number of differentially expressed genes between the Bayesian hierarchial model and the Tseng Empirical Bayes one are related to different factors, namely normalisation method and specification of prior information. In the Bayesian Hierarchical method, the Normalisation step is performed inside the model through a multi slide linear Normalisation (ANOVA). In the Empirical Bayesian approach, data are normalised outside the model, through a loess Normalisation performed separately for each array. When incorporating the normalisation into the model, the likelihood is based on single channel expression measures over replicates, while with an external normalisation, the likelihood is based on an empirical relative measure of 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 Normalisation step into the model (3.4.2) all the genes emerged down regulated in the previous analysis were no more differentially expressed.

Using the Bayesian Hierarchical modelling with loess normalisation (3.4.2) 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 down regulated.

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 normalisation procedures adopted.

The Empirical Bayesian model and the full Bayesian one insert prior information on variability from the calibration experiment in different way. In the first the prior distribution for the variance of the normalised 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 (3.16). 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 *et al.*, 1972) and has optimality proprieties in a frequentist point of view.

The full Bayesian hierarchical model inserts information from self self experiment at the normalised log-ratio level for each gene, as well as at the single channel intensity level (figure 3.3).

The gene specific log-ratio (τ_g) probability density has informative distribution on its parameters $\mu_{\tau_g}, \sigma_{\tau_g}$ (equation 3.7). 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 3.4). 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 3.2, in our data Bayesian posterior estimates of gene-specific variances tend to be larger than the empirical Bayes esti-

mates. The reader can also appreciate that the distribution of log-ratios (figure 3.1) 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: as an illustration the figure 3.7 compare the empirical log ratio to the posterior estimates.

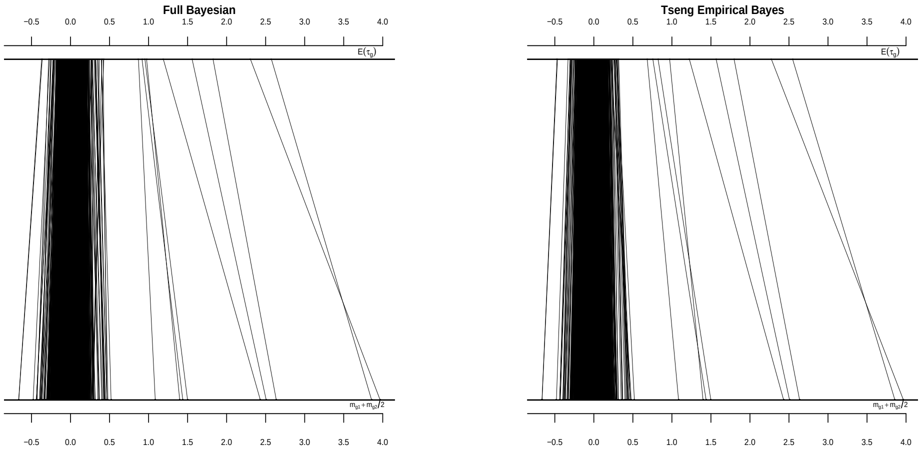

Figure 3.7: Shrinkage effect: both the Bayesian hierarchical model (up) and the Tseng's Empirical Bayes one (bottom) show a narrower range for the posterior mean of τ_g with respect to the empirical log ratios $\frac{m_{g1}+m_{g2}}{2}$; for the negative genes the shrinkage effect is stronger for the Bayesian model than for the Empirical Bayes one.

In conclusion, we showed how information from calibration experiments can be utilised to improve inference on differentially expressed genes in comparative experiments. We can point out that calibration experiment is a good answer to the problem of gene-specific variability estimate and let 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.

BIBLIOGRAPHY

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(6)**, 5009-5019.

Comander, J., Natarajan, S., Gimbrone, M. and García-Cardeña, G. (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. (2004) Efficient variance modelling for differential analysis of replicated gene expression data, *Bioinformatics*, to be published.

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(2)**, 335-347.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001) Empirical Bayes Analysis of a Microarray Experiment, *JASA* **96**, 1151-1160.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulations using multiple sequences, *Statistical Science*, **7**, 457-511.

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

Kerr, M.K., Afshari, C.A., Bennett, L., ,Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2002) Statistical analysis of gene expression microarray experiment with replication., *Statistica Sinica*, **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. (2003) Bayesian Modelling of Differential Gene Expression, *submitted*.

McLachlan, G.J. and Basford, K.E. (1988) Mixture Models. *New York, Marcel Dekker, Inc.*

McLahan, G. and Peel,D. (2000) Finite Mixture Models, *New York, Wiley*.

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

Rocke, D.M. and Durbin, B. (2001) A model for Measurement Error for Gene Expression Data, *Journal of Computational Biology*, **8(6)**, 557-569.

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

Speed, T. (2003) (eds) Statistical Analysis of Gene Expression Microarray Data, *Wiley*.

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

Tanner, M.A. (1996) Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions, *Springer*.

Tao, H., Bausch, C., Richmond, C., Blattner, F.R. and Conway, T. (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media, *Journal of Bacteriology*, **181**, 6425-6440.

Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalisation, models of variations and assessment of gene effects, *Nucleic Acids Research*, **29**, 2549-2557.

Tusher, V.G. and Tibshirani, R. and Chu, G. (2001) Significance analysis of microarray applied to the ionizing radiation response, *PNAS*, **98(9)**, 5116-5121.

Wit, E. and McClure, J. (2004) Statistics for microarrays, *Wiley*.

Wu, H., Kerr, M.K., Cui, X. and Churchill, G.A. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments In Parmigiani, G. *et al* (eds), *The analysis of gene expression data: methods and software*, Springer, New York, NY, pp. 313-341.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalisation for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, **30(4)**, e15.

4 Calibration experiment for the analysis of microarray data: an application on un-and-LPS stimulated human leukocyte model.*

4.1 Abstract

One of the difficulties when analysing expression measures obtained by cDNA/oligonucleotide arrays is how to model the variance function for the whole set of genes. Several studies had showed the poor accuracy of models assuming a global variability measure. Many approaches to modelling gene-specific variance have been proposed. We take advantage of calibration experiment to analyse microarray data of LPS- and un-stimulated human leukocytes. In such experiments aliquots of the same RNA sample were labelled with Cy3 and Cy5 fluorescent dyes and co-hybridised to the microarray. From these calibration experiments, conjointly with replicates, it is possible to estimate the gene-specific variance to be incorporated in comparative experiments on the same specimens. We used a Bayesian hierarchical model to identify differentially expressed genes, taking into account the variability at gene level through calibration experiments. Our data on LPS-inducible gene expression profile both identified novel genes (e.g. IFI30, MLSN1, CFL2, AXIN1) suggesting new targets of study in order to better understand the pathophysiology of sepsis and inflammatory disease and confirmed the involvement of many cytokines and chemokines (IL-1b, IL-1RA, MIP-1a, -1b, -2b, -3a).

*An extract from this chapter with the same title was submitted to *Journal of Thrombosis and Haemostasis* in collaboration with Betti Giusti, Simona Toti, Annibale Biggeri.

4.2 Introduction

DNA microarray analysis has become a wide used technology for the study of gene expression profile on a genomic scale (Schena *et al.*, 1995), (Schena *et al.*, 1996). Experiments using DNA microarray allow the monitoring of expression levels for thousands of genes simultaneously. The basic strategy for microarray study is to retrotranscribe to cDNAs and differentially label with two fluorescent molecules (Cy3 and Cy5) RNA from two sources, a control and an experimental sample. The labelled cDNAs are co-hybridised on the microarray slide and the microarray is scanned. By using dedicated softwares, the intensities of the emission peaks of both Cy3- and Cy5-labelled targets in each gene spot are quantified. Finally, the log2 of the ratio of the normalised intensities is calculated for each gene-expression value to indicate its relative expression in the test versus the control state (Quackenbush, 2001).

