Genet. Sel. Evol. 39 (2007) 633–650
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007029

Available online at:
www.gse-journal.org

Original article

Analysis of the real EADGENE data set: Comparison of methods and guidelines for data normalisation and selection of differentially expressed genes (*Open Access publication*)

Florence JAFFRÉZIC[a*], Dirk-Jan DE KONING[b], Paul J. BOETTCHER[c], Agnès BONNET[d], Bart BUITENHUIS[e], Rodrigue CLOSSET[f], Sébastien DÉJEAN[g], Céline DELMAS[h], Johanne C. DETILLEUX[i], Peter DOVČ[j], Mylène DUVAL[h], Jean-Louis FOULLEY[a], Jakob HEDEGAARD[e], Henrik HORNSHØJ[e], Ina HULSEGGE[k], Luc JANSS[e], Kirsty JENSEN[b], Li JIANG[e], Miha LAVRIČ[j], Kim-Anh LÊ CAO[g,h], Mogens Sandø LUND[e], Roberto MALINVERNI[c], Guillemette MAROT[a], Haisheng NIE[l], Wolfram PETZL[m], Marco H. POOL[k], Christèle ROBERT-GRANIÉ[h], Magali SAN CRISTOBAL[d], Evert M. VAN SCHOTHORST[n], Hans-Joachim SCHUBERTH[o], Peter SØRENSEN[e], Alessandra STELLA[c], Gwenola TOSSER-KLOPP[d], David WADDINGTON[b], Michael WATSON[p], Wei YANG[q], Holm ZERBE[m], Hans-Martin SEYFERT[q]

[a] INRA, UR337, Jouy-en-Josas, France (INRA_J); [b] Roslin Institute, Roslin, UK (ROSLIN); [c] Parco Tecnologico Padano, Lodi, Italy (PTP); [d] INRA, UMR444, Castanet-Tolosan, France (INRA_T); [e] University of Aarhus, Tjele, Denmark (AARHUS); [f] University of Liège, Liège, Belgium (ULg2); [g] Université Paul Sabatier, Toulouse, France (INRA_T); [h] INRA, UR631, Castanet-Tolosan, France (INRA_T); [i] Faculty of Veterinary Medicine, University of Liège, Liège, Belgium (ULg1); [j] University of Ljubljana, Slovenia (SLN); [k] Animal Sciences Group Wageningen UR, Lelystad, The Netherlands; [l] Wageningen University and Research Centre, Wageningen, The Netherlands (WUR); [m] Ludwig-Maximilians-University, Munich, Germany; [n] RIKILT-Institute of Food Safety, Wageningen, The Netherlands (WUR); [o] University of Veterinary Medicine, Hannover, Germany; [p] Institute for Animal Health, Compton, UK (IAH); [q] Research Institute for the Biology of Farm Animals, Dummerstorf, Germany

(Received 10 May 2007; accepted 6 July 2007)

[*] Corresponding author: florence.jaffrezic@jouy.inra.fr

Abstract – A large variety of methods has been proposed in the literature for microarray data analysis. The aim of this paper was to present techniques used by the EADGENE (European Animal Disease Genomics Network of Excellence) WP1.4 participants for data quality control, normalisation and statistical methods for the detection of differentially expressed genes in order to provide some more general data analysis guidelines. All the workshop participants were given a real data set obtained in an EADGENE funded microarray study looking at the gene expression changes following artificial infection with two different mastitis causing bacteria: *Escherichia coli* and *Staphylococcus aureus*. It was reassuring to see that most of the teams found the same main biological results. In fact, most of the differentially expressed genes were found for infection by *E. coli* between uninfected and 24 h challenged udder quarters. Very little transcriptional variation was observed for the bacteria *S. aureus*. Lists of differentially expressed genes found by the different research teams were, however, quite dependent on the method used, especially concerning the data quality control step. These analyses also emphasised a biological problem of cross-talk between infected and uninfected quarters which will have to be dealt with for further microarray studies.

quality control / differentially expressed genes / mastitis resistance / microarray data / normalisation

1. INTRODUCTION

Microarray analyses have been highlighted as an area of high priority within the European Animal Disease Genomics Network of Excellence (EADGENE), to study host-pathogen interactions in animals. Microarrays give the possibility to study the changes of expression of thousands of genes simultaneously depending on the pathogen.

A large variety of methods for normalising and analysing microarray data has, however, been proposed in the literature, and there is still no clear consensus about which analysis process is recommended. The aim of this joint research work was to review the methods and software packages used by the EADGENE partners and to provide some general guidelines for further analyses. To achieve this goal, a real data set was distributed among the workshop participants. The real data was provided by an EADGENE funded microarray study looking at the gene expression changes following artificial infection of cows with two different mastitis causing bacteria: *Escherichia coli* and *Staphylococcus aureus*. The effect of artificial infection was tested over time in 12 dairy cows using three udder quarters in each cow for different time points following infection and one for the control sample. The study included two species of bacteria as well as several time-points, resulting in a true analytical challenge (48 microarrays in total). The EADGENE partners who provided the data were RIBFA and the Roslin Institute.

In this paper three main steps of microarray data analysis will be discussed: data quality control, normalisation and statistical methods for the detection of differentially expressed genes. For each of these steps, the techniques used by the workshop participants will be presented and compared.

