Statistical tools to uncover genetic architecture of tick resistance using high-density oligonucleotide gene expression microarrays

Fernando Flores Cardoso[1,2], Poliana Fernanda Giachetto[3], Michel Eduardo Beleza Yamagishi[3], Roberto Hiroshi Higa[3]

[1]Embrapa Southern Region Animal Husbandry, PO Box 242, Bagé, RS, 96401-970, [2]National Council for Scientific and Technological Development (CNPq) Scholar, [3]Embrapa Agriculture Informatics, PO Box 6041, Campinas, SP, CEP 13083-970, Brazil. E-mail: fcardoso@cppsul.embrapa.br

Short title: Statistical tools for expression arrays

Introduction

Gene expression analysis using oligonucleotide microarrays are useful to identify genes that contribute to economically important phenotypic variation in farm animals, including, for instance, parasite resistance. In these experiments, the effect of a treatment, condition or genotype on gene expression (transcript abundance) is simultaneously measured for thousands of genes, facilitating the identification of gene regulatory networks and immunologic pathways (LOCKHART et al., 1996).

The processing of samples to assess gene expression involves several steps, from the RNA isolation, its transcription into cDNA and labeling with fluorescent dye, followed by hybridization on a microarray and scanning to obtain images of the fluorescence intensity, which will, at the end, be used for statistical analysis and significance test.

The aim of this paper is to present procedures and tools used for quality control and statistical analysis of high-density microarray (Affymetrix GeneChip®) to search for genes that may be differentially expressed and that may provide clues to unravel parasite resistance.

Data quality control and normalization

The GeneChip® Bovine Genome Array system uses 25-oligonucleotide sequences as probes with a gene or target sequence being represented typically by a set of 11 probes. It is assumed that these sequences are uniquely associated to a single gene, and have relatively uniform hybridization characteristics along the target. Each probe in a set, called a perfect match (PM), is paired with a second probe that has the same sequence except for a single base change at the 13th or middle position, called mismatch (MM) (CRAIG et al., 2003).

There are several steps and procedures embedded in the experiment useful to verify the quality of the RNA sample and experimental procedures (AFFYMETRIX, 1999). Quality control assumes that for a particular experiment and tissue, most genes have similar pattern of expression, while just few genes change their expression profile; therefore, arrays should have comparable metrics in the evaluated criteria. These criteria, implemented using the R/affy and R/affyqcreport Bioconductor packages (GENTLEMAN et al., 2004), include screening array images for artifacts, raw data distributions consistency, correlations between arrays, similar percentage of detected genes, background and noise values and scale factors to standardize the average fluorescence intensity. Moreover, control probes signals for image alignment, cDNA labeling, hybridization and internal genes are also checked.

Normalization is a process for reducing non-biological sources of variation across different arrays. For this end, the most popular procedure is the robust multiarray average (RMA) of Irizarry et al. (2003). It is based on PM values only and consists of four steps: a background adjustment, quantile normalization,

log2-transformation and finally summarization of the multiple corrected probes intensities to a single signal value. In addition, two statistical quality parameters are derived from the RMA procedure: RLE (relative log expressions) and NUSE (normalized unsealed standard errors) (BRETTSCHNEIDER et al., 2008), which can be obtained by the RMA/Express software (http://rmaexpress.bmbolstad.com).

Differential expression significance tests

The differentially expressed genes identification is also based on a multi-state procedure, which can be pursued using R/maanova package (WU et al., 2003). We start with a two-step single-gene mixed model analysis (WOLFINGER et al., 2001). The first step removes the main effects from all nuisance factors averaged over all genes and can be written in general as:

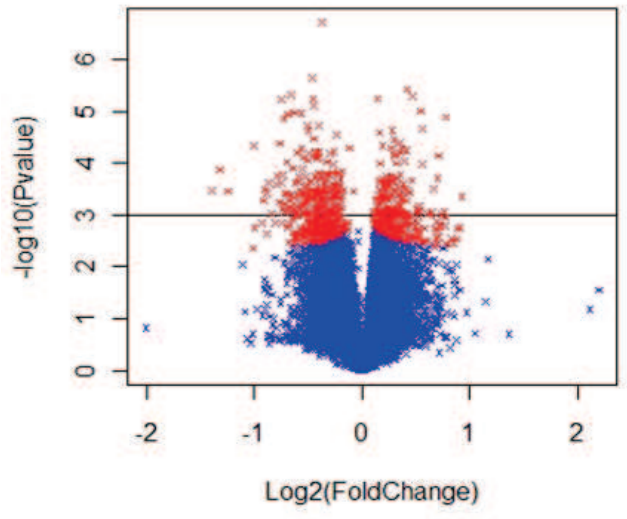
$$Y_{ijk} = \mu + T_i + A_{j(i)} + \varepsilon_{ijk},$$