In addition to pre analytical factors (i.e. cell harvesting conditions, biological variation, mRNA quality), poor reproducibility in microarray data analysis has been attributed to variety of factors including printed spot quality, hybridisation and differential incorporation of fluorescent nucleotides (Brenner *et al.*, 2000), (Hedge *et al.*, 2000), (Goryachev *et al.*, 2001), (Wildsmith *et al.*, 2001), (Schroeder *et al.*, 2002). Researchers typically adjust or normalise the data to correct for two common experimental biases: 1) fluorescent background, or fluorescence that is not due to fluorescent cDNA hybridisation; and 2) fluorochrome-specific differences.

Typically, background bias is removed by subtracting fluorescence outside the spot from fluorescence inside the spot, assuming that background is homogeneous across the local area. To adjust for bias due to fluorochrome-specific differences, the simplest normalisation method is to set the mean total intensities of Cy3 and Cy5 signals of one microarray to be equivalent, assuming that the expression patterns of most genes do not change between cellular states (Quackenbush, 2001). Reverse labelling of the samples is also used to ensure that results are not biased as a result of gene-specific preferential incorporation of the dyes (Kerr 2001). A challenging technical consideration in microarray data analysis is the identification and singling out of statistical models to distinguish non biological from biological variability in the data. In microarray literature the first popular method of analysis was introduced by Newton in 2001. It analyses each single slide and identifies three cut off limits to define differentially expressed genes; however, this approach does not permit to analyse more than one slide at the same time. Several methods were proposed to analyse different arrays at the same time. In particular, some authors (Tusher *et al.* (2001), Lönnstedt and Speed (2002), Efron *et al.* (2001)) propose multi slides approaches, to evaluate the significance of gene effects taking into account gene specific variability through replicates. A further source of information on gene specific variability can be obtained from calibration experiment. In such experiments the probes hybridised on the two channels come from the same population (self-self experiment). Tseng *et al.* (2001) proposed the statistical model to use information from replicates and calibration experiment in an Empirical Bayes perspective. In the present paper we use a Fully Bayesian model (Blangiardo *et al.*, 2004) and compare the result to whom obtained by Tseng model. Data are obtained from un-stimulated and lipopolysaccharide LPS-stimulated human leukocytes. Although few comprehensive study of gene expression in LPS stimulate peripheral blood mononuclear cells has been reported, many LPS inducible genes have been investigated and information is available in the literature. Due to the great number of information available on this experimental model, we could better and easily validate results obtained by microarray data analysis.

4.3 Materials and Methods

Cell Preparations Peripheral blood mononuclear cells (PBMC), obtained from the EDTA anti-coagulated peripheral blood of 20 healthy volunteers, were prepared by density gradient centrifugation (Ficoll-Hypaque, Nycomed Pharma AS, Oslo, Norway), extensively washed and counted (Neri Serneri *et al.*, 1992). For the experiments of self-self hybridisation, after PBMC isolation, total RNA was extracted from the pellets by RNeasy Maxi Kit (QIAGEN GmbH, Germany) according to the manufacturer protocol. During isolation a DNase treatment was performed with the Qiagen RNase-free DNase set (Qiagen). Mononuclear cells

(107 cells/mL) were incubated in RPMI 1640 (Gibco, Grand Island, NY) containing 100 mg/mL gentamcin at 37°C in a humidified atmosphere with 5% CO2. Mononuclear cells unstimulated or stimulated with LPS (10 mg/mL final concentration, Sigma Aldrich Corporation, St. Louis, MO, USA) were incubated at 37°C for 3 hours and then washed twice with phosphate-buffered saline (PBS). Total RNA was then extracted from the pellets.

4.3.1 Microarray experiments

We used poly-L-lysine (Sigma) coated arrays of 14000 genes (70 mer oligonucleotides - Human Array-Ready Oligo set, Operon Technologies). Oligonucleotides were spotted with a 32 pins arrayer (GeneMachines OmniGrid) in 50-52% humidity at 21°C and cross-linked by UV irradiation at 65 mJ/cm2 (Stratalinker model 1800 UV Illuminator, Stratagene). Microarray were made up 22x21 spots grid, for a total of 14784 spots. The 14784 spots included 13971 oligonucleotides representing each one different gene, 29 negative controls (mixtures of oligonucleotide from other organisms), 2 positive controls (a mixture of all the human oligonucleotides) and 872 blanks (only printing solution). After substraction of absent spots due to failure during the printing procedure, 13282 (89.8%) out of 14784 spots were available. For self-self experiments, equal amounts of total RNA from mononuclear cells of different control subjects were pooled and subdivided in aliquots of 20 mg. For stimulated/unstimulated experiments equal amounts of total RNA from stimulated or un-stimulated cells were pooled and subdivided in aliquots of 20 mg. Total RNA was reverse transcribed with oligo (dT)12-18 (Gibco Brl), amino-allyl-dUTP (Sigma), and Superscript II Rnase H enzyme, then hydrolyzed, purified by using Microcon-30 column (Millipore) and labelled with NHS-Cyanine dyes (Cy3 and Cy5, Amersham Biosciences). For each experiment, the two probes (one labelled with Cy3 and one with Cy5) were purified by using QIAquick PCR purification Kit (Qiagen), mixed and hybridised on the array. Sixty microliters of hybridisation mixture was heated at 95°C for 2 min and applied under a 22x50 mm Lifterslip (Erie Scientific). Slides were sealed in CMT chambers (Corning). The setting and the subsequent hybridisation of microarray were performed according to

`http://cmgm.stanford.edu/pbrown` and
`http://www.microrrays.org/protocols.htlm`.
After an overnight incubation (15-18 h) at 65°C, arrays were washed and scanned by using a 4000B Scanner (Axon), at 10 mm spatial resolution and a 33% of laser power. Each hybridisation (3 self-self hybridisation and 2 unstimulated/LPS stimulated hybridisation) produces a pair of 16-bit images, which were processed using the GenePix Pro 4.1 software (Axon Instruments, Union City, CA). Spot intensity, referred to throughout the paper, is the median foreground (F) pixel intensity after the median background (B) pixel intensity subtraction for the Cy3 fluorescence (median FCy3-532 - median BCy3-532) and Cy5 fluorescence (median FCy5-635- median BCy5-635).

4.3.2 Quality control

Poorly spotted genes, expressing weak or distorted signals, were automatically discarded by GenePix Pro 4.1 software and manually by visual inspection. Besides applying the quality control present in GenePix Pro 4.1, with the aim of evaluating the spot quality and to eliminate the spots that present a low quality, we used several visual and analytical controls (Simon *et al.*, 2003). 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. Viceversa, spots with a large signal for one channel and low signal for the other are not eliminated, but modified to become analysable. In fact, on one hand is not recommended to penalized the information on the spot that gives an important contribution for the channel with a large signal; on the other hand comparing the log ratio from different arrays this spot could produce distort conclusions about the expression, based on a difference between the log ratio that depends almost wholly from the low signal channel. To avoid these two problems, for a spot with a "large signal" (defined over 500) and a "low signal" (defined lower than 200), we forced the low intensity signal to 200. After the identification of each feature within the microarray by using the specific GAL file and the application of the quality criteria of GenePix Pro 4.1 software and the visual inspection, the analysable spots resulted 80%, 87% and 90% as concerned the 3 calibration experiments, and 83%, 87%, for the 2 experiments of comparison of LPS stimulated and un-stimulated PBMC. After the other steps of quality control, the analysable features ranged between 35% and 65%. The 2887 genes analysed in the experiment were that not missing in all the 5 arrays.

