The EADGENE Microarray Data Analysis Workshop (Open Access publication)

Dirk-Jan de Koning, Florence Jaffrezic, Mogens Sandø Lund, Michael Watson, Caroline Channing, Ina Hulsegge, Marco H. Pool, Bart Buitenhuis, Jakob Hedegaard, Henrik Hornshøj, et al.

Genet. Sel. Evol. 39 (2007) 621–631
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007028

Available online at:
www.gse-journal.org

Original article

The EADGENE Microarray Data Analysis Workshop (*Open Access publication*)

Dirk-Jan DE KONING[a*], Florence JAFFRÉZIC[b], Mogens Sandø LUND[c], Michael WATSON[d], Caroline CHANNING[a], Ina HULSEGGE[e], Marco H. POOL[e], Bart BUITENHUIS[c], Jakob HEDEGAARD[c], Henrik HORNSHØJ[c], Li JIANG[c], Peter SØRENSEN[c], Guillemette MAROT[b], Céline DELMAS[f], Kim-Anh LÊ CAO[f,g], Magali SAN CRISTOBAL[f], Michael D. BARON[h], Roberto MALINVERNI[i], Alessandra STELLA[i], Ronald M. BRUNNER[j], Hans-Martin SEYFERT[j], Kirsty JENSEN[a], Daphne MOUZAKI[a], David WADDINGTON[a], Ángeles JIMÉNEZ-MARÍN[k], Mónica PÉREZ-ALEGRE[k], Eva PÉREZ-REINADO[k], Rodrigue CLOSSET[l], Johanne C. DETILLEUX[l], Peter DOVČ[m], Miha LAVRIČ[m], Haisheng NIE[n], Luc JANSS[c]

[a] Roslin Institute, Roslin, UK
[b] INRA, UR337, Jouy-en-Josas, France
[c] University of Aarhus, Tjele, Denmark
[d] Institute for Animal Health, Compton, UK
[e] Animal Sciences Group Wageningen UR, Lelystad, The Netherlands
[f] INRA, UMR444, Castanet-Tolosan, France
[g] Université Paul Sabatier, Toulouse, France
[h] Institute for Animal Health, Pirbright, UK
[i] Parco Tecnologico Padano (PTP), Lodi, Italy
[j] Research Institute for the Biology of Farm Animals, Dummerstorf, Germany
[k] University of Córdoba, Córdoba, Spain
[l] University of Liege, Liege, Belgium
[m] University of Ljubljana, Ljubljana, Slovenia
[n] Wageningen University and Research Centre, Wageningen, The Netherlands

(Received 10 May 2007; accepted 3 July 2007)

Abstract – Microarray analyses have become an important tool in animal genomics. While their use is becoming widespread, there is still a lot of ongoing research regarding the analysis of microarray data. In the context of a European Network of Excellence, 31 researchers representing 14 research groups from 10 countries performed and discussed the statistical analyses

* Corresponding author: DJ.dekoning@bbsrc.ac.uk

of real and simulated 2-colour microarray data that were distributed among participants. The real data consisted of 48 microarrays from a disease challenge experiment in dairy cattle, while the simulated data consisted of 10 microarrays from a direct comparison of two treatments (dye-balanced). While there was broader agreement with regards to methods of microarray normalisation and significance testing, there were major differences with regards to quality control. The quality control approaches varied from none, through using statistical weights, to omitting a large number of spots or omitting entire slides. Surprisingly, these very different approaches gave quite similar results when applied to the simulated data, although not all participating groups analysed both real and simulated data. The workshop was very successful in facilitating interaction between scientists with a diverse background but a common interest in microarray analyses.

gene expression / two colour microarray / statistical analysis

1. INTRODUCTION

The recent development of high throughput gene-expression technologies, such as microarrays, has given rise to a plethora of new research hypotheses and possibilities. Extensive reviews are available about the application [3], design [6], and analysis [12] of microarray studies. In livestock, microarrays have been proposed to study gene-expression in the parasite (Malaria [13]; Trypanosomosis [8]) as well as host response following infection (*e.g. Mycobacterium paratuberculosis* infection in cattle [7]; *Eimeria* infection in poultry [11]). Other applications in livestock include the evaluation of the effects of diet on gene expression in beef cattle [4] and gene expression differences related to differences of muscling in pigs [5].