2. MATERIALS AND METHODS

2.1. Presentation of the data

2.1.1. Comparison of E. coli vs. S. aureus *elicited mastitis in cows using transcriptomic profiling*

The outcome of an udder infection (mastitis) is influenced by the species of the infecting bacteria. Coliform bacteria, *e.g. E. coli*, tend to cause acute infections with severe inflammatory symptoms, while others, like *S. aureus* often result in chronic infections with less severe symptoms. The molecular causes underpinning these differences in host pathogen interactions are largely unknown. Here, we established a strictly controlled animal model to allow for a systematic analysis of the different immune responses elicited by *E. coli vs. S. aureus*, using strains of both pathogen species previously isolated from field cases of mastitis. Healthy heifers were infected in the fourth month of their first lactation. None of the cows had suffered a previous udder infection and their somatic cell counts were well below 100 000 cells per mL of milk.

Three trials were conducted, each comprising four animals. First, 500 CFU of our asseverated *E. coli* strain 1303 were infected into udder quarters at time 0, 12 and 18 h. The fourth quarter was kept as a control. The animals were culled after 24 h and sampled. All animals showed signs of acute clinical mastitis by 12 h after challenge: increased somatic cell count (SCC), decreased milk yield, leucopenia, fever and udder swelling. Quantitative RT-PCR analysis revealed that the expression of Toll-like receptor (TLR) 2, TLR4 and beta-defensin-encoding genes was greatly enhanced in the 24 h infected quarters, while the relative mRNA copy numbers remained low in the uninfected control quarters, which is coherent with the microarray results presented below. Secondly, animals infected with 10 000 CFU of the *S. aureus* strain 1027 in a similar scheme over 24 h (n = 4) showed no or only modest clinical signs of mastitis. No evidence of alteration in TLR or beta-defensin-encoding indicator genes for activated innate immune defense was found. In the third trial, four animals were infected with the *S. aureus* pathogen. For each of them (i) two quarters were infected at time 0, (ii) a third quarter at time 60 h, and animals

were killed after 72 h. Hence, there were two quarters per animal with *S. aureus* inoculated for 72 h, one quarter with the pathogen inoculated for 12 h and again one control quarter. *S. aureus* caused clinical symptoms and increased expression of the TLR and beta-defensin-encoding indicator genes in this third group of animals, infected over 72 h (n = 4).

Assignment of the animals to become inoculated with *E. coli* or *S. aureus* was completely at random and arbitrary. The three trials were conducted at three different days. Inter-animal transmission can be excluded, thanks to proper handling of the inoculates. The identity of the pathogens were verified from re-isolates of milk samples. In addition to the classical microbiological verification, strain identity was verified using diagnostic digests of pathogen residential plasmids as criteria.

The clinical and qRT-PCR data proved that the *E. coli* infected animals all developed symptoms of acute mastitis, earlier than 24 h after infection. *S. aureus* pathogens, however, needed more time to elicit not only clear infection related symptoms of mastitis, but also the activation of the immune defense within the udder. We also noted a clear host-individual influence in this regards. Samples from all these udder quarters were carefully asseverated and stored in liquid nitrogen, for subsequent DNA-microarray analyses.

The microarray experiment was carried out using the Bovine 20K array (ARK-Genomics). A common reference design was used and the reference sample was made up of all 48 RNA samples. The reference sample was labelled with Cy3 and the treatment with Cy5 on each microarray slide. All samples were collected in Hannover (Germany) by Holm Zerbe, Hans-Joachim Schuberth, and Wolfram Petzl, and had been validated by Hans-Martin Seyfert in Dummerstorf (Germany). The samples were shipped to the Roslin Institute for transcriptome profiling by Elizabeth Glass and Kirsty Jensen.

The Bovine 20K microarray was subdivided in 48 blocks, with 12 rows and 4 columns. Each of the 48 resulting blocks was printed with its own unique print-tip (*i.e.* there are 48 print-tips). Each block consisted of 30 sub-grid rows and 30 sub-grid columns. Almost all (19 705) features were printed in duplicate within the same block, 324 printed 4 times and 2 printed 12 times. Annotations were provided by Mark Fell of the Roslin Institute and were distributed among the workshop participants. The microarrays were scanned and data were extracted using Bluefuse (http://www.cambridgebluegnome.com/bluefuse.htm). Bluefuse does not provide an estimate of the background intensity, and therefore no further background correction was possible on these data.

2.2. Normalisation of the data

2.2.1. Data quality control

Several quality control procedures were used by the authors and Table I presents an overview of these techniques. Most of the teams used the spot quality indicators provided by the scanning software (Bluefuse) to make decisions about excluding spots from the analysis. There are several indicators of quality provided by the Bluefuse software: (a) the probability that a clone is expressed in the tissue studied (PON) with a value between 0 and 1; (b) a manual quality flag from A (good) to E (bad); (c) a compound 'confidence' quality indicator between 0 and 1; and (d) a binary quality indicator that is 0 (bad) or 1 (good). The simplest approaches were to remove spots with manual flags or with Bluefuse flag values equal to D or E because their confidence levels were lower than 0.30 (meaning a poor quality of spot). In more sophisticated approaches, raw data were visualised using R-LimmaGUI [15] to check the overall quality by several criteria, such as M boxplots, M-A plots, and Cy5-Cy3 scatter plots. INRA_T pointed out, using simple descriptive statistics that array BTK2-74 was different from the other slides given the mean, minimum and maximum, and should be deleted from the analysis. M-A plots of the raw data were atypical and showed a clear 'fishtail' pattern for low intensity spots, where the log-ratios (M) diverged, as shown in Figure 1. This indicated relatively noisy data due to many spots with low intensities. ROSLIN therefore proposed to add 2^8 to all the channel intensities. IDL deleted spots with intensities above 65 000 (oversaturated spots) or with values within the experimental error, *i.e.* spots smaller than 400 [8]. AARHUS suggested a quality weighting of the data [9] by down-weighting the spots with low quality based on Bluefuse 'Confidence' or 'P ON' measurements. For all teams, data were log2 transformed and the log-ratio between Cy5 and the reference Cy3 was considered as the observed intensity.