4.3.3 Normalisation

Normalisation procedure is a very important step to eliminate the multiple sources of variations introduced during the microarray fabrication and hybridisation processes. We do not describe the different aspects of this procedure, but refer to the complete review on microarray normalisation methods of Yang *et al.* (2002). We performed two different types of normalisation: for each slide a local A-dependent normalisation (loess), considering all the genes present on the array, is used for Newton and Tseng model. For our Bayesian hierarchical model, the normalisation was part of the modelling.

4.4 Models

To identify the differentially expressed genes we applied three different methods. The first (Newton *et al.*, 2001) works only on comparative arrays and analyses them separately. We introduce it in the work as a standard analysis of comparative experiments. We considered a gene as differentially expressed if it emerges in both the comparative arrays (p=0.01 and p=0.1). The software is freely available in R (SMA library, `www.R-project.org`).
The second approach was introduced by Tseng *et al.* (2001) and treats for the first time the calibration experiment. It is an empirical Bayesian approach. Gene specific variance of comparative experiment is estimated by replicates of comparative arrays weighted by observed variances from calibration experiment. We assume a gene is differentially expressed if the posterior credibility interval (p=0.05) does not include 0. The software is freely available on request at
`http://www.pitt.edu/ctseng`.

The method we use was introduced in Blangiardo *et al.* (2004). It is a full Bayesian method, so all the parameters (included the gene specific variance) have a prior distribution with parameters estimated from calibration experiment. To define a gene as differentially expressed we consider the posterior credibility interval (p=0.05). The WinBugs code is freely available on request (`blangiar@ds.unifi.it`).

4.4.1 Real-Time Polymerase Chain Reaction (RT-PCR)

For RT-PCR, 2 μ g of the same total RNAs used for comparative experiments was reverse-transcribed using Moloney Murine Leukemia Virus (M-MLV) transcriptase (Gibco BRL) and random hexamer primers (Amersham). In order to quantify the transcribed PLAUR and GRM4 gene, we performed TaqMan RT-PCR (PE Applied Biosystems) on an ABI Prism 7700 instrument. VIC-labelled human GAPDH (assay-on-demand $\#4326317E$) and FAM-labelled human PLAUR (assay-on-demand $\#Hs00182181\_m1$) and GRM4 (assay-on-demand $\#Hs00168265\_m1$) TaqMan predeveloped assays (Applied Biosystems) were used. Expression of PLAUR and GRM4 genes was normalised to GAPDH and displayed as fold-change relative to the unstimulated sample used as the calibrator.

4.5 Results

4.5.1 Microarray

Genes found differentially expressed with at least one of the three methods are shown in Table 4.1: Newton model identified 7 differentially expressed genes, Full Bayesian hierarchical model found 26 differentially expressed genes, while Tseng one found 44 differentially expressed genes. All the 7 genes emerged from Newton analysis were identified also by the other two models (Table 4.1). 22 out of 26 genes identified by Full Bayesian model were characterized also by Tseng model (table 4.1). As concerned our model, expression levels of 2 (IFI30 and PRKAG2) of 26 genes were decreased in LPS stimulated leukocytes in comparison with those in un-stimulated PBMC (table 4.1). Conversely, 24 transcripts were over-expressed in LPS stimulated PBMC

(Table 4.1). As concerns Tseng model, expression levels of 19 of 44 genes were decreased and 25 of 44 were increased in LPS stimulated leukocytes (Table 4.1). The LPS induced transcripts identified by both the two models mainly consist of genes encoding proteins 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; cytoskeleton protein such as vimentin and cofillin 2; and plasminogen activator inhibitor type 2 (PAI-2) (table 4.1).

4.5.2 Validated expression of selected genes by RT-PCR

In order to validate the expression of genes identified by only one model, we arbitrarily selected 2 differentially expressed transcripts (PLAUR for our model and GRM4 for Tseng's model) and evaluated them in the same total RNAs used for comparative microarray experiments by RT-PCR. RT-PCR analysis confirmed that PLAUR mRNA was up-regulated by LPS stimulus [fold-change 4.3 (4.5-4.1)], whereas the down-regulated GRM4 mRNA according to Tseng's model, but not to the other models resulted equally expressed [fold-change 1.3 (1.2-1.4)] in LPS- and un-stimulated PBMC.

		Newton	Bayesian Hierarchical Model		Empirical Bayesian model	
ID	Symbol	M	Post Mean	CrI	Post Mean	CrI
2064	VIM	-	0.32	0.04,0.06	0.43	0.12,0.76
2563	TAC1	-	-	-	0.28	0.06,0.50
2890	PRKCG	-	0.42	0.15,0.70	0.48	0.09,0.87
12183	MAN2B2	-	-	-	-0.30	-0.54,-0.07
14623	IFI30	-	-0.40	-0.65,-0.08	-0.70	-1.10,-0.23
23672	LRP6	-	-	-	-0.40	-0.72,-0.05
24030	SLC31A2	-	-	-	-	-
42500	ARL5	-	-	-	0.36	0.06-0.65
43265	MLSN1	-	0.39	0.10,0.71	0.48	0.23,0.74
47584	KCNS3	-	-	-	-0.20	-0.48,0.00
68879	BPM4	-	-	-	-0.30	-0.54,-0.05
73817	SCYA3/MIP-1A	3.96	2.30	2.00,2.61	3.96	3.50,4.45
75948	SCYA20/MIP-3A	1.43	0.92	0.64,1.19	1.45	0.95,1.93
75703	SCYA4/MIP-1B	2.50	1.56	1.27,1.86	2,50	2.04,2.96
75716	SERPINB2/PAI-2	2.43	1.19	0.91,1.47	2.43	2.12,2.73
76095	IER3	-	0.87	0.56,1.17	1.08	0.55,1.65
76722	CEBPD	-	-	-	-0.30	-0.58,-0.00
78452	SLC20A1	-	-	-	-0.30	-0.5,-0.01
81134	IL1RN	1.49	0.96	0.64,1.26	1.50	1.21,1.79
89690	GRO3/MIP-2B	-	0.98	0.68,1.27	1.40	1.09,1,74
100015	HAB1	-	-	-	-0.50	-0.83,-0.16
103931	DKF2P434B	-	0.28	0.01,0.55	0.41	0.06,0.72
105958	PLXND1	-	-	-	-0.20	-0.47,-0.01
118463	PNPLA2	-	-	-	-0.30	-0.5,-0.03
126256	IL1b	3.85	2.57	2.28,2.86	3.85	3.57,4.14
138263	-	-	0.30	0.01,0.59	-	-
166204	PHF1	-	-	-	0.25	0.02,0.50
169301	-	-	0.40	0.11,0.65	0.52	0.20,0.86
171185	P38IP/FAM48A	-	0.31	0.04,0.6	0.44	0.09,0.76
171731	SLC14A1	-	-	-	-	-
175038	ARMC8	-	-	-	-	-
178078	GRM4	-	-	-	-0.40	-0.75,-0.10
179657	PLAUR	-	0.37	0.09,0.68	-	-
180141	CFL2	-	0.43	0.16,0.69	-	-
184434	AXIN1	-	0.40	0.10,0.67	-	-
184711	DKFZp434B2016	-	-	-	-0.40	-0.67,-0.14
184776	RPL23A	-	-	-	-0.40	-0.85,-0.05
195453	RPS27	-	0.32	0.02,0.63	0.45	0.11,0.77
198951	JUNB	-	-	-	0.37	0.02,0.71
240122	CDC14B	-	0.34	0.06-0.63	0.47	0.23-0.71
251928	NPIP	-	-	-	-0.30	-0.54,-0.05
259842	PRKAG2	-	-0.40	-0.67,-0.08	-0.70	-1.00,-0.31
266902	NTF5	-	0.32	0.03,0.60	0.44	0.05,0.84
270062	DKFZp586D924	-	-	-	-0.40	-0.63,-0.16
272205	FLJ0034	-	-	-	-0.30	-0.57,-0.10
272801	FLJ20464	-	0.36	0.09,0.62	0.47	0.22,0.72
272802	ANKMY1	-	-	-	0.29	0.03,0.55
274431	LTBP3	-	0.33	0.04,0.59	0.43	0.13,0.71
274535	SCYA3LI	2.63	1.82	1.54,2.11	2.68	2.00,3.31
278976	BIN2	-	-	-	-0.30	-0.54,-0.05
279886	RANBP9	-	-	-	-0.30	-0.60,-0.05

Table 4.1: Differentially expressed genes: Posterior Credibility Intervals at 95%.