In a recent review, Allison *et al.* outline the areas of consensus and outstanding questions with regards to microarray analysis [2]. Some points of consensus regarding data analysis as presented by those authors [2] were the following: (1) many methods exist for the pre-processing (normalisation, *etc.*) of two-colour microarrays, but there is no clear winner and none were discussed in detail; (2) using fold-change alone as a test for differential expression is inefficient; (3) false discovery rate is a good alternative to conventional multiple testing; and (4) unsupervised classification is overused and should be validated using re-sampling techniques. The most relevant outstanding questions were [2] the following: (1) the best image processing algorithm; (2) the evaluation of data quality; and (3) the assessment of intersections between sets of findings within and between experiments.

Given the lack of consensus in many areas, especially for the two-colour arrays that are abundant in livestock research, we organised a workshop on the analysis of microarrays. Conferences dealing with the statistical analyses of

microarrays using common sets of data have been successfully organised annually in the United States since 2000 [10] (http://www.camda.duke.edu/). These conferences have been large scale events, attracting 250 or more participants. In contrast, the present workshop was limited to 35 participants to maximise interaction and focussed on microarray experiments in the context of the genetics of host-pathogen interaction in livestock. The workshop was organised through the EC-funded network of excellence (NoE) EADGENE (European Animal Disease Genetics Network of Excellence for Animal Health and Food Safety; http://www.eadgene.info/).

2. WORKSHOP GOALS

The main aim of the workshop was to bring together scientists from within the EADGENE network with an interest in microarray analyses and to facilitate interaction and future collaboration between these scientists. In order to focus the discussions, the workshop was organised around two sets of data, real and simulated, that were distributed among the participants prior to the workshop. The methods of analysis, the interpretation of results and how to use the (quite complex) real experimental design were left to the participants. This was advantageous as it led to very different approaches by different groups. The diversity of approaches was a major contributor to our ability to identify outstanding questions in the treatment of microarray data.

The statistical aspects of a microarray study include the design of the study, the quantification of the hybridisation intensities, the pre-processing and normalisation of data, the inference and classification of results, the biological interpretation and finally the validation of differentially expressed genes as well as other follow-up studies. For practical reasons this workshop only dealt with the following aspects of microarray analysis: (1) some of the pre-processing of the raw microarray intensities (mainly quality control); (2) normalisation of the microarray data; (3) the detection of differentially expressed genes; (4) the clustering and classification analyses of the differentially expressed genes as well as the biological interpretation (real data only).

The workshop format allowed comparison of results for a real microarray experiment that was relevant to the remit of EADGENE as well as simulated data with known parameters, which facilitated a comparison of performance between groups. However, it must be stressed that the interaction among scientists, facilitated through common data sets, was the main objective.

Table I. Overview of workshop participants and acronyms used to describe the groups throughout the four articles.

Acronym	Affiliation	Group size	Real data	Simulated data
AARHUS	University of Aarhus, Tjele, Denmark	6	x	
CDB	University of Córdoba, Spain	3		x
IAH_C	Institute for Animal Health, Compton, UK	1	x	
IAH_P	Institute for Animal Health, Pirbright, UK	1		x
IDL	Animal Sciences Group Wageningen UR, Lelystad, NL	2	x	x
INRA_J	INRA, Jouy-en-Josas, France	3	x	x
INRA_T	INRA, Castanet-Tolosan, France	(8, 6)[a]	x (8)	x (6)
ULg	University of Liege, Liege, Belgium	2 x 1[b]	x	
PTP	Parco Tecnologico Padano, Lodi, Italy	3	x	
RIBFA	Research Institute for the Biology of Farm Animals, Dummerstorf, Germany	5[c]	x	
ROSLIN	Roslin Institute, Roslin, UK	4	x	x
SLN	University of Ljubljana, Slovenia	2	x	x
WUR	Wageningen University and Research Centre, NL	2	x	

[a] This group included six members for the analysis of simulated data (four from INRA-Station d'amélioration génétique des animaux and two from INRA-Laboratoire de génétique cellulaire) and eight members for the analysis of real data (four from INRA-Station d'amélioration génétique des animaux, three from INRA-Laboratoire de génétique cellulaire and one from the Paul Sabatier University).