2.2.2. Correction for spatial and intensity-dependent bias

Normalisation of the data is a two-step process including first a correction for spatial bias, and second a correction for intensity-dependent bias. Correction for spatial bias was usually carried out separately for each block (print-tip) of each array, by either subtracting the median for each block, or by subtracting the corresponding row and column means (RC correction, excluding control spots) [1]. The intensity dependent bias was removed by either block-Loess correction [14], or by a global Loess correction [17]. Two levels for each of

Table I. Overview of the methods used by each team for data quality control and normalisation.

Team	Quality Control (QC)	Normalisation	Softwares
ROSLIN	(1) Genespring. (2) 2^8 added to all spot intensities. Three plots per slide: MA, print-tip box-plot and spatial plot. Summary statistics were used.	Four normalisations: global and local dye and spatial correction.	Genespring Bioconductor (Limma and Marray) with changes.
AARHUS	Diagnostic plots. Weights for data quality were used based on 'Confidence' and 'P-ON' weights.	Print-tip Loess correction was used with different weights.	Bioconductor Limma.
INRA_T	Clones with quality control flag A, B or C (n = 6948) were kept. Slide BTK2-74 was omitted.	Arrays and genes were mean centered.	R.
IDL	Would use background information but it was not provided here with Bluefuse. Spot edits, oversaturated spots and spots smaller than 400 were removed.	Print tip Loess. Heterogeneous variance correction [8].	Limma http://www.asgbioinformatics.wur.nl
ULg1		2-step Wolfinger procedure. Random effect model with a fixed dye effect, a random print-tip effect and an interaction term.	SAS®
SLN		Samples were normalised to uninfected group.	(1) Genespring. (2) Orange http://www.ailab.si/orange
IAH	Blank, auto_excl and man_excl spots were removed. Replicate spots were averaged.	Median normalisation for each slide. Scale normalisation between arrays.	Co-Express programmed in R.
INRA_J	Manual flags were considered as missing (man_excl).	Global Loess and print-tip correction. Replicate spots were considered as independent observations.	R.
ULg2	'DNA', 'blank', 'buffer', 'nothing', 'light reference' were deleted. Spots were deleted if quality is bad (E) for 25 consecutive spots.	Global Loess.	R, Maanova.
WUR	M box-plots across slides. Slide BTK2-74 was omitted. Background (BG) was estimated and spots where Signal/BG <2 were deleted (from 20 k to 8.7 k genes).	Global Loess. 'Rikilt' normalisation.	Genemaths XT Limma.
PTP	Quality measures were used to delete the worst spots. Non-relevant spots were also deleted.	Global Loess, Mixed model across slides (Wolfinger 2-step procedure [16]).	SAS®