4.6 Discussion

Several studies have provided information about expression and mechanisms of action of LPS-inducible gene products, therefore it represents a good experimental design to evaluate the ability of a statistical analysis model to identified differentially expressed genes. In many experiments, the genes present on the array (target) cover a very large percentage of the genome studied. (i.e. yeast, microarray with a specific pattern of genes). However, the data analysis of high density microarray for screening purpose needs more attention. In fact, in this perspective, mRNAs from specific cells or tissues are hybridised to array with a high number of genes of which only a defined percentage is expressed.

In our study, in fact, the genes to be analysed after quality control ranged between 35% and 65%: this may be due to 1) no expression of specific genes present on our array in PBMC; 2) low sensitivity of the scanner instrument acquiring the fluorescence emissions of genes with low expression in PBMC. Results from single-slide Newton analysis are based on a global relative standard deviation (coefficient of variation) for the whole set of genes. Using replicates originates an increase of accuracy identifying differentially expressed genes. In addition, using a calibration experiment allows adding further information at gene level. In Tseng perspective, for each gene information from calibration as well as from comparative is needed. For this reason, we excluded from the analysis the genes missing in one of the datasets. Under a full Bayesian point of view, the information is inserted on the hyperparameters and can be estimated also for the genes not present in calibration experiment. From our data emerge that Newton model is the most conservative and call differentially expressed only the genes with the more strong differences: confirm of this is that all the 7 genes were identified also by Tseng and full Bayesian model. As concerns the two models using calibration experiments, 22 genes were common to the two models: 22/26 (84.6%) in our model and 22/44 (50%) in Tseng model of the characterized genes, respectively. In particular, among the genes identified only by Tseng model there were 17/22 (77.2%) genes significantly down-regulated. This can find an explanation mainly in the different normalisation procedure (Further details on modelling comparisons are found in Blangiardo *et al.*, 2004). In the present work, the validation experiments by real time PCR confirmed the significant differential expression of PLAUR gene (identified only by full Bayesian model), but did not confirm the significant differential expression of GRM4 gene (identified only by Tseng model). Then, validation analysis on PLAUR and GRM4 genes shows an agreement with full Bayesian results. In this study we could evaluate the gene expression of 2887 genes that were not missing in the 5 analysed arrays. Among the differentially expressed genes, many cytokine and chemokine genes were identified to be highly inducible by LPS stimulus as expected. The well-known proinflammatory cytokine IL1b was expressed at higher levels in LPS-stimulated PBMC than resting cells confirming the central role in the initiation of systemic response. Many chemokines resulted up regulated simultaneously in LPS stimulated cells: MIP-1 a, MIP-1 b, MIP-2 b and CCL3L1 for the CC subfamily, and MIP-3 a for the CXC subfamily. As concerns MIP-1 a, MIP-1 b, MIP-2 b and MIP-3 a our data are in accord to other data from the literature obtained by serial analysis of gene expression (SAGE) in LPS stimulated human monocytes. The gene expression profile of further genes identified by full Bayesian model is in agreement with data available in the literature. In fact, according to our previous data (Pepe *et al.*, 1997), the expression of PAI-2 mRNA was increased by the stimulation with LPS. PAI-2, a member of the serine protease inhibitor (SERPIN) superfamily is a rapidly inducible inflammatory mediator with characteristics of an early response gene (Schwartz 1992). Also vimentin gene showed an increased expression in LPS-stimulated leukocyte. This result is in keeping with the work of Mor-Vaknin and colleagues (2003) establishing that vimentin is secreted by macrophages into the extracellular space in response to proinflammatory signaling pathways. Among differentially expressed genes identified by full Bayesian model there were genes for which no previous evidence was available about their response to LPS stimulus such as IFI30, MLSN1, NTF5 and CFL2 or with unknown function such as FLJ20464 and P38IP. These findings stimulate further studies to evaluate the role of these genes in the activation of PBMC by LPS stimulation.

BIBLIOGRAPHY

Blangiardo, M., Toti S., Lagazio, C., Giusti, B. and Biggeri, A. Investigating gene-specific variance via Bayesian hierarchical modelling, in *"Statistical Modelling: Proceedings of the 19th International Workshop on Statistical Modelling, Florence (Italy), 4-8 July, 2004", Florence University Press*, 81-85.

Brenner S., Johnson M., Bridgham J., Golda G., Lloyd D.H., Johnson D., Luo S., McCurdy S., Foy M., Ewan M., Roth R., George D., Eletr S., Albrecht G., Vermaas E., Williams S.R., Moon K., Burcham T., Pallas M., DuBridge R.B., Kirchner J., Fearon K., Mao J., and Corcoran K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays, *Nature Biotechnology* **18**, 630-634.

Goryachev, A.B., Macgregor, P.F. and Edwards, A.M. (2001) Unfolding of microarray data, *Journal of Computational Biology* **8**, 443-461.

Hegde, P., Qi R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Snesrud, E., Lee, N. and Quackenbush, J. (2000) A concise guide to cDNA microarray analysis, *Biotechniques* **29**, 548-50, 552-556.

Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**, 817-837.

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2002) Statistical analysis of gene expression microarray experiment with replication., *Statistica Sinica*, **12**, 203-217.

Kikuchi, H., Hossain, A., Yoshida, H. and Kobayashi, S. (1998) Induction of cytochrome P-450 1A1 by omeprazole in human HepG2 cells is protein tyrosine kinase-dependent, and is not inhibited by naphthoflavone, *Archives of Biochemestry and Biophysics*, **358**,351-358.

Kirschning, C.J., Wesche, H., Ayres, T.M. and Rothe, M. (1998) Human toll-like receptor-2 confers responsiveness to bacterial lipopolysacharide, *J Exp Med* **188**, 2091-2097.

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

Mor-Vaknini, N., Punturieri, A., Sitwala, K. and Markovitz, D.M. (2003) Vimentin is secreted by activated macrophages, *Nature Cell Biology* **5**, 59-63.

Neri Serneri, G.G., Abbate, R., Gori, A.M., Attanasio, M., Martini, F., Giusti, B., Dabizzi, P., Poggesi, L., Modesti, P.A., Trotta, F., Ristagno, C., Boddi, M. and Gensini, G.F. (1992) Transient intermittent lymphocyte activation is responsible for the instability of angina, *Circulation* **86**, 790-797.

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

Pepe G, Giusti B, Attanasio M, Gori A.M, Comeglio P, Martini F, Gensini G.F, Abbate R, Neri Serneri G.G. (1997) Tissue factor and plasminogen activator inhibitor type 2 expression in human stimulated monocytes is inhibited by heparin, *Seminars of Thrombosis and Hemostasis* **23**, 135-141.

Poltorak, A., He, X., Smirnova, I., Liu, M.J., Van Huffel, C., Du, X., Birdwell, D., Alejos, E., Silva, M., Galanos, C., Freudenberg, M., Ricciardi Castagnoli, P., Layton, B., Beutler, D. (1998) Detective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene, *Science* **282**, 2085-2088.