[b] Two members from different groups within the University of Liege analysed and presented independently.

[c] This group included two members from Ludwig-Maximilian University in München, and one member of the University of Veterinary Medicine in Hannover.

3. THE WORKSHOP PARTICIPANTS

The data was analysed by 42 participants, representing 14 research groups from 11 EADGENE partners. During a 3-day workshop, attended by 31 participants, all groups presented and discussed their findings. The details of the different groups as well as their acronym and group sizes are presented in Table I. While all participants had shared interests through their involvement in EADGENE, they had varying levels of experience in the analyses of microarray data and different interests in taking part.

Some participants were routinely involved in the analyses of microarrays in their own institutes while others were using this workshop to gain 'hands-on' experience with the analyses of microarray data. Some groups had developed sophisticated tools to deal with a specific aspect of microarray analyses and used the workshop to demonstrate or test-drive their approach. Because the

real data was from a mastitis experiment, some participants had a particular interest in this disease and its study *via* microarray analyses.

The detailed results on the analyses of the real data are given by Jaffrézic *et al.* [9] for the quality control, the normalisation and statistical testing and Sørensen *et al.* [14] for the multiple gene analyses. The detailed results of the simulated data analyses are presented by Watson *et al.* [15].

4. OVERVIEW OF DATA

4.1. Real data

The real data consisted of 48 microarrays from an artificial infection experiment in dairy cattle with several time points and two different infectious agents: *Escherichia coli* and *Staphylococcus aureus*.

For further details on the experimental procedures see Jaffrezic *et al.* [9].

The microarray experiment was carried out using the Bovine 20K array (ARK-Genomics: http://www.ark-genomics.org/). A reference design, without dye-swap, was used and the reference sample was made up of a pool of all 48 RNA samples. The resulting microarrays were scanned and data were extracted using BlueFuse (BlueGnome, http://www.cambridgebluegnome.com/bluefuse.htm). No further adjustments or normalisations were made to these data prior to distribution to the participants. The distributed data included an automated annotation of the microarray provided by Mark Fell (Roslin Institute).

4.2. Simulated data

The microarray data were simulated using Simage [1] (http://bioinformatics.biol.rug.nl/websoftware/simage/simage_start.php). This provides a menu-driven interface in which the user can define gene effects as well as numerous noise factors. Using "Simage-R Parameter" we estimated summary statistics from a randomly selected microarray slide from the real data and used this to simulate 10 slides of a direct comparison (A *versus* B) where every second slide had treatments reversed for dyes. Although the parameter settings for the simulated data were derived from a real microarray, a lot of noise was added to really test the QC and normalisation approaches of the various groups. From the 2400 genes that were simulated, 624 were differentially expressed (264 up regulated from A to B and 360 down regulated).

4.3. Differences between real and simulated data

While the simulated data provided a simple A *versus* B comparison, the real data had the components of time and type of bacteria that were used to infect the cows. The researcher could ask different questions: *e.g.* which genes are differentially expressed following infection with a specific pathogen? At what time post infection is the differential expression most prominent? What genes are only differentially expressed following infection with one pathogen and not the other? Although unintended as a discussion area for the workshop, the real data did illustrate a problem in experimental design, which was not discovered by a pilot experiment, namely the dependence in gene expression between the udder quarters in the real data.