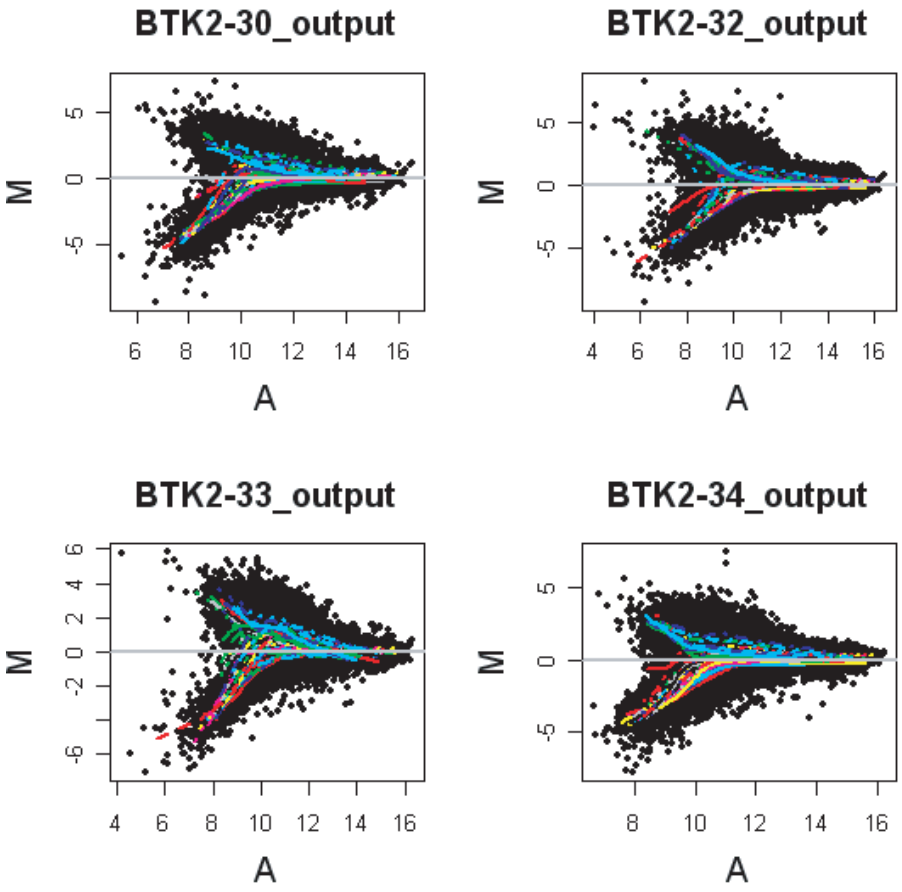

Figure 1. The "fishtail" appearance of M-A plots for the raw data for slides 1–4. Lines are Loess curves for each of the 48 print-tips. Control spots were omitted.

the two normalisation steps were examined by ROSLIN to check whether these steps should be global (*i.e.*, chip-wide) or local (*i.e.*, print-tip). The choice was informed by comparisons of summary measures of M-A plots, spatial heat diagrams and print-tip box-plots for the raw data and all four normalisations. The local spatial bias (RC correction) and local intensity-bias (MA normalisation) were found here to perform consistently well regarding the spatial plot in the F-test of differences between blocks in M values, the M-A plot in the F-test of a block MA correction *versus* a chip-wide MA correction, and the print-tip box-plots in the mean inter-quartile range of M. This local RC-MA normalisation

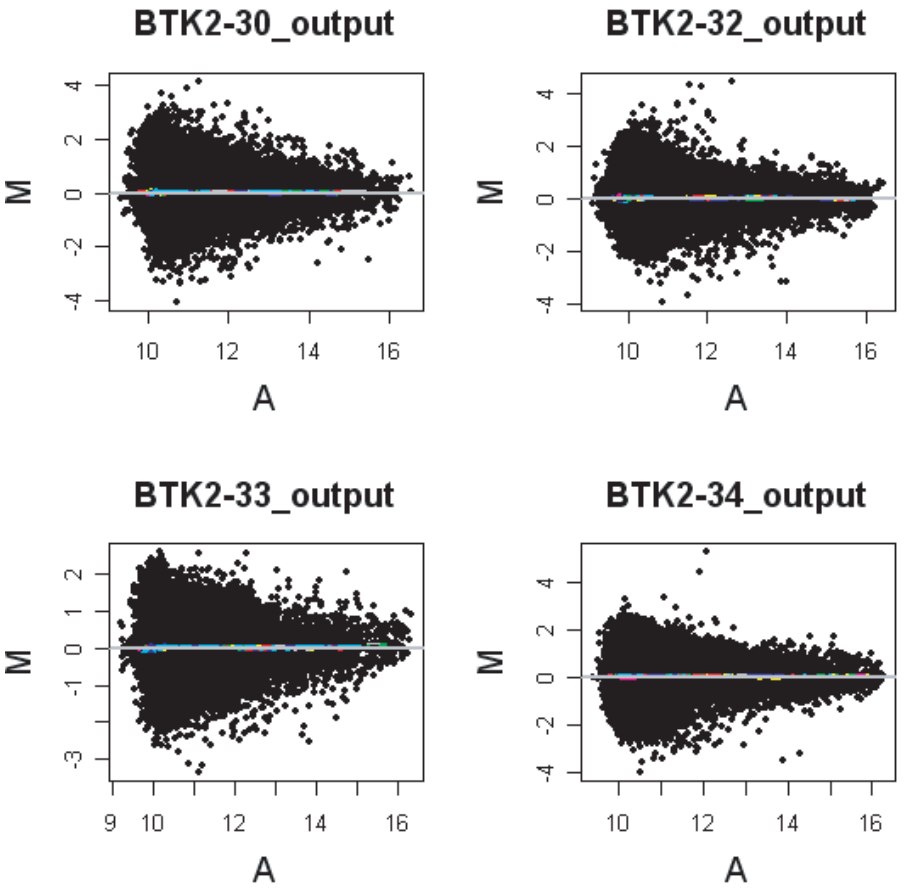

Figure 2. M-A plots after normalisation for ROSLIN team with a print-tip Loess correction for slides 1 to 4.

was therefore chosen by this team for normalising the data. Figure 2 shows the corresponding M-A plots after normalisation. Since *E. coli* and *S. aureus* samples were hybridised always at the same channel and against a common reference, the setup of this experiment requires no dye swap effect correction, which is often a source of experimental noise.

Another possible approach for data normalisation is to use an ANOVA model. This approach was used here by two teams (ULg1 and PTP) using a two-stage mixed-model approach [5] with the Proc Mixed SAS® procedure. In the first stage, initial models were fitted to each array separately to take account of the experimental systematic effects on the base-2 logarithm of the

pixel values. The model included a fixed dye effect, a random print-tip effect and an interaction between dye and print-tip effects. Effects of print-tip were considered as random because of the manufacturing variation expected between print-tips. Residuals obtained from this model were then analysed to find differentially expressed genes.