Quackenbush, J. (2001) Computational analysis of microarray data, *Nature Review Genetics* **2**, 418-427.

Qureshi, S.T., La riviere, L., Leveque, G., Clermont, S., Moore, K.J., Gros, P., and Malo, D. (1999) Endotoxin-tolerant mice have mutations in Toll-like receptor 4 (Trl4), *Journal of Experimental Medicine* **189**, 615-625.

Raetz, C.R., Ulevitch, R.J., Wright, S.D., Sibley, C.K., Ding, A., Nathan, C.F. (1991) Gram-negative endotoxin: an extraordinary lipid with pround effets on eukaryotic signal transduction, *FASEB Journal* **5**, 2652-2660.

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

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. and Davis, R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes, *Proc Natl Acad Sci USA*, **96**, 10614-10619.

Schroeder, B.G., Peterson, L.M. and Fleischmann, R.D. (2002)Improved quantitation and reproducibility in Mycobacterium tuberculosis DNA microarray, *Journal of Molecular Microbiology Biotechnology* **4**, 123-126.

Shalon, D., Smith, S.J., and Brown, P.O. (1996) A DNA microarray system for analysing complex DNA samples using two-color fluorescent probe hybridisation, *Genome Research* **6**, 639-645.

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

Tusher, V.G. and Tibshirani, R. and Chu, G. (2001) Significance analysis of microarray applied to the ionizing radiation response, *PNAS*, **98(9)**, 5116-5121.

Wildsmith, S.E,. Archer, G.E., Winkley, A.J., Lane, P.W., Bugelski, P.J. (2001) Maximization of signal derived from cDNA microarray, *Biotechniques* **30**, 202-208.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalisation for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, **30(4)**, e15.

Yang, R.B., Mark, M.R., Gray, A., Huang, A., Xie, M.H., Zhang, M., Goddard, A., Wood, W.I., Gurney, A.L., and Godowski, P.J. (1998) Toll-like receptor-2 mediates lipopolysaccharide-induced cellular signalling, *Nature* **395**, 284-288.

CONCLUSIONS

The first two chapter present the main different approaches to model variability in microarray data. If the first chapter offers a wider spectrum of possible modelling, the second focus the attention on a subgroup of more similar methods, compared in terms of results on variability estimates and differential expression. From a methodological point of view we considered:

- the modifications of *t* test, the simplest methods to stable the denominator of the *t* statistic, introduced under a Parametric or not Parametric point of view, from a frequentist or Bayesian perspective;
- Mixed ANOVA models, which permits to decompose the variance in several components;
- two components error model, which parameterise separately low intensities from medium-high ones;
- mixture model approach, which limits the number of parameters and gains strength from the closer genes in terms of variance.

In the second paper we limit the analysis to the modified *t* tests, applied to the study of two public datasets (TCDD comparative experiment and E-Coli calibration dataset).

They show a different power of identification of differentially expressed genes: SAM analysis seems the least conservative methods (264 called genes) and includes all the genes found differentially expressed by the other approaches. On the other hand, the Empirical Bayes approach is the most conservative one, found only 61 differentially expressed genes. In general, all the five methods seem to give a good capability to recognise false positive.

In the third paper we have sorted out how normalisation and variability modelling can originate different results in gene expression. In particular, the normalisation into the model and a Full Bayesian approach produces more conservative estimates, with respect to external normalisation or an Empirical Bayesian approach, but the estimates are more robust if the normalisation is changed; it is up on the researcher to choose the approach to use, depending also on the confidence he has on the data to be analyse. Anyway, we retain that normalisation procedures used by different methods should be carefully evaluated before comparing the approaches to distinguish which part of the differences in expression is due to the effective power of the methodology and which is for normalisation technique.

Nevertheless, we can point out that calibration experiment is a good answer to the problem of variability estimate and let include prior information both working in a fully Bayesian framework and in an Empirical Bayesian one. For this reason, it works well when considering 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. So, a calibration experiment can be seen as a investment for the future experiments on the same tissue, cellular line or species.

The last paper find a biological base of the results obtained by the Full Bayesian model. Our data on LPS-inducible gene expression profile both identified novel genes (e.g. IFI30, MLSN1, CFL2, AXIN1) suggesting new targets of study in order to better understand the pathophysiology of sepsis and inflammatory disease and confirmed the involvement of many cytokines and chemokines (IL-1b, IL-1RA, MIP-1a, -1b, -2b, -3a).

COLORED FIGURES

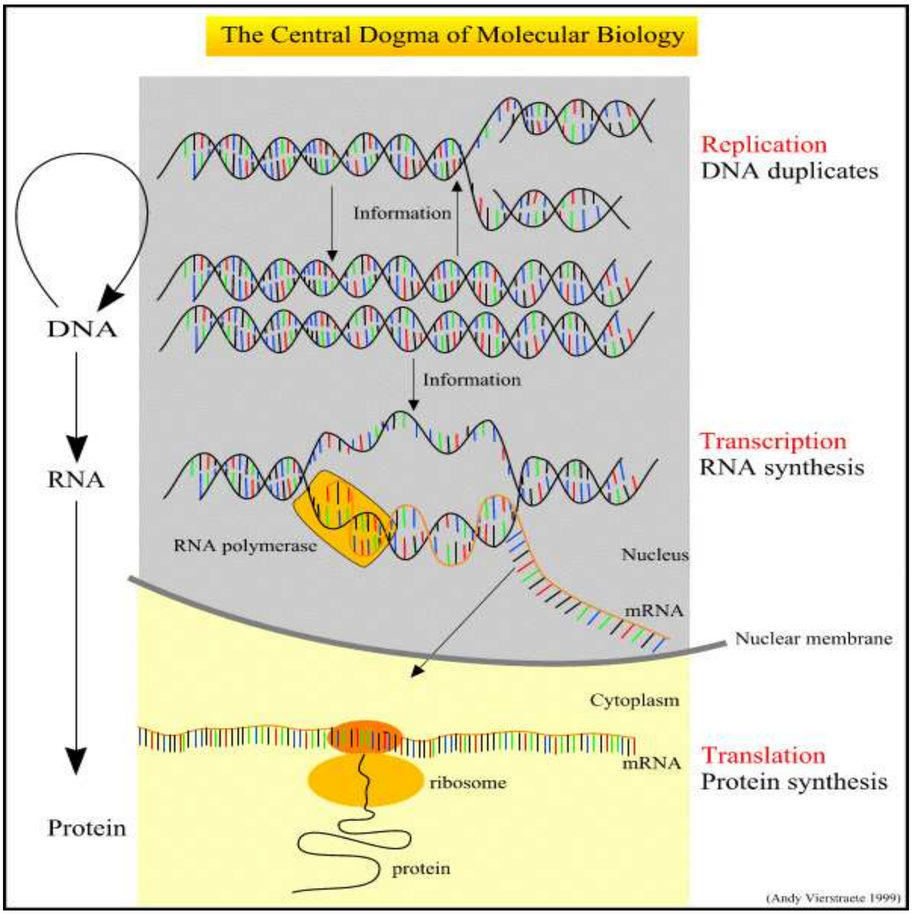

Figure 4.1: The central dogma of molecular biology.