The results from the workshop may at first glance seem contradictory: the real data results were quite different between groups (both in numbers of differentially expressed genes and gene order) [9] while the results from the simulated data indicate that most of the approaches gave good and comparable results [15]. It could be argued that methods that give similar results for simulated data should give similar results for real data, if the two sets of data are comparable. The difference between the results from real and simulated data is most likely due to differences in the expected statistical power to detect differentially expressed genes between the two sets of data: while the real data consisted of 48 microarrays, any comparison between two time points within an infection had only four microarrays contributing to each time point, while the contrast was indirect *via* a reference design. The simulated data consisted of 10 microarrays for a direct A *versus* B comparison making it more powerful than the comparisons within the real data. It could be argued retrospectively that for comparison of methods the real data had too little power and too many possible scenarios to be tested, while the simulated data had too much power to reveal subtle differences between methods. This emphasises the benefits of this kind of workshop, since this finding was only apparent after combining and contrasting the approaches and results of the different groups, and the observation will be fed forward into future workshops.

In the simulated data only two levels of differential expression were simulated: one for up-regulated genes and one for the down-regulated genes. Even so, the mixture model distributions of test statistics showed that the various noise contributions produced symmetrical distributions for the up and down regulated genes. The simulated data was notably different from real data [15], but still allowed valuable comparisons of approaches to analysis. The simulated data did confirm that when the power of an experiment is high, many of the specific differences between the methods that are applied may become

less crucial, provided that they deal adequately with high levels of technical bias or noise. At the same time, many microarray experiments have moderate to low power and hence comparison of methods on the basis of real data has considerable merit.

5. QUESTIONS, CONSENSUS, AND RECOMMENDATIONS

5.1. Quality Control (QC)

The approaches presented during the workshop showed most divergence at the QC stage. In terms of QC of the real data, several groups used the spot quality indicators provided by the scanning software (Bluefuse) to make decisions about excluding spots from the analysis. Other groups indicated that they would normally take account of background intensities for quality control but 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 for the real data. One group re-estimated background from the data provided and used ratios between signal and inferred background to exclude bad spots while another group excluded spots on the basis of absolute intensity (mainly for the simulated data). Further to omitting bad spots based on quality indicators or (relative) intensity, some groups omitted entire slides from further analyses based on QC criteria. As an alternative to using quality indicators to include or omit spots from further analyses, one group used quality indicators as statistical weights in both normalisation and analysis of the microarrays.

The different approaches for QC led to some groups omitting no spots at all while other groups omitted many spots and even entire slides (up to two or three slides for the simulated data). Removing spots from subsequent analyses often renders the statistical model unbalanced and reduces the degrees of freedom, and hence the power of the test for a single gene. The effect of removing spots on normalisation, shrinkage of gene variance and multiple testing may counterbalance the loss of power, but these effects are less predictable. Therefore, approaches that utilise all spots but account for different quality of spots deserve further attention.

Another point of discussion was when to apply the QC: It was argued that outlier spots can only be identified after normalisation has taken account of spatial effects but the counter argument was that spots with saturated intensity measurements would bias the normalisation and should be removed before normalisation. It was suggested that QC should be applied at several stages of

the analyses but this was not implemented by any of the groups. Although QC was widely debated during the workshop, we did not define a 'best practise' for QC, although we can make a recommendation to evaluate the effect of various levels of spot editing. Again, the benefits of the workshop are shown by the unexpected identification of QC and data editing as critical factors for discussion and further study. Many publications concentrate on the statistical analysis of simulated or established datasets. While comparison showed that the statistics are generally well understood or accepted; how real experimental data is pre-processed remains a matter for further study.

5.2. Normalisation, significance testing and multi-gene analyses

For the normalisations, many groups removed intensity related bias (when the relationship between average intensity and the ratio between the two colour intensities is non-linear) by LOWESS (or LOESS) regression. One group did additional spatial smoothing while few groups included across-slide normalisation. The latter is emphasised in the results for the simulated data [15] as a way of making slides more comparable, especially for the noisy data in this study. The gene expression contrasts were mainly estimated using linear models or mixed linear models with various approaches to shrink the gene variance prior to significance testing. To address the multiple testing issues, most groups used some variant of the false discovery rate (FDR) but there were also some standard and novel approaches based on mixture distributions. The main problem of comparing gene lists between groups is that it could not be determined what stage of the different analysis pipelines caused the results to differ.