It has to be emphasised that all the genes were used in the normalisation procedures presented above, based on the underlying assumption that most of the genes are not differentially expressed and that the observed differences are only due to technical artefacts. This assumption has to be checked for every experiment and may sometimes not be verified, especially when using dedicated chips.

2.2.3. Software packages used for the data normalisation

Four teams (ROSLIN, AARHUS, IDL, WUR) used the Bioconductor package Limma – Linear Models for Microarray Analysis [13] in R for data normalisation. A bioinformatics pipeline was developed by IDL to handle both data normalisation and detection of differentially expressed genes accessible at http://www.ASGbioinformatics.wur.nl. The SAS® software was used for normalisation using an ANOVA model.

2.3. Methods used to find differentially expressed genes

Three main biological questions were investigated on this data set: which genes are differentially expressed (1) between the two types of infection (*E. coli* and *S. aureus*); (2) over time within each bacteria; and (3) across time and bacteria. Table II presents an overview of the statistical methods used by each team to find differentially expressed genes.

2.3.1. ANOVA approach with different variance models

Three teams (ROSLIN, AARHUS, IDL) used for this part of the analyses the Bioconductor R package Limma [13], which allows complex designs and provides robust t- and F-statistics for differential gene expression by the use of empirical Bayes methods (eBayes) for shrinking the residual variances of genes towards their approximate median value. This approach is based on an inverse chi-square prior on the variances [12]. The linear model used here accounted for within-array replicate spots and included the effects of time and

Table II. Overview of the methods and software packages used by each team for the detection of differentially expressed genes.

Team	Comparison	Single gene analysis Statistical method	Softwares
ROSLIN	Uninfected quarters, between infections and across time within each infection.	(1) Genespring. (2) Limma + FDR.	(1) Genespring. (2) Bioconductor Limma with eBayes correction.
AARHUS	Infected at different times *vs.* uninfected for each separate experiment.	Limma + FDR.	Bioconductor Limma with eBayes correction.
INRA_T	*E. coli* infected *vs.* unifected.	Fisher test with FDR.	R.
IDL	For both *E. coli* and *S. aureus*: infected at different times *vs.* uninfected.	Limma FDR + Fold change cut-off.	Limma https://www.asgbioinformatics.wur.nl
ULg1	Infected *vs.* uninfected. *E. coli vs. S. aureus*	SAS®, Wolfinger mixed model, FDR correction.	SAS®
SLN	Infected *vs.* uninfected.	2-fold change. Anova with FDR.	(1) Genespring. (2) Orange http://www.ailab.si/orange
IAH	*S. aureus vs. E. coli*	Derived from clusters and differences over time.	Co-Express programmed in R.
INRA_J	Different time points within infection or same time point between infections.	Anova model, Structural mixed model for variances, Time course with EDGE.	'SMVar' function programmed in R EDGE [3].
ULg2	For *S. aureus*: infected *vs.* uninfected.	Bayesian analysis of variance.	R, BAMarray http://www.bamarray.com
WUR	Infection and time combined.	2-way ANOVA. 1-way ANOVA.	TIGR [11].
PTP	Within and between infections at different times.	Linear mixed model: bacteria effect, and time as a categorical or continuous variable.	SAS®.

challenging bacteria. Differentially expressed genes between types of infection were tested based on the robust t-statistics and differential expression of genes over the different time points used a moderated F-test. Another approach also based on an ANOVA model but with a different variance model was used by INRA_J. It is based on a structural mixed model on the variances [7] and is implemented in R in the 'SMVar' function. Here, a fixed condition effect and a random gene effect were considered to model the log of the variances. Two other teams (WUR and ULg2) used TIGR Multiple Experimental Viewer v4.0 [11] and the BAMarray software [6] for Bayesian analysis of variance, respectively. In the latter approach, genes are clustered into groups of equal variances and data are rescaled to satisfy the equal variance assumption. Then, a hierarchical Bayesian model is used to synthesise information across all genes simultaneously, and estimated effects for genes unlikely to be differentially expressed are shrinked to zero to enhance patterns of interest.

2.3.2. Models for time-course study

In the ANOVA models presented above, observations were assumed to be independent, which was not the case in this time-course study since measurements were made at different time points for each animal. Three longitudinal approaches were proposed here to take these correlations into account and find differentially expressed genes over time in the two infections.

The first approach was performed by PTP using the Mixed procedure of SAS®. A gene-by-gene analysis was performed on the residuals obtained from the normalisation process. The effects included in this linear mixed model were the following: a fixed bacteria effect, a non-parametric mean curve by fitting the time effect as a qualitative variable or a parametric function of time (linear and quadratic regression on time), and the interactions between bacteria and these time effects. A linear random regression model was considered (using a random cow effect and an interaction with time). A quadratic random regression model was also investigated but did not converge. For each gene-specific model, custom hypothesis tests were constructed to determine whether gene expression was different between healthy and infected quarters, or between quarters infected with *E. coli* and *S. aureus* at different times.

The second longitudinal approach considered in this workshop by INRA_J was based on the Edge package [3]. In this gene-specific model, the population average time curve was modelled using a natural cubic spline function and the correlation structure was fitted with a random intercept. Two biological questions can be addressed with this approach. First, is the effect for each gene

constant for each infection. Second, is the expression pattern over time, *i.e.* the average time curve, for each gene the same in the two infections.