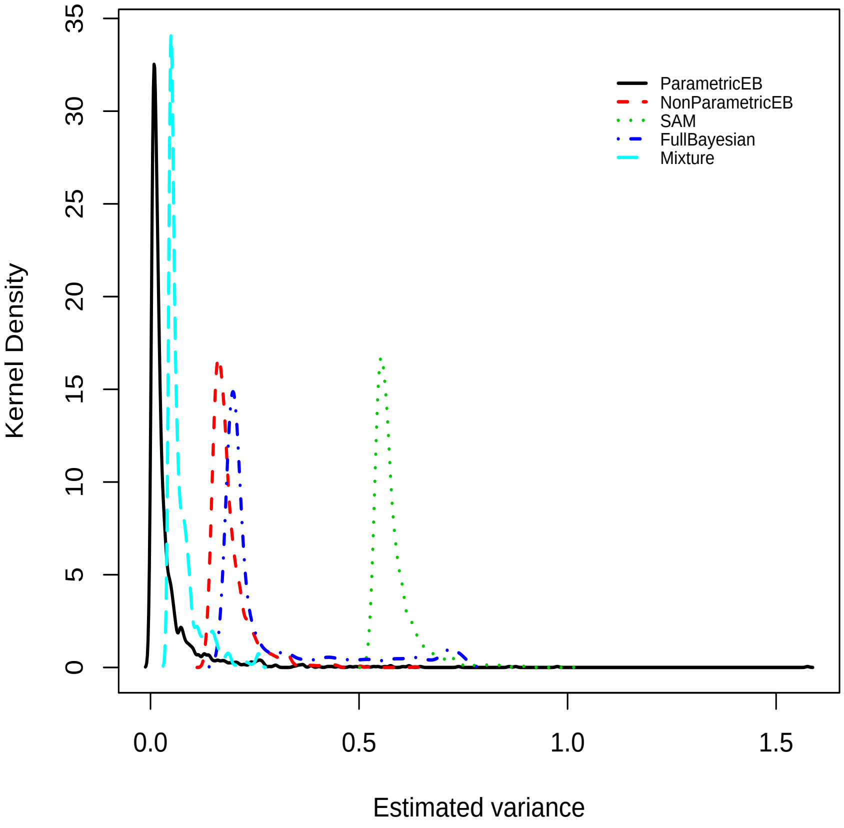

Figure 4.2: Kernel density for estimated variability (TCDD dataset): the EB curve is the closest to 0, but it has the longest right tail; on the other extreme the mixture curve has the narrowest distribution, but is centered around 0.25; Non Parametric Empirical Bayes and SAM show a similarity, but are centered on different values, while Full Bayesian approach has a larger distribution and a longer tail.

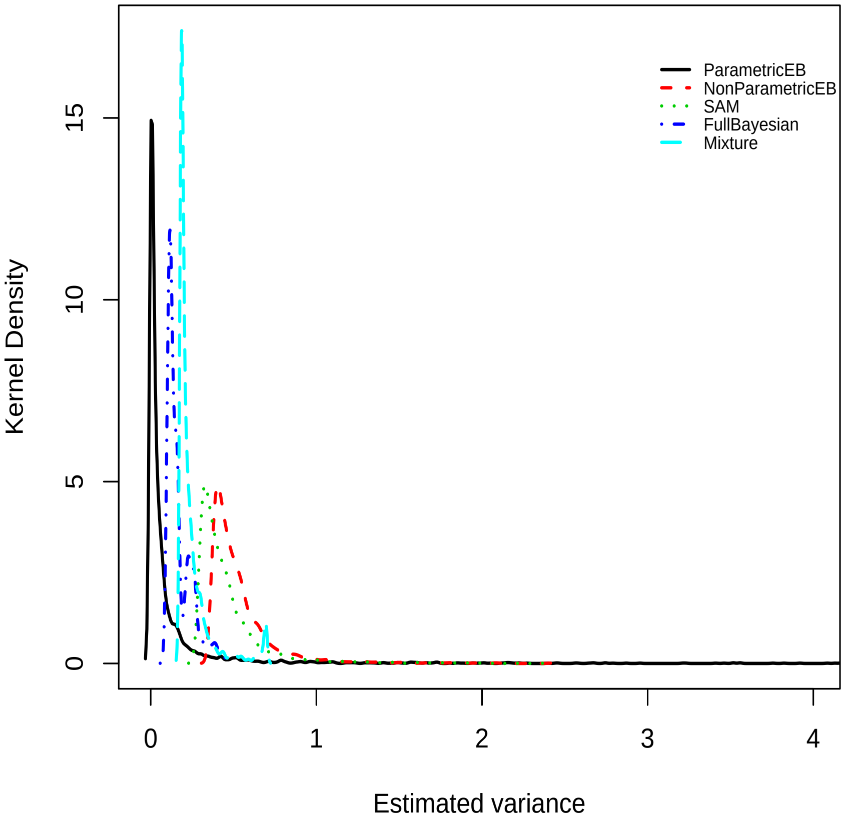

Figure 4.3: Kernel density for estimated variability (E-Coli dataset): the EB curve is the closest to 0, but it has the longest right tail; on the other extreme the mixture curve has the narrowest and distribution, but is centered around 0.25; the full Bayesian model presents a lower variability and a narrower distribution than SAM and non Parametric Empirical Bayesian one.

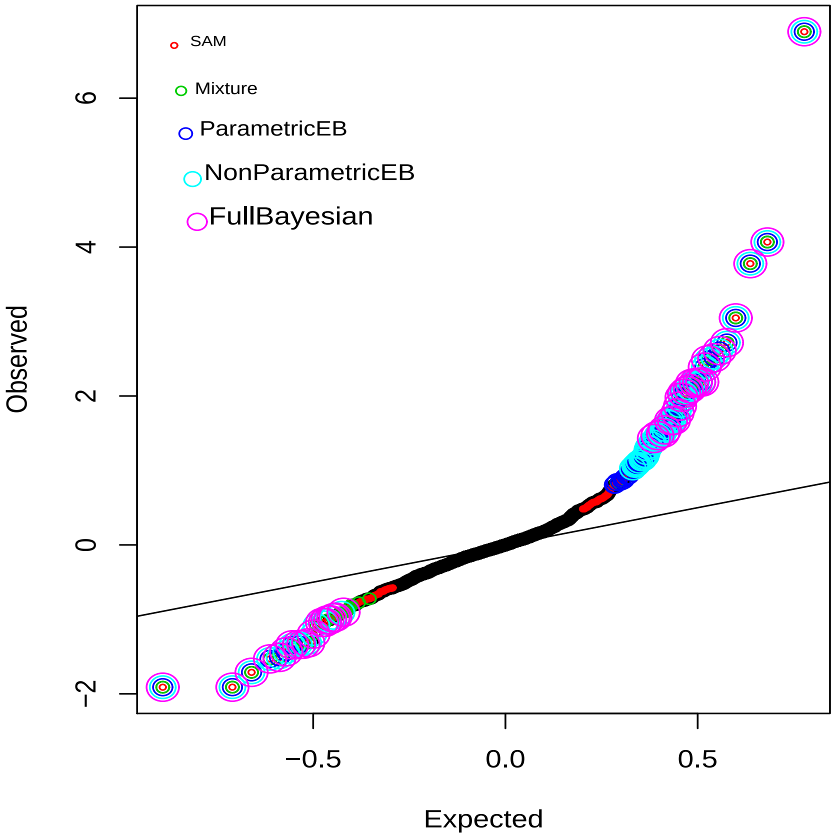

Figure 4.4: Differentially expressed genes by SAM plot (TCDD dataset): the observed relative difference (t_g) are plotted versus expected relative difference under H_0 hypothesis (mean of t_g calculated on permutated data). The solid line is the 45° line (observed equal to expected). Genes highlighted with different color and size are the differentially expressed by the 5 approaches. Moving to the extreme values, the agreement within the methodologies increases.