The only three approaches that performed very poorly in detecting differentially expressed genes in the simulated data was one approach based on fold-changes only, and two using ANOVA combined with the lowest level of normalisation – chip median correction. Because of the prominent print-tip effects in the simulated microarray data, failure to account for these will result in many spurious effects when using only chip-median correction and/or analysing fold-changes only [15].

With regards to the recent review by Allison *et al.* [2], the workshop echoed the points regarding the current lack of agreement in pre-processing (although the main differences were in QC rather than normalisation), in particular that fold-changes are not good criteria, and that the FDR, as well as some other novel multiple testing approaches, provide an attractive alternative to conventional multiple testing strategies. We did not address the outstanding questions

on image processing algorithms because we only used results from a single processing algorithm for the workshop.

The multi-gene analyses were too diverse for a meaningful comparison although some trends are described by Sørensen *et al.* [14]. Those analyses that were aimed at assigning biological meaning to the differentially expressed genes were hampered by the limited annotation that was available for the clones on the microarray. This will improve over time with the ongoing annotation of the cow genome sequence.

5.3. Recommendations

While the participants agreed that the workshop had been very useful, we also debated recommendations for potential future workshops on the same topic. The following recommendations were made: (1) Provide different levels of pre-processed data for different analyses. You can provide raw image files to compare image processing algorithms, while you also provide normalised data to compare different models to obtain gene lists or to compare different clustering approaches. Likewise, when comparing bioinformatics tools for the biological interpretation of microarray results, you provide a pre-set common gene list, preferably for a model species with good bioinformatics resources. (2) Ask participants to analyse specific contrasts or scenarios. For the present workshop, we gave real data on 48 slides as well as experimental details, but the participants could decide on what part of the data they would use and what contrasts they would estimate. (3) Simulate microarrays that are more similar to real data and if possible, include a range of gene effects and variances as well as a correlation structure among genes. (4) Because of limited presentation time at the workshop, some details of the analyses were missed, in particular analyses that were initially done, but not carried further. One option is to have a pre-meeting participant survey that includes what approaches were tested and how they performed. Other ways of having a more uniform reporting structure among groups may also benefit the comparisons.

6. CONCLUSIONS

The workshop succeeded in its main aim of sharing expertise and experience among statisticians and biologists using microarrays in livestock research. At least one group used it as a starting point to re-visit their own data analysis pipeline in the view of analyses and expectations of other groups. Furthermore, the three companion papers with details on the various analyses and results will

provide pointers for colleagues in the wider community regarding the options available for microarray analyses [9, 14, 15].

While a direct comparison of results between groups remained challenging, it was extremely useful to discuss microarray analyses on the basis of two common data sets. In many conferences or workshops, participants only present their own data and hence any conclusions about methods cannot be separated from the experiments to which these methods were applied. The joint analyses of the same data that was done during the workshop will also have added value to the original experiment. Furthermore, the workshop is not an endpoint but the starting point of new collaborations among researchers analysing microarray data but also between these researchers and biologists that ultimately give meaning to the results.

ACKNOWLEDGEMENTS

The organisers acknowledge local support from the University of Aarhus colleagues, in particular Karin Smedegard. We greatly appreciate the data contributed to the workshop by Hans-Martin Seyfert, the analysis done by Kirsty Jensen and all the work included of their colleagues. For the simulated data, we acknowledge early access to Simage thanks to Ritsert Jansen and his group. The workshop was organised and financially supported by EADGENE (EU Contract No. FOOD-CT-2004-506416).

REFERENCES

[1] Albers C.J., Jansen R.C., Kok J., Kuipers O.P., van Hijum S.A., SIMAGE: simulation of DNA-microarray gene expression data, BMC Bioinformatics 7 (2006) 205.

[2] Allison D.B., Cui X., Page G.P., Sabripour M., Microarray data analysis: from disarray to consolidation and consensus, Nat. Rev. Genet. 7 (2006) 55–65.

[3] Butte A., The use and analysis of microarray data, Nat. Rev. Drug Discov. 1 (2002) 951–960.