In the last approach, an ANOVA was performed by INRA_T on the expression value for each *E. coli* clone, with the time factor as an explanatory variable (4 levels: 0, 6, 12 and 24 h). A standard Fisher test was used to test the effect of time on each gene. After selection of the differentially expressed genes over time, a clustering approach based on smoothed expression curves [4,10] was used to find clusters of genes with similar expression profiles. This second step is presented in the post-analysis paper.

2.3.3. Correction for multiple tests

Regarding the correction for multiple tests, all teams used the classical Benjamini and Hochberg [2] correction at a 5% False Discovery Rate (FDR) threshold, either using R functions or the SAS® Multitest procedure.

3. RESULTS

Although various methods were applied for normalising and analysing the data, it was reassuring to see that most of the teams found similar biological results. First, it was found that the largest number of differentially expressed genes was obtained when comparing samples from udder tissue challenged for 24 h with *E. coli* to non-challenged tissue. In contrast, challenging with *S. aureus* did not result in a dramatic transcriptional response. Second, quite a large number of differentially expressed genes were detected at time zero between the two groups of infections. This showed a cross-talk between udder quarters or an invasion by immune cells from the infected quarters, since all udders were collected simultaneously at the end of exposure. We will present here the results obtained with the Bioconductor Limma package (ROSLIN in Tab. II) which was used by many teams and was shown to perform well for differential gene expression analysis.

The uninfected quarters from the *E. coli* infection exhibited differential expression in 402 clones representing 359 genes compared to the *S. aureus* uninfected quarters. The most up-regulated genes included metallothioneins and lipopolysaccharide binding protein, indicating that an immune response has been triggered in the uninfected quarters. Furthermore, the MHC class II invariant chain molecule, CD74, was down-regulated, suggesting that the cell populations present in the mammary gland quarter had altered in response to the infection of neighbouring quarters. Considerable overlap was observed in

the gene lists from the *E. coli* uninfected quarters and the 331 genes declared to be differentially expressed at 6 h post *E. coli* infection. More than 600 clones, representing 538 genes exhibited differential expression at 12 h post infection, and the number of differentially expressed genes reached a maximum at 24 h post infection when the transcription of 1190 genes was affected. Many of the most up-regulated genes at this time point are associated with the influx of neutrophils into the infected gland, including S100 calcium binding proteins A8, A9 and A12, colony stimulating factor 3 and several chemoattractants for neutrophils, *e.g.* interleukin 8 and chemokine (C-X-C motif) ligand 1 and 2.

No gene was identified as being significantly differentially expressed during *S. aureus* infection using the cut-off FDR value of 0.05. This lack of statistical support principally results from high levels of variation between the biological replicates. This may be an artefact of the experimental procedure; large mammary gland samples were collected for RNA extraction which may have comprised variable amounts of *S. aureus* infected tissue, because the bacteria causes a localised infection. Therefore the gene lists were expanded to include those genes with t-test p-values less than or equal to 0.01 and a fold change greater than or equal to 1.5. At 6 h post *S. aureus* infection, 154 genes exhibited differential expression. The most highly up-regulated gene was lactotransferrin, an antimicrobial protein secreted in milk. Interestingly, this gene was only observed to be up-regulated at 24 h post *E. coli* infection. At 12 and 24 h post infection 182 and 266 genes were declared differentially expressed, respectively. However, the number decreased to 97 by 72 h post infection. There was some overlap between the lists of differentially expressed genes during *E. coli* and *S. aureus* infections, including the up-regulation of superoxide dismutase and the down-regulation of interleukin 7. The analysis of the microarray data identified two putative genes that may be indicative of *S. aureus* infection. Leucine rich repeat kinase 2 was down-regulated at all 4 time points during *S. aureus* infection but not during *E. coli* infection. In addition, a clone (AJ814901) whose sequence currently only matches EST was also down-regulated during *S. aureus* infection.

Various comparisons of lists of differentially expressed genes found by the EADGENE teams were performed. We focussed mainly on the comparison of the differentially expressed genes found between time 0 and 24 h for the *E. coli* infection, which exhibited the largest transcriptional response. It was found that although all the teams found a large number of differentially expressed genes between these two time points, the lists of genes were still dependent on the method chosen. Figure 3 gives the Venn diagram for the differentially expressed genes found by three of the teams. They used different data quality

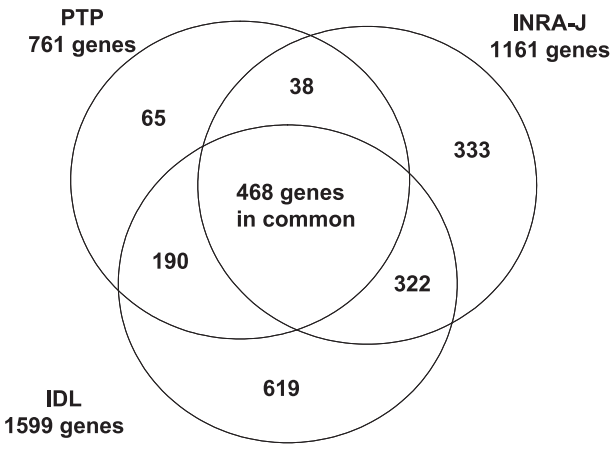

Figure 3. Venn diagram for the lists of differentially expressed genes found for *E. coli* between times 0 and 24 h after infection at a 5% Benjamini and Hochberg threshold for IDL, INRA_J and PTP teams. Normalisation and analyses methods used by these teams are presented in Tables I and II.