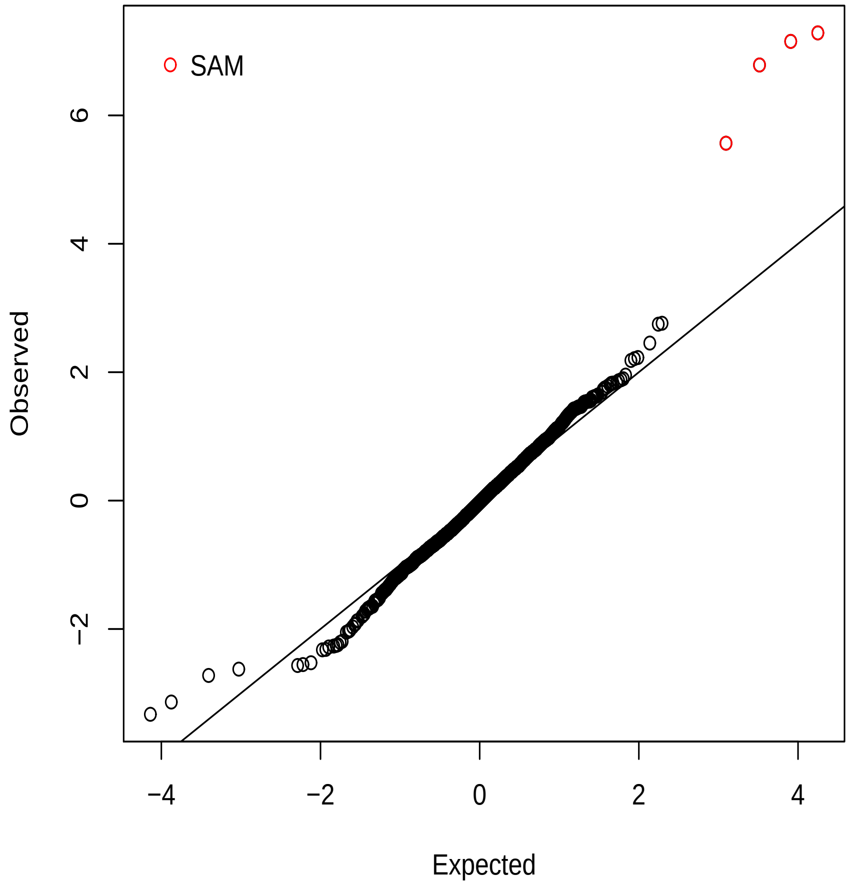

Figure 4.5: Differentially expressed genes by SAM plot (E-Coli dataset): the observed relative difference (t_g) are plotted versus expected relative difference under H_0 hypothesis (mean of t_g calculated on permutated data). The solid line is the $45°$ line (observed equal to expected). Genes highlighted with different color and size are the differentially expressed by the 5 approaches. All the points are spread on the bisector and only 4 are identified as significant by SAM analysis.

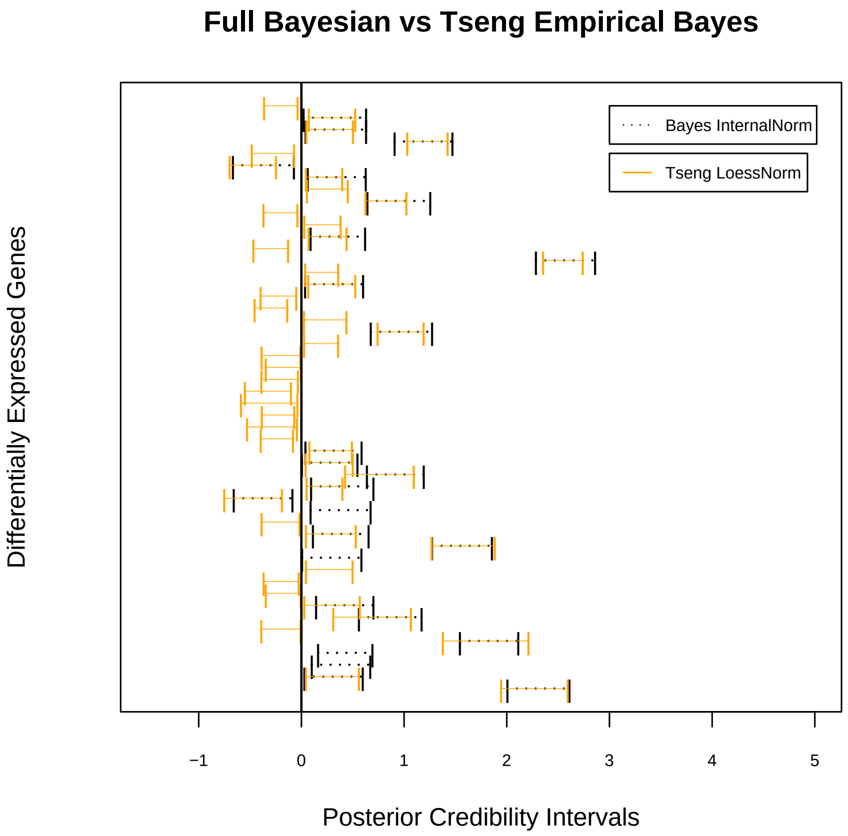

Figure 4.6: Posterior credibility intervals for differentially expressed genes. We compared the Bayesian hierarchical approach and the Empirical Bayesian one in terms of posterior credibility interval: the first method finds 26 *differentially expressed genes (dotted line), while the other one finds* 46 *differentially expressed genes (dashed line). The comparison between the two approach shows a shrinkage effect for the Fully Bayesian model: all the CI are shifted to zero.*

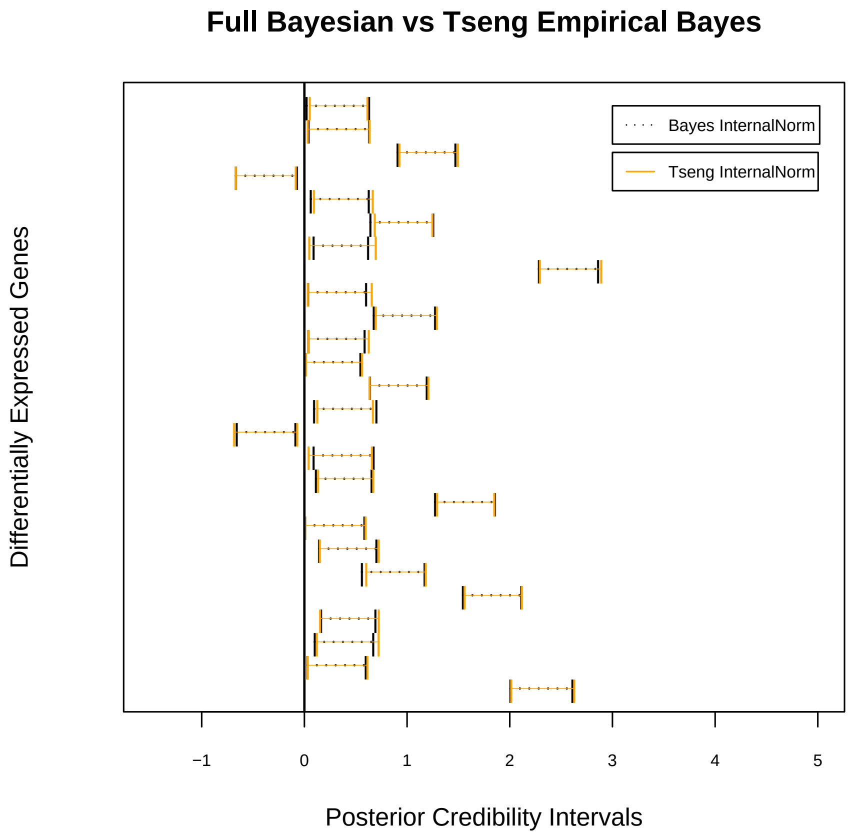

Figure 4.7: Posterior confidence intervals for differentially expressed genes: Full Bayesian model vs Tseng model with internal normalisation. The number of differentially expressed genes is 26 for both the models. The genes emerged down regulated by Tseng model with loess normalisation disappear.