[4] Byrne K.A., Wang Y.H., Lehnert S.A., Harper G.S., McWilliam S.M., Bruce H.L., Reverter A., Gene expression profiling of muscle tissue in Brahman steers during nutritional restriction, J. Anim. Sci. 83 (2005) 1–12.

[5] Cagnazzo M., te Pas M.F., Priem J., de Wit A.A., Pool M.H., Davoli R., Russo V., Comparison of prenatal muscle tissue expression profiles of two pig breeds differing in muscle characteristics, J. Anim. Sci. 84 (2006) 1–10.

[6] Churchill G.A., Fundamentals of experimental design for cDNA microarrays, Nat. Genet. 32 (Suppl.) (2002) 490–495.

[7] Coussens P.M., Jeffers A., Colvin C., Rapid and transient activation of gene expression in peripheral blood mononuclear cells from Johne's disease positive cows exposed to *Mycobacterium paratuberculosis* in vitro, Microb. Pathog. 36 (2004) 93–108.

[8] El Sayed N.M., Hegde P., Quackenbush J., Melville S.E., Donelson J.E., The African trypanosome genome, Int. J. Parasitol. 30 (2000) 329–345.

[9] Jaffrézic F., de Koning D.J., Boettcher P.J., Bonnet A., Buitenhuis B., Closset R., Déjean S., Delmas C., Detilleux J.C., Dovč P., Duval M., Foulley J.-L., Hedegaard J., Hornshoj H., Hulsegge I., Janss L., Jensen K., Jiang L., Lavrič M., Lê Cao K.-A., Lund M.S., Malinverni R., Marot G., Nie H., Petzl W., Pool M.H., Robert-Granié C., San Cristobal M., van Schotshorst E.M., Schuberth H.-J., Sørensen P., Stella A., Tosser-Klopp G., Waddington D., Watson M., Yang W., Zerbe H., Seyfert H.-M., Analysis of the real EADGENE data set: Comparison of methods and guidelines for data normalisation and selection of differentially expressed genes., Genet. Sel. Evol. 39 (2007) 633–650.

[10] Johnson K., Lin S., Call to work together on microarray data analysis, Nature 411 (2001) 885.

[11] Min W., Lillehoj H.S., Kim S., Zhu J.J., Beard H., Alkharouf N., Matthews B.F., Profiling local gene expression changes associated with *Eimeria maxima* and *Eimeria acervulina* using cDNA microarray, Appl. Microbiol. Biotechnol. 62 (2003) 392–399.

[12] Quackenbush J., Microarray data normalization and transformation, Nat. Genet. 32 (Suppl.) (2002) 496–501.

[13] Rathod P.K., Ganesan K., Hayward R.E., Bozdech Z., DeRisi J.L., DNA microarrays for malaria, Trends Parasitol. 18 (2002) 39–45.

[14] Sørensen P., Bonnet A., Buitenhuis B., Closset R., Déjean S., Delmas C., Duval M., Glass L., Hedegaard J., Hornshoj H., Hulsegge I., Jaffrézic F., Jensen K., Jiang L., de Koning D.J., Lê Cao K.-A., Nie H., Petzl W., Pool M.H., Robert-Granié C., San Cristobal M., Lund M.S., van Schotshorst E.M., Schuberth H.-J., Seyfert H.-M., Tosser-Klopp G., Waddington D., Watson M., Yang W., Zerbe H., Analysis of the real EADGENE data set: Multivariate approaches and post analysis, Genet. Sel. Evol. 39 (2007) 651–668.

[15] Watson M., Pérez-Alegre M., Baron M.D., Delmas C., Dovč P., Duval M., Foulley J.-L., Garrido-Pavón J.J., Hulsegge I., Jaffrézic F., Jiménez-Marín Á., Lavrič M., Lê Cao K.-A., Marot G., Mouzaki D., Pool M.H., Robert-Granié C., San Cristobal M., Tosser-Klopp G., Waddington D., de Koning D.J., Analysis of a simulated microarray dataset: Comparison of methods for data normalisation and detection of differential expression., Genet. Sel. Evol. 39 (2007) 669–683.