control procedures and different normalisation and analyses methods: the IDL team used a print-tip Loess normalisation and the Limma package, INRA_J used a global Loess normalisation and the structural model for variances, and PTP used the global Loess and a 2-step mixed model approach with SAS®. It was found that 468 genes were detected in common for these three teams, and 790 genes were detected in common for IDL and INRA_J. It is interesting to also note that IDL and PTP teams, despite using very different approaches, found 658 genes in common among the 761 genes detected by PTP. When focussing only on the 500 most differentially expressed genes found by the three teams, only 206 genes were found in common for the three approaches, as shown in Figure 4. A larger consistency in the ranking of the genes could have been expected, especially between IDL and INRA_J which used Limma and SMVar shrinkage approaches, respectively. In fact, both teams found here only 272 genes in common among the first 500, although the two methods were found in previous studies [7] to provide very similar results in the detection of differentially expressed genes. The main difference in the analyses performed by these two teams was in the data quality control step. On the contrary, 323 were found in common between IDL and PTP teams, who used very different statistical approaches for the detection of differentially expressed genes but a more similar approach for data quality control, with the removal of oversaturated and low quality spots. The data quality control step therefore appears here

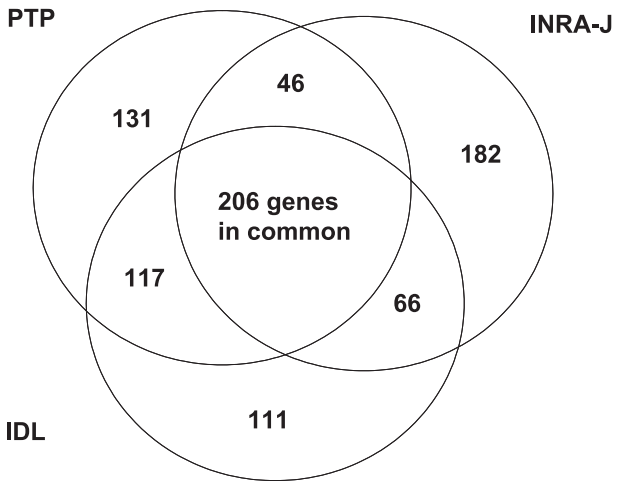

Figure 4. Venn diagram for the 500 first differentially expressed genes found for *E. coli* between times 0 and 24 h after infection for IDL, INRA_J and PTP teams.

to be essential for microarray data analysis but a consensus for best practice still has to be reached.

It has to be emphasised that this study presented the number of genes that were declared to be differentially expressed by statistical methods, at a 5% Benjamini-Hochberg threshold. A biological validation still has to be performed, however, for most of these genes to be able to differentiate between the true positives and the false positives.

4. DISCUSSION

Quality control of the data proved, in this workshop, to be an important first step. Simple summary statistics for each slide, as well as print-tip box-plots, MA and spatial plots can be used for quality checks. Recommendation would be to delete spots that were flagged as low quality, or to perform some quality weighting [9]. It was found here that one of the slides was of very poor quality and had to be deleted from the analysis. Bluefuse does not use a measure of background intensity from pixels around the spot, nor does it make an explicit estimate from within-spot pixels, and therefore the background intensities were not provided.

Following quality control, the normalisation step is divided into two steps: a correction for spatial bias and a correction for intensity-dependent bias. The first step is performed on each block (print-tip) for each array separately either

by subtracting the median for each block, subtracting the corresponding row and column means (excluding control spots) or by including a random array block effect in the Wolfinger *et al.* [16] two-step mixed model approach. The second step is performed using either a local or global Loess correction. Using various quality control plots, one of the teams found that the local Loess correction was the most adapted for this data set. Many teams used the Bioconductor R package Limma to perform this normalisation. More diversity was observed among the teams for the data quality control step than for the normalisation step.

Two main approaches were used for the statistical analysis of these data. The first approach was based on ANOVA models and allowed the detection of differentially expressed genes using two by two comparisons with robust t-tests. The construction of these robust t-statistics was based either on the eBayes Limma shrinkage [13] or on a structural mixed model on variances [7] – SMVar function in R. These analyses provided lists of genes that were differentially expressed within each infection at different time points, as well as between the two infections. The second approach was based on longitudinal models and took into account the correlations between measurements involved in this time-course study. For this second set of analyses, a random regression model was used with SAS® in a two-step mixed model approach, and the EDGE package developed by Dabney *et al.* [3] was applied to these data. These analyses allowed the detection of genes that had a pattern of expression changing over time or that differed for both infections. Since these data come from a longitudinal study, it is advisable here to use the latter approaches that take into account correlation between measurements rather than the ANOVA based models which assume independence of the observations. Correction for multiple tests was performed by all the teams using Benjamini and Hochberg [2] FDR approach at a 5% threshold.

Although various quality control procedures, data normalisation and analysis methods were used, all the teams generally obtained the same main biological results. In fact, all participants found that most of the differentially expressed genes were found between the uninfected group and quarters that had been challenged by the *E. coli* pathogens for 24 h. On the contrary, very little transcriptional response was observed for the *S. aureus* infection.