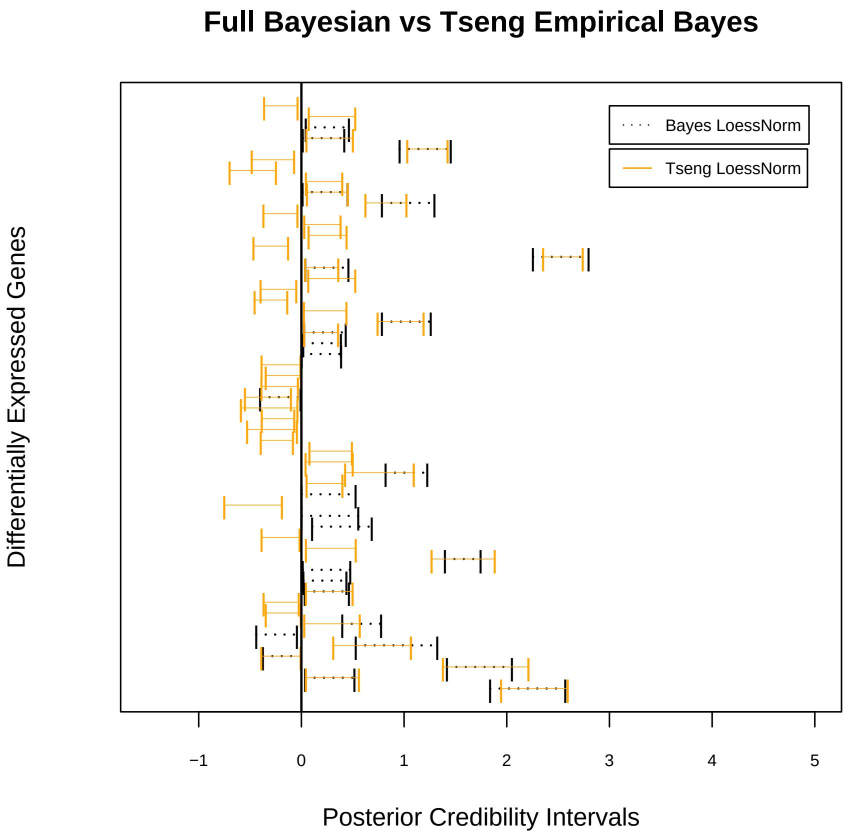

Figure 4.8: Posterior confidence intervals for differentially expressed genes: Full Bayesian model with loess external normalisation vs Tseng model with loess external normalisation.

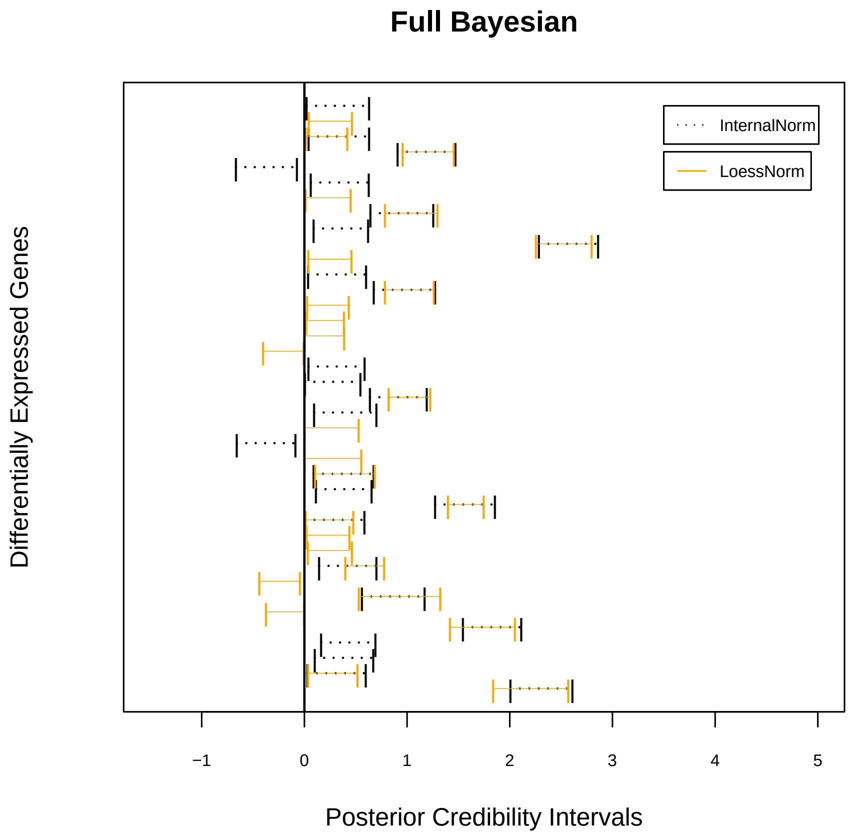

Figure 4.9: Posterior confidence intervals for differentially expressed genes: Full Bayesian model with loess external normalisation vs Full Bayesian model with internal normalisation.

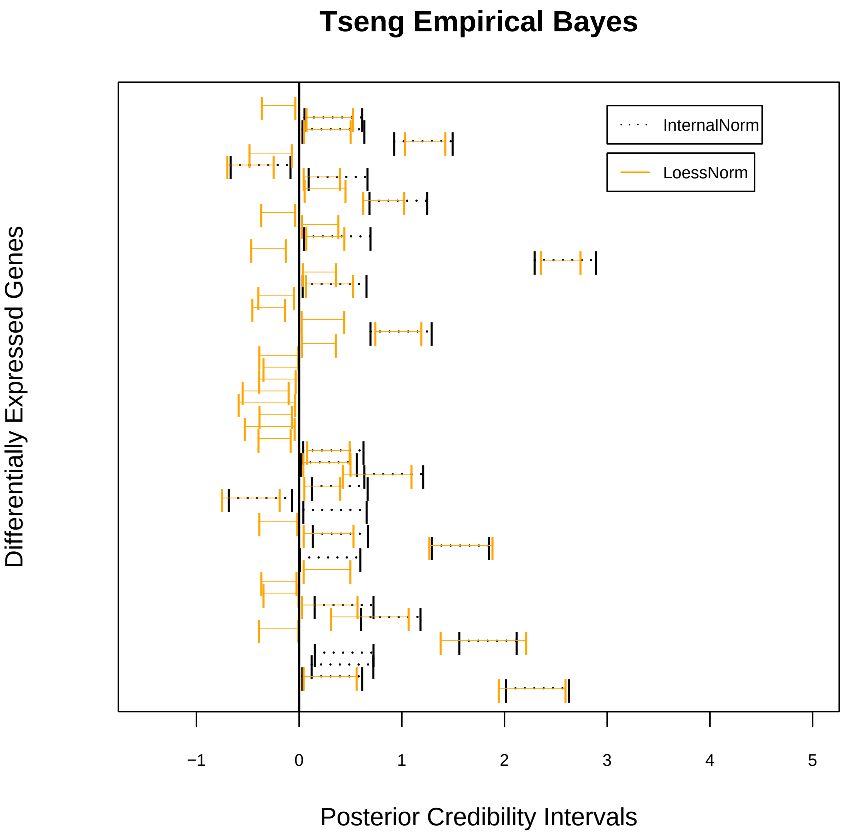

Figure 4.10: Posterior confidence intervals for differentially expressed genes: Tseng model with loess external normalisation vs Tseng model with internal normalisation.