where Y_{ijk} is the normalized intensity value for treatment (T) *i* ,array (A) *j* within T_i, whereas μ is a constant. Note that no genetic effects are modeled in this step. All gene-specific variation is then left in the error term ε_{ijk}. In the second step, the residuals from the model above, denoted as r_{ijk} become the response variable for a "gene model," which can be written as:

$$r_{ijk} = G_k + (TG)_{ik} + e_{ijk},$$

where G_k is the main effect of gene *k* and $(TG)_{ik}$ is the interaction between treatment *i* and gene *k*, which is of primary interest measuring the treatment effects on each gene. Random terms are $A_{j(i)}$, ε_{ijk} and e_{ijk}, with the respective variances, σ^2_A, σ^2_ε, and $\sigma^2_{e(k)}$, where the *k* subscript indicates heterogeneous error variances for each gene.

For significance test, combining information across genes is desirable due to the relatively small data points number obtained for each individual gene. Here, we recommend the F_S test statistic constructed base on an error variance estimator that can borrow information across genes using the James–Stein shrinkage concept (CUI et al., 2005).



Because the F_S distribution under the null hypothesis does not follow a standard form, R/maanova provides permutation method to calculate the significance of this statistical test. Finally, to control the false discovery rate (FDR) adjusted p-value can be obtained by the Storey (2002) method. The side figure shows a volcano plot generated by R/maanova, depicting the level of significance in the y-axis and the amount of expression change represented by the Log2(FoldChange) in x-axis. In this plot, each x is a gene and red coloring indicates significance under the F_S permutation adjusted test.

Annotation

The result of differential expression analysis is a list of probes with their respective fold change values and statistical significance (p-value). The next step is to identify the genes mapped by the probes and then search for functional information, allowing the researcher to select the most interesting genes to investigate the mechanisms that underlie the biological processes or phenotypes under study. When using commercial arrays from Affymetrix, NimbleGen, Agilent and Illumina, there are packages for the annotation process available at the Bionconductor repository, which search information in public databases such as GO (http://www.geneontology.org), KEGG (http://www.genome.ad.jp/kegg/kegg2.html) and Expasy (http://www.expasy.org.br).

In addition to those packages, there are web tools, such as Blast2GO (CONESA et al., 2005; GÖTZ et al., 2008) and DAVID (HUANG et al., 2009), which can also be used. Besides the annotation, these tools allow a more systematic functional interpretation of the differentially expressed genes set, obtained by enrichment analysis. Such analysis helps to reduce the dimensionality of data, identifying the gene ontology terms and/or pathways that are overrepresented in the gene set. The functional enrichment analysis requires, besides the gene set, a list of all elements that make up the background population of the analysis, for example, all valid probes used in testing for differential expression. Different kind of statistical tests can be used for functional enrichment analysis, being the most common Binomial test, Fisher's exact test, and Z-score testAlternatively, the microarray manufacturer's provides, on the company website, a file with the probes annotation. However, usually this annotation is not updated, making the use of the tools above mentioned, among others, the best choice for the annotation process.

Cluster analyses

Clustering is an exploratory data analysis which has been used for a long time in research areas such as image processing and pattern recognition. Basically, it consists on dividing a set of objects into subsets (called clusters), such that two objects in the same cluster are more similar to each other, according to a pre-specified criterion.

In microarray gene expression analysis, clustering analysis is used to group genes having similar expression profiles. Due to the large number of genes simultaneously analyzed, even though having found those differentially expressed, clustering analysis can be used to group co-expressed genes, helping on the identification of functional relationship among them as well as on the reduction of the amount of information to be analyzed (BRUN et al., 2005). Clustering analysis can also be used to group samples (arrays) based on their expression profile similarity. In this case, result can reveal new sub-classes related to the investigated phenotype (e.g. a new sub-type of disease). In addition, sample clustering can also be helpful for data quality control. Two arrays corresponding to biological replicates that do not group together can indicate problems in the experimental design, pre-processing and/or hybridization step (WIT and MCCLURE, 2004).

Usually, clustering analysis can be divided into three steps: (i) pre-processing; (ii) clusters identification; and (iii) clusters validation. In step (i), the variables used by the clustering algorithm are selected and/or normalized; and the similarity measure to be used is chosen. In step (ii) the clustering algorithm and the respective parameters to be used (e.g. k-means, hierarchical clustering and self organizing maps – SOM) are run in order to build the clusters (EVERITT et al., 2001). Finally, in step (iii), techniques for cluster validation, like Dunn index, silhouette width and/or connectivity index (HANDL et al., 2005) are used to evaluate the clusters obtained in step (ii). As the output of a clustering analysis is highly dependent of the choices made in steps (i) and (ii), the output of the step (iii) provides an indication of the appropriateness of those choices.