It can be argued, however, that the robustness observed here concerning the biological results may be due to the extremely large transcriptional response with the *E. coli* infection. These conclusions may therefore not be generalised to other experiments with only small transcriptional changes. Moreover, several methods used here such as the shrinkage approaches (Limma, SMVar,

BAMarray) were designed to improve the performance under high-noise, low replicate, small-change settings. The *E. coli* data, which exhibited a very large transcriptional response, may therefore not allow pointing out the subtle differences between these various methods.

All the teams pointed out the heterogeneity between the two uninfected groups which should have been comparable but exhibited an unexpectedly large number of differentially expressed genes. This observation raised an important biological and experimental design problem about cross-talking between udder quarters. This issue will be studied more thoroughly by the EADGENE biologists in further experiments.

A comparison of the lists of differentially expressed genes found by the workshop participants was performed for *E. coli* between times 0 and 24 h. Due to the various methods used for normalising and analysing the data, the lists were not exactly similar. It was reassuring, however, to find that even using two very different approaches, (1) normalisation by print-tip Loess and analysis with Limma in R; and (2) global Loess and Wolfinger's two-step mixed model approach in SAS®, the lists of differentially expressed genes still remained quite similar.

Here all participants had the same raw data set to analyse. Comparison of methods may have been easier, however, if each step had been evaluated separately: first, data quality control, then normalisation on a common previously cleaned data set and finally detection of differentially expressed genes on a common previously normalised set of data. Criteria to compare procedures for data quality control is still an open and essential issue for microarray data analysis.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Caroline Channing, Karin Smedegard and WP1.4 for organising the workshop, Zerbe *et al.* for providing the real data sets and EADGENE for financial support (EU Contract No. FOOD-CT-2004-506416).

REFERENCES

[1] Baird D., Johnstone P., Wilson T., Normalization of microarray data using a spatial mixed model analysis which includes splines, Bioinformatics 20 (2004) 3196–3205.

[2] Benjamini Y., Hochberg Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 85 (1995) 289–300.

[3] Dabney A.R., Leek J.T., Monsen E., Storey J.D., Edge manual, Department of Biostatistics, University of Washington, http://faculty.washington.edu/jstorey/edge, 2006.
[4] Déjean S., Martin P.G.P., Baccini A., Besse P., Clustering time series gene expression data using smoothing spline derivatives, EURASIP J. Bioinformatics Syst. Biol. (2007) ID 70561.
[5] Gibson G., Wolfinger R.D., Gene expression profiling using mixed models, in: Saxton A.M. (Ed.), Genetic analysis of complex trait using SAS®, SAS® User Press, Cary NC, 2004, Chap. 11, pp. 251–278.
[6] Ishwaran H., Rao J.S., Kogalur U.B., BAMarray™: Java software for Bayesian analysis of variance for microarray data, BMC Bioinformatics 7 (2006) 59.
[7] Jaffrézic F., Marot G., Degrelle S., Hue I., Foulley J.-L., A structural mixed model for variances in differential gene expression studies, Genet. Res. 89 (2007) 19–25.
[8] Pool M.H., Hulsegge B., Janss L.L.G., Background bias on cDNA micro-arrays, EAAP, Uppsala, Sweden, 2005.
[9] Ritchie M.E., Diyagama D., Neilson J., van Laar R., Dobrovic A., Holloway A., Smyth G.K., Empirical array quality weights for microarray data, BMC Bioinformatics 7 (2006) 261, http://www.biomedcentral.com/1471-2105/7/261
[10] Robert-Granié C., Baccini A., Besse P., Déjean S., Ferré P.J., Liaubet L., Martin P.G.P., San Cristobal M., Kinetics analysis of microarray data using semiparametric mixed models, 8th World Congress on Genetics Applied to Livestock Production, Belo-Horizonte, Brazil, August 13–18, 2006.
[11] Saeed A.I., Sharov V., White J., Li J., Liang W., Bhagabati N., Braisted J., Klapa M., Currier T., Thiagarajan M., Sturn A., Snuffin M., Rezantsev A., Popov D., Ryltsov A., Kostukovich E., Borisovsky I., Liu Z., Vinsavich A., Trush V., Quackenbush J., TM4: a free, open-source system for microarray data management and analysis, Biotechniques 34 (2003) 374–378.
[12] Smyth G.K., Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Statist. Appl. Genet. Mol. Biol. 3 (2004) 3.
[13] Smyth G.K., Limma: linear models for microarray data, in: Gentleman R.C., Carey V.J., Dudoit S., Irizarry R., Huber W. (Eds.), Bioinformatics and Computional Biology using R and Biocondutor, Springer, New York, 2005, pp. 397–420.
[14] Smyth G.K., Speed T., Normalization of cDNA microarray data, Methods 31 (2003) 265–273.
[15] Wettenhall J.M., Smyth G.K., LimmaGUI: A graphical user interface for linear modeling of microarray data, Bioinformatics 20 (2004) 3705–3706.
[16] Wolfinger R.D., Gibson G., Wolfinger E.D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules R.S., Assessing gene significance from cDNA microarray expression data via mixed models, J. Comp. Biol. 8 (2001) 625–637.
[17] Yang Y., Dudoit S., Luu P., Peng V., Ngai J., Speed T., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. 30 (2002) e15.