Final remarks

The microarray technology allows to integrate genetics and physiology in the study of relevant issues related to parasite resistance.

Quality of experimental design and implementation in all steps is essential to soundness of the results obtained and inferences derived.

The nature of microarray studies is prospective, providing insights for the associations between groups of genes and physiological traits, links between the genome and the biological processes involved in the manifestation of the phenotype and may generate new hypotheses to be tested in subsequent studies, using more specific or targeting approaches.

References

AFFYMETRIX. **Affymetrix microarray suite user guide**, version 4. ed. Santa Clara CA, 1999.

BRETTSCHNEIDER, J.; COLLIN, F.; BOLSTAD, B. M.; SPEED, T. P. Quality assessment for short oligonucleotide microarray data. **Technometrics**, v. 50, n. 3, p. 241-264, 2008.

BRUN, M.; JOHNSON, C. D.; RAMOS, K. S. Clustering: revealing intrinsic dependencies in microarray data. In: DOU-GHERTY, E. R.; SHMULEVICH, I.;CHEN, J.; WANG, J. (Ed.). **Genomic signal processing and statistics**. New York: Hindawi Publishing Corporation, 2005. p. 129-162.

CONESA, A.; GOTZ, S; GARCÍA-GÓMEZ, J. M.; TEROL, J.; TALÓN, M.; ROBLES, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674-3676, 2005.

CRAIG, B. A.; BLACK, M. A.; DOERGE, R. W. Gene expression data: the technology and statistical analysis. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 8, n. 1, p. 1-28, 2003.

CUI, X. G.; HWANG, J. T. G.; QIU, J.; BLADES, N. J.; CHURCHILL, G. A. Improved statistical tests for differential gene expression by shrinking variance components estimates. **Biostatistics**, v. 6, n. 1, p. 59-75, 2005.

EVERITT, B. S.; LANDAU, S.; LEESE, M. **Cluster analysis**. 4. ed. London: Arnold, 2001. 237 p.

GENTLEMAN, R. C; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J.; HORNIK, K.; HOTHORN, T.; HUBER, W.; IACUS, S.; IRIZARRY, R.; LEISCH, F.; LI, C.; MAECHLER, M.; ROSSINI, A. J.; SAWITZKI, G.; SMITH, C.; SMYTH, G.; TIERNEY, L.; YANG, J. Y.; ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology**, v. 5, n. 10, p. R80, 2004.

GÖTZ, S.; GARCÍA-GÓMEZ, J. M.; TEROL, J.; WILLIAMS, T. D.; NAGARAJ, S. H.; NUEDA, M. J.; ROBLES, M.; TALÓN, M.; DOPAZO, J.; CONESA, A. High-throughput functional annotation and data mining with the Blast2GO suite. **Nucleic Acids Research**, v. 36, n. 10, p. 3420-3435, 2008.

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. **Bioinformatics**, v. 21, n. 15, p. 3201-3212, 2005.

HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. **Nature Protocols**, v. 4, n. 1, p. 44-57, 2009.

IRIZARRY, R. A.;HOBBS, B.; COLLIN, F.; BEAZER-BARCLAY, Y. D.; ANTONELLIS, K. J.; SCHERF, U.; SPEED, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. **Biostatistics**, v. 4, n. 2, p. 249-264, 2003.

LOCKHART, D. J.; DONG, H. L.; BYRNE, M. C.; FOLLETTIE, M. T.; GALLO, M. V.; MITTMANN, M.; WANG, C.; KOBAYASHI, M.; HORTON, H.; BROWN, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nature Biotechnology**, v. 14, n. 13, p. 1675-1680, 1996.

STOREY, J. D. A direct approach to false discovery rates. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 64, n. 3, p. 479-498, 2002.

WIT, E.; MCCLURE, J. **Statistics for microarrays; design, analysis and inference**. 1. ed. Hoboken: John Wiley, 2004. 278 p.

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. **Journal of Computational Biology**, v. 8, n. 6, p. 625-637, 2001.

WU, H.; KERR, M.; CUI, X.; CHURCHILL, G. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In: PARMIGIANI, G.; GARRET, E. S.; IRIZARRY, R. A.; ZEGER, S. L. (Eds.) **The analysis of gene expression data: methods and software**. New York: Springer, 2003. p. 341

.